Robustly Estimating Changes in Image Appearance

Michael J. Black

Xerox Palo Alto Research Center, 3333 Coyote Hill Rd., Palo Alto, CA 94304
E-mail: black@parc.xerox.com

David J. Fleet

Xerox Palo Alto Research Center, 3333 Coyote Hill Rd., Palo Alto, CA 94304
Department of Computing and Information Science, Queen's University, Kingston, Canada
E-mail: fleet@parc.xerox.com

and

Yaser Yacoob

Computer Vision Laboratory, University of Maryland, College Park, MD 20742
E-mail: yaser@cs.umd.edu

We propose a generalized model of image "appearance change" in which brightness variation over time is represented as a probabilistic mixture of different causes. We define four generative models of appearance change due to: 1) object or camera motion; 2) illumination phenomena; 3) specular reflections; and 4) "iconic changes" which are specific to the objects being viewed. These iconic changes include complex occlusion events and changes in the material properties of the objects. We develop a robust statistical framework for recovering these appearance changes in image sequences. This approach generalizes previous work on optical flow to provide a richer description of image events and more reliable estimates of image motion in the presense of shadows and specular reflections.

Key Words: optical flow, mixture models, outliers, probabilistic models, illumination change, specularities, iconic change.

1. INTRODUCTION

As Gibson noted, the world is made up of surfaces that "flow or undergo stretching, squeezing, bending, and breaking in ways of enormous mechanical complexity" ([21], page 15). These events result in a wide variety of changes in the "appearance" of objects in a scene. While motion and illumination changes are examples of common scene events that result in *appearance change*, numerous other events occur in nature that cause changes in appearance. For example, the color of objects can change due to chemical processes (e.g., oxidation), objects can change state (e.g., evaporation, dissolving), or objects can undergo radical changes in structure (e.g., exploding, tearing, rupturing, boiling). In this paper we formulate a robust statistical framework for representing certain classes of appearance changes. In so doing we have three primary goals. First, we wish to "explain" appearance changes in an image sequence as resulting from a "mixture" of causes. Second, we wish to locate where particular types of appearance change are taking place in an image. And, third, we want to provide a framework that generalizes previous work on motion estimation.

The estimation of motion in image sequences is a difficult problem that involves pooling noisy measurements to make reliable estimates. This assumes some *model* of the image variation within a region. For example, it is commonly assumed that the brightness within a region is conserved through time, that a single motion is present, and that the motion can be described by a low-order polynomial. For natural scenes, this model is a crude approximation that fails to capture many kinds of appearance change such as those mentioned above.

When our models of the scene are violated, we have two choices, namely, formulate more realistic models or adopt robust statistical techniques to cope with the violations of the assumptions. In general we should pursue the former while recognizing its limitations. For example, "better" models may require that more parameters be estimated which may be undesirable since even simple

models may be underconstrained. Furthermore, although "simple" models used for optical flow are typically linear in the unknown parameters and admit closed form or efficient iterative solutions, significantly "better" models may be non-linear and computationally prohibitive. Finally, in natural scenes, any models we formulate will be approximate and certain appearance changes will remain unmodeled. These unmodeled image variations require us to maintain a robust statistical formulation even as our models improve.

In this paper we pursue a strategy of both constructing more realistic models of appearance change and formulating the problem using robust statistical techniques. Specifically, we discuss the use of four generative models to "explain" the classes of appearance change illustrated in Figure 1. A change in "form" is modeled as the motion of pixels in one image to those in the next image; that is, an image at time $t+1$ can be explained by warping the image at time t using this image motion. Our framework uses a layered representation to model multiple motions in a region due to occlusion and limited forms of transparency.

Illumination variations may be global, occurring throughout the entire image due to changes in the illuminant. They may also be local like the cast shadow of the hand that appears in Figure 1 (upper right). In this paper we model illumination change simply as a smooth function that amplifies/attenuates image contrast. By comparison, specular reflections (Figure 1, lower right) are typically local, especially near regions of high surface curvature, and can be modeled, in the simplest case, as a near saturation of image intensity.

The fourth class of events considered in this paper is iconic change [9]. We use the word "iconic" to indicate changes that are "pictorial." These are *systematic* changes in image appearance that are not readily explained by physical models of motion, illumination, or specularity. A simple example is the blinking of the eye in Figure 1 (lower left). Examples of physical phenomena that give rise to iconic change include occlusion, disocclusion, changes in surface materials, and motions of non-rigid objects. In this paper we consider iconic changes to be object specific and we "learn" models of the iconic structure for particular objects using eigenspace techniques [38].

These different types of appearance change commonly occur together with natural objects, for example, with articulated human motion or the textural motion of plants, flags, water, etc. We employ a probabilistic mixture model formulation [30] to recover the various types of appearance change and to perform a soft assignment, or classification, of pixels to causes. This is illustrated in Figure 2. In natural speech, the appearance change of a mouth between frames can be great due to the appearance/disappearance of the teeth, tongue, and mouth cavity. While changes around the mouth can be modeled by a smooth deformation (image $t+1$ warped to approximate image t) the large disocclusions are best modeled as an iconic change (taken here to be a linear combination of learned basis images).

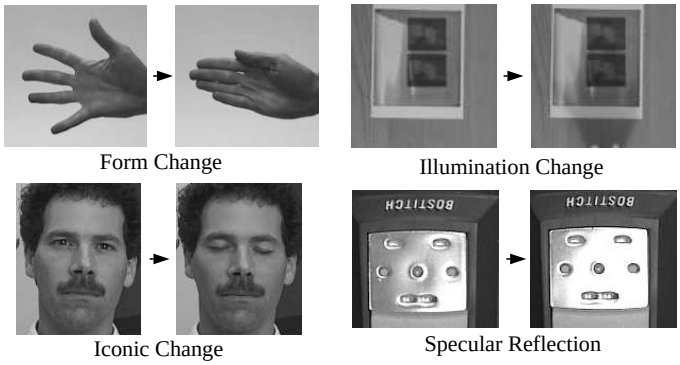

FIG. 1. Four classes of appearance change (explained in text).

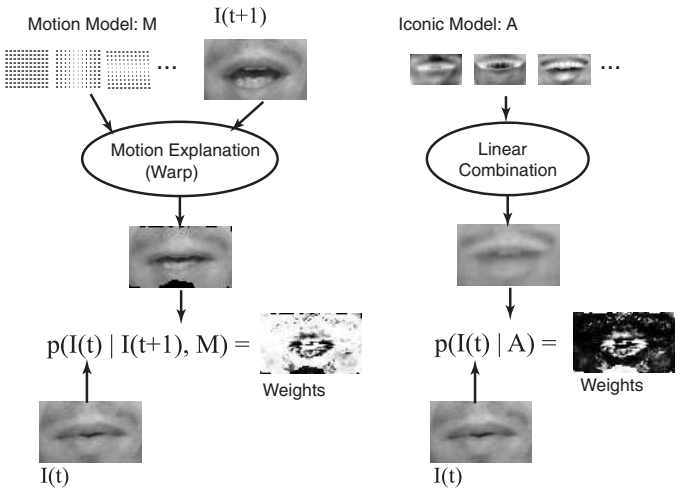

FIG. 2. Two generative models of an image at time t. Motion is represented by a parameterized deformation from the image at time $t+1$ to the image at time t. Iconic change is represented by a linear combination of learned basis images. The "weights" represent the probability that the pixels in $I(t)$ were generated (or are explained) by each of the models.

Both deformation and iconic change can be viewed as generative models and our goal is to estimate the parameters of these models. We define the probability of observing the image at time t given each of these "causes". Given this formulation, the Expectation-Maximization (EM) algorithm [16, 30] is used to iteratively compute maximum likelihood estimates for the deformation and iconic model parameters as well as the probabilities that pixels at time t are explained by each of the causes. These probabilities are the "weights" in Figure 2 and they provide a soft assignment of pixels to causes.

Changes in image appearance that are not modeled well by iconic change, deformation, illumination variations, or specularities are considered to be *outliers* [24]. To repre-

sent them explicitly, the mixture-model contains an *outlier layer* that receives high weights for pixels that are unexplained by any of the models [26]. This helps to ensure robustness when violations of the models occur. The outlier layer also helps us to identify regions where our models fall short of explaining the appearance change, and therefore require improvement. Below we describe our mixture-model formulation and a collection of appearance-change models that generalize the notion of brightness constancy used in estimating optical flow.

2. CONTEXT AND PREVIOUS WORK

Previous work in image sequence analysis has focused on the measurement of optical flow using conservation assumptions and smooth models of the optical flow field [4]. One common assumption, referred to as brightness constancy, is that the image brightness $I(\mathbf{x}, t)$ at a pixel $\mathbf{x} = [x, y]$ and time t can be represented by a deformation of the image at time $t + 1$:

$$I(\mathbf{x}, t) = I(\mathbf{x} - \mathbf{u}(\mathbf{x}), t + 1), \tag{1}$$

where $\mathbf{u}(\mathbf{x}) = (u(\mathbf{x}), v(\mathbf{x}))$ represents the horizontal and vertical displacement of the pixel. Although useful in many contexts, it is well-known that brightness constancy is often violated by shadows, global illumination changes, specular reflections, and the occlusion or disocclusion of surfaces. In the remainder of this section we review approaches for making optical flow estimation robust to changes such as these and relate these to our formulation of appearance change.

Image Preprocessing

One approach to coping with violations of brightness constancy has been to preprocess the image to extract image properties whose deformations through time provide a more reliable measure of the desired flow field. Common approaches include band-pass filtering and contrast gain normalization to remove smooth illumination variations, or the extraction of image features such as edges or regions to achieve robustness with respect to even more significant appearance changes [1, 32, 39]. Fleet and Jepson [19] proposed the use of local phase information (from the output of bandpass filters). Phase is stable with respect to smooth variations in illumination and smooth geometric deformations between frames. Moreover, locations of phase instability can be detected and therefore ignored, making the subsequent estimation of optical flow more robust [20].

Robust Estimation

Image preprocessing, although useful, will not always account for the full range of ways in which image brightness may change. Models of brightness change will be an approximation to the true physical processes and hence will be violated. These violations can be viewed as statistical outliers, and suggest the need for robust estimation.

The assumption of smooth optical flow fields, although useful in many situations, is also violated often in practice. Models of smoothness have been applied in image patches using regression techniques or through the propagation of local information using regularization techniques. In particular, with linear parameterized models, the optical flow field is often represented by a low-order polynomial (constant, affine, or planar) [5, 8, 14, 19, 36], and is estimated by collecting hundreds or thousands of constraints over an image region and using regression methods or other search techniques. These approaches can recover accurate motion estimates when the motion model is a good approximation to the image motion. In real scenes, however, simple motion models are often inappropriate either because the motion is more complex or there are multiple objects moving with different velocities.

Early attempts at robust optical flow estimation involved least-squares regression followed by outlier detection and rejection, and then re-estimation of the motion for the remaining image pixels [25]. Black and Anandan [8] introduced robust statistical techniques (M-estimation) [24] to compute a dominant motion while automatically down-weighting outliers. Multiple motions can be computed in a region by successively applying robust estimation techniques to the outliers [8, 37]. Bab-Hadiashar and Suter [3] developed a robust approach using the Least Median of Squares (LMedS) technique [35] to estimate the dominant motion in a region. These robust methods can typically cope with a small number of motions within a region but not with general flow fields. Other methods add further robustness by allowing regions to vary in size [10], or by regularizing flow both within and between image regions [28].

Layered Models

The robust estimation techniques above typically assume a single dominant motion within a given region. Layered models relax this assumption and estimate multiple motions in a region. Darrell and Pentland [15] introduced the idea of estimating global motions in layers and presented an optimization scheme using ideas from robust statistics. Wang and Adelson [42] also formulated a model that groups coherent velocity estimates into layers but their approach did not exploit the layered model to directly estimate motion from images. Jepson and Black [26] assumed that the motion in the scene could be represented by a mixture of distributions and used the Expectation Maximization (EM) algorithm to decompose the motion into a fixed number of layers. These layered approaches, and the EM algorithm in particular, have become popular methods for motion estimation [2, 28, 43, 44, 45].

One issue with layered models concerns the estimation of the appropriate number of layers. A number of authors

have used a minimum description length criterion to strike a balance between accurate encoding of the motion and the number of layers needed to represent it [2, 15, 27].

With parameterized models, the estimation of the motion of a given surface may be adversely affected by distant, and quite unrelated, image points. These distant motions can act as "leverage points" [35] that pull the solution away from the desired local motion. A spatial smoothness constraint can be added to the computation of the weights that assign pixels to layers [27, 44]. This may reduce the effect of leverage points by encouraging layers to have spatially coherent support.

Specialized Spatial Models

Parameterized approaches may perform poorly when the spatial variation of the image motion is more complex than a low-order polynomial. To handle complex motions with concise models, Black *et al.* [13] proposed "learning" linear parameterized models from training examples using principal component analysis (PCA). Similarly, Fleet *et al.* [18] modeled motion features, such as dynamic occlusion edges and moving bars, using linear combinations of steerable basis flow fields. These linear models constrain the interpretation of image motion, and are used in the same way as translational or affine motion models. Similar approaches have been used in modeling the deformations between individual faces in a database of face images [6, 17, 23, 33, 40, 41].

Generalizing Brightness Constancy

Much of the recent work in motion estimation has focused on achieving increased robustness in the presence of unmodeled appearance changes. In this paper we take the approach of explicitly modeling many of these events and hence extend the notion of "constancy" to more complex types of appearance change.

One motivation for this is our interest in recognizing complex non-rigid and articulated motions, such as human facial expressions. Previous work in this area has focused on analyzing the image motion of face regions such as mouths [12]. But image motion alone does not capture appearance changes such as the systematic appearance/disappearance of the teeth and tongue during speech and facial expressions. For machine recognition we would like to be able to model these intensity variations.

Our framework extends several previous approaches that generalize the brightness constancy assumption. Mukawa [31] extended the brightness constancy assumption to allow illumination changes that are a smoothly varying function of the image brightness. In a related paper, Negahdaripour [34] proposed a general linear brightness constraint

$$I(\mathbf{x}, t) = m(\mathbf{x}, t) I(\mathbf{x} - \mathbf{u}(\mathbf{x}), t+1) + c(\mathbf{x}, t) \qquad (2)$$

where $m(\mathbf{x}, t)$ and $c(\mathbf{x}, t)$ allow for multiplicative and additive deviations from brightness constancy and are assumed to be constant within an image region.

Another generalization of brightness constancy was proposed by Nastar *et al.* [33]. Treating image intensity I as the height of a surface in 3D XYI-space, they proposed a physically-based approach for finding the deformation from an XYI surface at time t to the XYI surface at $t+1$. This allows for a general class of smooth deformations between frames, including both multiplicative and additive changes to intensity.

One variation on the general form of (2) is the the use of object-specific models of image brightness [7, 22, 23, 41]. Hager and Belhumeur [22] used principal component analysis to find a set of orthogonal basis images, $\{B_j(\mathbf{x})\}_{j=1}^{n}$, that spanned the ensemble of images of an object under a wide variety of illuminant directions. They constrained deviations from brightness constancy to lie in the subspace of illumination variations, giving the constraint

$$I(\mathbf{x}, t) = I(\mathbf{x} - \mathbf{u}(\mathbf{x}; \mathbf{m}), t+1) + \sum_{j=1}^{n} b_j \, B_j(\mathbf{x}), \qquad (3)$$

where $\mathbf{u}(\mathbf{x}; \mathbf{m})$ is a parameterized (affine) model of image motion. The authors estimated the motion coefficients $\mathbf{m}$ and the subspace coefficients $b_1 \ldots b_n$.

This is similar to our model of illumination variation but does not allow mixtures of multiple causes within a region. These approaches are also related to the eigentracking work of Black and Jepson [11] in which subspace constraints were used to help account for iconic changes in appearance while an object was being tracked.

3. MIXTURE MODEL OF APPEARANCE CHANGE

The approach presented here recasts a number of the above approaches in a probabilistic mixture model framework [30]. We propose a set of generative models that can be used to construct or explain an image. Unlike the approaches above, the mixture model framework decomposes the appearance change into multiple causes. It also performs a soft assignment of pixels to the different models while allowing for outliers, i.e., pixels that are not well explained by any one model.

In particular, we assume that the image $I(\mathbf{x}, t)$ at location $\mathbf{x}$ at time t is generated, or explained, by one of n causes I_{C_i}, $i = 1, \ldots, n$. The causes, $I_{C_i}(\mathbf{x}, t; \mathbf{a}_i)$, can be thought of as overlapping *layers* and are simply images that are generated given a vector of parameters $\mathbf{a}_i$. We will consider four causes below, namely, motion (I_M), illumination variations (I_L), specular reflections (I_S), and iconic (pictorial) changes (I_P). A fifth cause (I_O) will represent outliers.

Given n of the above causes, the probability of observing the image $I(\mathbf{x}, t)$ is a mixture model [30] given by

$$p(I(\mathbf{x},t) \mid \{\mathbf{a}_j, \sigma_j\}_{j=1}^{n}) = \sum_{i=1}^{n} w_i(\mathbf{x})\, p_i(I(\mathbf{x},t) \mid \mathbf{a}_i, \sigma_i). \quad (4)$$

The $w_i(\mathbf{x})$ are "ownership probabilities." They specify the relative probabilities that the different models account for the appearance change at pixel $\mathbf{x}$. At each pixel, these probabilities sum to unity; that is, $\sum_i w_i(\mathbf{x}) = 1$. In practice, we use a single outlier model while we may employ any number of motion, illumination, specularity, or iconic models to explain the image region.

The dependence of $w_i(\mathbf{x})$ on image location allows for the fact that the appearance change at different pixels will often be explained by different causes that vary across the image. This generalizes the more common formulation in which mixture probabilities π_i replace the w_i in (4), and are given as the average weights over an image region (or over an ensemble of independent samples from the distribution); that is, $\pi_i = \sum_{\mathbf{x}} w_i(\mathbf{x})/N$. In our formulation, the causes provide parametric models over the entire image region, while the weights represent the relative likelihoods $p_i(I(\mathbf{x},t)|\, \mathbf{a}_i, \sigma_i)$ of each cause at every pixel.

Finally, the σ_i in (4) are scale parameters that are used to control a form of deterministic annealing in the estimation of the parameters (to be discussed below).

Robustness in the current framework occurs in two ways, namely, with the use of an outlier layer and with the form of the likelihood function used. In contrast to a Gaussian mixture formulation, the component probabilities used here for the generative models of appearance change, $p_i(I(\mathbf{x},t)|\, \mathbf{a}_i, \sigma_i)$, are defined to be

$$p_i(I(\mathbf{x},t)|\, \mathbf{a}_i, \sigma_i) = \frac{2\sigma_i^3}{\pi(\sigma_i^2 + \Delta I_{C_i}^2)^2}, \quad (5)$$

where $\Delta I_{C_i} = I(\mathbf{x},t) - I_{C_i}(\mathbf{x},t;\, \mathbf{a}_i)$. This is a t-distribution of degree 3 centered at $I_{C_i}(\mathbf{x},t;\, \mathbf{a}_i)$ with standard deviation σ_i [29]. Simply put, the probability that an image pixel at time t is explained by each cause is a function of the difference between the observed intensity and that predicted by the model. This likelihood function (Figure 3) has the properties that it falls off more rapidly than a Gaussian distribution and has heavier tails. This reflects our expectation that the residuals ΔI_{C_i} contain outliers [24]. The fact that the likelihood drops rapidly will have the effect of forcing large residuals for a given model to be accounted for by other models, thereby helping to separate the explanation of image data into distinct causes.

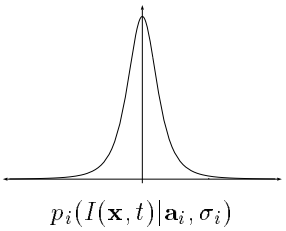

FIG. 3. A robust likelihood function, p_i.

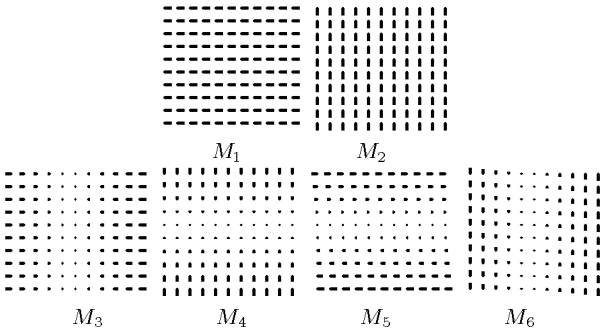

FIG. 4. Affine flow basis set.

3.1. Sources of Appearance Change

In what follows, we describe the four generative models of appearance change and the outlier model in more detail.

Motion

Motion is a particularly important type of appearance change that is modeled by

$$I_M(\mathbf{x}, t;\, \mathbf{a}_M) = I(\mathbf{x} - \mathbf{u}(\mathbf{x};\, \mathbf{a}_M),\, t+1).$$

The image at time t is generated by warping the image at time $t+1$ by a flow field $\mathbf{u}(\mathbf{x};\, \mathbf{a}_M)$. We use a parametric description of optical flow in which the motion in an image region is modeled as a linear combination of k basis flow fields $\{M_j(x)\}_{j=1}^{k}$:

$$\mathbf{u}(\mathbf{x};\, \mathbf{a}_M) = \sum_{j=1}^{k} m_j\, M_j(\mathbf{x}). \quad (6)$$

where $\mathbf{a}_M = [m_1, \ldots, m_k]$ is the vector of parameters to be estimated.

For the experiments in Section 5 we use an affine flow model. For an image region about pixel (x_c, y_c), the affine model is given by

$$u(x,y) = m_0 + m_1(x - x_c) + m_2(y - y_c), \quad (7)$$
$$v(x,y) = m_3 + m_4(x - x_c) + m_5(y - y_c), \quad (8)$$

Equivalently, we can express affine motion as in (6) with an explicit set of constant and linear basis flow fields, as shown in Figure 4.

Illumination Variations

Illumination changes may be global, resulting from changes in the illuminant, or local, as a result of shadows

FIG. 5. Linear illumination-change basis images.

cast by objects in the scene. The mixture formulation allows both of these types of variation to be modeled, where the (ownership) weights in the mixture indicate where illumination variations have occurred in the image.

With simple changes in illumination, an image at time t can be written as a scaled version of the image at time $t+1$, i.e., $[1 + L(\mathbf{x}; \mathbf{a}_L)]\, I(\mathbf{x}, t+1)$ where $1 + L(\mathbf{x}; \mathbf{a}_L)$ is the scaling function parameterized by $\mathbf{a}_L$. The change in appearance is then $L(\mathbf{x}, \mathbf{a}_L)\, I(\mathbf{x}, t+1)$. If we allow for motion as well as illumination change, then the change in image appearance can be written as

$$I_{L,M}(\mathbf{x}, t; \mathbf{a}_M, \mathbf{a}_L) = L(\mathbf{x}; \mathbf{a}_L)\, I(\mathbf{x} - \mathbf{u}(\mathbf{x}; \mathbf{a}_M), t+1). \quad (9)$$

This states that the illumination change is a scaled version of the motion-compensated image at time $t+1$. When estimating the parameters $\mathbf{a}_L$ here we assume that the motion $\mathbf{u}(\mathbf{x}; \mathbf{a}_M)$ is known and fixed.

We take $L(\mathbf{x}; \mathbf{a}_L)$ to be a parametric model, expressed as a weighted sum of basis images. For example, in the case of linear spatial variation, L is given by

$$L(\mathbf{x}; \mathbf{a}_L) = l_1 + l_2(x - x_c) + l_3(y - y_c) = \sum_{i=1}^{3} l_i\, L_i(\mathbf{x})$$

where (x_c, y_c) is the center of the relevant image region, $\mathbf{a}_L = [l_1, l_2, l_3]$ are the model parameters, and $L_i(\mathbf{x})$ denote the basis images, shown for the linear model in Figure 5.

Specularity Model

Specularities are typically local and result in near saturation of image brightness. While more sophisticated models of specularities may be formulated, we have experimented with a simple model which works well in practice:

$$I_S(\mathbf{x}, t; \mathbf{a}_S) = s_1 + s_2(x - x_c) + s_3(y - y_c) = \sum_{i=1}^{3} s_i\, S_i(\mathbf{x})$$

where S_i are the same linear basis images as in Figure 5 and $\mathbf{a}_S = [s_1, s_2, s_3]$. Note that unlike the illumination model, the specularity term is independent of the image.

Iconic Change

In addition to the generic types of appearance change above, there are image appearance changes that are specific to particular objects or scenes. Systematic changes in appearance exhibit spatial or temporal structure that can be modeled and used to help explain appearance changes in image sequences. Recall the example of human mouths in Figure 2. As people talk, their lips deform smoothly but there are also changes that cannot be characterized as smooth deformation, such as the appearance and disappearance of the teeth as the mouth opens and closes (Figure 6).

As with the models above, we use a linear, parametric model of iconic change. However, here we learn the appropriate model from the individual frames of a training image sequence using principal component analysis. This is described in Section 6; for now it is sufficient to think of the iconic model, like the specularity model, as a linear combination of basis images $A_i(\mathbf{x})$

$$I_P(\mathbf{x}, t; \mathbf{a}_P) = \sum_{i=1}^{q} a_i\, A_i(\mathbf{x}), \quad (10)$$

where $\mathbf{a}_P = [a_1, \ldots, a_q]$ is the vector of scalar values to be estimated.

Outlier Model

For the outlier layer, we adopt a simple model in which image intensity is uniformly distributed between the minimum and maximum intensity values; that is, the model can generate (explain) any pixel with uniform probability. Chosing σ_O such that it satisfies

$$p_O(I(\mathbf{x}, t)|\sigma_O) = \frac{2\sigma_O^3}{\pi(\sigma_O^2 + (2.5\sigma_O)^2)^2} = 1/256$$

gives the likelihood of an outlier.

4. EM-ALGORITHM

We seek a maximum likelihood estimate of the global model parameters $\mathbf{a}_1, \ldots, \mathbf{a}_n$ and the ownership probabilities, $w_1(\mathbf{x}), \ldots, w_n(\mathbf{x})$ that yield a soft assignment of pixels to models. If the parameters of the different models are known, then we can compute the probability that pixel $\mathbf{x}$ belongs to cause i. These probabilities, referred to as ownership weights, are given by [30]

$$w_i(\mathbf{x}, \sigma_i) = \frac{p_i(I(\mathbf{x}, t)|\, \mathbf{a}_i, \sigma_i)}{\sum_{j=1}^{n} p_j(I(\mathbf{x}, t)|\, \mathbf{a}_j, \sigma_j)} \quad (11)$$

These ownership weights force every pixel to be explained by some combination of the different causes. As the σ_j go

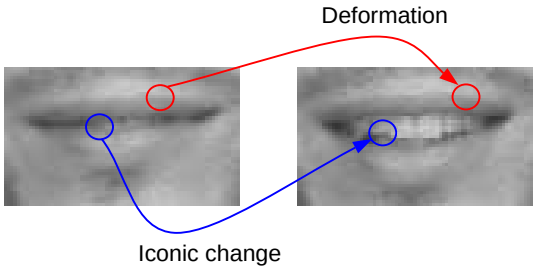

FIG. 6. Object-specific deformation and iconic change.

to zero, the likelihood function approaches a delta function. Therefore, for small values of σ_j the weights will tend towards zero or one.

The maximum likelihood estimate of the model parameters, given the ownership weights, satisfies [30]

$$\sum_{\mathbf{x}} \sum_{i=1}^{n} w_i(\mathbf{x}, \sigma_i) \frac{\partial}{\partial \mathbf{a}_i} \log p_i(I(\mathbf{x}, t) | \mathbf{a}_i, \sigma_i) = 0 \quad (12)$$

where the derivative of the log likelihood is given by

$$\frac{\partial \log p_i(I(\mathbf{x}, t) | \mathbf{a}_i, \sigma_i)}{\partial \mathbf{a}_i} = \Psi(\Delta I_{C_i}, \sigma_i) \frac{\partial I_{C_i}(\mathbf{x}, t; \mathbf{a}_i)}{\partial \mathbf{a}_i} \quad (13)$$

where

$$\Psi(\Delta I_{C_i}, \sigma_i) = \frac{-4 \Delta I_{C_i}}{\sigma_i^2 + \Delta I_{C_i}^2}, \quad (14)$$

and $\Delta I_{C_i} = I(\mathbf{x}, t) - I_{C_i}(\mathbf{x}, t; \mathbf{a}_i)$ for the i^{th} model. Note the similarity between the derivative of the log likelihood used here in Figure 7 and the shape of the influence functions of common robust M-estimators [24]. In M-estimation this shape has the effect of reducing the influence of outliers on the maximum likelihood estimate. The Ψ-function here has the same effect.

In the case of Gaussian mixtures with linear models, the model parameters can be computed in closed form given the ownership weights. However, with the robust likelihood function and the nonlinear models used here, we incrementally compute the $\mathbf{a}_i$ satisfying (12). Briefly, we replace $\mathbf{a}_i$ with $\mathbf{a}_i + \delta \mathbf{a}_i$ where $\delta \mathbf{a}_i$ is an incremental update. We approximate (12) by its first order Taylor expansion, simplify, and solve for $\delta \mathbf{a}_i$ using gradient ascent. We then update $\mathbf{a}_i \leftarrow \mathbf{a}_i + \delta \mathbf{a}_i$. This algorithm is similar to that described in [8, 11] for the robust estimation of optical flow.

Estimation of large image motions requires a coarse-to-fine process in which the images are represented with a Gaussian pyramid. The $\mathbf{a}_M$ are updated at a coarse level and then projected to the next finer level where they are used to warp the image at time $t+1$ towards the image at t, thereby incrementally reducing the difference between the images (see [5, 8] for details).

The EM algorithm [16] alternates between solving for the weights, $w_i(\mathbf{x}, \sigma_i)$, given an estimate of the parameters, $\mathbf{a}_1 ... \mathbf{a}_n$ (the Expectation step), and then updating the parameters, $\mathbf{a}_1 ... \mathbf{a}_n$, with the weights held fixed (the Maximization step). It should be noted that although the EM algorithm works well in practice, its validity with finite mixtures of t-distributions remains unclear.

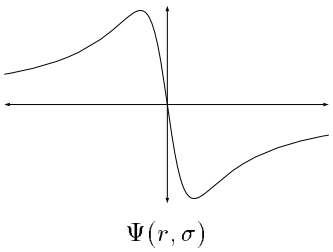

FIG. 7. $\Psi(r, \sigma)$ (the derivative of the log likelihood).

Each model has an associated value of σ_i which determines what residual values are considered to be outliers. A common approach for improving the stability of the estimation process and for avoiding local maxima is to use a deterministic annealing scheme in which the values of σ_i start at a high value and are lowered to the value that gives the desired outlier rejection properties. Here these values are determined empirically. For all the experiments in this paper the value of σ_i began at 45.0 and was lowered by a factor of 0.95 at each iteration of the optimization to a minimum of 10.0. These same values of σ were used for all the models.

The effect of σ on the interaction between the models is interesting to consider. For high values of σ, the likelihood function falls off slowly and hence models tend to "share" the explanation of pixels; that is, the $w_i(\mathbf{x})$ are close to $1/n$. When the residual errors, $\Delta I_{C_i} = I(\mathbf{x}, t) - I_{C_i}(\mathbf{x}, t; \mathbf{a}_i)$, result in likelihoods that are lower than p_O, then the normalization in (11) has the effect of shifting the weight to the outlier layer. At the beginning of the annealing process the high value of σ means that the outlier probability is much smaller than the generative model would suggest; that is, smaller than $p_O = 1/256$. In this case the outlier layer accounts for few if any of the pixels. As σ is annealed the outlier probability monotonically increases towards $1/256$ and more pixels are accounted for by that layer.

5. GENERIC APPEARANCE CHANGE

This section presents examples of generic appearance changes that are common in natural scenes, namely, motion, illumination variations, and specularities.

5.1. Multiple Motions

We begin with an experiment involving multiple motions within a region. Figure 8 shows a person moving behind a plant. We assume that there are two affine motions present and solve for them using the robust mixture formulation. The figure shows the weights for the foreground layer ($w_{M_1}(\mathbf{x})$) and the background layer ($w_{M_2}(\mathbf{x})$) where white indicates a weight near 1.0 and black near 0.0. The outlier layer receives high weight in regions that border occlusion boundaries. This simple model of layers does not account for the appearance/disappearance of image pixels and hence these regions are assigned automatically to the outlier layer.

5.2. Shadows

We next consider a mixture of motion and illumination variation (Figure 9). The appearance variation between Figures 9*a* and *b* includes both global motion and an illu-

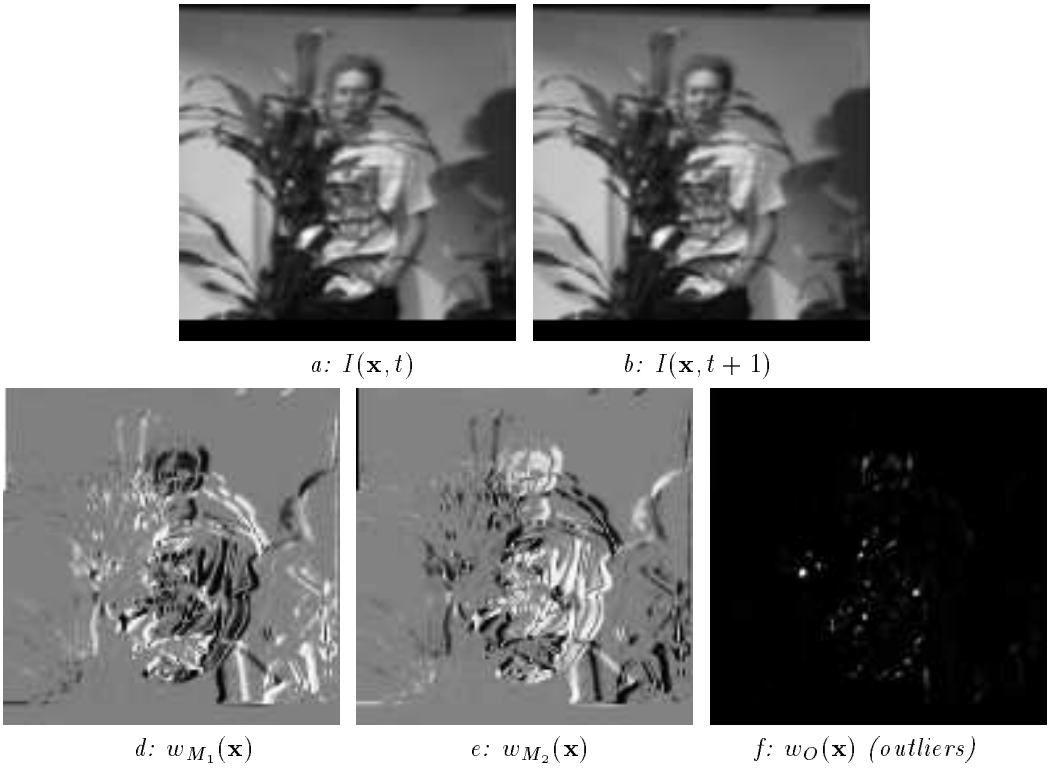

FIG. 8. Multiple Motion Experiment (see text).

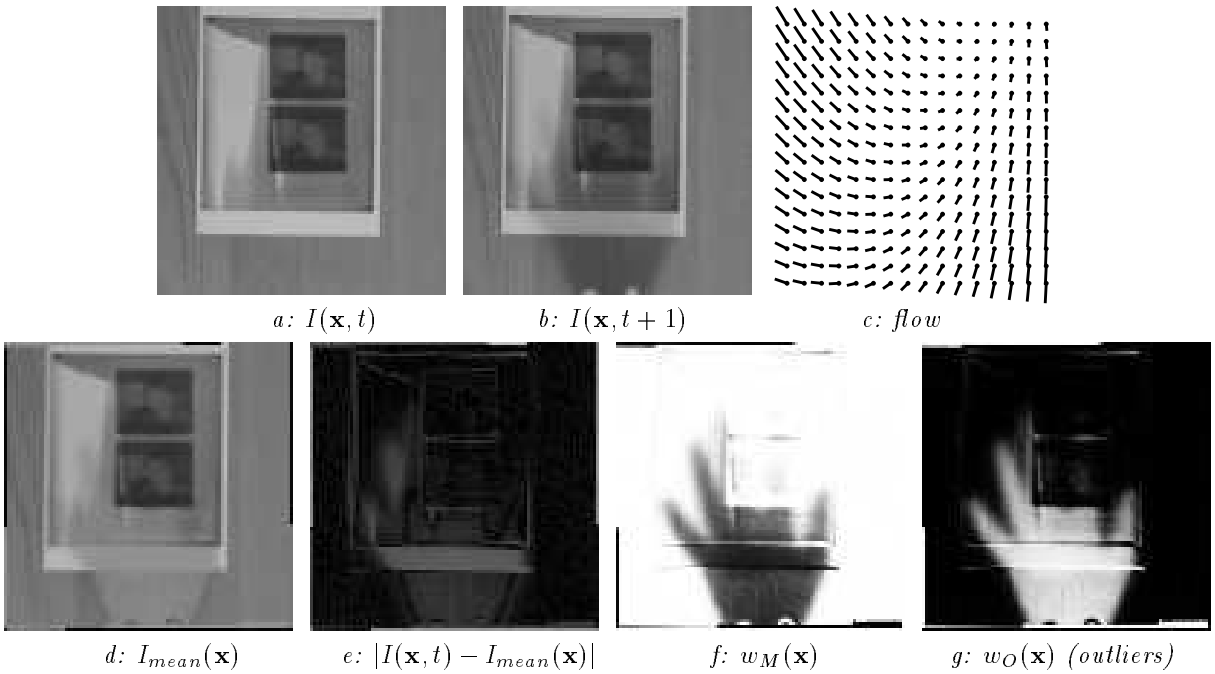

FIG. 9. Illumination Experiment 1 (cast shadow of a hand). Appearance change estimated using a single motion layer with outliers (see text).

mination change caused by the shadow of a hand in frame $t+1$. The true motion field contains expansion due to the motion of the background. Figure 9 shows the results of assuming just a single motion within the region. A three level pyramid is used in the coarse-to-fine estimation and the motion is computed using the affine model presented in Section 3.

The result of the estimation process is a mixture model for the image at time t based on the image at time

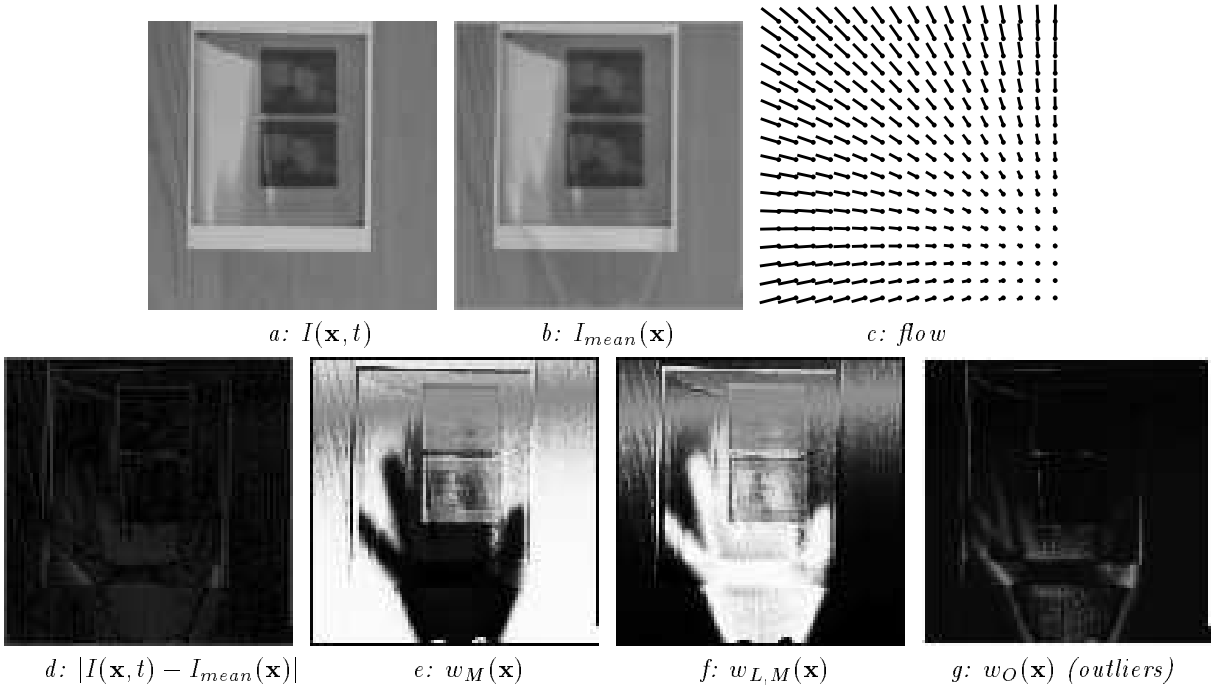

FIG. 10. Illumination Experiment 2. Appearance change estimated using a mixture of motion and illumination change models (see text).

$t+1$ and the appearance change parameters. The mixture model forms a generative model from which we can sample reconstructions of the image at time t. We can also compute the expectation $E[I(\mathbf{x},t) \mid \{\mathbf{a}_j, \sigma_j\}_{j=1}^n]$ as a way of illustrating the generative model. Think of drawing samples (images) from the generative model. At a given pixel $\mathbf{x}$, with probability $w_M(\mathbf{x})$, the intensity sample is $I_M(\mathbf{x},t;\mathbf{a}_M)$ while with probability $w_O(\mathbf{x})$ the intensity sample is drawn from the outlier distribution. The expected value of the outlier model is 128. The expected image, I_{mean}, from the generative model is therefore $I_{mean}(\mathbf{x}) = w_M(\mathbf{x})I_M(\mathbf{x}) + w_O(\mathbf{x})128$. For example, Figure 9*d* shows this "mean reconstruction" image (at time t) that results from the mixture of the deformation (from time $t+1$) and the outlier layer. Note that the uniform expected intensity of the outlier model means that the outlier pixels corresponding to the shadowed hand region appear roughly as a uniform gray.

The absolute difference between this model image and the actual image at time t is shown in Figure 9*e*, and Figure 9 *f, g* shows the weights for the single motion layer and the outlier layer. Note first that, while this robust formulation of the motion-only model is able to detect the correct outliers, the recovered optical flow is inaccurate. The large number of unmodeled intensity changes pull the solution away from the true motion. Outlier maps like this, with large numbers of outliers, provide a clear indication that the model fails to explain the appearance changes caused by the shadow, and that a richer class of models is required.

If, instead, we allow a mixture of the affine motion model (I_M) and the linear illumination model ($I_{L,M}$), we see an improvement in the image motion. We estimate the ownership weights $w_M(\mathbf{x})$ and $w_{L,M}(\mathbf{x})$ that assign pixels to the models and the motion parameters $\mathbf{a}_M$ and illumination parameters $\mathbf{a}_L$ as described in the previous section. Figure 10*b* shows the mean reconstruction of the image at time t, which is now much closer to the actual image in Figure 10*a*. This image is given by $I_{mean}(\mathbf{x}) = w_M(\mathbf{x})I_M(\mathbf{x}) + w_{L,M}(\mathbf{x})I_{L,M}(\mathbf{x}) + w_O(\mathbf{x})128$. Figures 10*d-f* show the reconstruction error, the weight images $w_M(\mathbf{x})$ and $w_{L,M}(\mathbf{x})$, and the outlier image. The motion weights $w_M(\mathbf{x})$ are near 1 (white) when the appearance change is captured by motion alone. Where there is illumination change as well as motion, in the region of the hand, the weights $w_M(\mathbf{x})$ are near 0 (black) and weights $w_{L,M}(\mathbf{x})$ are near 1. The gray regions indicate weights near 0.5 which are equally well described by the two models.

The outlier layer (Figure 10*g*) indicates which pixels had appearance changes that were not well explained by either model. Compared with the motion-only model in Figure 9*f*, the motion+illumination model exhibits far fewer outliers, most of which now occur around the boundary of the shadow. Our simple illumination model only accounts for a linear illumination change while the actual shadow fades non-linearly at the edges of the hand. Thus the boundary regions are not well explained by the illumination change model. To better account for local variations in illumination one could replace the linear model L with a regularized model of the illumination variation (see [43] for regularization in a mixture-model framework).

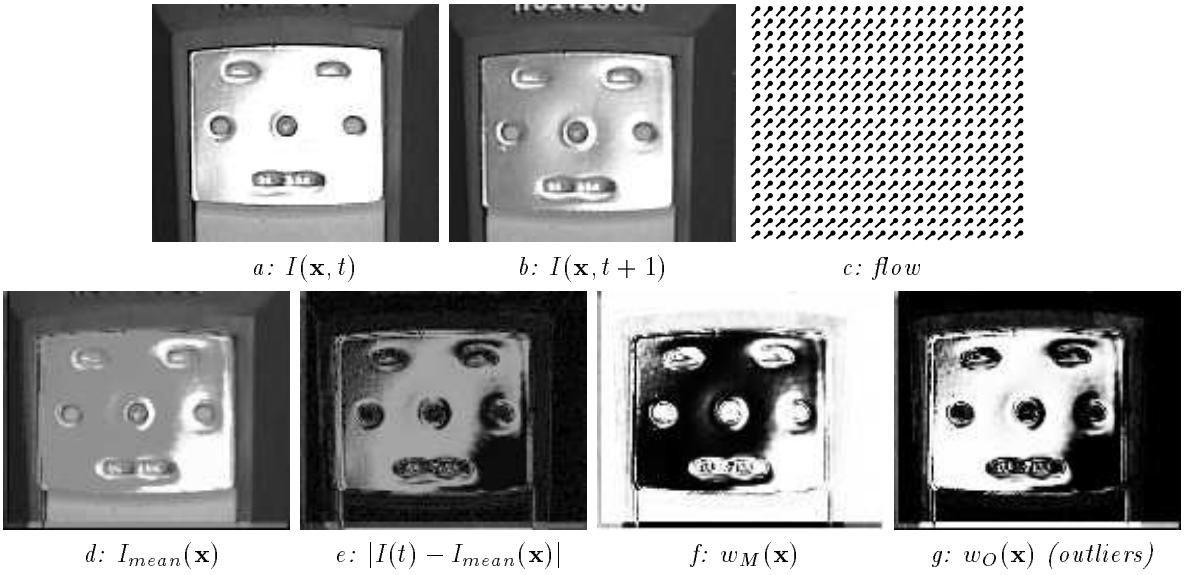

FIG. 11. Specularity Experiment 1 (a moving stapler). Appearance change estimated using a single motion layer with outliers (see text).

Finally, note that there is a significant difference between the flow fields computed using these two different models as shown in Figures 9*c* and Figure 10*c*. The motion is Figure 10*c* is qualitatively correct. Explicitly accounting for the illumination change thus results in a more accurate representation of the true motion.

5.3. Specularities

Consider the example in Figure 11 in which a stapler with a prominent specularity on the metal plate is moved. We first apply a single affine motion model with outliers to explain the appearance change. A four level pyramid was employed to capture the large motion between frames; other parameters remained unchanged. The recovered motion is relatively accurate despite the fact that very few pixels were actually used in computing it. Examining the motion layer weights, $w_M(\mathbf{x})$ in Figure 11*f*, we see that the motion layer receives high weight in the uniform regions which provide few constraints on the motion. The outlier layer, by comparison captures the majority of the metal plate where the specularity occurs. As above, this indicates that the structure of the appearance change in this region could not be explained by the single motion model.

We next model this situation using a mixture of motion (I_M) and specularity (I_S) models. The simplified model of specularities assumes that some regions of the image at time t can be modeled as a warp of the image at time $t+1$ while others are best modeled as a linear brightness function.

The estimated flow field is shown in Figure 12*c*. The mean reconstructed image, computed from the mixture of the motion and the linear brightness models, is shown in Figure 12*b*; this is given by $I_{mean}(\mathbf{x}) = w_M(\mathbf{x})I_M(\mathbf{x}) + w_S(\mathbf{x})I_S(\mathbf{x}) + w_O(\mathbf{x})128$. The reconstruction error is shown in Figure 12*d*. The ownership weights for the two model components, along with the weights for the outlier layer are shown in Figures 12*e,f,g*. Note how the weights in Figure 12*e* are near zero for the motion model where the specularity changes significantly. The weights also show that the outlier layer (Figure 12*g*) no longer accounts for the majority of the specularity. The region of specularity in the lower right corner of the metal plate is similar in both frames and hence is "shared" by both models.

5.4. Combining Models of Appearance Change

We consider a final example of appearance change that combines all the generic models. Figure 13 shows two frames from a sequence in which a pair of scissors moves rigidly and casts a shadow on a stationary, roughly planar, surface. The change in orientation of the scissors with respect to the light source causes a significant specular reflection. Four appearance models, plus the outlier model, are required to explain the change between this pair of images. The parameters were the same as in all the other experiments and a three level pyramid was used for motion estimation.

The ownership weights corresponding to each of the models are shown in Figure 13 and include the motion of the background ($w_{M_1}(\mathbf{x})$), the illumination change caused by the shadow cast on background ($w_{L,M_1}(\mathbf{x})$), the motion of the scissors ($w_{M_2}(\mathbf{x})$), and the specular reflection ($w_S(\mathbf{x})$). Difference images help illustrate which parts of the image are accounted for by each of the models. Notice that the motion model, M_1, accounts for much of the background but the area where the scissors cast a shadow has lower probability of being explained by that model. This same region is accounted for by the illumination model as can be seen in the weights $w_{L,M_1}(\mathbf{x})$. Notice that in dark regions of the image the illumination model can account for

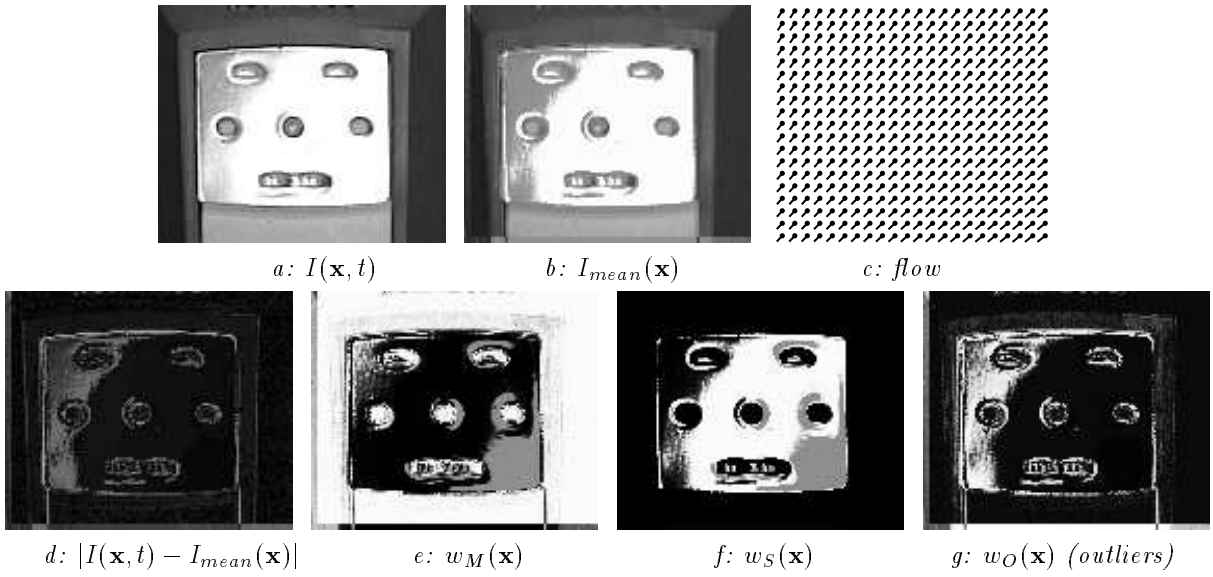

FIG. 12. Specularity Experiment 2. Appearance changed accounted for using a mixture of motion and specularity models (see text).

the appearance nearly as well as the simple motion model; this is due to the multiplicative nature of the illumination term.

The motion, M_2, accounts for motion of the hand and portions of the scissors. The blades of the scissors exhibit a large change due specular reflection and this is accounted for by the specularity model as can be seen in the weights $w_S(\mathbf{x})$. Additionally, the outlier layer, $w_O(\mathbf{x})$, largely accounts for the regions around the edges of the scissors and hand which correspond to regions of occlusion or disocclusion for which we do not have a generative model. Finally, the expected image, $I_{mean}(\mathbf{x})$, provides a reasonable reconstruction of the image $I(\mathbf{x},t)$; the difference image corresponding to the mean illustrates that more pixels are well modeled with the mixture model than with any of the individual models alone.

This example raises a number of interesting issues. For this experiment, we manually selected the number and type of models to employ. Ideally we would like to determine the models automatically but to do so will require us to model the prior probabilities of observing the different types of appearance change in typical image sequences. Appropriate prior models will be required to choose among competing hypotheses. As the number of models increases so does the danger of over parameterization and computational instability. Here a notion of spatial locality of the causes (modeled as a prior probability) may be useful (cf. [43, 44]).

6. EXPERIMENTS: ICONIC CHANGE

Unlike the generic illumination and reflection events in the previous section, here we consider image appearance changes that are specific to particular objects or scenes. Following previous work on eigen-based representations of image structure and image motion [7, 9, 13, 18, 22, 23, 41], we learn parameterized models of motion and iconic structure from examples. We then use these in our mixture model framework to explain motion and iconic change in human mouths.

6.1. Learned Iconic Model

To capture the iconic change in domain-specific cases, such as the mouths in Figure 14, we construct a low-dimensional model of the p images in the training set using principal component analysis (PCA). For each $s = n \times m$ training image we construct a 1D column vector by scanning the pixels in the standard lexicographic order. Each 1D vector becomes a column in an $s \times p$ matrix B. We use singular value decomposition to decompose B as

$$B = A\Sigma_a V_a^T. \tag{15}$$

Here, A is an orthogonal matrix of size $s \times p$, the columns of which represent the principal directions in the training set. Σ_a is a diagonal matrix with singular values $\lambda_1, \lambda_2, \ldots, \lambda_p$ sorted in decreasing order along the diagonal.

Because there is a significant amount of redundancy in the training sequence, the effective rank of B will be much smaller than p. Accordingly, the first few columns of A provide a basis that spans the majority of the structure in B. Here we express the i^{th} column of A as a 2D basis image, $A_i(\mathbf{x})$, so that we can approximate images like those in the training set as

$$I_P(\mathbf{x},t;\mathbf{a}) = \sum_{i=1}^{q} a_i A_i(\mathbf{x}), \tag{16}$$

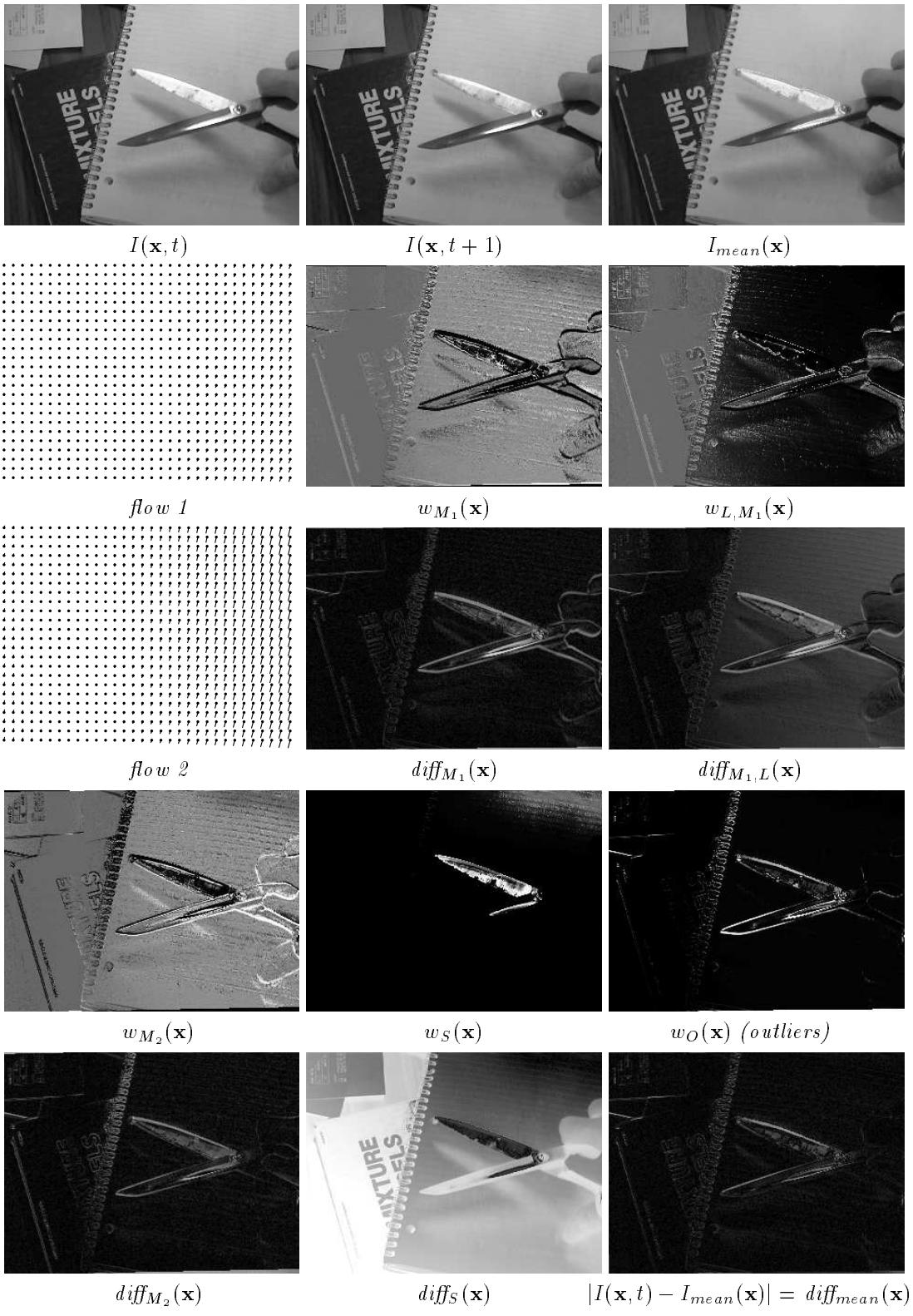

FIG. 13. Combination of two motions, a shadow, and a specular reflection; see text.

FIG. 14. Example frames from training sequences of facial expressions (anger, joy, sadness).

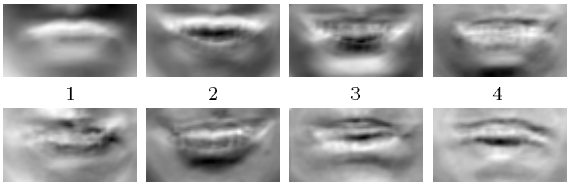

FIG. 15. First eight basis appearance images, $A_1(\mathbf{x}), \ldots, A_8(\mathbf{x})$, for the facial expression experiment.

where $\mathbf{a} = [a_1, \ldots, a_q]$ is the vector of scalar values to be estimated and $q < p$.

Figure 14 shows samples of mouth images taken from a training set of approximately 500 images. The training set included image sequences of different subjects performing the facial expressions "joy," "anger," and "sadness." The faces of each subject were stabilized with respect to the first frame in the sequence using a planar motion model [12]. The mouth regions were extracted from the stabilized sequences and PCA was performed. The first 11 basis images account for 85% of the variance in the training data and the first eight of these are shown in Figure 15.

6.2. Learned Deformations

We learn a domain-specific model for the deformation component of the appearance change in much the same way using PCA [13]. We first compute image motion for each training sequence using the brightness constancy assumption and a robust optical flow algorithm [8]. The training set consists of a set of p optical flow fields. For images with $s = n \times m$ pixels, each flow field contains $2s$ quantities (i.e., the horizontal and vertical flow components at each pixel). For each flow field we place the $2s$ values into a column vector by scanning $u(\mathbf{x})$ and then $v(\mathbf{x})$ in lexicographic order. The resulting p vectors become the columns of a $2s \times p$ matrix F.

As above we use PCA to decompose F as $F = M \Sigma_m V_m^T$. Flow fields like those in the training set can then be approximated as

$$\mathbf{u}(\mathbf{x}; \mathbf{m}) = \sum_{j=1}^{k} m_j \, M_j(\mathbf{x}),$$

where $k < p$, and $M_j(\mathbf{x})$ denotes the j^{th} column of M interpreted as a 2D vector field. Note that this model is conceptually equivalent to the affine models used above except that it is tailored to a domain-specific class of motions.

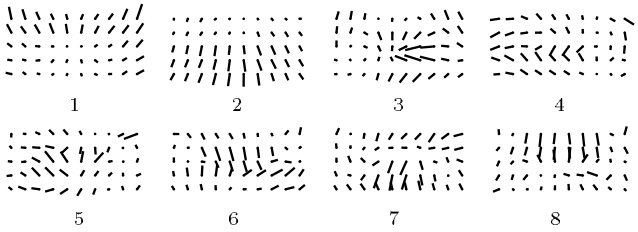

FIG. 16. First eight basis flow fields, $M_1(\mathbf{x}), \ldots, M_8(\mathbf{x})$ for the facial expression mouth motion.

Figure 16 shows the first eight basis flow fields recovered for this training set. The first 11 basis flow fields account for 85% of the variance in the training set.

6.3. Mixture of Motion and Iconic Change

We model appearance change of a mouth as a mixture of the learned motion and iconic models. We performed a number of experiments with image sequences of subjects who were not present in the training set. In our experiments we used 11 basis vectors for both motion and iconic models. We estimated the parameters for deformation $\mathbf{a}_M$, iconic change $\mathbf{a}_P$, the ownership weights, $w_M(\mathbf{x})$ and $w_P(\mathbf{x})$, and the outlier weights between each consecutive pair of frames using a four-level pyramid and the EM-algorithm as described earlier.

Figure 17 shows two consecutive frames from a smiling sequence; notice the appearance of teeth between frames. The motion model, $I_M(\mathbf{x}, t; \mathbf{a}_M)$, captures the deformation around the mouth but cannot account for the appearance of teeth. The recovered flow field is shown in Figure 17*c* and one can see the expansion of the mouth. The iconic model, I_P, on the other hand, does a reasonable job of recovering an approximate representation of the image at time t (Figure 17*d*). The iconic model however does not capture the brightness structure of the lips in detail. This behavior is typical; the iconic model is an approximation to the brightness structure so, if the appearance change can be described as a smooth deformation, then the motion model will likely do a better job of explaining this structure.

The behavior of the mixture model can be seen in the weights (Figures 17*i* and 17*j*). The weights for the motion model, $w_M(\mathbf{x})$, are near zero in the region of the teeth, near one around the high contrast boarder of the lips, and near 0.5 in the untextured skin region which is also well modeled by the iconic approximation I_P.

Figure 17*g* is the expected image given the generative model. Note how this image resembles the original image in Figure 17*a*. Also notice that the iconic model fills in around the edges of the stabilized image where no information was available for warping the image.

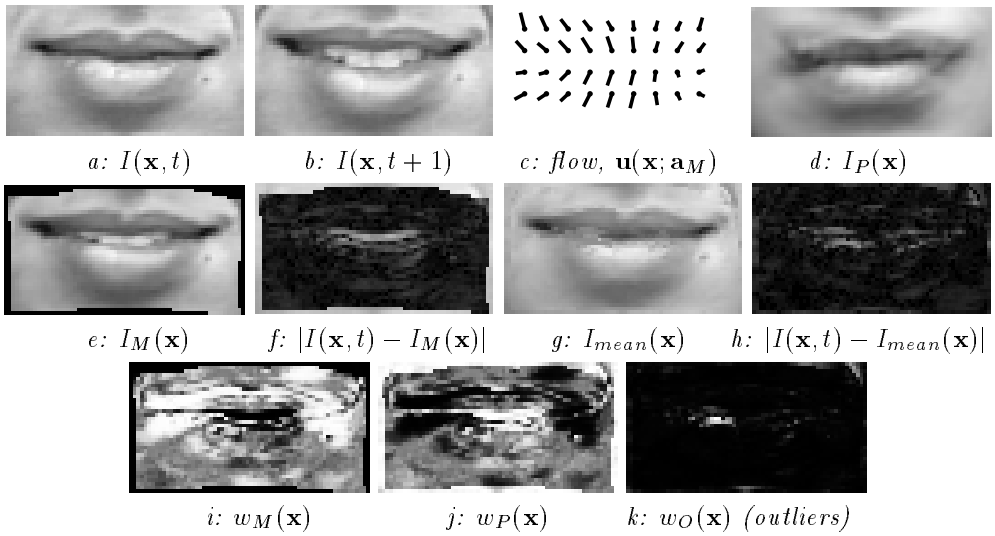

FIG. 17. Facial Expression Experiment. Appearance changed modeled as a mixture of motion and iconic change (see text).

Not all the changes in the image could be accounted for by the two models. There is a change on the lower lip that is due to specular reflection. This specularity was apparently not captured in the learned iconic model and since no specularity model was included here, those pixels are best explained as outliers (Figure 17*k*).

7. FUTURE DIRECTIONS

A research issue that warrants further work is the use of priors on the collection of models that would enable one to prefer some explanations over others. Without the use of priors, a mixture model with several sources of appearance change may produce several equally likely explanations. The probabilistic formulation here should facilitate such an approach.

As in Section 5.4, we may expect more than one instance of each type of appearance change within an image region. In this case we will need to estimate the number of instances of each appearance model that are required. There has been recent work on this topic in the area of multiple motion estimation [2, 28, 44].

A related issue is the use of spatial smoothness in the modeling of appearance change. In place of the parameterized models we might substitute regularized models of appearance change with priors on their spatial smoothness. In a mixture model framework for motion estimation, Weiss [43, 44] has shown how to incorporate regularized models and smoothness priors on the ownership weights.

8. CONCLUSIONS

Appearance changes in image sequences result from a complex combination of events and processes, including motion, illumination variations, specularities, changes in material properties, occlusions, and disocclusions. In this paper we propose a robust statistical framework that models these variations as a probabilistic mixture of causes. To illustrate these ideas, we have proposed some simple generative models.

Unlike previous work, the approach allows us to pull apart, or factor, image appearance changes into different causes and to locate where in the image these changes occur. Moreover, multiple, competing, appearance changes can occur in a single image region. We have implemented and tested the method on a limited suite of image sequences with different types of appearance change.

One way to view this work is as a generalization of current work in the field of motion estimation to richer models of appearance change that allow one to relax the brightness constancy assumption. We expect that more complex models of illumination variation and iconic change can be accommodated by the framework and we feel that it presents a promising direction for research in image sequence analysis.

REFERENCES

1. P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision*, 2:283–310, 1989.

2. S. Ayer and H. Sawhney. Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding. In *Fifth International Conference on Computer Vision*, pages 777–784, Boston, MA, 1995.

3. A. Bab-Hadiashar and D. Suter. Robust optical flow computation. *Internation Journal of Computer Vision*, 29(1):59–77, 1998.

4. J. L. Barron, D. J. Fleet, and S. Beauchemin. Performance of optical flow techniques. *Internation Journal of Computer Vision*, 12(1):43–77, 1994.

5. J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In G. Sandini, editor, *Proc. of Second European Conference on Computer Vision, ECCV-92*, volume 588 of *LNCS-Series*, pages 237–252. Springer-Verlag, May 1992.

6. D. Beymer. Feature correspondence by interleaving shape and texture computations. In *Proc. Computer Vision and Pattern Recognition, CVPR-96*, pages 921–928, San Francisco, June 1996.
7. D. Beymer and T. Poggio. Image representations for visual learning. *Science*, 272:1905–1909, June 1996.
8. M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, January 1996.
9. M. J. Black, D. J. Fleet, and Y. Yacoob. A framework for modeling appearance change in image sequences. In *Proceedings of the International Conference on Computer Vision*, pages 660–667, Mumbai, India, January 1998.
10. M. J. Black and A. D. Jepson. Estimating optical flow in segmented images using variable-order parametric models with local deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10):972–986, October 1996.
11. M. J. Black and A. D. Jepson. EigenTracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1):63–84, 1998.
12. M. J. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25(1):23–48, 1997.
13. M. J. Black, Y. Yacoob, A. D. Jepson, and D. J. Fleet. Learning parameterized models of image motion. In *Proc. Computer Vision and Pattern Recognition, CVPR-97*, pages 561–567, Puerto Rico, June 1997.
14. M. Bober and J. Kittler. Robust motion analysis. In *Computer Vision and Pattern Recognition, CVPR-94*, pages 947–952, Seattle, WA, 1994.
15. T. Darrell and A. Pentland. Cooperative robust estimation using layers of support. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):474–487, May 1995.
16. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society Series B*, pages 1–38, 1977.
17. T. Ezzat and T. Poggio. Facial analysis and synthesis using image-based models. In *International Conference on Automatic Face and Gesture Recognition*, pages 116–121, Killington, Vermont, 1996.
18. D. J. Fleet, M. J. Black, and A. D. Jepson. Motion feature detection using steerable flow fields. In *Proc. Computer Vision and Pattern Recognition, CVPR-98*, pages 274–281, Santa Barbara, CA, June 1998.
19. D.J. Fleet and A.D. Jepson. Computation of component image velocity from local phase information. *International Journal of Computer Vision*, 5:77–104, 1990.
20. D.J. Fleet and A.D. Jepson. Stability of phase information. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15:1253–1268, 1993.
21. J. J. Gibson. *The Ecological Approach to Visual Perception.* Houghton Mifflin, Boston, MA, 1979.
22. G. D. Hager and P. N. Belhumeur. Real-time tracking of image regions with changes in geometry and illumination. In *Proc. Computer Vision and Pattern Recognition, CVPR-96*, pages 403–410, San Francisco, June 1996.
23. P. Hallinan. *A deformable model for the recognition of human faces under arbitrary illumination.* PhD thesis, Harvard University, Cambridge, MA, August 1995.
24. F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions.* John Wiley and Sons, New York, NY, 1986.
25. M. Irani, B. Rousso, and S. Peleg. Detecting and tracking multiple moving objects using temporal integration. In G. Sandini, editor, *Proc. of Second European Conference on Computer Vision, ECCV-92*, volume 588 of *LNCS-Series*, pages 282–287. Springer-Verlag, May 1992.
26. A. Jepson and M. J. Black. Mixture models for optical flow computation. In Ingmer Cox, Pierre Hansen, and Bela Julesz, editors, Partitioning Data Sets: With Applications to Psychology, Vision and Target Tracking, pages 271–286, DIMACS Workshop, April 1993. AMS Pub., Providence, RI.
27. S. X. Ju. *Estimating image motion in layers: The Skin and Bones model.* PhD thesis, University of Toronto, 1999.
28. S. X. Ju, M. J. Black, and A. D. Jepson. Skin and bones: Multilayer, locally affine, optical flow and regularization with transparency. In *Proc. Computer Vision and Pattern Recognition, CVPR-96*, pages 307–314, San Francisco, June 1996.
29. K.L. Lange, R.J.A. Little, and J.M.G.taylor. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896, 1989.
30. G.J. McLachlan and K.E. Basford. *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker Inc., N.Y., 1988.
31. N. Mukawa. Estimation of shape, reflection coefficients and illuminant direction from image sequences. In *Proceedings of the International Conference on Computer Vision*, pages 507–512, Osaka, Japan, 1990.
32. D. W. Murray and B. F. Buxton. *Experiements in the Machine Interpretation of Visual Motion.* MIT Press, Cambridge, MA, 1990.
33. C. Nastar, B. Moghaddam, and A. Pentland. Generalized image matching: Statistical learning of physically-based deformations. In B. Buxton and R. Cipolla, editors, *European Conf. on Computer Vision, ECCV-96*, volume 1064 of *LNCS-Series*, pages 589–598, Cambridge, UK, 1996. Springer-Verlag.
34. S. Negahdaripour. Revised definition of optical flow: Integration of radiometric and geometric cues for dynamic scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(9):961–979, September 1998.
35. P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection.* John Wiley & Sons, New York, 1987.
36. H. Sawhney, S. Ayer, and M. Gorkani. Model-based 2D&3D dominant motion estimation for mosaicing and video representation. In *Fifth International Conference on Computer Vision*, pages 583–590, Boston, MA, 1995.
37. H. S. Sawhney and S. Ayer. Compact representations of videos through dominant and multiple motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):814–831, 1996.
38. M. Turk and A. Pentland. Face recognition using eigenfaces. In *Proc. Computer Vision and Pattern Recognition, CVPR-91*, pages 586–591, Maui, June 1991.
39. A. Verri and T. Poggio. Motion field and optical flow: Qualitative properties. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 490–498, 1989.
40. T. Vetter. Learning novel views to a single face image. In *International Conference on Automatic Face and Gesture Recognition*, pages 22–27, Killington, Vermont, 1996.
41. T. Vetter, M. J. Jones, and T. Poggio. A bootstrapping algorithm for learning linear models of object classes. In *Proceedings of IEEE conference on Computer Vision and Pattern Recognition*, pages 40–47, Puerto Rico, June 1997.
42. J. Y. A. Wang and E. H. Adelson. Representing moving images with layers. *IEEE Transactions on Image Processing*, 3(5):625–638, September 1994.

43. Y. Weiss. Smoothness in layers: Motion segmentation using nonparametric mixture estimation. In *Proceedings of IEEE conference on Computer Vision and Pattern Recognition*, pages 520–526, Puerto Rico, June 1997.

44. Y. Weiss and E. H. Adelson. A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models. In *Proc. Computer Vision and Pattern Recognition, CVPR-96*, pages 321–326, San Francisco, June 1996.

45. A. Yuille, T. Yang, and D. Geiger. Robust statistics, transparency and correspondence. Technical Report 90–7, Harvard Robotics Laboratory.

ACKNOWLEDGMENTS

We thank Allan Jepson and Yair Weiss for their comments, Jeffrey Cohn for use of the facial expression sequences, and David Tyler for pointing out the relationship between the likelihood function we used and Student's t-distribution along with properties of the EM algorithm with mixtures of t-distributions. DJF is grateful for financial support from NSERC Canada, the Xerox Corporation, and an Alfred P. Sloan research fellowship. We also thank the anonymous reviewers for their thorough comments on this manuscript.