A Framework for Modeling Appearance Change in Image Sequences

Michael J. Black* **David J. Fleet†** **Yaser Yacoob‡**

* *Xerox Palo Alto Research Center,* 3333 Coyote Hill Road, Palo Alto, CA 94304
† *Dept. of Computing & Information Science,* Queen's University, Kingston, Ont. K7L 3N6
‡ *Computer Vision Laboratory,* University of Maryland, College Park, MD 20742
black@parc.xerox.com fleet@qucis.queensu.ca yaser@cs.umd.edu

Abstract

Image "appearance" may change over time due to a variety of causes such as 1) object or camera motion; 2) generic photometric events including variations in illumination (e.g. shadows) and specular reflections; and 3) "iconic changes" which are specific to the objects being viewed and include complex occlusion events and changes in the material properties of the objects. We propose a general framework for representing and recovering these "appearance changes" in an image sequence as a "mixture" of different causes. The approach generalizes previous work on optical flow to provide a richer description of image events and more reliable estimates of image motion.

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" ([9], 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 (eg., oxidation), objects can change state (eg., evaporation, dissolving), or objects can undergo radical changes in structure (eg., exploding, tearing, rupturing, boiling). In this paper we formulate a general framework for representing appearance changes such as these. 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.

We propose 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. An image at time $t+1$ can be explained by warping the image at time t using this image motion.

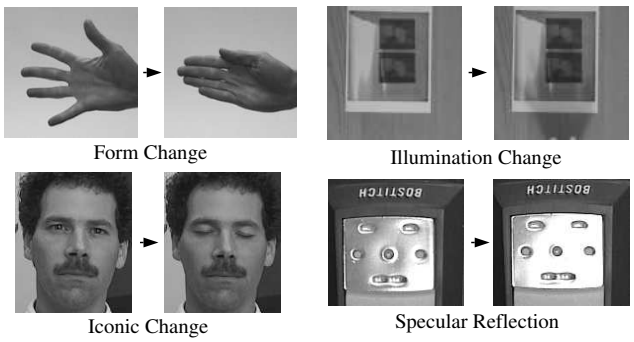


Figure 1: Examples of appearance change.

Illumination variations (Figure 1, upper right), may be global, occurring throughout the entire image due to changes in the illuminant, or local as the result of shadowing. Here we model illumination change as a smooth function that amplifies/attenuates image contrast. By comparison, specular reflections (Figure 1, lower right) are typically local 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 [6]. 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 the iconic structure for particular objects.

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 [14] 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

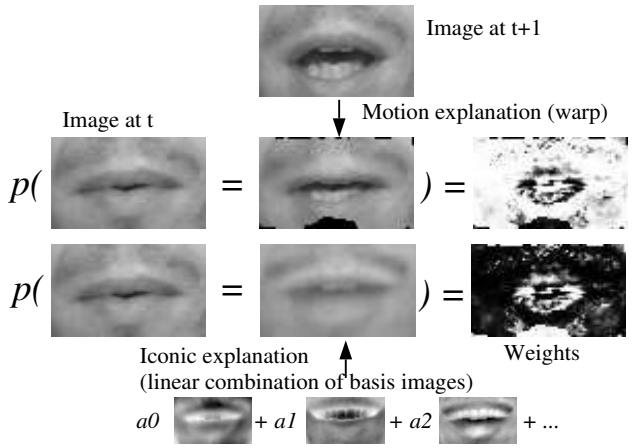


Figure 2: Object specific appearance change between a images at times t and $t+1$ is modeled as a mixture of motion and iconic change (see text).

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). We use the EM-algorithm [14] to iteratively compute maximum likelihood estimates for the deformation and iconic model parameters as well as the posterior 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.

Below we describe this mixture-model formulation and some simple 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 the brightness constancy assumption. The assumption states that the image brightness $I(\vec{x}, t)$ at a pixel $\vec{x} = [x, y]$ and time t is a simple deformation of the image at time $t+1$:

$$I(\vec{x}, t) = I(\vec{x} - \vec{u}(\vec{x}), t+1), \quad (1)$$

where $\vec{u}(\vec{x}) = (u(\vec{x}), v(\vec{x}))$ represents the horizontal and vertical displacement of the pixel. This model is applied in image patches using regression techniques or locally using regularization techniques. The recovered image motion can be used to "warp" one image towards the other.

While optical flow is an important type of image appearance change it is well known that it does not capture all the important image events. One focus of recent work in motion estimation is to make it "robust" in the presence of these unmodeled changes in appearance (ie. violations of the brightness constancy assumption) [3]. The approach here is quite different in that we explicitly model 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 image motion of face regions such as the mouth [5]. 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 [15] extended the brightness constancy assumption to allow illumination changes that are a smoothly varying function of the image brightness. In a related paper, Negahdaripour and Yu [17] proposed a general linear brightness constraint

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

where $m(\vec{x}, t)$ and $c(\vec{x}, t)$ are used to account 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.* [16]. Treating the image as 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, as does the general constraint in (2).

A number of authors have proposed more general linear models of image brightness [2, 10, 11, 18]. For example, Hager and Belhumeur [10] use principal component analysis (PCA) to find a set of orthogonal basis images, $\{B_j(\vec{x})\}$, that spans the ensemble of images of an object under a wide variety of illuminant directions. They constrain deviations from brightness constancy to lie in the subspace of illumination variations, giving the constraint

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

where $\vec{u}(\vec{x}; \vec{m})$ is a parameterized (affine) model of image motion. The authors estimate the motion parameters $\vec{m} = [m_1, \ldots, m_k]$ and the subspace parameters $b_1 \ldots b_n$. Hallinan [11] proposed a model that included both a model of illumination variation and a learned deformation model (EigenWarps). These approaches are also related to the

eigentracking work of Black and Jepson [4] in which subspace constraints are used to help account for iconic changes in appearance while an object is being tracked.

In [6] we extended these general linear brightness models by allowing spatially varying explanations for pixels

$$I(\vec{x},t) \quad = \quad w_0(\vec{x})I_{\text{Motion}}(\vec{x}) + w_1(\vec{x})I_{\text{Iconic}}(\vec{x}).$$

The terms $w_i(\vec{x})$ are spatially varying "weights" between zero and one that indicate the extent to which a pixel can be explained, or modeled, by the individual causes.

The approach presented here casts the above models in a probabilistic mixture model framework. The models above can be thought of as different generative models that can be used to construct or explain an image; in a sense, they embody different "constancy" assumptions. Unlike the approaches above, however, the mixture model framework factors appearance change into multiple causes and performs a soft assignment of pixels to the different models

3 Mixture Model of Appearance Change

Mixture models [14] have been used previously in motion analysis for recovering multiple motions within an image region [1, 13, 19]. The basic goals are to estimate the parameters of a set of models given data generated by multiple causes and to assign data to the estimated models. Here we use this idea to account for co-occurring types of appearance change. Within some image region R we may expect a variety of appearance changes to take place between frames.

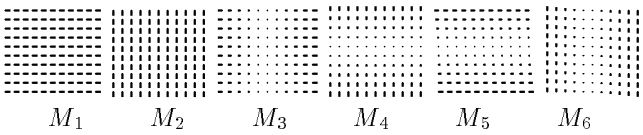
In particular, we assume that a pixel $I(\vec{x},t)$ at location $\vec{x} \in R$ and time t is generated, or explained, by one of n causes I_{C_i}, $i = 1,\ldots,n$. The causes, $I_{C_i}(\vec{x},t;\vec{\alpha}_i)$, can be thought of as overlapping "layers" and are simply images that are generated given some parameters $\vec{\alpha}_i$. We will consider four causes below namely: motion (I_{C_1}), illumination variations (I_{C_2}), specular reflections (I_{C_3}), and iconic changes (I_{C_4}). Given these causes, the probability of observing the image $I(\vec{x},t)$ is then

$$p(I(\vec{x},t)|\vec{\alpha}_1,\ldots,\vec{\alpha}_n,\sigma_1,\ldots,\sigma_n) = \sum_{i=1}^{n} \pi_i p_i(I(\vec{x},t)|\vec{\alpha}_i,\sigma_i)$$

where the π_i are mixture proportions [14] which we take to be $1/n$ for each i indicating that each cause is equally likely. The $\vec{\alpha}_i$ are parameters of model I_{C_i} for which we seek a maximum likelihood estimate and the σ_i are scale parameters. Here we make the very crude assumption that the causes are independent.

In contrast to the traditional mixture of Gaussians formulation, the component probabilities, $p_i(I(\vec{x},t)|\vec{\alpha}_i,\sigma_i)$, are defined to be

$$p_i(I(\vec{x},t)|\vec{\alpha}_i,\sigma_i) = \frac{2\sigma^3}{\pi(\sigma^2 + (I(\vec{x},t) - I_{C_i}(\vec{x},t;\vec{\alpha}_i))^2)^2}.$$



M_1 M_2 M_3 M_4 M_5 M_6

Figure 3: Affine flow basis set.

L_1 L_2 L_3

Figure 4: Linear illumination-change basis images.

This is a robust likelihood function (Figure 5) the tails of which fall off more sharply than those of a normal distribution. This reflects our expectation that the residuals $I(\vec{x},t) - I_{C_i}(\vec{x},t;\vec{\alpha}_i)$ contain outliers [12].

Below we define the individual sources of appearance change.

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

$$I_{C_1}(\vec{x},t;\vec{m}) = I(\vec{x} - \vec{u}(\vec{x};\vec{m}),\, t+1)\ .$$

This represents the image at time $t+1$ warped by a flow field $\vec{u}(\vec{x};\vec{m})$. 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 $M_j(x)$:

$$\vec{u}(\vec{x};\vec{m}) \ = \ \sum_{j=1}^{k} m_j\, M_j(\vec{x}). \tag{4}$$

where $\vec{\alpha}_1 = \vec{m} = [m_1,\ldots,m_k]$ is the vector of parameters to be estimated. An affine basis set, shown in Figure 3, is used for the experiments in Section 5.

Illumination Variations: Illumination changes may be global as a result of changes in the illuminant, or local as the result of shadows cast by objects in the scene. The mixture formulation allows both of these types of variation to be modeled.

We adopt a simple model of illumination variation

$$I_{C_2}(\vec{x},t;\vec{l}) = L(\vec{x};\vec{l})\, I(\vec{x} - \vec{u}(\vec{x};\vec{m}),\, t+1), \tag{5}$$

which states that the illumination change is a scaled version of the motion-compensated image at time $t+1$. When estimating the parameters $\vec{\alpha}_2 = \vec{l}$ we assume that the motion $\vec{u}$ is known and fixed.

We take $L(\vec{x};\vec{l})$ to be a parametric model, expressed as a weighted sum of basis images. For example, in the case of

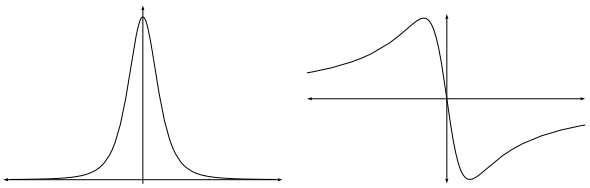


Figure 5: A robust likelihood p_i and ψ (the derivative of the log likelihood).

linear spatial variation, L is given by

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

where (x_c, y_c) is the center of the relevant image region, $\vec{l} = [l_1, l_2, l_3]$ are the model parameters, and $L_i(\vec{x})$ denote the basis images, like those for the linear model in Figure 4.

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:

$$I_{C_3}(\vec{x},t;\vec{s}) \;=\; s_1 + s_2(x-x_c) + s_3(y-y_c) \;=\; \sum_{i=1}^{3} s_i\, S_i(\vec{x})$$

where S_i are the same linear basis images as in Figure 4 and $\vec{\alpha}_3 = \vec{s}$.

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 with the models above, we use a parametric model of iconic change. However, here we learn the appropriate model by constructing a linear, parametric model of 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

$$I_{C_4}(\vec{x},t;\,\vec{a}) \;=\; \sum_{i=1}^{q} a_i A_i(\vec{x}), \tag{6}$$

where $\vec{\alpha}_4 = \vec{a} = [a_1, \ldots, a_q]$ is the vector of scalar values to be estimated.

4 EM-Algorithm

We seek a maximum likelihood estimate of the parameters $\vec{\alpha}_1, \ldots, \vec{\alpha}_n$ and a soft assignment of pixels to models. If the parameters of the models are known, then we can compute the posterior probability, $w_i(\vec{x}, \sigma_i)$, that pixel $\vec{x}$ belongs to cause i. This is given by [14]

$$w_i(\vec{x},\,\sigma_i) \;=\; \frac{p_i(I(\vec{x},t)|\vec{\alpha}_i,\sigma_i)}{\sum_{j=1}^{n} p_j\,(I(\vec{x},t)|\vec{\alpha}_j,\sigma_j)}\;. \tag{7}$$

These ownership weights force every pixel to be explained by some combination of the different causes. As the σ go to zero, the likelihood function approaches a delta function hence, for small values of σ, the weights will tend towards zero or one.

The maximum likelihood estimate [14] of the parameters is defined in terms of these ownership weights and can be shown to satisfy

$$\sum_{\mathbf{x}\in R}\sum_{i=1}^{n} w_i(\vec{x},\sigma_i)\frac{\partial}{\partial\vec{\alpha}_i}\log\, p_i(I(\vec{x},t)|\vec{\alpha}_i,\sigma_i) = 0 \tag{8}$$

where $\partial \log\, p_i(I(\vec{x},t)|\vec{\alpha}_i,\sigma_i)/\partial\vec{\alpha}_i =$

$$\psi(I(\vec{x},t) - I_{C_i}(\vec{x},t;\vec{\alpha}_i),\sigma_i)\frac{\partial}{\partial\vec{\alpha}_i} I_{C_i}(\vec{x},t;\vec{\alpha}_i), \tag{9}$$

and where

$$\psi(r,\sigma) = \frac{-4r}{\sigma^2 + r^2} \tag{10}$$

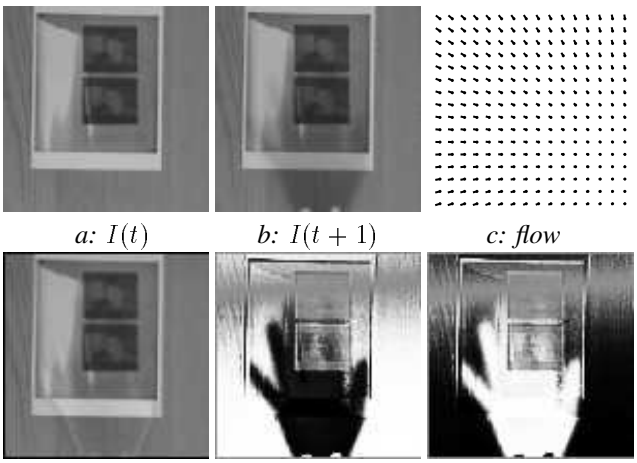
is a robust influence function [12] (Figure 5) that reduces the effect of "outliers" on the maximum likelihood estimate.

In the case of mixtures of Gaussian densities, the parameters can be computed in closed form. In the case of the robust likelihood function we incrementally compute the $\vec{\alpha}_i$ satisfying (8). Briefly, we replace $\vec{\alpha}_i$ with $\vec{\alpha}_i + \delta\vec{\alpha}_i$ where $\delta\vec{\alpha}_i$ is an incremental update. We approximate (8) by its first order Taylor expansion, simplify, and solve for $\delta\vec{\alpha}_i$. We then update $\vec{\alpha}_i \leftarrow \vec{\alpha}_i + \delta\vec{\alpha}_i$ and repeat until convergence.

The EM algorithm alternates between solving for the weights given an estimate of the I_{C_i} (the Expectation step), and then updating the parameters with the weights held fixed (the Maximization step). A continuation method is used to lower the value of σ during the optimization to help avoid local maxima. 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 algorithm is embedded within a coarse-to-fine process that first estimates parameters at a coarse spatial resolution and then updates them at successively finer resolutions.

As in [13] we can add an explicit "outlier layer" with a fixed likelihood

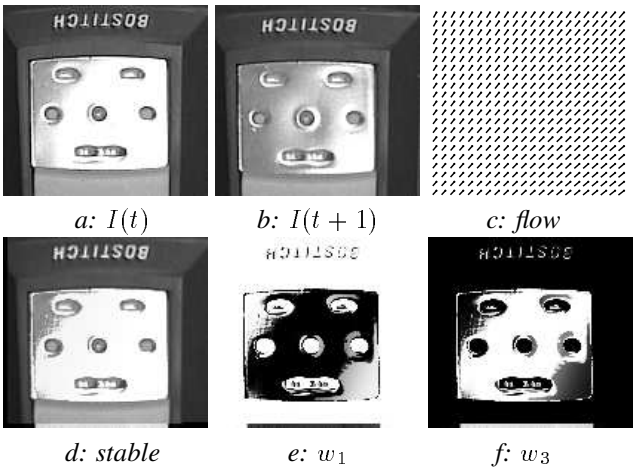
$$p_0 = \frac{2\sigma^3}{\pi(\sigma^2 + (2.5\sigma)^2)^2}.$$



a: $I(t)$ *b:* $I(t+1)$ *c: flow*

d: stable *e:* w_1 *f:* w_2

Figure 6: Illumination Experiment (cast shadow of a hand).



a: $I(t)$ *b:* $I(t+1)$ *c: flow*

d: stable *e:* w_1 *f:* w_3

Figure 7: Specularity Experiment (a moving stapler).

This term is used only in the normalization in Equation (7) which is performed over $i = 0, \ldots, n$. Residual errors greater than 2.5σ will have weights lower than the outlier layer and which will be reduced further by the normalization.

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 Shadows

We first consider a mixture of motion and illumination variation (Figure 6). In this experiment we use a mixture of just two models: the affine motion model (I_{C_1}) and the linear illumination model (I_{C_2}). We estimate the ownership weights $w_1(\vec{x})$ and $w_2(\vec{x})$ that assign pixels to the models and the motion parameters $\vec{\alpha}_1$ and illumination parameters $\vec{\alpha}_2$ as described in the previous section. 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 appearance variation between Figures 6*a* and *b* includes both global motion and an illumination change caused by a shadow of a hand in frame $t+1$. The estimated motion field (Figure 6*c*) contains some expansion as the background surface moved towards the camera. Figures 6*e* and *f* show the weight images $w_1(\vec{x})$ and $w_2(\vec{x})$ in which the shadow region of the hand is clearly visible. The motion weights $w_1(\vec{x})$ are near 1 (white) when the appearance change is captured by motion alone. When there is illumination change as well as motion, the weights $w_1(\vec{x})$ are near 0 (black). The gray regions indicate weights near 0.5 which are equally well described by the two models.

We can produce a "stabilized" image using the weights:

$$I_{Stable}(\vec{x}) = w_1(\vec{x}) I_{C_1}(\vec{x}, t; \vec{\alpha}_1) \; + \; w_2(\vec{x}) I_{C_2}(\vec{x}, t; \vec{\alpha}_2).$$

The stabilized image is shown in Figure 6*d*; note the shadow has been removed and the image is visually similar to $I(\vec{x}, t)$.

The illumination model only accounts for a globally linear illumination change while the actual shadow fades smoothly at the edges of the hand. To account for local variations in illumination one could replace the linear model L with a regularized model of the illumination variation (see [19] for regularization in a mixture-model framework).

5.2 Specularities

Consider the example in Figure 7 in which a stapler with a prominent specularity on the metal plate is moved. We model this situation using a mixture of motion (I_{C_1}) and specularity (I_{C_3}) models. This 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.

A four level pyramid was employed to capture the large motion between frames; other parameters remained unchanged. The estimated flow field is shown in Figure 7*c*. The stabilized image, using motion and the estimated linear brightness model is shown in Figure 7*d*. Note how the weights in Figures 7*e* and *f* are near zero for the motion model where the specularity changes significantly. 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.

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. First we show how parameterized models of image motion and iconic structure can be learned from examples. We then use

Figure 8: Example frames from training sequences of facial expressions (anger, joy, sadness).

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 8, 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 assume that the number of training images, p, is less than the number of pixels, s, and we use singular value decomposition (SVD) to decompose B as

$$B \;=\; A \Sigma_a V_a^T \;. \tag{11}$$

Here, A is an orthogonal matrix of size $s \times s$, the columns of which represent the principal component 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 rank of B will be much smaller than p. Thus if we express the i^{th} column of A as a 2D basis image $A_i(\vec{x})$, then we can approximate images like those in the training set as

$$I_{C_4}(\vec{x}, t; \vec{a}) \;=\; \sum_{i=1}^{q} a_i A_i(\vec{x}) \;, \tag{12}$$

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

Figure 8 shows samples of mouth images taken from a training set of approximately 500 images. The training set included image sequences of a variety 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 [5]. 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 9.

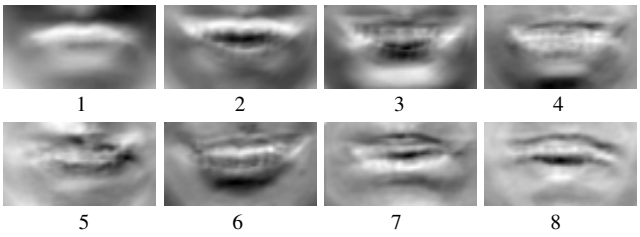


Figure 9: First eight basis appearance images, $A_1(\vec{x}), \ldots, A_8(\vec{x})$, for the facial expression experiment.

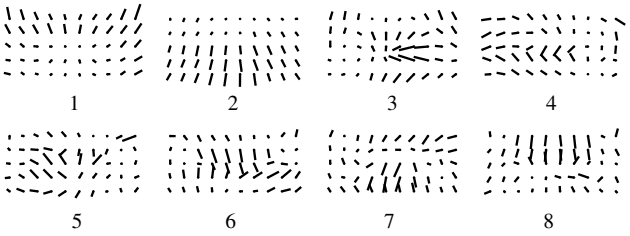


Figure 10: First eight basis flow fields, $M_1(\vec{x}), \ldots, M_8(\vec{x})$ for the facial expression mouth motion.

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 (see [7]). We first compute image motion for each training sequence using the brightness constancy assumption and a robust optical flow algorithm [3]. 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 $u(\vec{x})$ and then $v(\vec{x})$ 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

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

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

Figure 10 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 $\vec{\alpha}_1 = \vec{m}$, iconic change $\vec{\alpha}_4 = \vec{a}$, and the ownership weights, w_1 and w_4 between each consecutive pair of frames using the EM-algorithm as described earlier with a four-level pyramid.

Figure 11 shows two consecutive frames from a smiling sequence; notice the appearance of teeth between frames. The motion model, $(I_{C_1}(\vec{x}, t; \vec{\alpha}_1))$, does a good job of capturing the deformation around the mouth but cannot account for the appearance of teeth. The recovered flow field is shown in Figure 11*d* and one can see the expansion of the mouth. The iconic model, I_{C_4}, on the other hand, does a reasonable job of recovering an approximate representation of the image at time t (Figure 11*c*). 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 11*g* and 11*h*). The weights for the motion model, $w_1(\vec{x})$, 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_{C_4}.

Figure 11*f* is the "stabilized" image using both motion and iconic models $(w_1(\vec{x}) I_{C_1}(\vec{x}, t; \vec{\alpha}_1) + + \, w_4(\vec{x}) I_{C_4}(\vec{x}, t; \vec{\alpha}_4))$ Note how the stablized image resembles the original image in Figure 11*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.

6.4 Discussion

Our motivation in exploring image deformation and iconic change is to address a general theory of appearance change in image sequences. While optical flow characterizes changes that obey brightness constancy, it is only one class of appearance change. Occlusion/disocclusion is another class in which one surface *progressively* covers or reveals another. While optical flow and occlusion/disocclusion have been studied in detail, other types of appearance variations have not. In particular, with complex objects such as mouths, many of the appearance changes between frames are not image deformations that conserve brightness.

One could ask: "Why model image deformation"? While all image changes might be modeled by iconic change this does not reflect the natural properties of objects (their "structural texture" [9]) and how they change. Motion is a natural category of appearance change that is important to model and recover.

One could also ask: "Why model iconic change"? While optical flow methods exist that can ignore many appearance changes that do not obey brightness constancy, it is important to account for, and therefore model, these image changes. Iconic change may be important for recognition. For example, we postulate that the systematic appearance/disappearance of teeth should be a useful cue for aiding speech and expression recognition. In addition, we believe that the temporal change of some objects may not be well modeled as image deformation. For example, bushes and trees blowing in the wind exhibit spatiotemporal texture that might best be modeled as a combination of motion and iconic change.

7 Future Directions

The experiments here have focused on pairs of causes. A natural extension of the work would be to combine all four types of appearance change in a single mixture formulation. Towards this end, a research issue that warrants further work is the use of priors on the collection of models that enable one to prefer some explanations over others.

Additionally, 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 [1, 20].

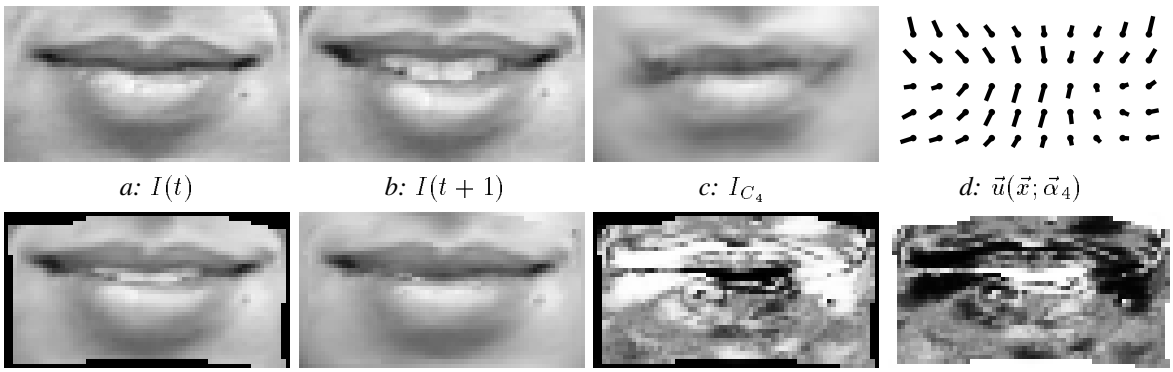
A related issue is the use of spatial smoothness in the modeling of appearance change. In place of the parameterized models we might substitute regularized models of appearance change with priors on their spatial smoothness. In a mixture model framework for motion estimation, Weiss [19, 20] has shown how to incorporate regularized models and smoothness priors on the ownership weights.

Another outstanding research issue concerns the learning and use of domain-specific models when more than one domain of interest exists. When one has several domain-specific models the problems of estimation, indexing, and recognition become much more interesting (cf. [7]).

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 framework that models these variations as a mixture of causes. To illustrate the 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 variety of image sequences with



a: $I(t)$ *b:* $I(t+1)$ *c:* I_{C_4} *d:* $\vec{u}(\vec{x};\vec{\alpha}_4)$

e: I_{C_1} *f:* $w_1 I_{c_1} + w_4 I_{c_4}$ *g:* w_1 *h:* w_4

Figure 11: Facial Expression Experiment.

different types of appearance change.

One way to view this work is as a generalization of current work in the field of motion estimation. The framework presented here is more general than previous approaches which have relaxed 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.

Acknowledgements. We thank Allan Jepson for his comments and Jeffrey Cohn for the facial expression sequences.

References

[1] S. Ayer and H. Sawhney. Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding. *ICCV*, pp. 777–784, 1995.

[2] D. Beymer and T. Poggio. Image representations for visual learning. *Science*, Vol. 272, pp. 1905–1909, June 1996.

[3] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *CVIU*, 63(1):75–104, Jan. 1996.

[4] M. J. Black and A. D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *ECCV*, pp. 329–342, 1996.

[5] M. J. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motions. *ICCV*, pp. 374–381, 1995.

[6] M. J. Black, Y. Yacoob, and D. J. Fleet. Modelling appearance change in image sequences. *3rd Int. Workshop on Visual Form*, Capri, Italy, May 1997.

[7] M. J. Black, Y. Yacoob, A. D. Jepson, and D. J. Fleet. Learning parameterized models of image motion. *CVPR*, pp. 561–567, 1997.

[8] T. Ezzat and T. Poggio. Facial analysis and synthesis using image-based models. *Int. Conf. on Auto. Face and Gesture Recog.*, pp. 116–121, 1996.

[9] J. J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, MA, 1979.

[10] G. D. Hager and P. N. Belhumeur. Real-time tracking of image regions with changes in geometry and illumination. *CVPR*, pp. 403–410, 1996.

[11] P. Hallinan. *A deformable model for the recognition of human faces under arbitrary illumination*. PhD thesis, Harvard Univ., Aug. 1995.

[12] 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.

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

[14] G.J. McLachlan and K.E. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker Inc., N.Y., 1988.

[15] N. Mukawa. Estimation of shape, reflection coefficients and illuminant direction from image sequences. *ICCV*, pp. 507–512, 1990.

[16] C. Nastar, B. Moghaddam, and A. Pentland. Generalized image matching: Statistical learning of physically-based deformations. *ECCV*, pp. 589–598, 1996.

[17] S. Negahdaripour and C. Yu. A generalized brightness change model for computing optical flow. *ICCV*, pp. 2–11, 1993.

[18] T. Vetter, M. J. Jones, and T. Poggio. A bootstrapping algorithm for learning linear models of object classes. *CVPR*, pp. 40–47, 1997.

[19] Y. Weiss. Smoothness in layers: Motion segmentation using nonparametric mixture estimation. *CVPR*, pp. 520–526, 1997.

[20] Y. Weiss and E. H. Adelson. A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models. *CVPR*, pp. 321–326, 1996.