Learning Parameterized Models of Image Motion

Michael J. Black* **Yaser Yacoob†** **Allan D. Jepson‡** **David J. Fleet§**

* Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304
† Computer Vision Laboratory, University of Maryland, College Park, MD 20742
‡ Department of Computer Science, University of Toronto, Toronto, Ontario M5S 1A4
§ Department of Computing and Information Science, Queen's University, Kingston, Ontario K7L 3N6
black@parc.xerox.com, yaser@cs.umd.edu, jepson@vis.toronto.edu, fleet@qucis.queensu.ca

Abstract

A framework for learning parameterized models of optical flow from image sequences is presented. A class of motions is represented by a set of orthogonal basis flow fields that are computed from a training set using principal component analysis. Many complex image motions can be represented by a linear combination of a small number of these basis flows. The learned motion models may be used for optical flow estimation and for model-based recognition. For optical flow estimation we describe a robust, multi-resolution scheme for directly computing the parameters of the learned flow models from image derivatives. As examples we consider learning motion discontinuities, non-rigid motion of human mouths, and articulated human motion.

1 Introduction

Parameterized models of optical flow address the problems of motion *estimation* and motion *explanation*. They aid in estimation by enforcing strong constraints on the spatial variation of the image motion within a region. Because these methods pool hundreds or thousands of motion constraints in a region to estimate a much smaller number of model parameters, they generally provide accurate and stable estimates of optical flow. Likewise, the small number of parameters provides a concise description of the image motion that can be used for explanation or recognition. For example, parameterized flow models have been used to recognize facial expressions from motion [7].

There are two main problems with parameterized motion models. First, many image regions contain multiple image motions because of moving occlusion boundaries, transparency, reflections, or independently moving objects. A great deal of work has been devoted to extending parameterized models to cope with these situations. The second problem is that parameterized models make strong assumptions about the spatial variation of the image motion within a region. Common motion models based on low-order polynomials (e.g. affine motion) have limited applicability to complex natural scenes.

Examples of complex motions include motion discontinuities, non-rigid motion, articulated motion, and motion "texture". It may be impractical to devise and use explicit mathematical models of the motion in these cases. Therefore, here we "learn" models of optical flow from examples. Given a training set of flow fields (see 1), we use principal component analysis (PCA) to learn a set of basis flow fields that can be used to approximate the training data (Fig. 2). Individual flow fields are then represented as a linear combination of the basis flows (Fig. 3). In this paper we apply this approach to motion boundaries, the motion of a human mouth, and the motion of human legs while walking.

To compute optical flow with a learned model we *directly* estimate the coefficients of the linear combination of basis flows from derivatives of image intensity. These coefficients are estimated using a robust, coarse-to-fine, gradient-based algorithm. This provides a flow field that is consistent with the learned model and is optimal under the assumption of brightness constancy. In this way one can estimate complex optical flow fields more quickly and reliably than with conventional techniques. Moreover, if the model provides a good description of the spatiotemporal variation of image intensity, then one can also use the estimated coefficients of the model for subsequent recognition/interpretation of the image motion.

2 Related Work

Much of the recent work on learning parameterized models of image deformation has occurred in the face recognition literature to model the deformations between the faces of different people [3, 9, 11, 13, 18]. Correspondences between different faces were obtained either by hand or by an optical flow method, and were then used to learn a lower-dimensional model. In some cases this involved learning the parameters of a physically-based deformable object [13]. In other cases a basis set of deformation vectors was obtained (e.g., see the work of Hallinan [11] on learning "Eigen-Warps"). These methods have not been applied to the mod-

eling of image motion in natural scenes.

Related work has focused on learning the deformation of curves or parameterized curve models [2, 16]. Sclaroff and Pentland [16] estimated modes of deformation for sillohettes of non-rigid objects. They interpolated a sparse set of correspondences between sillohette boundaries in consecutive frames to produce a basis set of flows, much like those learned in this paper. The basis was then used to warp the original images for synthesis and view interpolation. Unlike our approach, they did not learn the basis flows from optical flow data, and did not use them to estimate image motion.

In addition to optical flow estimation, we are interested in the use of parameterized models for motion-based recognition. Black and Yacoob [7] modeled the motion of a human face and facial features using parameterized flow models (planar, affine, and affine+curvature). They showed how simple models could represent a rich variety of image motions, and how the motion parameters could be used to recognize facial expressions. However, their motion models were hand-coded. In this paper we show how appropriate models of facial feature motion can be learned.

Another application examined below is the learning of motion models for the detection of motion discontinuities. This application is similar to modeling step edges in static scenes by learning a parameterized model from examples of edges [14]. It differs from previous attempts to detect motion discontinuities that applied edge detectors to optical flow, checked for bimodality in local flow estimates, or used energy-based methods [4, 15, 17].

3 Learning Parameterized Flow Models

Learning a parameterized model for a particular class of motions requires that we have a "training set" of flow fields containing representative samples of the class. For relatively simple classes such as motion discontinuities we can generate this training set synthetically. For more complex motions of natural objects we will need to estimate the image motion for training sequences. Since training is done off-line, we can afford to use a computationally expensive robust optical flow algorithm [5].

In either case, the training set from which we learn a model of image motion is a set of p optical flow fields. For images with $s = n \times m$ pixels, each flow field contains $2s$ quantities (i.e., horizontal and vertical elements of the flow at each pixel). For each flow field we place the $2s$ values into a vector of length $2s$ by scanning the horizontal elements of the flow, $u(x, y)$ in standard lexicographic order, followed by the vertical elements, $v(x, y)$. This gives us p vectors that become the columns of a $2s \times p$ matrix F.

Principal Components Analysis (PCA) of F can then be used to compute a low-dimensional model for the spatial structure of the flow fields. Toward this end, the Singular

Figure 1: Discontinuity training set. Left: model for generating synthetic flow fields. Right: samples from the training set. The horizontal component is shown above the vertical component. Black denotes pixels moving left or up or u and v respectively. White denotes motion right or down.

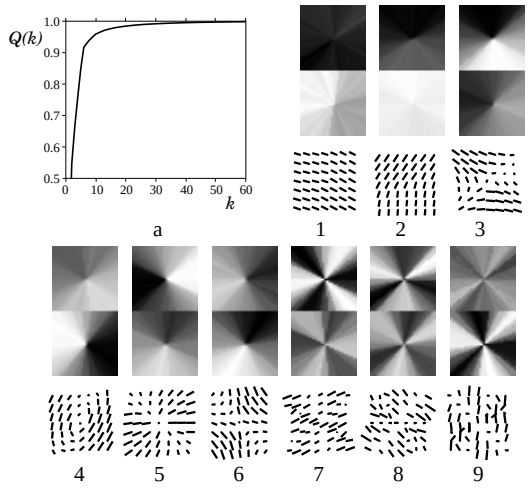

Figure 2: (a) The fraction of the variance in the training set accounted for by the first k principal components. (1-9) The first nine basis flows depicted as in Fig. 1, along with corresponding vector fields.

Value Decomposition (SVD) of F can be written as

$$F = M\Sigma V^T, \tag{1}$$

where $M = [\vec{m}_1, \vec{m}_2, \ldots, \vec{m}_p]$ is a $2s \times p$ matrix. The columns, $\vec{m}_i$, form an orthonormal basis for the range of F, Σ is a $p \times p$ diagonal matrix containing the singular values $\lambda_1, \lambda_2, \ldots, \lambda_p$ sorted in decreasing order along the diagonal, and V^T is a $p \times p$ orthogonal matrix. We can approximate a given flow field, $\vec{f}$, by a linear combination of the first k basis elements in M

$$\vec{f}_k = \sum_{i=1}^{k} a_i \vec{m}_i. \tag{2}$$

where the a_i are the parameters of the model to be estimated. Let $\vec{u}(\vec{x}; \vec{a}) = (u(x, y), v(x, y))$ denote the flow field that corresponds to the linear approximation, $\vec{f}_k$, where $\vec{x} = (x, y)$ and $\vec{a} = (a_1, a_2, \ldots, a_k)^T$.

The quality of the approximation provided by the first k columns of M is easily characterized in terms of the fraction of the variance of the training set that is accounted for by the

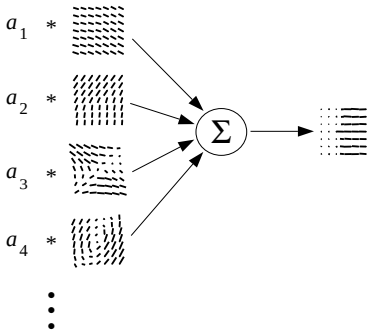

Figure 3: A motion field discontinuity can be represented and detected with a linear combination of a small number of the basis motions (cf. [16]).

selected components. This fraction is given by

$$Q(k) = \left(\sum_{i=1}^{k} \lambda_i^2 \right) / \left(\sum_{i=1}^{p} \lambda_i^2 \right) . \qquad (3)$$

If the singular values, λ_i, rapidly decrease to zero as i increases then $Q(k)$ rapidly increases towards 1, and a low-dimensional linear model provides an accurate approximation of the flow.

3.1 Example: Motion Discontinuities

For illustration, we applied this approach to learn a parameterized model of motion discontinuities. First, a synthetic training set of 200 flow fields was constructed. Each contained a motion discontinuity through the center of a 32×32 pixel region (see Fig. 1). The orientation, θ, and the translational motions on either side of the boundary, $\vec{u}_0$ and $\vec{u}_1$, were chosen randomly.

We then computed the SVD of the training set. The fraction of the variance accounted for by the first k components, namely $Q(k)$, rapidly approaches 1 (see Fig. 2a). Despite the variability of the input flow fields, nine basis vectors account for 95% of the variance in the training set. These basis flows are shown in Figure 2(1-9). Note that the basis set can also approximate translational motion since the random training data contains flow fields in which $\vec{u}_0$ is close to $\vec{u}_1$. Note the similarity between the basis vectors for a motion discontinuity and those learned for an intensity edge in [14].

4 Direct Estimation of Motion Parameters

Given a learned set of basis flows, we now consider the problem of estimating the optical flow in an arbitrary image region, R, using the parameterized model. Our goal is to find the coefficients $\vec{a}$ that produce a flow field satisfying the brightness constancy assumption

$$I(\vec{x} + \vec{u}(\vec{x}; \vec{a}), t+1) \; = \; I(\vec{x}, t) \quad \forall \vec{x} \in R. \qquad (4)$$

Equation (4) states that the image, I, at frame $t+1$ is a warped version of the image at time t.

To recover the parameters we formulate an objective function to be minimized, namely

$$E(\vec{b}; \vec{a}) = \sum_{\vec{x} \in R} \rho(I(\vec{x} + \vec{u}(\vec{x}; \vec{a} + \vec{b}), t+1) - I(\vec{x}, t), \, \sigma). \quad (5)$$

Given an estimate, $\vec{a}$, of the motion parameters (initally zero), the goal is to estimate the update, $\vec{b}$, that minimizes (5). Here, σ is a scale parameter and $\rho(\cdot, \sigma)$ is a robust error norm applied to the residual error $r(\vec{x}, \vec{a} + \vec{b}) = I(\vec{x} + \vec{u}(\vec{x}; \vec{a}), t+1) - I(\vec{x}, t)$.

Large residual errors, r, may be caused by changes in image appearance that are not accounted for by the learned flow model. The influence of these "outliers" can be reduced through the use of an appropriate robust error norm ρ. For the experiments below we take ρ to be

$$\rho(r, \sigma) \quad = \quad r^2 / (\sigma^2 + r^2) \; ,$$

which was used successfully for flow estimation in [5].

To minimize (5) we first linearize about the update vector $\vec{b}$ to give the approximate objective function $\tilde{E}(\vec{b}; \vec{a}) =$

$$\sum_{\vec{x} \in R} \rho(\vec{u}(\vec{x}; \vec{b}) \cdot \vec{\nabla} I(\vec{x} + \vec{u}(\vec{x}; \vec{a}), t+1) + r(\vec{x}, \vec{a}), \, \sigma), \quad (6)$$

where $\vec{\nabla} I(\vec{x} + \vec{u}(\vec{x}; \vec{a}), t+1) = [I_x, I_y]^T$ represents the partial derivatives of the image at time $t+1$ warped by the current motion estimate $\vec{u}(\vec{x}; \vec{a})$.

The particular optimization scheme is a straightforward extension of that used by Black and Anandan [5] for estimating optical flow with affine and planar motion models. This involves a coarse-to-fine iteration strategy, where the motion parameters $\vec{a}_j$ determined at a coarser scale are used in the estimation of $\tilde{E}(\vec{b}; \vec{a}_{j+1})$ at the next finer scale. The motion parameters, $\vec{a}_j$, from the coarse level are used in (6) to warp the image at time $t+1$ towards the image at time t. The basis flows at a coarse scale are simply smoothed and subsampled versions of the basis flows at the next finer scale. These coarse-scale basis vectors may deviate slightly from orthogonality but this is not significant given our optimization scheme.

At each scale a coordinate descent procedure is used to minimize $\tilde{E}(\vec{b}; \vec{a}_j)$. To deal with the non-convexity of the objective function, the robust scale parameter, σ, is initially set to a large value and then slowly reduced. For the experiments below, σ is lower from $25\sqrt{2}$ to $15\sqrt{2}$ by a factor of 0.95 at each iteration. Upon completion of a fixed number of descent steps (or when a convergence criterion is met), the new estimate for the flow coefficients is taken to be $\vec{a}_j + \vec{b}$. At the finest scale $\vec{a}_{j+1} = \vec{a}_j + \vec{b}$ is accepted as the solution for the flow parameters, otherwise $\vec{a}_{j+1} = 2(\vec{a}_j + \vec{b})$ is provided to the next finer scale as the initial guess (the factor of 2 reflects the doubling of the pixel resolution).

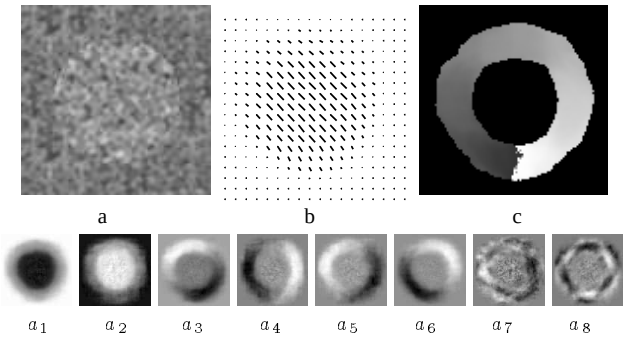

Figure 4: Moving disk example. (a) Disk moves one pixel down and to right over stationary background. (b) Estimated image motion. (c) Motion-discontinuity orientation can be computed from two orthogonal basis flows.

Note that in (6) that the gradient term does not depend on $\vec{b}$. This avoids the need to rewarp the image and recompute the image gradient at each descent step. In fact, the image gradient in (6) can be pre-multiplied by the basis flows since these quantities will not change during the minimization of $\tilde{E}(\vec{b}; \vec{a}_j)$. Hager and Belhuemer [10] used this fact for real-time affine tracking.

5 Experimental Results

We now present experiments to illustrate the use of learned models in two different applications. First, the models are used to estimate dense optical flow. Second, learned motion models are applied to a specific object in a known location. We consider examples of human mouths and legs where it is assumed that regions of interest have been found by tracking of the face or torso (see [7, 12]).

5.1 Motion Discontinuities

The learned motion-discontinuity model is applied to a textured moving disk in Fig. 4. Nine basis vectors were used and the motion coefficients were estimated in 32×32-pixel regions centered on each pixel in the image. The motion of the center pixel in each region is used to produce the dense flow field in Fig. 4(b). The coefficients of the orthogonal basis flows can be used to compute the orientation of the motion boundary at every pixel. The result is illustrated by the gray-scale encoding of orientation in Fig. 4(c). The images at the bottom of the figure show the value of the coefficients at each pixel.

Figure 5 shows the application of the motion discontinuity model to a natural image sequence. The camera is translating to the right, yielding a roughly translational vector field. The learned model, with nine basis vectors, was applied at every fourth pixel in the image. The estimated flow vectors from the 4×4 pixel block in the center of each patch are used to produce a dense flow field with a motion estimate at every pixel. The horizontal component of the flow

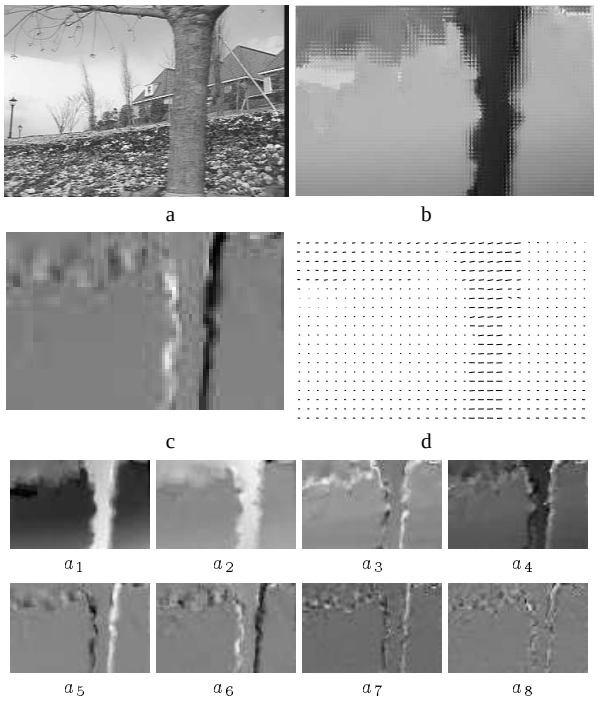

Figure 5: Flower-garden sequence. (a) First image. (b) Estimated horizontal flow (darker pixels denote greater leftward motion). (c) Detected motion boundary (white=occlusion, black=disocclusion). (d) Estimated flow field.

is shown in Fig. 5(b) and the vector field is shown in Fig. 5(d).

The detection of motion discontinuities here was straightforward. To detect a vertical occlusion/disocclusion boundary, we generated a synthetic occlusion flow field and projected it onto the basis set. The coefficients of this prototype occlusion boundary are then correlated with the coefficients estimated in each image region. A high correlation indicates the presence of a vertical occlusion boundary (shown as "white" in Fig. 5(c)) and a negative correlation indicates a disocclusion boundary ("black" in Fig. 5(c)).

5.2 Non-Rigid Motion

Black and Yacoob [7] described a method for recognizing human facial expressions from the coefficients of a parameterized model. They modeled the face as a plane and used its motion to stabilize the image sequence. The motion of the eyebrows and mouth were estimated relative to this stabilized face using a seven parameter model (affine plus a vertical curvature term). While this hand-coded model captures sufficient information about feature deformation to allow recognition of facial expressions, it it does not capture the variability of human mouths observed in natural speech.

Here we learn a parameterized model of mouth motion from examples. We collected four 150 image training sequences of a single speaker. The sequences contain natural

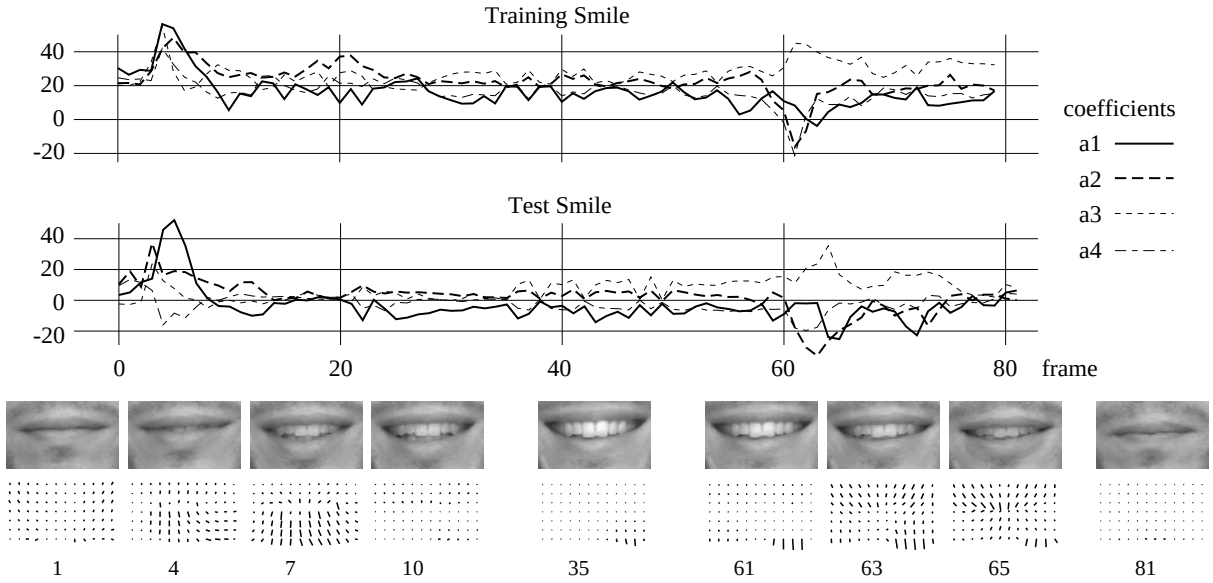

Figure 8: Smile experiment. Coefficients a_1, a_2, a_3, and a_4 are plotted over 80 frames for smile expressions in one training sequence and in the test sequence. Selected images and the corresponding estimated flow field are shown. Numbers under the images and flow fields correspond to frame numbers on the graphs.

Figure 6: Example frames from the 600 image training set.

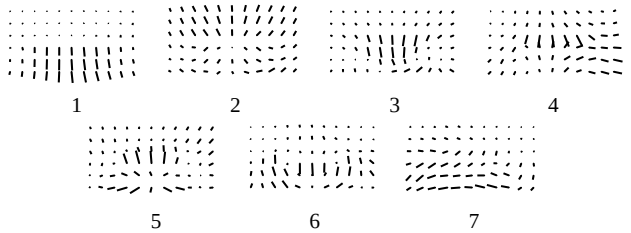

Figure 7: Basis flows for non-rigid mouth motion.

speech, smiling, and a test word which was repeated three times (see Fig. 6). Unlike the previous example, we did not have ground-truth optical flow from which to learn a model of mouth motion. Instead, we used the optical flow method in [5] to estimate dense flow fields between consecutive pairs of frames. It should be noted that the estimation of mouth motion is difficult since the lips are not highly textured, they deform and move large distances between frames, and the appearance/disappearance of teeth, tongue, and mouth cavity violates the brightness constancy assumption (see [8]). We also note that estimation of the dense training flow takes twice as long to compute as the direct estimation using the learned models.

Since the image motion of the mouth is highly constrained, the optical flow structure in the 600 training flow fields can be modeled by a small number of principal component flow fields. In this case, 90% of the variance in the training flow fields is accounted for by the first seven components (shown in Fig. 7). In contrast the seven-parameter model in [7] only accounted for 62% of the variance.

We evaluate the learned model with a 150-image test sequence in which the subject smiles and speaks the word from the training set. A sample of the images from the smile portion of the sequence are shown in Fig. 8. Below each image is the estimated flow using the learned 7-parameter model. The value of the first four coefficients of the model at each frame are plotted above the images. Notice the similarity between the training smile and the test smile. Similar plots were used for recognition in [7].

Figure 9 shows every second frame corresponding to the test utterance. Speech, unlike expression, is characterized by large, rapidly changing motions. Without a highly constrained model such as the one learned here, it can be difficult to estimate motions of this kind. The same word was uttered three times in the training set and once in the test set. If the model is accurately capturing the motion of the lips then the estimated coefficients of each utterance should be similar. The plots of selected coefficients (a_1, a_4, a_5, and a_6) are shown at the top of Fig. 9. While the plots appear to be highly correlated, further studies with a range of speakers are required to determine whether these motion coefficients

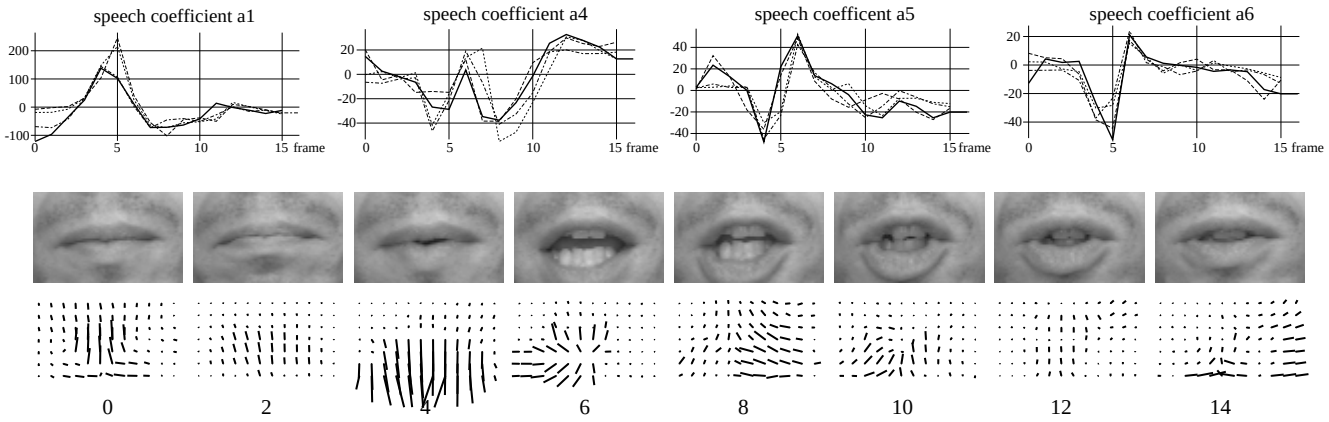

Figure 9: Speech experiment. The plots shows four coefficients (a_1, a_4, a_5, or a_6) against four seperate utterances of the same word, three from the training sequences (dotted curves) and one from the test sequence (solid curve). Below are sample images from the test sequence with corresponding flow fields.

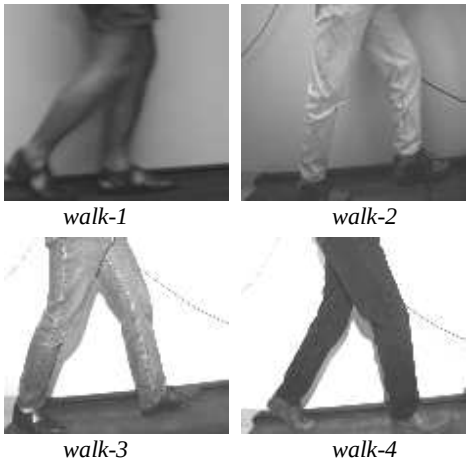

Figure 10: Articulated human motion. Top row: images from training sequences. Bottom row: test sequences.

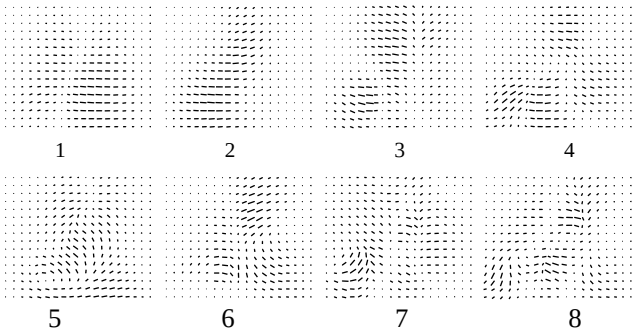

Figure 11: Basis flow fields for the walking sequences.

are useful for automated speech understanding.

5.3 Articulated Motion

Like mouths, the articulated motion of human limbs can be large, varied, and difficult to model. We assume that the subject is viewed from the side (though the approach can be extended to cope with other views) and that the image sequence has been stabilized with respect to the torso. Two training and two test sequences (Fig. 10) of a subject walking on a treadmill were acquired with slightly different lighting conditions, viewing position, and speed of activity.

SVD was performed on the 350-image training set. The first nine basis vectors account for 90% of variance in the training data and are used in our experiments (see Fig. 11.) Note that the first component essentially encodes the scissors-like expansion/contraction of the legs (cf. [2]).

Figure 12 shows results of motion estimation using a nine-parameter learned model for a 200-image training sequence (Walk-2) and a 200-image test sequence (Walk-4). Each sequence contains approximately seven complete cycles of the motion. Note the similarity of the two plots for the first coefficient (a_1). The magnitude of the parameter a_2 varies between the two sequences but is consistent within a sequence. Further experimentation with additional subjects will be necessary to determine the feasibility of activity recognition based on these parameters.

6 Conclusion

We presented a framework for learning parameterized models of image motion. Parameterized models provide strong constraints on the spatial variation of the flow within an image region and provide a concise description of the motion in terms of a small number of parameters. The framework described here extends parameterized flow methods to more complex motions that can be approximated as a linear combination of basis flow fields. It is important to note that the coefficients of the motion models are estimated directly from the image derivatives and do not require the prior computation of dense image motion.

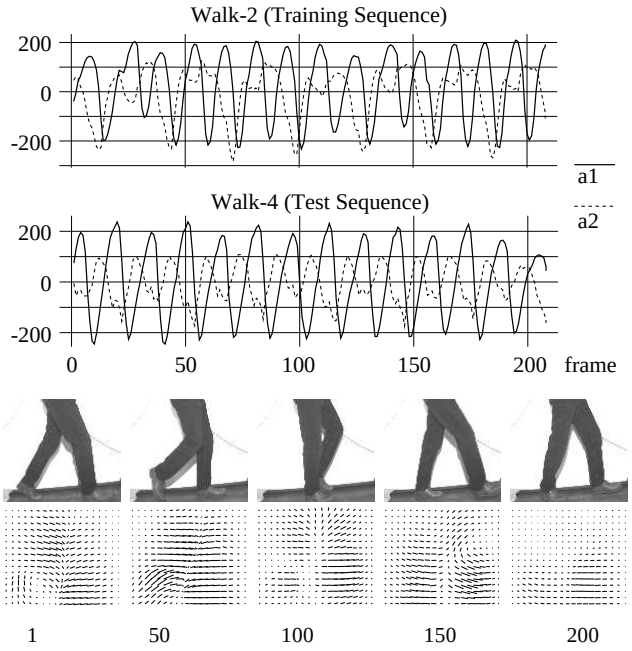

Figure 12: Plots of the first 2 motion coefficients for one training and one test sequence. Below: images and estimated flow for every 50th frame in the test sequence.

The methods can be used to learn generic flow models that can be applied at every image location in the way that current affine models are employed. In particular we are exploring the representation and recognition of motion features, such motion discontinuities and moving bars, and their relationship to the detection of static image features such as edges and line.

The approach can also be used to learn object-specific models (e.g. mouth motion) that are applied in specific image regions, and which may be useful for motion-based recognition. Alignment of these models with the image is important and it may be possible to refine this alignment automatically (see [6]).

A number of other research issues remain unanswered. Learned models are particularly useful in situations where optical flow is hard to estimate, but in these situations it is difficult to compute reliable training data. This problem is compounded by the sensitivity of PCA to outliers. PCA also gives more weight to large motions making it difficult to learn compact models of motions with important structure at multiple scales. Future work will explore non-linear models of image motion, robust and incremental learning, and models of motion texture.

Acknowledgements. We are grateful to Prof. Mubarak Shah for first suggesting to us the application of PCA to optical flow fields. DJF thanks NSERC Canada and Xerox PARC for their financial support.

References

[1] J. Barron, D. Fleet, and S. Beauchemin. Performance of optical flow techniques. *IJCV*, 12(1), 1994.

[2] A. Baumberg and D. Hogg. Learning flexible models from image sequences. *ECCV'94*, pp. 299–308.

[3] D. Beymer. Feature correspondence by interleaving shape and texture computations. *CVPR'96*, pp. 921–928.

[4] M. J. Black and P. Anandan. Constraints for the early detection of discontinuity from motion. *AAAI'90*, pp. 1060–1066.

[5] 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.

[6] M. J. Black and A. D. Jepson. EigenTracking: Robust matching and tracking of articulated objects using a view-based representation, to appear, *IJCV*.

[7] M. J. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motions. *ICCV'95*, pp. 374–381.

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

[9] 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.

[10] G. D. Hager and P. N. Belhumeur. Real-time tracking of image regions with changes in geometry and illumination. *CVPR'96*, pp. 403–410.

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

[12] S. Ju, M. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated motion. *Int. Conf. Auto. Face and Gesture Recog.*, pp. 38–44, 1996.

[13] C. Nastar, B. Moghaddam, and A. Pentland. Generalized image matching: Statistical learning of physically-based deformations. *ECCV'96*, pp. 589–598.

[14] S. Nayar, S. Baker, and H. Murase. Parametric feature detection. *CVPR'96*, pp. 471–477.

[15] S. A. Niyogi. Detecting kinetic occlusion. *ICCV'95*, pp. 1044–1049.

[16] S. Sclaroff and A. Pentland. Physically-based combinations of views: Representing rigid and nonrigid motion. *Work. Motion of Non-rigid & Articulated Objects*, pp. 158–164, 1994.

[17] A. Spoerri and S. Ullman. The early detection of motion boundaries. *ICCV'87*, pp. 209–218.

[18] T. Vetter. Learning novel views to a single face image. *Int. Conf. Auto. Face & Gesture Recog.*, pp. 22–27, 1996.