Cardboard People: A Parameterized Model of Articulated Image Motion

Shanon X. Ju* **Michael J. Black†** **Yaser Yacoob‡**

* Department of Computer Science, University of Toronto, Toronto, Ontario M5S 1A4 Canada
† Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304
‡ Computer Vision Laboratory, University of Maryland, College Park, MD 20742.
juxuan@vis.toronto.edu, black@parc.xerox.com, yaser@umiacs.umd.edu

1 Introduction

In this paper we extend the work of Black and Yacoob [5] on tracking and recognition of human facial expressions to the problem of tracking and recognizing the articulated motion of human limbs. We make the assumption that a person can be represented by a set of connected planar patches: the *cardboard person model* illustrated in Figure 1. In the case of faces, Black and Yacoob [5] showed that a planar model could well approximate the motion of a human head and that it provides a concise description of the optical flow within a region. This motion can be estimated robustly and it can be used for recognition.

To extend the approach in [5] to track articulated human motion we approximate the limbs as planar regions and recover the motions of these planes while constraining the motion of the connected patches to be the same at the points of articulation. To recognize articulated motion we will need to know the relative motion of each of the limbs. Given the computed motions of the thigh and calf, for example, we can solve for the relative motion of the calf with respect to the thigh. We posit that this relative image motion of the limbs is sufficient for recognition of human activity.

The tracking of human motion using these parameterized flow models is more challenging than the previous work on facial motion tracking. In the case of human limbs, the motion between frames can be very large with respect to the size of the image region, the deformations of clothing as a person moves make tracking difficult, and the human body is frequently self-occluding and self-shadowing. Additionally, facial motion recognition need only work over a relatively narrow range of views while we should be able to recognize human activities from a wider set of views (front, back, side, etc.). These differences between facial motion and general articulated human motion will require us to extend the previous methods in a number of ways. In this paper we focus on the problem of tracking the limbs of a person using articulated planar patches. At the end of the paper we analyze the performance of the current approach, discuss how it might be extended, and present some thoughts on the future of the method.

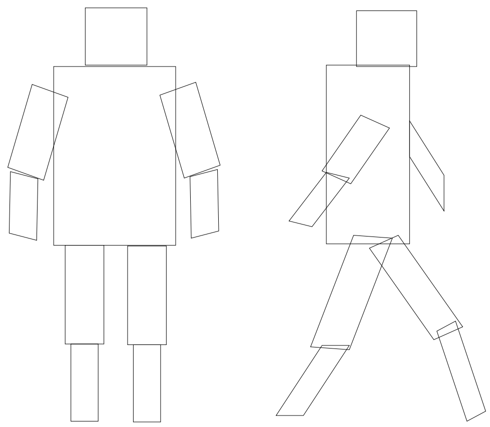

Figure 1: The cardboard person model. The limbs of a person are represented by planar patches.

2 Previous Research

Many approaches to tracking the movement of humans have focused on detecting and tracking the edges of the figure in the images. These methods typically attempt to match the projection of a detailed articulated 3D body model to the edge data [9, 10, 16]. A number of authors have extended active contour models to model articulated motion [6, 12, 13, 20]. For example, Baumberg and Hogg [2] track the outline of a moving body using a modal-based flexible shape model which captures the considerable outline variations in the human silhouette during movement. Stick-figure models of humans have also been matched to image data [1, 14]. These methods are typically only applied to humans viewed from the side.

The above methods do not explicitly use image motion to track and recognize activity. Pentland and Horowitz [15], however, describe the fitting of a 3D physically-based articulated model to optical flow data. Parts of a person are described as superquadrics with constraints on the articulated motion of the parts. In contrast, Wang *et al.* [19] use a 3D articulated model of a human leg to constrain the optical flow and they recover the motion of the articulated parts *directly* from changing image brightness without first com-

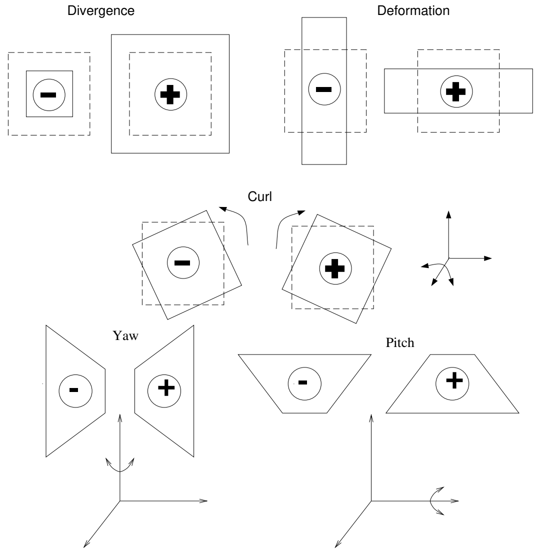

Figure 2: Divergence ($a_1 + a_5$), deformation ($a_1 - a_5$), curl ($-a_2 + a_4$), image yaw (a_6) and image pitch (a_7).

puting flow.

The approaches above typically require 3D models of the body. Furthermore, edges often play the central role in tracking and motion estimation. In this paper, we propose a parametrized motion model for tracking body parts. This model shifts the focus of tracking from edges to the intensity pattern created by each body part in the image plane. The tracking employs a 2D model-based approach for enforcing inter-part motion consistency for recovery thus simplifying the tracking and reducing the computations. We further develop an approach for viewer-based motion recognition of human activity and provide preliminary results.

3 Motion Estimation of a Rigid Object

The image motion of a rigid planar patch of the scene can be described by the following eight-parameter model:

$$u(x,y) = a_0 + a_1 x + a_2 y + a_6 x^2 + a_7 xy, \quad (1)$$
$$v(x,y) = a_3 + a_4 x + a_5 y + a_6 xy + a_7 y^2, \quad (2)$$

where $\mathbf{a} = [a_0, a_1, a_2, a_3, a_4, a_5, a_6, a_7]$ denotes the vector of parameters to be estimated, and $\mathbf{u}(\mathbf{x}, \mathbf{a}) = [u(x,y), v(x,y)]^T$ are the horizontal and vertical components of the flow at image point $\mathbf{x} = (x, y)$. The coordinates (x, y) are defined with respect to a particular point. Here this is taken to be the center of the patch but could be taken to be at a point of articulation.

The assumption of brightness constancy for a given patch and the planar motion model gives rise to the optical flow constraint equation

$$\nabla I \cdot \mathbf{u}(\mathbf{x}, \mathbf{a}_s) + I_t = 0, \ \forall \mathbf{x} \in \mathcal{R}_s \quad (3)$$

where $\mathbf{a}_s$ denotes the planar model for patch s, $\mathcal{R}_s$ denotes the points in patch s, I is the image brightness function and t represents time. $\nabla I = [I_x, I_y]$, and the subscripts indicates partial derivatives of image brightness with respect to the spatial dimensions and time at the point $\mathbf{x}$.

We use this constraint equation in the next section to solve for the motions of the patches. These parameters will be used to interpret the motion within each region. Various, low-level, interpretations of the motion parameters are shown in Figure 2.

4 Estimating Articulated Motion

For an articulated object, we assume that each patch is connected to only one preceding patch and one following patch, that is, the patches construct a chain structure (see Figure 3). For example, a "thigh" patch may be connected to a preceding "torso" patch and a following "calf" patch. Each patch is represented by its four corners. Our approach is to simultaneously estimate the motions, $\mathbf{a}_s$, of all the patches. We minimize the total energy of the following equation to estimate the motions of each patch (from 0 to n)

$$E = \sum_{s=0}^{n} E_s = \sum_{s=0}^{n} \sum_{\mathbf{x} \in \mathcal{R}_s} \rho(\nabla I \cdot \mathbf{u}(\mathbf{x}, \mathbf{a}_s) + I_t, \sigma) \quad (4)$$

where we take ρ to be an error norm with a redescending influence function.

Equation 4 may be ill-conditioned due to the lack of sufficient brightness variation within the patch. The articulated nature of the patches provides an additional constraint on the solution. This articulation constraint is added to Equation 4 as follows

$$E = \sum_{s=0}^{n} \left(\frac{1}{|\mathcal{R}_s|} E_s + \lambda \sum_{\mathbf{x} \in \mathcal{A}_s} ||\mathbf{u}(\mathbf{x}, \mathbf{a}_s) - \mathbf{u}(\mathbf{x}, \mathbf{a}')||^2\right), \quad (5)$$

where $|\mathcal{R}_s|$ is the number of pixels in patch s, λ controls relative importance of the two terms, $\mathcal{A}_s$ is the set of articulated points for patch s, $\mathbf{a}'$ is the planar motion of the patch which is connected to patch s at the articulated point $\mathbf{x}$, and $|| \cdot ||$ stands for the norm function. The use of a quadratic function for the spatial coherence term indicates that no outlier is allowed.

Instead of using a constraint on the image velocity at the articulation points, we can make use of the distance between a pair of points. Assuming $\mathbf{x}'$ is the corresponding image point of the articulated point $\mathbf{x}$, and $\mathbf{x}'$ belongs to the patch connected to patch s at point $\mathbf{x}$ (see Figure 3), Equation 5 can be modified as

$$E = \sum_{s=0}^{n} \left(\frac{1}{|\mathcal{R}_s|} E_s + \lambda \sum_{\mathbf{x} \in \mathcal{A}_s} ||\mathbf{x} + \mathbf{u}(\mathbf{x}, \mathbf{a}_s) - \mathbf{x}' - \mathbf{u}(\mathbf{x}', \mathbf{a}')||^2\right) \quad (6)$$

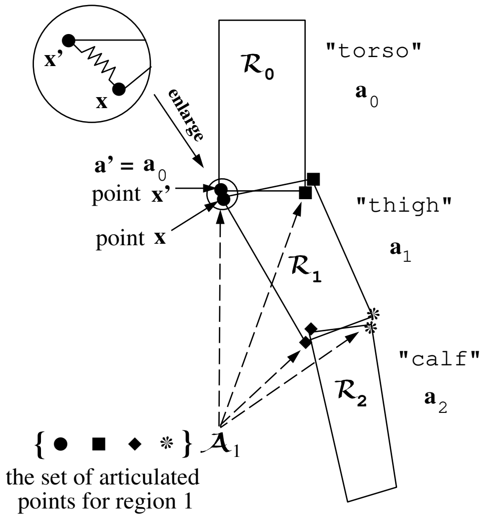

Figure 3: The "chain" structure of a three-segment articulated object.

This formulation has the advantage that the pair of articulated points, $\mathbf{x}$ and $\mathbf{x}'$, will always be close to each other at any time. The second energy term (the "smoothness" term) in Equation 6 can also be considered as a spring force energy term between two points (Figure 3).

We minimize Equation 6 using the simple gradient descent scheme with a continuation method [4, 7]. This involves in taking derivatives of the equation with respect to each of the planar motion parameters. At each step, we take into account both the optical flow constraints within the patch and the motion parameters of the connected patches.

For the experiments in this paper we take ρ to be

$$\rho(x, \sigma) = \frac{x^2}{\sigma^2 + x^2}, \tag{7}$$

which is the robust error norm used in [4]. As the magnitudes of residuals $\nabla I \cdot \mathbf{u}(\mathbf{x}, \mathbf{a}_s) + I_t$ grow beyond a point, their influence on the solution begins to decrease and the value of $\rho(\cdot)$ approaches a constant.

The value σ is a scale parameter, which effects the point at which the influence of outliers begins to decrease. In order to automatically estimate the value of σ, we assume that the residuals can be modeled by a mixture of two Gaussian distributions: one is to model the object, the other is to model the outliers. Since $\frac{\sigma}{1.4826}$ is the median value of the absolute values of a one-dimensional normal distribution [17], the robust estimation of σ from residuals can be defined as:

$$\sigma_{est} = 1.4826 \, median_{\mathbf{x}} |\nabla I \cdot \mathbf{u}(\mathbf{x}, \mathbf{a}_s) + I_t| \tag{8}$$

Equation 6 is minimized using continuation method that begins with a large σ and lowers it gradually [4, 7]. We define $\sigma_t = \sigma_{est}$, that is, the σ at iteration t is equal to σ_{est} which is computed from Equation 8 given current motion estimate $\mathbf{a}_s$. This σ_t is adjusted so that

$$\sigma_t \in \lfloor r_f \sigma_{t-1}, r_s \sigma_{t-1} \rfloor \cap \lfloor \sigma_{min}, \sigma_{max} \rfloor,$$

where r_f and r_s are the fastest and the slowest annealing rate respectively, and $\sigma_{-1} = \sigma_{max}$. The effect of this procedure is that initially almost no data are rejected as outliers then gradually the influence of outliers is reduced. The value of σ_{max} is $10\sqrt{3}$, and σ_{min} is $2\sqrt{3}$. The annealing rate r_s is 0.97, and r_f is 0.9. These parameters remain fixed for the experiments in this paper.

To cope with large motions, a coarse-to-fine strategy is used in which the motion is estimated at a coarse level then, at the next finer lever, the image at time $t+1$ is warped towards the image at time t using the current motion estimate. The motion parameters are refined at this level and the process continues until the finest level.

4.1 Computing the relative motions

The planar motions estimated from the Equation 6 are absolute motions. In order to recognize articulated motion, we need to recover the motions of limbs which are relative to their preceding (parent) patches. We define

$$\mathbf{u}(\mathbf{x} + \mathbf{u}(\mathbf{x}, \mathbf{a}_{s-1}), \mathbf{a}_s^r) = \mathbf{u}(\mathbf{x}, \mathbf{a}_s) - \mathbf{u}(\mathbf{x}, \mathbf{a}_{s-1}), \tag{9}$$

where $\mathbf{a}_s^r$ is the relative motion of patch s, $\mathbf{u}(\mathbf{x}, \mathbf{a}_s) - \mathbf{u}(\mathbf{x}, \mathbf{a}_{s-1})$ is the relative displacement at point $\mathbf{x}$, and $\mathbf{x} + \mathbf{u}(\mathbf{x}, \mathbf{a}_{s-1})$ is the new location of point $\mathbf{x}$ under motion $\mathbf{a}_{s-1}$. A planar motion has eight parameters, therefore four different points of patch s are sufficient to solve $\mathbf{a}_s^r$ given the linear equations 9. In our experiments, we use the four corners of the patches.

4.2 Tracking the articulated object

In the first frame, we interactively define each patch by its four corners. For each patch, the first two corners are defined as the articulated points, whose corresponding points are the last two corners of its preceding patch. This definition of articulated points shows that two connected patches share one common "edge". Once the "chain" structure is defined, the object is automatically tracked thereafter. Tracking is achieved by using the articulated motion between two frames to predict the location of each patch in the next frame. We update the location of each of the four corners of each patch by applying its estimated planar motion to it.

5 Experimental Results

In this section we illustrate the performance of the tracking algorithm on several image sequences of lower body human movement. We focus on "walking" (on a treadmill, for simplicity) and provide the recovered motion parameters for two leg parts during this cyclic activity. Notice that during "walking" the upper body plays only a minor role in recognition (it can, however, be appreciated that the movement of the torso and the arms can be used in determining heading, speed of "walking" and clues regarding the positions of lower body parts). To facilitate the use of our gradient-based flow estimation approach, we use a 99Hz video-camera to capture a few cycles of "walking"

Each sequence contains 500 to 800 frames. All the parameters used in the motion estimation algorithm were exactly the same in all the experiments. In particular, for each pair of images, 30 iterations of gradient descent were used at each level, and 3 levels were used in the coarse-to-fine strategy. The value of λ is 0.005.

Figures 4, 6, and 8 demonstrate three "walking" sequences taken from different view-points. The left column in each figure shows three input images some frames apart, the right column shows the tracking of two parts (the "thigh" and "calf"). Various motion parameters for these sequences are shown in Figures 5, 7, and 9. The first row in Figures 5 and 7 shows the horizontal and vertical translation (left most graph, dashed line is the vertical translation) and "curl" (right graph) for the "thigh". The second row shows the graphs for the "calf." In Figure 9 the "curl" graphs are replaced by the "deformation" and "divergence" and "image pitch". These graphs are only meant to provide an idea about the effectiveness of our tracking model and its ability to capture meaningful parameters of the body movement.

In Figures 5 and 7 it is clear that the horizontal translation and "curl" parameters capture quite well the cyclic motion of the two parts of the leg. The translation of the "calf" is relative to that of the "thigh" and therefore it is significantly smaller. On the other hand, the rotation (i.e., "curl") is more significant at the "calf". Notice that Figures 5 and 7 are qualitatively quite similar despite the difference in viewpoint. In Figure 9 the translations are smaller than before but still disclose a cyclic pattern. The "deformation," "divergence," and pitch capture the cyclic motion of the "walking away" on the treadmill. Notice that the pitch measured at the two parts is always reversed since when the "thigh" rotates in one direction the "calf" is bound to be viewed to be rotating in a opposite way.

In summary, the reported experiments show that the image motion models are capable of tracking articulated motion quite accurately over long sequences and recovering a meaningful set of parameters that can feed into a recognition system. For related work see [8].

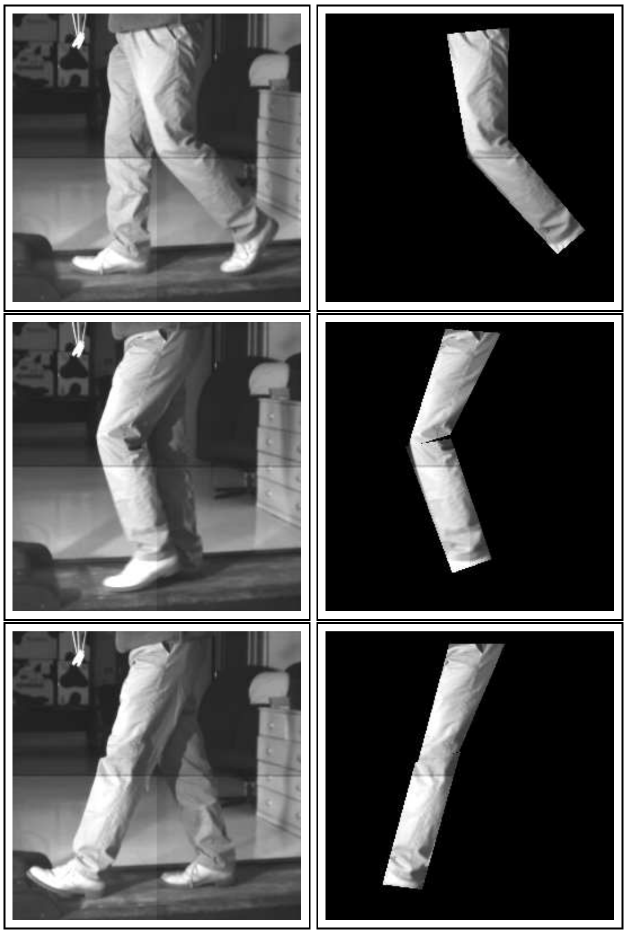

Figure 4: Walking parallel to the imaging plane. Three frames shown twenty frames apart.

6 Recognition of Movement

The goal of recognition of human movement encompasses answering: When does the activity begin and end? What class does the observed activity most closely resemble? What is the period (if cyclical) of the activity?

Seitz and Dyer [18] proposed an approach for determining whether an observed motion is periodic and computing its period. Their approach is based on the observation that the 3D points of an object performing affine-invariant motion are related by an affine transformation in their 2D motion projections. Once a period is detected, a matching of a single cycle of the motion to known motions can, in principal, provide for the recognition of the activity.

Our approach to recognition takes advantage of the economy of the parameterized motion models in capturing the range of motions and deformations of each body part. In the absence of shape cues, we employ a viewer-centered representation for recognition. Let ${C_v}^{i_j}(t)$ denote the temporal curve created by the motion parameter a_j of patch i viewed at angle v (where $j \in a_0, ..., a_7$). We make the observation that the following transformation does not change the nature of the activity represented by ${C_v}^{i_j}(t)$

$$ {D_v}^{i_j}(t) = S_i * {C_v}^{i_j}(t + T_i) \quad (10) $$

where ${D_v}^{i_j}(t)$ is the transformed curve. This transformation captures the translation, T_i, of the curve and the scal-

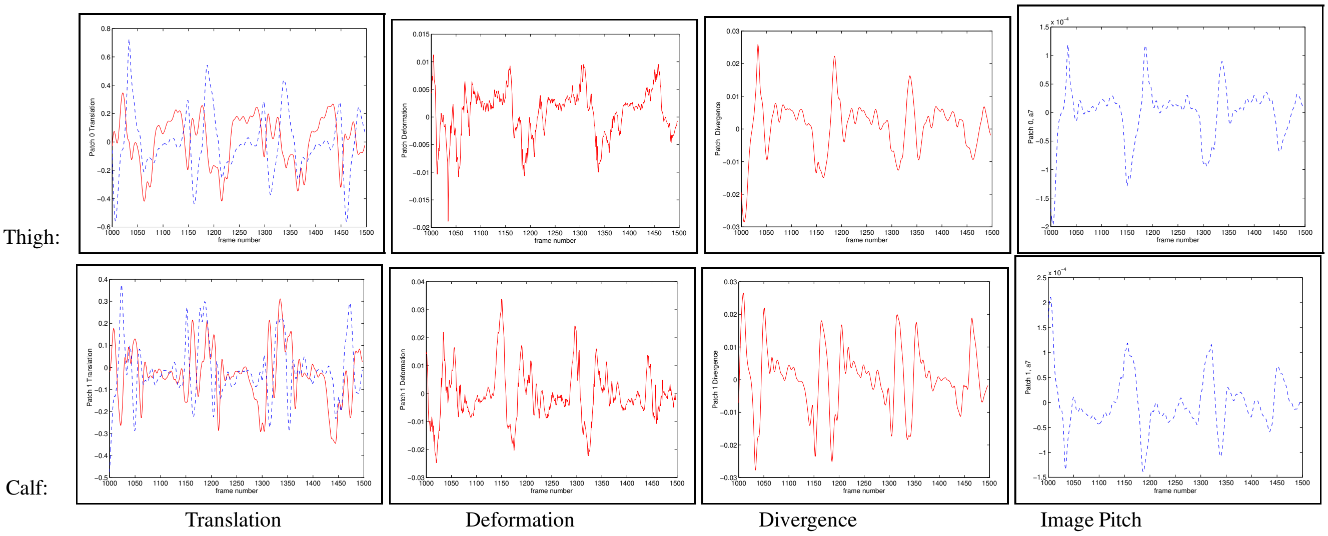

Figure 9: Tracking results for Figure 8

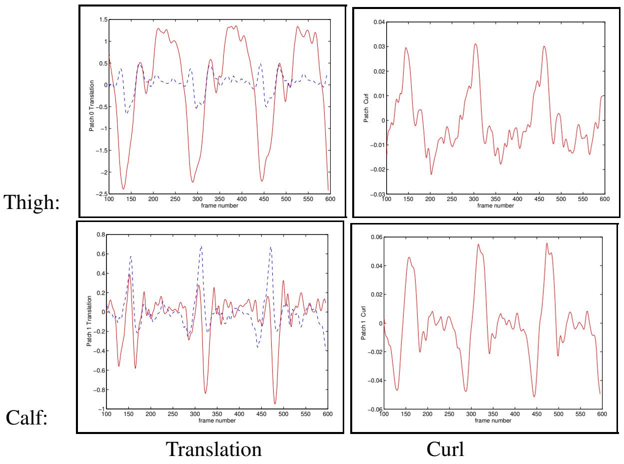

Figure 5: Motion parameters for walking parallel to the imaging plane (Figure 4).

ing, S_i, in the magnitude of the image-motion measured for parameter a_j. The scaling of the curve allows accounting for different distances between the human and the camera (while the viewing angle is kept constant) and accounts for the physiological variation across humans. Notice that this transformation does not scale the curve in the temporal dimension since the nature of the activity changes due to temporal scaling (e.g., different speeds of "walking" can be captured by this scaling). This temporal scaling can be expressed as an affine transformation

$$D_v{}^{ij}(t) = S_i * C_v{}^{ij}(\alpha_i t + T_i) \qquad (11)$$

where $\alpha_i > 1.0$ leads to a linear speed up of the activity and $\alpha_i < 1.0$ leads to its slow down.

The recognition of an activity can be posed as a matching problem between the curve created by parameter a_j over time and a set of known curves (corresponding to known activities) that can be subject to the above transformation. Recognition of an activity for some viewpoint v requires that a single affine transformation should apply to all parameters a_j, this can be posed as a minimization of the error (under some error norm)

$$E(v) = \sum_{j \in 0..7} \rho[D_v{}^{ij}(t) - (S_i * C_v{}^{ij}(\alpha_i t + T_i)), \sigma] \qquad (12)$$

Recognition over different viewpoints requires finding the minimum error between all views v, which can be expressed as

$$E = \min_v \sum_{j \in 0..7} \rho[D_v{}^{ij}(t) - (S_i * C_v{}^{ij}(\alpha_i t + T_i)), \sigma] \qquad (13)$$

Recognition over multiple body parts uses the inter-part hierarchy relationships to progressively find the best match. As demonstrated and discussed in Section 5, the motion parameters are stable over a wide range of viewpoints of the activity, so that they could be represented by a few principal directions.

Our formulation requires computing a *characteristic* curve $C_v{}^{ij}$ for each activity and body part viewed at angle v. Constructing this characteristic curve can be achieved by tracking the patch motions over several subjects and employing Principal Component Analysis (PCA) to capture the dominant curve components. Given an observed activity captured by $D^{ij}(t)$ (notice that the v is dropped since it is unknown), our algorithm determines the characteristic curve that minimizes the error function given in Equation 13

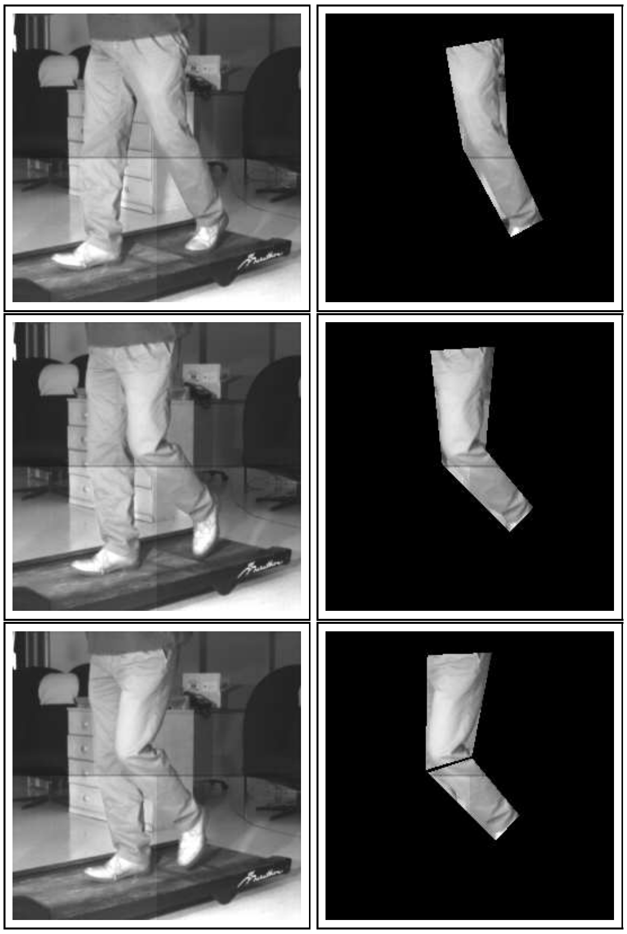

Figure 6: Walking 45 degress relative to the imaging plane.

by employing the recently proposed eigentracking approach [3] on the curves.

We are currently constructing these characteristic curves for several human activities. Davis [8] has independently proposed a somewhat similar model for learning and recognition of motion curves from multiple-views.

7 Discussion

The approach described here extends previous work on facial motion to articulated motion and shows promise for tracking and recognition of human activities. There are, however, a number of issues that still need to be addressed. First, the motion of human limbs in NTSC video (30 frames/sec) can be very large. For example, human limbs often move distances greater than their width between frames. This causes problems for a hierarchical gradient-based motion scheme such as the one presented here. To cope with large motions of small regions we will need to develop better methods for long-range motion estimation.

Unlike the human face, people wear clothing over their limbs which deforms as they move. The "motion" of the deforming clothing between frames is often significant and, where there is little texture on the clothing, may actually be the dominant motion within a region. A purely flow-based tracker such as the one here has no "memory" of what is being tracked. So if it is deceived by the motion of the clothing in some frame there is a risk that tracking will be lost. We are exploring ways of adding a template-style form of memory to improve the robustness of the tracking.

Self occlusion is another problem typically not present with facial motion tracking. Currently we have not addressed this issue, preferring to first explore the efficacy of the parameterized tracking and recognition scheme in the non-occlusion case. In extending this work to cope with occlusion, the template-style methods mentioned above may be applicable.

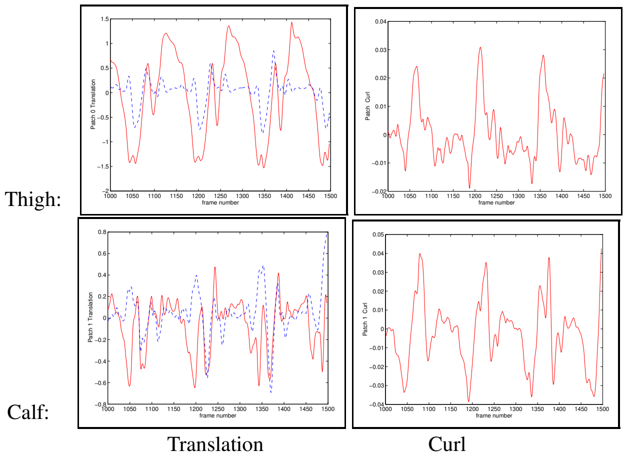

Figure 7: Motion parameters for walking 45 degress relative to the imaging plane (Figure 6).

8 Conclusion

We have presented a method for tracking articulated motion in an image sequence using parameterized models of optical flow. The method extends previous work on facial motion tracking [5] to more general animate motion. Unlike previous work on recovering human motion, this method assumes that the activity can be described by a the motion of a set of planar patches with constraints between the patches to enforce articulated motion. No 3D model of the person is required, features such as edges are not used, and the optical flow is estimated directly using the parameterized model. An advantage of the 2D parameterized flow models is that recovered flow parameters can be interprated and used for recognition as described in [5]. Previous methods for recognition need to be extended to cope with the cyclical motion of human activities and we have proposed a method for performing view-based recognition of human activities from the optical flow parameters. Our current work is focused on the automatic segmentation of articulated motion

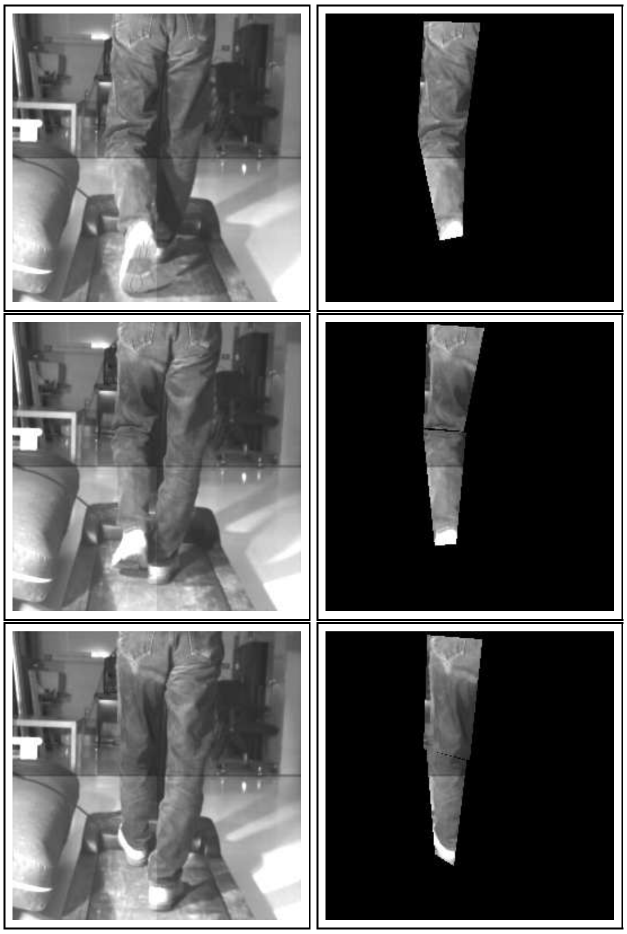

Figure 8: Walking perpendicular to the imaging plane.

into parts and the development of robust view-based recognition schemes for articulate motion.

References

[1] A. G. Bharatkumar, K. E. Daigle, M. G. Pandy, and J. K. Aggarwal, "Lower limb kinematics of human walking with the medial axis tranfromation," *IEEE Workshop on Motion of Non-rigid and Articulated Objects*, 1994, 70-76.

[2] A.M. Baumberg and D.C. Hogg, "An Efficient Method for Contour Tracking using Active Shape Models," *IEEE Workshop on Motion of Non-rigid and Articulated Objects,* 1994, 194-199.

[3] M.J. Black and A.D. Jepson, "EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation, " *ECCV,* 1996, 329-342.

[4] 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.

[5] M. J. Black and Y. Yacoob. Tracking and Recognizing Rigid and Non-rigid Facial Motions Using Local Parametric Models of Image Motions. *ICCV*, 1995, 374-381.

[6] A. Blake, M. Isard, and D. Reynard. Learning to track curves in motion. *IEEE Conf. Decision Theory and Control*, pp. 3788–3793, 1994.

[7] A. Blake and A. Zisserman. *Visual Reconstruction*. The MIT Press, Cambridge, Massachusetts, 1987.

[8] J. W. Davis, "Appearance-Based Motion Recognition of Human Actions," MIT Media Lab, M.S. Thesis Technical Report #387, 1996.

[9] D. Gavrila and L.S. Davis, "3D Model-Based Tracking of Humans in Action: A Multi-View Approach," To appear in *CVPR*, 1996, 73–80.

[10] L. Goncalves, E. Di Bernardo, E. Ursella and P. Perona, "Monocular Tracking of the Human Arm in 3D," *ICCV* , Boston, MA, 1995, 764-770.

[11] S. X. Ju, M.J. Black, and A. Jepson. Skin and bones: Multi-layer, locally affine, optical flow and regularization with transparency. *CVPR*, 1996, 307–314.

[12] I. A. Kakadiaris, D. Metaxas, and R. Bajcsy. Active part-decomposition, shape and motion estimations of articulated objects: A physics-based approach. *CVPR*, 1994, 980–984.

[13] C. Kervrann and F. Heitz. A hierarchical statistical framework for the segmentation of deformable objects in image sequences. *CVPR*, 1994, 724–728.

[14] S. A. Niyogi and E. H. Adelson, "Analyzing and recognizing walking figures in XYT," *CVPR*, 1994, 469-474.

[15] A. Pentland and B. Horowitz. Recovery of nonrigid motion and structure. PAMI, 13(7):730–742, July 1991.

[16] K. Rohr, "Towards Model-based Recognition of Human Movements in Image Sequences," CVGIP: Image Understanding, Vol. 59, 1994, 94-115.

[17] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. J. Wiley & Sons, NY, 1987.

[18] S.M. Seitz and C.R. Dyer, "Affine Invariant Detection of Periodic Motion," *CVPR*, 1994, 970-975.

[19] J. Wang, G. Lorette, and P. Bouthemy. Analysis of human motion: A model-based approach. In *7th Scandinavian Conf. Image Analysis*, Aalborg, Denmark, 1991.

[20] A. L. Yuille, P. W. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *Int. J. Computer Vision*, 8(2):99–111, 1992.