RECOGNIZING HUMAN MOTION USING PARAMETERIZED MODELS OF OPTICAL FLOW

MICHAEL J. BLACK
Xerox Palo Alto Research Center, Palo Alto, CA 94304, USA
YASER YACOOB
Computer Vision Laboratory, Center for Automation Research
University of Maryland, College Park, MD 20742, USA
AND
SHANON X. JU
Department of Computer Science, University of Toronto
Toronto, Ontario M5S 1A4, Canada

1. Introduction

The tracking and recognition of human motion is a challenging problem with diverse applications in virtual reality, medicine, teleoperations, animation, and human-computer interaction to name a few. The study of human motion has a long history with the use of *images* for analyzing animate motion beginning with the improvements in photography and the development of motion-pictures in the late nineteenth century. Scientists and artists such as Marey [12] and Muybridge [26] were early explorers of human and animal motion in images and image sequences. Today, commercial motion-capture systems can be used to accurately record the 3D movements of an instrumented person, but the motion analysis and motion recognition of an arbitrary person in a video sequence remains an unsolved problem. In this chapter we describe the representation and recognition of human motion using parameterized models of optical flow. A person's limbs, face, and facial features are represented as patches whose motion in an image sequence can be modeled by low-order polynomials. A robust optical flow estimation technique is used to recover the motion of these patches and the recovered motion parameters provide a rich, yet concise, description of the human motion which can be used to recognize human activities, gestures, and facial expressions.

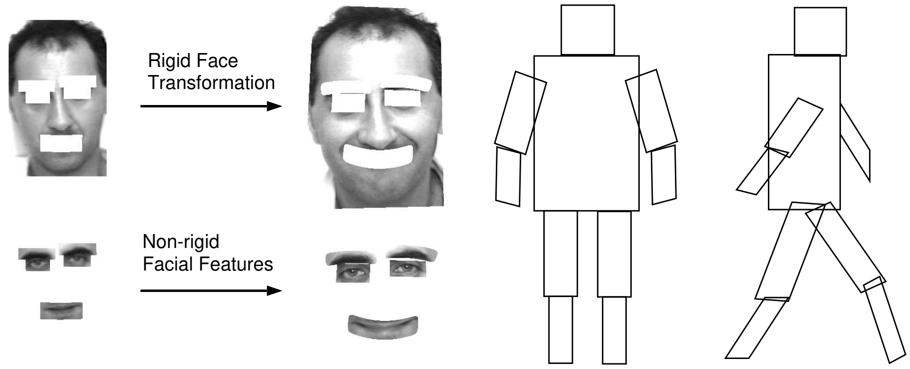


Figure 1. The cardboard person model. The limbs and face of a person are represented by planar patches. The motion of the face is further represented by the relative motions of the features.

In our representation we approximate the non-rigid motion of a person using a set of parameterized models of optical flow. This *cardboard-person* model is illustrated in Figure 1. While parameterized flow models (for example affine flow) have been used for representing image motion in rigid scenes, Black and Yacoob [7] observed that simple parameterized models could well approximate more complex motions if localized in space and time. Moreover, they showed that the motion of one body region (for example the face region) could be used to stabilize that body part in a warped image sequence. This allowed the image motions of related body parts (the eyes, mouth, and eyebrows) to be estimated *relative* to the motion of the face. Isolating the motions of these features from the motion of the face is critical for recognizing facial expressions using motion.

These parameterized motion models can be extended to model the articulated motion of the human limbs [18]. Limb segments can be approximated by planes and the motion of these planes can be recovered using a simple eight-parameter optical flow model. Constraints can be added to the optical flow estimation problem to model the articulation of the limbs and the relative image motions of the limbs can be used for recognition.

An example of tracking and recognizing facial motion is provided in Figure 2. Regions corresponding to parts of the face are located in the first image of the sequence. Then between pairs of frames, the image motion within the regions is computed robustly using a parameterized optical flow model. These models capture how the regions move and deform and the motion information is used to track the regions through the image sequence.

The motion of the regions between frames is described by a small set of parameters which can be used for recognition. Some of the information contained in these parameters is shown for the surprise expression in Figure

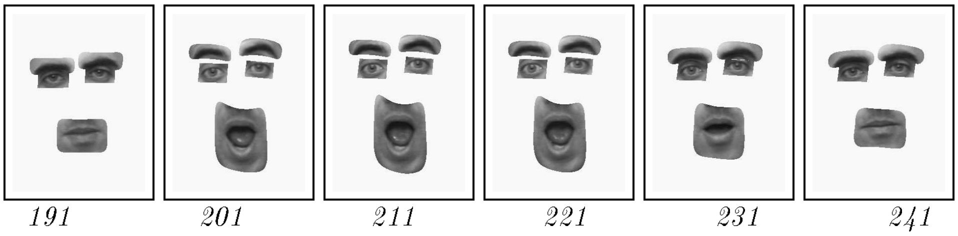


Figure 2. Surprise Experiment: facial expression tracking. Features every 10 frames.

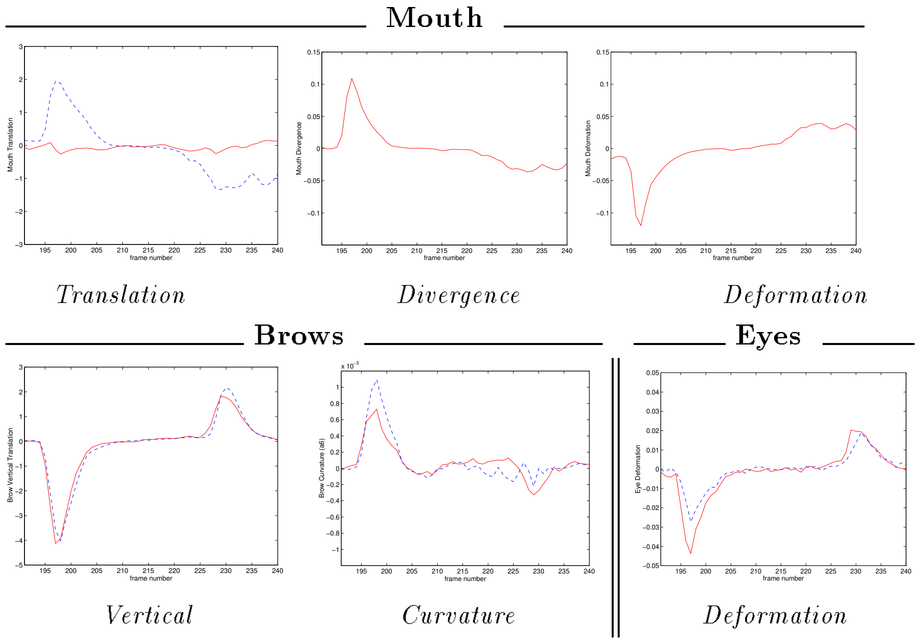


Figure 3. Motion parameters. For the mouth translation, the solid line indicates horizontal motion while the dashed line indicates vertical motion. For the eye and brows, the solid and dashed lines indicate left and right respectively.

3. During the initiation of the expression the mouth translates down, diverges, and deforms significantly. Simultaneously, the brows and eyes move upwards, the brows arch, and the eyes deform as they widen. The ending phase in this example shows a more gradual reversal of these parameters returning the face to the resting position.

It is important to note that these parameters only represent the motion of the region between two frames and that recognition is performed directly from the observed optical flow parameters. Additionally, the motion of facial features is estimated relative to the motion of the face which, in turn, is estimated relative to the motion of the torso. This relative, parameterized, motion information turns out to be a surprisingly rich representation which allows recognition of human motions without the use of complex 3D models or the matching of image features. Our experiments with facial and articulated motions indicate that the parameterized motion models can provide robust tracking and recognition over long image sequences.

In the following section we relate our cardboard-person model to other approaches. Then in Section 3 we present the optical flow models used for tracking and recognition. The robust estimation of the relative images motions of the body parts is described in Sections 4 and 5 and then Section 6 presents a few results which illustrate the tracking of faces and limbs in long image sequences. Section 7 presents recognition strategies for facial expressions, gestures, and cyclical movements such as walking. Finally in Sections 8 and 9 we examine the limitations of the current techniques and present some future directions.

2. Previous work

Human tracking in image sequences involves tracking the motion of a diverse set of body parts performing rigid, articulated, or deformable motions. These motions often exist simultaneously such as in facial expressions during head rotations or clothing deformations during locomotion. Approaches for tracking humans performing activities can be categorized according to being: *motion-based* versus *static*, *3D model-based* versus *2D image-based*, and *region-based* versus *boundary-based*.

Motion-based approaches consider the tracking of body parts as involving the recovery of the motion parameters between consecutive images [2, 14, 21, 28, 38]. These motion parameters can either be recovered directly from the spatio-temporal derivatives of the image sequence or from a dense optical flow field. Static approaches view body part tracking as the localization of a body part in a single image or pose estimation in 3D space [15, 16, 31, 40].

3D model-based approaches employ significant structural information about the body parts to recover their 3D pose in space [14, 15, 21, 28, 30] while 2D image-based approaches focus on the intensity (or color) distribution in the images to track the body parts possibly through employing 2D models of body part shape or motion [3, 7, 24, 31].

Region tracking integrates information over areas of the image to track the body part motion [4, 7, 24, 27, 37] while boundary tracking concentrates on the discontinuities in the projected image of the body part in motion [3, 8, 15, 16, 19, 31, 36, 39].

The limitations of each of the above categories are well known, but these limitations are exacerbated in the context of human motions. 3D motion recovery algorithms require a priori structure information (at least a coarse model) in the case of articulated objects or point correspondences for rigid objects [2]. On the other hand, image motions are not easily related to the multiple activities of body parts and their projection on the image plane. Recovering 3D models of the human body or its parts is generally difficult to achieve in an unconstrained environment due to the significant variability of human appearances (clothing, make up, hair styles, etc.). Additionally, 2D image-based tracking is complicated by the articulation and deformation of body parts and the dependence on the observation point of the activity. Boundary tracking allows focusing on information-rich parts of the image, these boundaries can occasionally be ambiguous, small in number, dependent on imaging conditions and may not sufficiently constrain the recovery of certain motions (e.g., rotation of a roughly cylindrical body part, such as a forearm, around its major axis). Region tracking employs considerably more data from images and thus can be more robust to ambiguous data, however, when regions are uniform multiple solutions may exist.

Generally, research on recognition of human activities has focused on one type of human body part motion and has assumed no coincidence of other motions. With the exception of [7], work on facial deformations (facial expressions and lip reading) has assumed that little or no rigid head or body motions are coincident [14, 13, 21, 24, 32, 36, 39, 40]. Articulated motion tracking work has assumed that non-rigid deformations are not coincident (e.g. clothing deformations during locomotion) [4, 3, 15, 16, 27, 28, 30, 31, 38]. Rigid motion recovery approaches such as [2] do not account for deformable and articulated motions (e.g., facial expressions and speech).

Recognition of human motion critically depends on the recovered representations of the body parts' activities. Most recognition work has employed well known techniques such as eigenspaces [22] dynamic time warping [15], hidden Markov models [25, 35], phase space [9, 23, 27], rule-based techniques [7, 39], and neural networks [32] (for a detailed overview see [10]).

In this chapter, we propose a 2D model-based framework for human part tracking using parametrized motion models of these parts. This framework shifts the focus of tracking from edges to the intensity pattern created by each body part in the image plane. It employs a 2D model-based viewer-centered approach to analyzing the data in image sequences. The approach enforces inter-part motion constraints for recovery which results in simplifying the tracking. We show that our paradigm provides a reasonable model of motion types prevalent in human activity. We further discuss viewer-centered motion recognition approaches of human activity that involve deformable motions (e.g., facial expressions), rigid motions (e.g., head gestures) and articulated motion (e.g., locomotion).

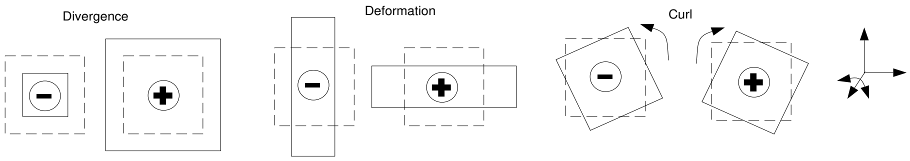


Figure 4. The figure illustrates the motion captured by the various parameters used to represent the motion of the regions. The solid lines indicate the deformed image region and the "–" and "+" indicate the sign of the quantity.

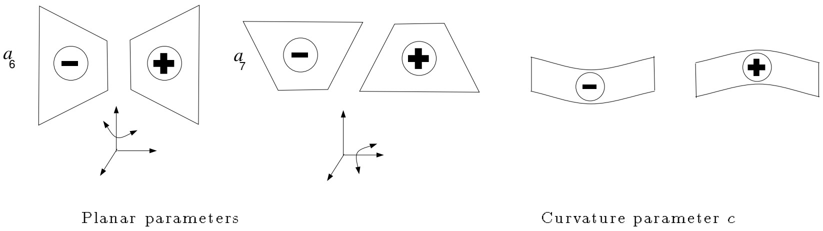


Figure 5. Additional parameters for planar motion and curvature.

3. Models of Image Motion

Parameterized models of image motion make explicit assumptions about the spatial variation of the optical flow within a region. Modeling the motion of a human body part involves making simplifying assumptions that approximate the image motion of that part. For example we assume that the limbs of the body and the face (excluding the face features) can be modeled as rigid planes. The image motion of a rigid planar patch of the scene can be described by the following eight-parameter model:

$$u(x, y) = a_0 + a_1 x + a_2 y + a_6 x^2 + a_7 xy, \tag{1}$$

$$v(x, y) = a_3 + a_4 x + a_5 y + a_6 xy + a_7 y^2, \tag{2}$$

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 region but could be taken to be at a point of articulation.

The parameters a_i have qualitative interpretations in terms of image motion. For example, a_0 and a_3 represent horizontal and vertical translation respectively. Additionally, we can approximately express *divergence* (isotropic expansion), *curl* (rotation about the viewing direction), and *deformation* (squashing or stretching) as combinations of the a_i (cf. [11, 20]). We define these quantities as

$$\text{divergence} \stackrel{\text{def}}{=} a_1 + a_5, \tag{3}$$

$$\text{curl} \stackrel{\text{def}}{=} -a_2 + a_4, \tag{4}$$
$$\text{deformation} \stackrel{\text{def}}{=} a_1 - a_5. \tag{5}$$

Note that these terms give qualitative measures that can be used to interprate the motion of a region. Translation, along with divergence, curl, and deformation, will prove to be useful for describing facial expressions and are illustrated in Figure 4. For example, eye blinking can be detected as rapid deformation, divergence, and vertical translation in the eye region.

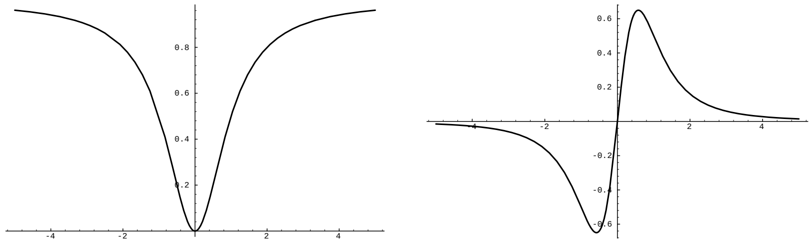
The parameters a_6 and a_7 roughly represent "yaw" and "pitch" deformations in the image plane respectively and are illustrated in Figure 5. While we have experimented with more complex models of rigid face motion, and Wang *et al.* [38] have used cylindrical models of limbs, we have found that the planar motion approximation is both simple and expressive enough to robustly represent qualitative rigid face and limb motions in a variety of situations.

To recognize the motion of the head using these parameters we need to know the head motion *relative* to the motion of the torso. Similarly, to recognize facial expressions from motion we need to know the motion of the eyes, mouth, and eyebrows relative to the motion of the face. In addition to isolating the motions of interest for recognition, this relative motion estimation allows us to estimate the motion of body parts that occupy small regions of the image; for example, facial features or fingers. The absolute motion of these regions in an image sequence may be very large with respect to their size making motion estimation difficult. The problem of estimating the motion of small parts like fingers is simplified if we know the motion of the torso, arm and hand.

For small regions of the image such as eyes and fingers, the planar model may not be necessary and the motion of these regions can be approximated by the simpler affine model in which the terms a_6 and a_7 are zero. The non-rigid motions of facial features such as the eyebrows and mouth, however, are not well captured by the rigid affine or planar models so we augment the affine model to account for the primary form of curvature seen in mouths and eyebrows. We add a new parameter c to the affine model

$$u(x, y) = a_0 + a_1 x + a_2 y \tag{6}$$
$$v(x, y) = a_3 + a_4 x + a_5 y + c x^2 \tag{7}$$

where c encodes curvature and is illustrated in Figure 5. This curvature parameter must be estimated in the coordinate frame of the face as described in [7]. As the experiments will demonstrate, this seven parameter model captures the essential image motion of the mouth and eyebrows.



$\rho(x, \sigma)$ $\psi(x, \sigma)$

Figure 6. Robust error norm (ρ) and influence function (ψ).

4. Parametric Motion Estimation

To estimate the motion parameters, $\mathbf{a}$, for a given patch we make the assumption that the brightness pattern within the patch remains constant while the patch may deform as specified by the model. This brightness constancy assumption gives rise to the optical flow constraint equation

$$\nabla I \cdot \mathbf{u}(\mathbf{x}, \mathbf{a}_s) + I_t = 0, \ \forall \mathbf{x} \in \mathcal{R}_s \tag{8}$$

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 $\mathbf{x}$.

Note that the brightness constancy assumption used to estimate the image motion is often violated in practice due to changes in lighting, specular reflections, occlusion boundaries, etc. It may also be violated because the motion model is only a rough approximation to the true motion; for example we model the face as a plane although it is not really rigid or planar.

Robust regression has been shown to provide accurate motion estimates in a variety of situations in which the brightness constancy assumption in violated [5]. To estimate the parameters $\mathbf{a}_s$ robustly we minimize

$$\sum_{\mathbf{x} \in \mathcal{R}_s} \rho(\nabla I \cdot \mathbf{u}(\mathbf{x}, \mathbf{a}_s) + I_t, \sigma), \tag{9}$$

for some robust error norm ρ where σ is a scale parameter. Violations of the brightness constancy assumption can be viewed as "outliers" [17] and we need to choose the function ρ such that it is insensitive to these gross errors.

For the experiments in this chapter we take ρ to be

$$\rho(x, \sigma) = \frac{x^2}{\sigma + x^2} \tag{10}$$

which is the robust error norm used in [5, 7, 18] and is shown in Figure 6. 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 function $\psi(x, \sigma)$ shown in Figure 6 is the derivative of ρ and characterizes the influence of the residuals [17]. The value σ is a scale parameter that effects the point at which the influence of outliers begins to decrease.

Equation (9) is minimized using a simple coordinate descent scheme with a continuation method that begins with a high value for σ and lowers it during the minimization (see [5, 7, 18] for details). The effect of this procedure is that initially no data are rejected as outliers then gradually the influence of outliers is reduced. 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 level, 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.

In the remainder of this chapter we use this robust estimation scheme for estimating face motion as described in [7] and for articulated motion as described in [18].

5. Estimating Relative Body Part Motion

The parametric motions of human body parts are inter-related as either *linked* or *overlapping* parts. Linked body parts, such as the "thigh" and "calf," share a joint in common and must satisfy an articulation constraint on their motion. The overlapping relation describes the relationship between regions such as the face and mouth in which the motion of the mouth is estimated relative to the motion of the face but is not constrained by it. These relationships lead to different treatments in terms of how the inter-part motions are estimated. These relations and the associated motion estimation techniques are described in this section and are illustrated with examples of facial motion estimation and articulated leg motion.

5.1. ESTIMATING THE RELATIVE MOTION OF OVERLAPPING REGIONS

To recover the motion of the face, we first estimate the planar approximation to the face motion. This motion estimate is then used to register the image at time $t + 1$ with the image at time t by warping the image at $t + 1$ back towards the image at t. Since the face is neither planar nor rigid this registration does not completely stabilize the two images. The residual motion is due either to the non-planar 3D shape of the head (its curvature and the nose for example) or the non-rigid motion of the facial features (cf.

work on plane+parallax models of motion in rigid scenes [33]). We have observed that the planar model does a very good job of removing the rigid motion of the face and that the dominant residual motion is due to the motion of the facial features. The residual motion in the stabilized sequence is estimated using the appropriate motion model for that feature (i.e., affine or affine+curvature). Thus stablizing the face with respect to the planar approximation of its motion between two images allows the relative motions of the facial features to be estimated.

The estimated parametric motion of the face and facial features estimated between two frames is used to predict the location of the features in the next frame. The face and the eyes are simple quadrilaterals which are represented by the image locations of their four corners. We update the location $\mathbf{x}$ of each of the four corners of the face and eyes by applying the planar motion parameters $\mathbf{a}_f$ of the face to get $\mathbf{x}' = \mathbf{x} + \mathbf{u}(\mathbf{x}, \mathbf{a}_f)$. Then the relative motion of the eyes locations is accounted for and the corners become $\mathbf{x}' + \mathbf{u}(\mathbf{x}', \mathbf{a}_{le})$ and $\mathbf{x}' + \mathbf{u}(\mathbf{x}', \mathbf{a}_{re})$ where $\mathbf{a}_{le}$ and $\mathbf{a}_{re}$ are the parameters estimated for the motions of the left and right eyes respectively. In updating the eye region we do not use the full affine model since when the eye blinks this would cause the tracked region to deform to the point where the eye region could no longer be tracked. Instead only the horizontal and vertical translation of the eye region is used to update its location relative to the face motion.

The curvature of the mouth and brows means that the simple updating of the corners is not sufficient for tracking. In our current implementation we use image masks to represent the regions of the image corresponding to the brows and the mouth. These masks are updated by warping them first by the planar face motion $\mathbf{a}_f$ and then by the motion of the individual features $\mathbf{a}_m$, $\mathbf{a}_{lb}$ and $\mathbf{a}_{rb}$ which correspond the mouth and the left and right brows respectively. This simple updating scheme works well in practice.

5.2. 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 7). 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)

$$E = \sum_{s=0}^{n} E_s = \sum_{s=0}^{n} \sum_{\mathbf{x} \in \mathcal{R}_s} \rho(\nabla I \cdot \mathbf{u}(\mathbf{x}, \mathbf{a}_s) + I_t, \sigma) \tag{11}$$

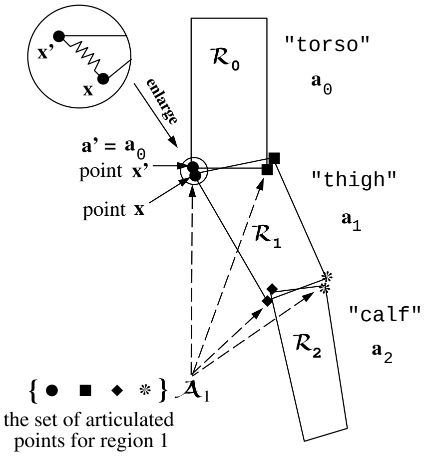


Figure 7. The "chain" structure of a three-segment articulated object.

where ρ is the robust error norm defined above.

Equation (11) 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 (11) as follows

$$E = \sum_{s=0}^{n} \Big(\frac{1}{|\mathcal{R}_s|} E_s + \lambda \sum_{\mathbf{x} \in \mathcal{A}_s} \|\mathbf{u}(\mathbf{x}, \mathbf{a}_s) - \mathbf{u}(\mathbf{x}, \mathbf{a}')\|^2 \Big), \tag{12}$$

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 $\|\cdot\|$ is the standard norm. The use of a quadratic function for the articulation constraint reflects that the assumption that no "outliers" are 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 7), Equation (12) can be modified as

$$E = \sum_{s=0}^{n} \Big(\frac{1}{|\mathcal{R}_s|} E_s + \lambda \sum_{\mathbf{x} \in \mathcal{A}_s} \|\mathbf{x} + \mathbf{u}(\mathbf{x}, \mathbf{a}_s) - \mathbf{x}' - \mathbf{u}(\mathbf{x}', \mathbf{a}')\|^2 \Big) \tag{13}$$

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 (13) can also be considered as a spring force energy term between two points (Figure 7).

The planar motions estimated from the Equation (13) 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

$$\mathbf{u}(\mathbf{x} + \mathbf{u}(\mathbf{x}, \mathbf{a}_{s-1}), \mathbf{a}_s^r) = \mathbf{u}(\mathbf{x}, \mathbf{a}_s) - \mathbf{u}(\mathbf{x}, \mathbf{a}_{s-1}), \tag{14}$$

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 (14). In our experiments we use the four corners of the patches.

6. Motion Estimation Results

In this section we illustrate the performance of the parameterized flow models on articulated, rigid and deformable body parts. Head motion and facial expressions are used in illustrating the rigid and deformable motions, respectively. For articulated motion we focus on "walking" (on a treadmill, for simplicity) and provide the recovered motion parameters for two leg parts during this cyclic activity.

The image sequence in Figure 8 illustrates facial expressions ("smiling" and "surprise") in conjunction with rigid head motion (in this case looming). The figure plots the regions corresponding to the face and the facial features tracked across the image sequence. The parameters describing the planar motion of the face are plotted in Figure 9 where the divergence due to the looming motion of the head is clearly visible in the plot of divergence. Notice that the rigid motions of the face are not visible in the plotted motions of the facial features in Figure 10. This indicates that the motion of the face has been factored out and that the feature motions are truly relative to the face motion. Analyzing the plots of the facial features reveals that a "smile" expression begins around frame 125 with an increase in mouth curvature followed by a deformation of the mouth. The curvature decreases between frames 175 and 185 and then a "surprise" expression begins around frame 220 with vertical eyebrow motion, brow arching, and mouth deformation.

Figures 11-14 demonstrate two "walking" sequences taken from different view-points. Each sequence contains 500 images and roughly three cycles of the activity. In Figures 11 and 13 the upper row shows three images from the sequences and the second row shows the tracking of two parts (the "thigh" and "calf"). Figures 12 and 14 show relevant recovered motions parameters over the entire 500 frame image sequences. The first rows in these figures

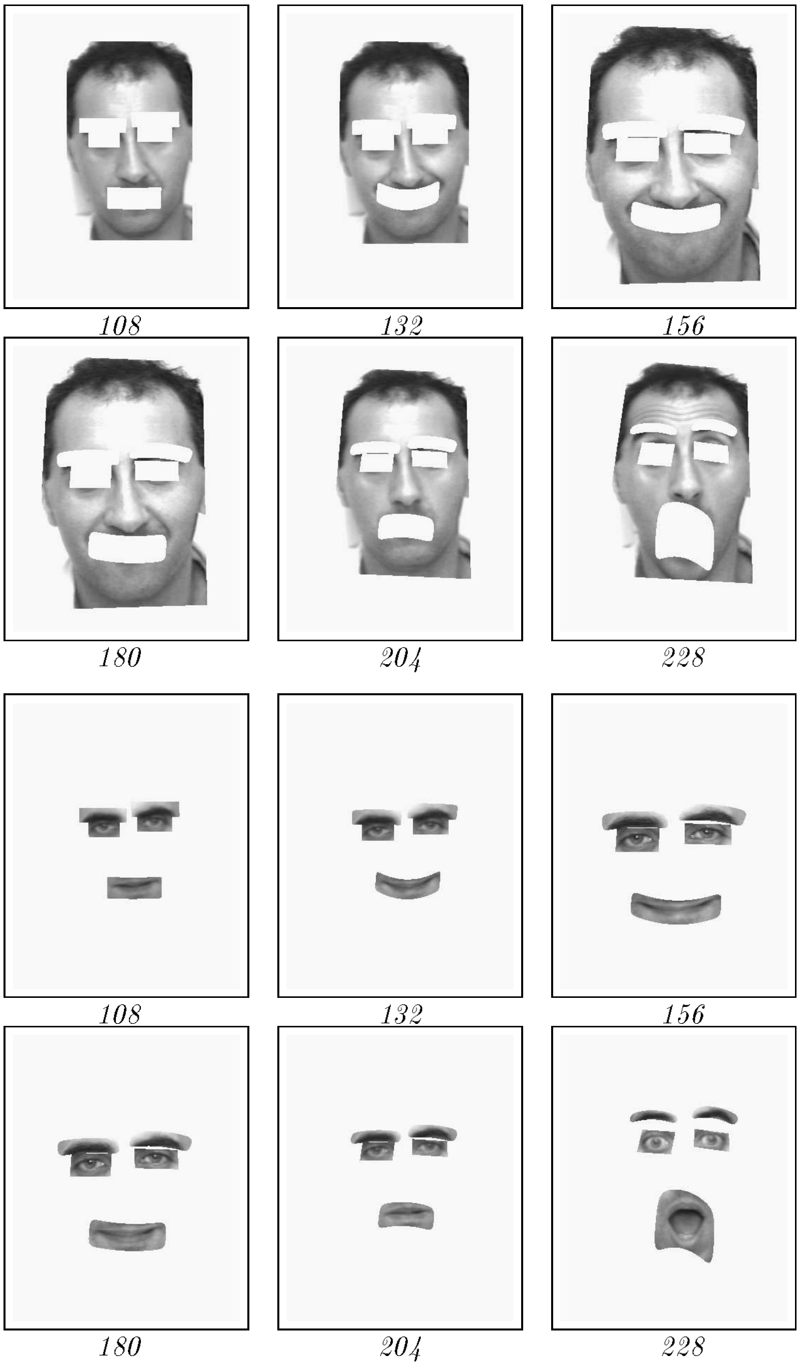


Figure 8. Looming Experiment. Facial expression tracking with rigid head motion (every 24 frames).

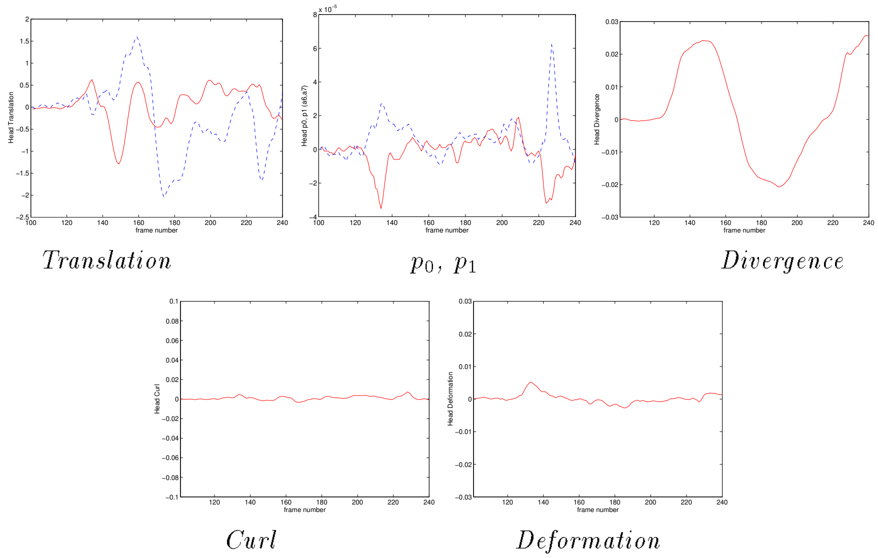


Figure 9. The Looming face motion parameters. Translation: solid = horizontal, dashed = vertical. Quadratic terms: solid = p_0, dashed = p_1.

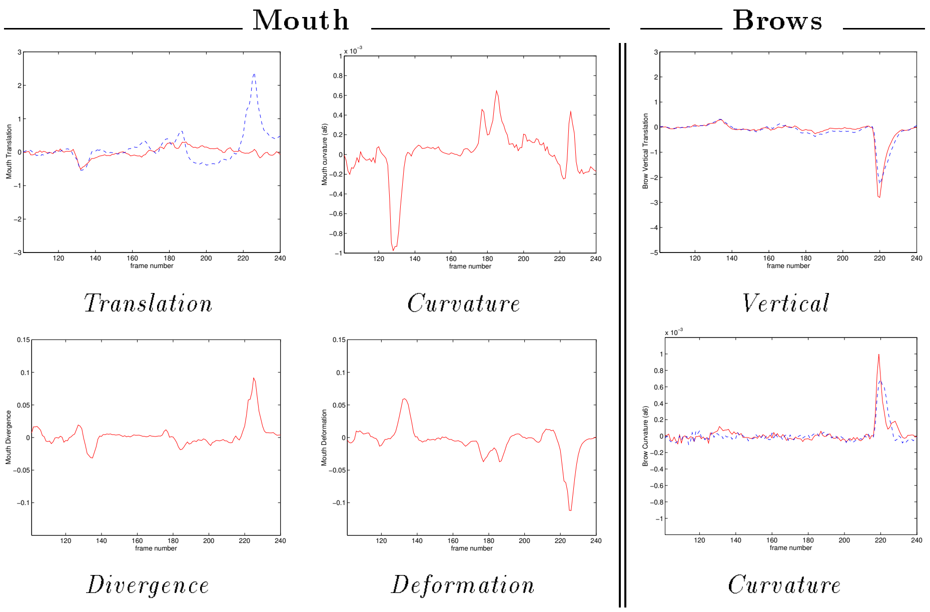


Figure 10. The Looming sequence. Mouth translation: solid and dashed lines indicate horizontal and vertical motion respectively. For the brows, the solid and dashed lines indicate left and right brows respectively.

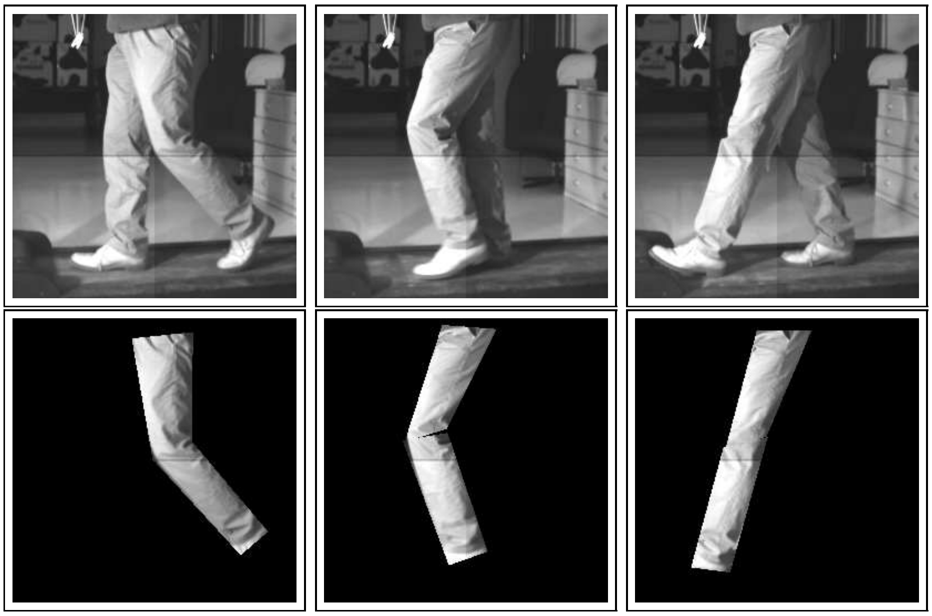
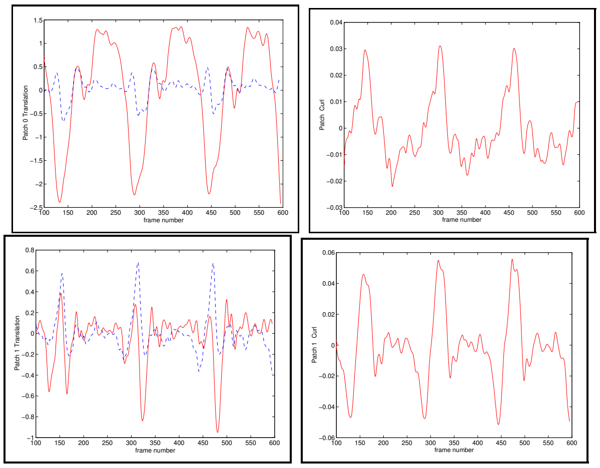


Figure 11. Walking parallel to the imaging plane.

Thigh:

Calf:



Translation Curl

Figure 12. Walking parallel to the imaging plane; motion parameters (translation and curl) over 500 frames. For translation, the dashed line indicates vertical translation while the solid line indicates horizontal translation.

show the motion of the thigh while the second rows show the motion of the calf. These graphs are only meant to demonstrate the effectiveness of our tracking model and its ability to capture meaningful parameters of the body movement.

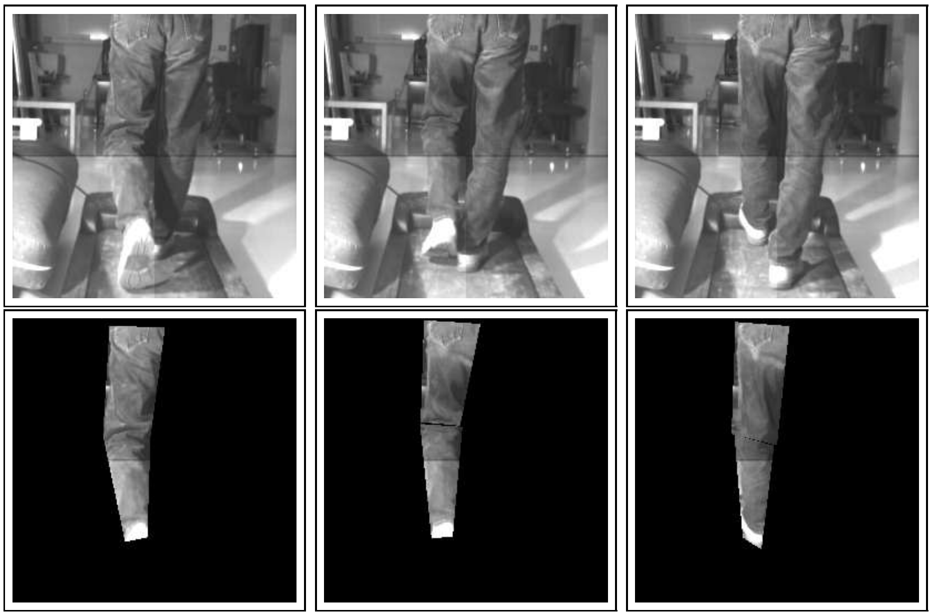


Figure 13. Walking perpendicular to the imaging plane.

In Figure 12 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". Note that the motion of these regions is described by a combination of translation and rotation because the motion is defined with respect to the center of the regions. A different choice of region center would result in different plots.

When a person is walking away from the camera as shown in Figure 14 the significant parameters which capture the cyclic walking motion are deformation, divergence, and "image pitch." Notice that the "image pitch" measured at the two parts is always reversed since when the "thigh" rotates in one direction the "calf" is viewed to be rotating in an opposite way.

In summary, the reported experiments show that the image motion models are capable of tracking rigid, deformable and articulated motion quite accurately over long sequences and recovering a meaningful set of parameters that can be exploited by a recognition system.

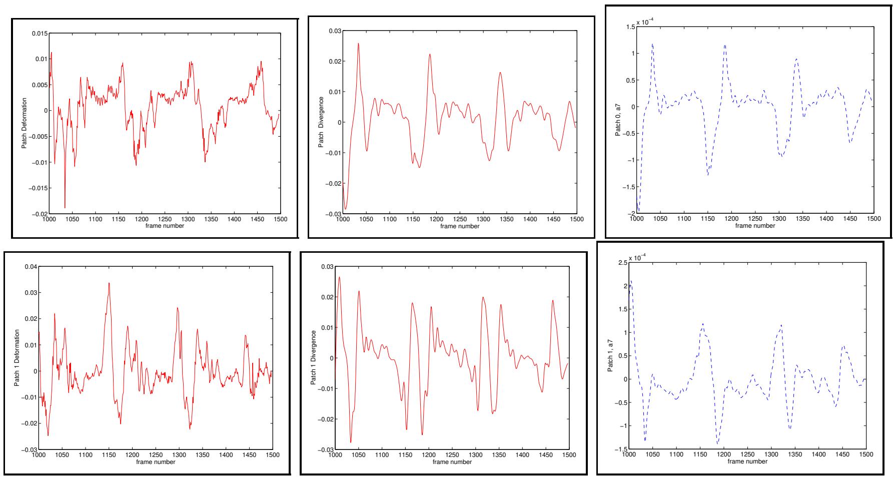
7. Recognition of Movement

Recognizing human movement requires answering:

– When does the activity begin and end?

Thigh:

Calf:



Deformation Divergence Image Pitch

Figure 14. Tracking results for Figure 13

- What class does the observed activity most closely resemble?
- What is the period (if cyclic) of the activity?

The answers to these questions involve spatio-temporal reasoning about a large parameter space. Our choice of parameterized motion models for tracking the diverse motions of human body parts dictates that recognition be considered as analysis of the spatio-temporal curves and surfaces that are created by the parameter values of each part. These curves are based on 2D image motion and therefore change with viewing angle. This leads us to formulate a view-based recognition strategy.

There are general issues to consider for a recognition approach:

- Cyclic actions require that enough cycles have been observed before recognition becomes possible. This leads to a focus on the coarse motions measured over a long sequence. Allmen and Dyer [1] proposed a method for detection of cyclic motions from their spatio-temporal curves and surfaces while Rangarajan and Shah [29] used a scale-space approach to match motion trajectories.
- The view-point of observing the human motion may affect the recognition of the activity (for an extreme case, consider recognition of human "walking" from a top view of the person). Campbell and Bobick [9] proposed a phase-space representation for recognition of human body motion from Moving Light Displays (MLD) capturing the full 3D articulations of the body parts. If only 2D motion is measured the viewpoint plays a critical role in recognition.
- Self-occlusions are quite prevalent in human movement. Capturing and representing these self-occlusions is a complex task even in the presence

of multiple cameras and availability of full 3D models. In our preliminary recognition work we do not capture and represent these self-occlusions, instead we focus on body parts that are visible throughout the activity.

Seitz and Dyer [34] 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 ${C_v}^{i_j}(t)$

$$ {D_v}^{i_j}(t) = S_i * {C_v}^{i_j}(t + T_i) \tag{15} $$

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

$$ {D_v}^{i_j}(t) = S_i * {C_v}^{i_j}(\alpha_i t + T_i) \tag{16} $$

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)

$$ E(v) = \sum_{j \in 0..7} \rho[{D_v}^{i_j}(t) - (S_i * {C_v}^{i_j}(\alpha_i t + T_i)), \sigma] \tag{17} $$

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

$$\min_{v} \sum_{j \in 0..7} \rho[D_v{}^{i_j}(t) - (S_i * C_v{}^{i_j}(\alpha_i t + T_i)), \sigma] \tag{18}$$

Recognition over multiple body parts uses the inter-part hierarchy relationships to progressively find the best match. As demonstrated and discussed in Section 6, 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 $C_v{}^{i_j}$ 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 $D^{i_j}(t)$ (notice that the v is dropped since it is unknown), our approach determines the characteristic curve that minimizes the error function given in Equation 18 by employing the recently proposed affine eigentracking approach [6] on the curves.

We are currently constructing these characteristic curves for several human activities. It is worth noting that, depending on the spatio-temporal complexity of the observed activity, simpler models could be used for recognition. For example in the case of facial expressions the activity can be simply captured by the model first proposed in [39] and used in [7]. Each expression was divided into three temporal segments: the *beginning*, *apex* and *ending*. Figure 15 illustrates qualitatively the different aspects of detecting and segmenting a "smile." In this Figure the horizontal axis represents the time dimension (i.e., the image sequence), the axis perpendicular to the page represents each one of the parameters relevant to a "smile" (i.e., a_3, *Divergence*, *Deformation*, and c) and the vertical axis represents the values of these parameters. This diagram is an abstraction to the progression of a "smile," therefore the parameter values are not provided. Notice that Figure 15 indicates that the change in parameter values might not occur at the same frames at either the beginning or ending of actions, but it is required that a significant overlap be detectable to label a set of frames with a "beginning of a smile" label, while the motions must terminate before a frame is labeled as an "apex" or an "ending."

The detailed development of the "smile" model is as follows. The upward and outward motion of the mouth corners results in a negative curvature of the mouth (i.e., the curvature parameter c is negative). The horizontal and overall vertical stretching are manifested by positive divergence (Div) and deformation (Def). Finally, some overall upward translation is caused by

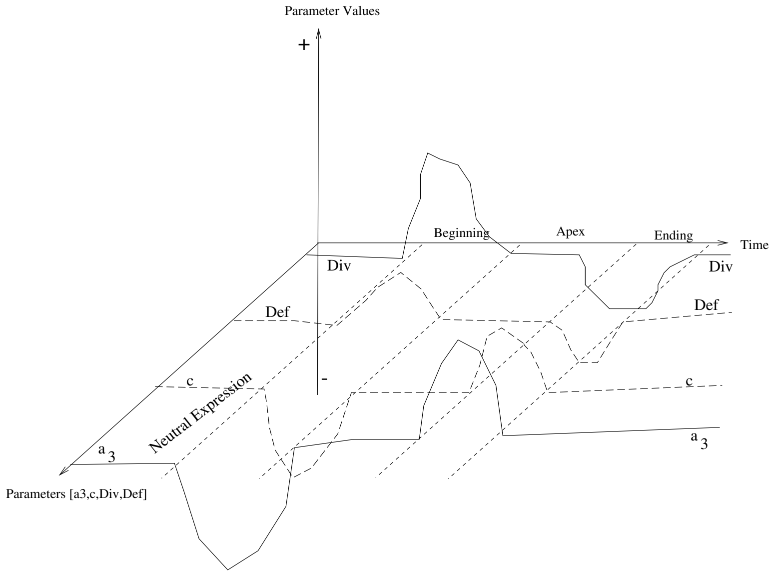


Figure 15. The temporal model of the "smile" expression.

the raising of the lower and upper lips due to the stretching of the mouth (a_3 is negative). Reversal of these motion parameters is observed during the ending of the expression.

The results of applying this recognition model to face expressions can be seen in Figure 16. Figure 16 shows the beginning of a "smile" expression while the head is rotating initially leftward and then rightward. The text that appears on the left side of each image represents a discrete interpretation of the underlying curves in terms of mid-level predicates which describe the facial motion [7]. Similarly, the text that appears on the right side represents the mid-level predicates of the head motion. The text below each image displays the recognized high-level description of the facial deformations and the head motions.

Figure 17 shows the recognition of head gestures based on the face motion recovery using a planar model. The gesture is recognized using the "curl" of the face. Other gestures were recognized in [25].

8. Discussion

We have demonstrated the use of parameterized optical flow methods for tracking and recognizing facial expressions and articulated motion. While the approach shows promise, there are a number of issues that still need to be addressed. First, the motion of human limbs in NTSC video (30

Figure 16. Four frames (four frames apart) of the beginning of a "smile" expression.

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 we have not addressed 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 17. Four frames (four frames apart) of the recognition of a head gesture signifying the expression of “more-or-less”.

9. Conclusion

We have presented a method for tracking non-rigid and articulated human motion in an image sequence using parameterized models of optical flow and have shown how this representation of human motion can support the recognition of facial expressions and articulated motions. Unlike previous work on recovering human motion, this method assumes that the activity can be described by the motion of a set of parameterized patches (e.g. affine, planar, etc.). In the case of facial motion, the motion of facial features is estimated relative to the motion of the face. For the articulated motion of limbs we add an additional articulation constraint between neighboring patches. 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 [7]. 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 non-rigid and articulated human motion into parts and the development of robust view-based recognition schemes for articulate motion.

References

1. M. Allmen and C.R. Dyer. Cyclic motion detection using spatiotemporal surfaces and curves. In *ICPR*, pages 365–370, 1990.
2. A. Azarbayejani, T. Starner, B. Horowitz, and A. Pentland. Visually controlled graphics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):602–604, June 1993.
3. A. Baumberg and D. Hogg. Learning flexible models from image sequences. In J. Eklundh, editor, *European Conf. on Computer Vision, ECCV-94*, volume 800 of *LNCS-Series*, pages 299–308, Stockholm, Sweden, 1994. Springer-Verlag.
4. A. G. Bharatkumar, K. E. Daigle, M. G. Pandy, and J. K. Aggarwal. Lower limb kinematics of human walking with the medial axis tranfromation. In *Proceedings of the Workshop on Motion of Non-rigid and Articulated Objects*, pages 70–76, Austin, Texas, November 1994.
5. 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.
6. M. J. Black and A. D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. In B. Buxton and R. Cipolla, editors, *European Conf. on Computer Vision, ECCV-96*, volume 1064 of *LNCS-Series*, pages 329–342, Cambridge, UK, 1996. Springer-Verlag.
7. M. J. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motions. In *Proceedings of the International Conference on Computer Vision*, pages 374–381, Boston, Mass., June 1995.
8. A. Blake and M. Isard. 3D position, attitude and shape input using video tracking of hands and lips. In *Proceedings of SIGGRAPH 94*, pages 185–192, 1994.
9. L.W. Campbell and A.F. Bobick. Recognition of human body motion using phase space constraints. In *Proceedings of the International Conference on Computer Vision*, pages 624–630, Boston, Mass., June 1995.
10. C. Cédras and M. Shah. Motion-based recognition: A survey. *Image and Vision Computing*, 13(2):129–155, March 1995.
11. R. Cipolla and A. Blake. Surface orientation and time to contact from image divergence and deformation. In G. Sandini, editor, *Proc. of Second European Conference on Computer Vision, ECCV-92*, volume 588 of *LNCS-Series*, pages 187–202. Springer-Verlag, May 1992.
12. François Dagognet. *Etienne-Jules Marey: A Passion for the Trace*. Zone Books, New York, 1992.
13. I. Essa, T. Darrell, and A. Pentland. Tracking facial motion. In *Proceedings of the Workshop on Motion of Non-rigid and Articulated Objects*, pages 36–42, Austin, Texas, November 1994.
14. I. A. Essa and A. Pentland. A vision system for observing and extracting facial action parameters. In *Proc. Computer Vision and Pattern Recognition, CVPR-94*, pages 76–83, Seattle, WA, June 1994.
15. D. Gavrila and L.S. Davis. 3d model-based tracking of humans in action: A multi-view approach. In *Proc. Computer Vision and Pattern Recognition, CVPR-96*, San Francisco, CA, June 1996.
16. L. Goncalves, E. Di Bernardo, E. Ursella, and P. Perona. Monocular tracking of

the human arm in 3D. In *Proceedings of the International Conference on Computer Vision*, pages 744–770, Boston, Mass., June 1995.

17. 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.
18. S. X. Ju, M. J. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated motion. In *International Conference on Automatic Face and Gesture Recognition*, pages 38–44, Killington, Vermont, 1996.
19. M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. In *Proc. First International Conference on Computer Vision*, pages 259–268, June 1987.
20. J. J. Koenderink and A. J. van Doorn. Invariant properties of the motion parallax field due to the movement of rigid bodies relative to an observer. *Optica Acta*, 22(9):773–791, 1975.
21. H. Li, P. Roivainen, and R. Forcheimer. 3-D motion estimation in model-based facial image coding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):545–555, June 1993.
22. N. Li, S. Dettmer, and M. Shah. Lipreading using eigensequences. In *International Workshop on Automatic Face and Gesture Recognition*, pages 30–34, Zurich, 1995.
23. J. Little and J. Boyd. Describing motion for recognition. In *International Symposium on Computer Vision*, pages 235–240, Miami, FL, November 1995.
24. K. Mase. Recognition of facial expression from optical flow. *IEICE Transactions*, E 74:3474–3483, 1991.
25. C. Morimoto, Y. Yacoob, and L.S. Davis. Recognition of head gestures using Hidden Markov Models. In *Proceedings of the International Conference on Pattern Recognition*, Vienna, Austria, 1996.
26. Eadweard Muybridge. *The Human Figure in Motion*. Dover Publications, Inc., New York, 1955.
27. S. A. Niyogi and E. H. Adelson. Analyzing and recognizing walking figures in xyt. In *Proc. Computer Vision and Pattern Recognition, CVPR-94*, pages 469–474, Seattle, WA, June 1994.
28. A. Pentland and B. Horowitz. Recovery of nonrigid motion and structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):730–742, July 1991.
29. K. Rangraajan, W. Allen, and M.A. Shah. Matching motion trajectories using scale-space. *Pattern Recognition*, 26(4):595–609, 1993.
30. J. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *Proceedings of the International Conference on Computer Vision*, pages 612–617, Boston, Mass., June 1995.
31. K. Rohr. Towards model-based recognition of human movements in image sequences. *CVGIP: Image Processing*, 59:94–115, 1994.
32. M. Rosenblum, Y. Yacoob, and L.S. Davis. Human emotion recognition from motion using a radial basis function network architecture. In *Proceedings of the Workshop on Motion of Non-rigid and Articulated Objects*, Austin, Texas, November 1994.
33. H. S. Sawhney. 3D geometry from planar parallax. In *Computer Vision and Pattern Recognition, CVPR-94*, pages 929–934, Seattle, WA, 1994.
34. S.M. Seitz and C.R. Dyer. Affine invariant detection of periodic motion. In *Proc. Computer Vision and Pattern Recognition, CVPR-94*, pages 970–975, Seattle, WA, June 1994.
35. T. Starner and A. Pentland. Visual recognition of American Sign Language using Hidden Markov Models. In *International Workshop on Automatic Face and Gesture Recognition*, pages 189–194, Zurich, 1995.
36. D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):569–579, June 1993.
37. S. Toelg and T. Pogio. Towards an example-based image compression architecture

for video-conferencing. Technical Report CAR-TR-723, Center for Automation Research, U. of Maryland, 1994.

38. J. Wang, G. Lorette, and P. Bouthemy. Analysis of human motion: A model-based approach. In *7th Scandinavian Conf. Image Analysis*, Aalborg, Denmark, 1991.
39. Y. Yacoob and L.S. Davis. Computing spatio-temporal representations of human faces. In *Proc. Computer Vision and Pattern Recognition, CVPR-94*, pages 70–75, Seattle, WA, June 1994.
40. A.. L. Yuille, D. S. Cohen, and P. W. Hallinan. Feature extraction from faces using deformable templates. In *Proc. Computer Vision and Pattern Recognition, CVPR-89*, pages 104–109, June 1989.