To Appear in ICCV-98, Mumbai-India, Subject to IEEE Copy-Rights

Learned Temporal Models of Image Motion

Yaser Yacoob and **Larry Davis**
Computer Vision Laboratory, University of Maryland, College Park, MD 20742

yaser/lsd@umiacs.umd.edu

Abstract

An approach for learning and estimating temporal-flow models from image sequences is proposed. The temporal-flow models are represented as a set of orthogonal temporal-flow bases that are learned using principal component analysis of instantaneous flow measurements. Spatial constraints on the temporal-flow are also developed for modeling the motion of regions in rigid and coordinated motion. The performance of these models is demonstrated on several long image sequences of rigid and articulated bodies in motion.

1 Introduction

Tracking the image motion of a human body in action is an exceptionally challenging computer vision problem. Even ignoring the fine structure of the hands, and assuming that the feet are rigidly connected to the calves and the hands to the forearms, a human body is composed of ten basic parts, many of which can move in quite independent ways. Natural human motions, such as walking, kicking, etc., are, of course, very constrained by factors including motion symmetries, static and dynamic balance requirements, gravity, etc. A physics-based approach to analysis of human motion might involve locating and tracking the limbs and extremities of the body under control of a mechanism that optimizes the tracking with respect to known physical constraints. This turns out to be a rather daunting enterprise, due to the difficulties of identifying body parts in natural video imagery and the challenges of developing efficient computational methods for modeling and enforcing such physical constraints.

An alternative approach is to develop "appearance-based" models for human motion, and use these models to control tracking of human motion. Examples of this approach include [3, 9]. The main challenges to such appearance-based methods are viewpoint dependence, dealing with appearance variability (due to changes in clothing, shadowing, body size and proportions between individuals), recognition in the presence of occlusion, etc.

In this paper we show how low-dimensional appearance-based models of articulated human motion can be recovered from observations of such motions, and how these models can be used to track the motions of other humans performing similar motions. We present some experimental evidence that suggests that the number of viewpoint-dependent appearance models that one would need to model a given motion is not overwhelming (see also the discussion in [7]), and also show how these models can be employed even in conditions when there is partial/full occlusion of some of the body parts (specifically, we demonstrate an ability to track both legs in motion from viewpoints in which one leg occludes part of the other).

The appearance models are created by applying a standard principal components analysis to time sequences of parametric models of body part motion. These observations are obtained using the "cardboard body" model introduced in [8] which employs the simple constraint that the motion of body parts must agree at the joints where those parts meet. Much of the analysis is carried out in a multi-temporal optical flow framework described in [10], which is crucial for analyzing time-varying images of humans since the instantaneous motions of body parts can span a broad spectrum of magnitudes, from sub-pixel to many pixels per frame. The flow models employed there were based on either constant flow or constant acceleration within the temporal integration window. For most human motions this assumption will not hold as the temporal window is enlarged. We propose, instead, using learned motion models to bridge the gap between traditional instantaneous flow estimation and multi-frame motion estimation. These learned motion models are then used in a spatio-temporally constrained image-motion formulation for simultaneous estimation of several rigid and non-rigid motions.

*The support of the Defense Advanced Research Projects Agency (ARPA Order No. C635), the Office of Naval Research (grant N000149510521) is gratefully acknowledged.

2 A Temporal Model for Image Motion

In the following we employ two temporal variables s and t. The global time t denotes time relative to the beginning of the image sequence while s denotes time relative to the time instant t. Let $I(x,y,t)$ be the image brightness at a point (x,y) at time t. The brightness constancy assumption of this point at a subsequent time s, $s = 1, ..., \mathrm{n}$, is given by

$$I(x,y,t) = I(x+\sum_{j=1}^{s} u(j), y+\sum_{j=1}^{s} v(j), t+s) \quad (1)$$

where $(u(j), v(j))$ is the horizontal and vertical instantaneous image velocity of the point (x,y) between frames $(t+j-1)$ and $(t+j)$. Let $(u,v) = (u(0), v(0))$ denote the instantaneous flow at time t. The special cases where $(u(j), v(j))$ are constant for all j or satisfy a constant acceleration model relative to t were considered in [10]:

$$u(j) = x_0 + x_1 j \quad (2)$$
$$v(j) = x_2 + x_3 j \quad (3)$$

throughout the period $\mathrm{n}\delta t$ (where n is the number of time instants and δt is the time increment-usually $\delta t = 1$). Such flow models are unlikely to hold over long intervals $n\delta t$. In the following, we develop a more general model in which the flow is a "learned" function of time $(u(j), u(j))$. Let the range of time over which temporal-flow (sequences of instantaneous flow) is estimated be $1, .., \mathrm{n}$. Expanding Equation (1) using a Taylor series approximation (assuming smooth spatial and temporal intensity variations) and dropping terms results in

$$0 = {I^s}_x(x,y,t)\sum_{j=1}^{s} u(j) + {I^s}_y(x,y,t)\sum_{j=1}^{s} v(j) + s{I^s}_t(x,y,t) \quad (4)$$

where I^s is the s-th frame (forward in time relative to I) of the sequence, and ${I^s}_x, {I^s}_y$ and ${I^s}_t$ are the spatial and temporal derivatives of image I^s relative to I.

Since Equation (4) is underconstrained for the recovery of $(u(j), v(j))$, the estimation of $(u(j), v(j))$ is ordinarily posed as an error minimization using a robust error norm, $\rho(\mathbf{x}, \sigma_e)$, that is a function of a scale parameter σ_e. Then, the error of the flow over a very small neighborhood, R, of (x,y) is,

$$E(u,v,s) = \sum_{(x,y)\in R} \rho({I^s}_x\sum_{j=1}^{s} u(j) + {I^s}_y\sum_{j=1}^{s} v(j) + s{I^s}_t, \sigma_e) \quad (5)$$

We have n equations of the form of Equation (5), one for each time instant. The *time-generalized* error is defined as

$$E_D(u,v) = \sum_{s=1}^{\mathrm{n}}\sum_{(x,y)\in R} \rho({I^s}_x\sum_{j=1}^{s} u(j) + {I^s}_y\sum_{j=1}^{s} v(j) + s{I^s}_t, \sigma_e) \quad (6)$$

Equation (6) gives equal weight to the error values at all subsequent time instants. Since it is expected that at each point (x,y) the accuracy of instantaneous motion estimation will temporally vary[2], we introduce a weight function $W(u,v,s)$ designed to minimize the influence of residuals of the relatively inaccurate time instants. Equation (6) now becomes

$$E_D(u,v) = \sum_{s=1}^{\mathrm{n}}\sum_{(x,y)\in R} \rho(W(u,v,s)({I^s}_x\sum_{j=1}^{s} u(j) + {I^s}_y\sum_{j=1}^{s} v(j) + s{I^s}_t), \sigma_e) \quad (7)$$

Since the weight function $W(u,v,s)$ could also reflect the degree of accuracy of the flow estimation, we redefine it to include a scaling parameter σ_w, $W(u,v,s,\sigma_w)$. The weighting function W was designed [10] to satisfy the following constraints:

- Take on values in the range $[0..c]$, c typically chosen as 1.0 for computational convenience.
- For a large σ_w, W approaches 1.0 regardless of (u,v) and s.
- Given σ_w, larger estimated flow (u,v) at point (x,y) leads to higher weights for the lower time instants of the error term ${I^s}_x\sum_{j=1}^{s} u(j) + {I^s}_y\sum_{j=1}^{s} v(j) + s{I^s}_t$, while a small flow leads to higher weights of the highest time instants.

The following Gaussian function, proposed in [10], satisfies the above requirements

$$W(u,v,s,\sigma_w) = e^{-(s-\frac{n}{(\alpha||(u,v)||^2+1.0)})^2/2{\sigma_w}^2} \quad (8)$$

where $||(u,v)||^2$ is the squared magnitude of the current flow estimate at (x,y), and α is a constant. Figure 1 qualitatively reflects the shape of the weighting function for a fixed σ_w. It illustrates the weighting as a function of time, s, and flow magnitude, $||(u,v)||$, at (x,y). Notice that when $||(u,v)||^2 << 1.0$ the maximal weight occurs at the highest time n, while higher values of $||(u,v)||^2$ lead to a maximal weight at lower

[2] This time-dependence is due to the possible existence of wide disparities in the flow field.

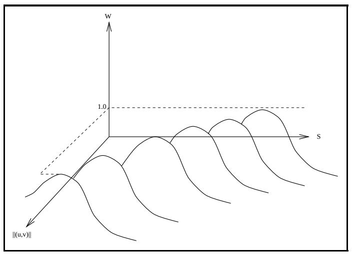


Figure 1: The weighting function as a function of s and flow magnitude $||(u,v)||$

time instants; specifically the Gaussian is centered at $\frac{n}{\alpha||(u,v)||^2+1.0}$. The scale parameter σ_w determines the width of the Gaussian, and the constants α and 1.0 can be changed to further shift the maximal weight location. The application of the weighting function in the estimation is as follows: in the first iteration, all time instants are given equal weight (1.0) by selecting a large σ_w. Afterwords, iteratively, the estimates are refined by decreasing σ_w.

This multi-temporal procedure is accompanied by a spatial coarse-to-fine strategy [2] that constructs a pyramid of the spatially filtered and sub-sampled images (for more information see [4]) and computes the flow initially at the coarsest level and then propagates the results to finer levels.

3 Learning Temporal-Flow Models

Temporal-flow models are constructed by applying principal component analysis to exemplar flow sequences. So, the functions of $(u(s), v(s))$ for $s = 1...\mathrm{n}$ are approximated by a linear combination of a *temporal-flow* basis-set of $1 \times 2 * \mathrm{n}$ vectors, U_i. The flow vector $(u(s), v(s))$ can be reconstructed as a component from $(u(\mathrm{n}), v(\mathrm{n}))$ using

$$\bar{e} = \sum_{i=1}^{q} c_i U_{i,j} \tag{9}$$

where $\bar{e}$, the temporal-flow vector, denotes the concatenation of $u(\mathrm{n})$ and $v(\mathrm{n})$ and c_i is the expansion coefficient of the U_i-th temporal-flow basis vector and q is the number of vectors used as the basis-set.

Equation (7) can now be expressed as:

$$\begin{aligned} E_D(u,v) &= \sum_{s=1}^{\mathrm{n}} \sum_{(x,y)\in R} \rho(W(u,v,s)([I^s{}_x \;\; I^s{}_y] \\ & [\sum_{j=1}^{s}\sum_{i=1}^{q} c_i U_{i,j}, \sum_{j=\mathrm{n}+1}^{\mathrm{n+s}}\sum_{i=1}^{q} c_i U_{i,j}]^T + sI^s{}_t), \sigma_e) \end{aligned} \tag{10}$$

where $[\,]^T$ is the transpose of the temporal-flow vector. Notice that the summation of the linear combination includes only the s values of u and v. Equation (10) essentially describes how image motion of a point (x, y) changes over time under the constraint of a temporal-flow basis-set.

Purely spatial constraints on image motions were recently proposed by Black et al. [6]. There, a low dimensional representation of the spatial distribution of image motions in a region was learned and used in recovering motion in image sequences. This spatial model provides only an instantaneous constraint on flow. In comparison, the temporal-flow models described above express how flow changes over time at (for the moment) a single point. In the subsequent section we explain how our temporal-flow model can be extended to include spatial constraints as well.

The temporal-flow basis-set is computed during a learning stage in which examples of the specific image-motions are subjected to principal component analysis. Specifically, let $(u^i(s), v^i(s))$ for $s = 1, ..., \mathrm{n}$ be the $i-$th instance (out of N instances) of an incremental flow series measured for an image point (x, y) at time instants $s = 1, ..., \mathrm{n}$. The estimation of $(u^i(s), v^i(s))$ can be carried out either using the multi-scale approach proposed in [10] or by direct two-frame flow estimation technique.

Let $\bar{e}^i$ be the vector obtained by concatenating $u^i(s)$ for $s = 1, ..., \mathrm{n}$ and $v^i(s)$ for $s = 1, ..., \mathrm{n}$. The set of vectors $\bar{e}^i$ can be arranged in a matrix A of N rows by $2 * \mathrm{n}$ columns. Matrix A can be decomposed using Singular Value Decomposition (SVD) as

$$A = U\Sigma V^T \tag{11}$$

where U is an orthogonal matrix of the same size as A representing the principal component directions in the training set. Σ is a diagonal matrix with singular values $\sigma_1, \sigma_2, ..., \sigma_\mathrm{N}$ sorted in decreasing order along the diagonal. The $\mathrm{N} \times \mathrm{N}$ matrix V^T encodes the coefficients to be used in expanding each column of A in terms of principal component directions. It is possible to approximate an instance of flow sequence $\bar{e}$ using the largest q singular values $\sigma_1, \sigma_2, ..., \sigma_q$, so that

$$\bar{e}^* = \sum_{l=1}^{q} c_l U_l \tag{12}$$

where $\bar{e}^*$ is the vector approximation, c_l are scalar values that can be computed by taking the dot product of $\bar{e}$ and the column U_l. In effect this amounts to projecting the vector $\bar{e}^*$ onto the subspace defined by the q basis vectors. The projection can also be viewed as a parameterization of the vector $\bar{e}$ in terms of the basis vectors U_l $(l = 1..q)$ where the parameters are the c_l's.

4 Parameterized Spatio-Temporal Image-Motion

Recently, it has been demonstrated that spatially parameterized flow models are a powerful tool for modeling instantaneous image motion ([5, 6, 8]). The temporal-flow learning and estimation algorithms can be extended to spatially parameterized models of image flow. In this section we describe the learned estimation of polynomial parameterized image motion models.

Recall that the flow constraint given in Equation (4) assumes constant flow over a small neighborhood around the point (x, y). Over larger neighborhoods, a more accurate model of the image flow is provided by low-order polynomials [1]. For example, the planar motion model [1] is an approximation to the flow generated by a plane moving in 3-D under perspective projection. The model is given by

$$\begin{aligned} U(x,y) &= a_0 + a_1 x + a_2 y + a_6 x^2 + a_7 xy \quad (13)\\ V(x,y) &= a_3 + a_4 x + a_5 y + a_6 xy + a_7 y^2 \quad (14) \end{aligned}$$

where a_i's are constants and (U, V) is the instantaneous velocity vector. The affine model is the special case where $a_6 = a_7 = 0$ and generally holds when the region modeled is not too large or subject to significant perspective effects. Equation (14) can be written in matrix form as

$$[UV]^t = \mathbf{X}\mathbf{P}^T \quad (15)$$

where

$$\mathbf{X}(x,y) = \begin{bmatrix} 1 & x & y & 0 & 0 & 0 & x^2 & xy & 0 \\ 0 & 0 & 0 & 1 & x & y & xy & y^2 & x^2 \end{bmatrix},$$

$$\mathbf{P} = \begin{bmatrix} a_0 & a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & 0 \end{bmatrix}^T$$

To exploit the economy of parameterized models, we re-formulate the temporal-flow models to learn the temporal evolution of the parameters of the planar model as opposed to only the flow values. Specifically, consider the parameters a_i to be a function of s (similar to the flow formulation), so that

$$\mathbf{P}(s) = [a_0(s) \ \ a_1(s) \ \ a_2(s) \ \ a_3(s) \ \ a_4(s) \ \ a_5(s) \ \ a_6(s) \ \ a_7(s) \ \ 0]^T$$

where $\mathbf{P}(s)$ is the image motion parameters computed between time instant $s-1$ and s.

Equation (7) can be rewritten as

$$E_D(u,v) = \sum_{s=1}^{\mathrm{n}} \sum_{x,y \in R} \rho(W(u,v,s)([I^s{}_x I^s{}_y]\mathbf{X}[\sum_{j=1}^{s}\mathbf{P}(j)]^T + sI^s{}_t), \sigma_e) \quad (16)$$

where R denotes the region over which the planar motion model is applied. Notice that the term $\sum_{j=1}^{s}\mathbf{P}(j)$ requires proper region registration between time instants. $\mathbf{P}(j)$, $j = 1, ..., \mathrm{n}$, can be represented by a linear combination of basis vectors in a manner similar to the temporal-flow representation developed earlier. Each basis vector, L_i is a vector of size $8 * \mathrm{n}$ since it generates the eight parameters for each time instant s. We can extract $\mathbf{P}(j)$ from the following elements of vector $\bar{e}$ which is the following sum of temporal-flow bases,

$$\bar{e} = [e]_{r=1,\ldots,8*n} = \sum_{i=1}^{q} c_i L_{i,r} \quad (17)$$

where c_i is the expansion coefficient of the L_i temporal-parameter basis vector. Equation (16) can now be rewritten as

$$E_D(u,v) = \sum_{s=1}^{\mathrm{n}} \sum_{(x,y)\in R} \rho(W(u,v,s)([I^s{}_x I^s{}_y]\mathbf{X}$$

$$[\sum_{j=1}^{s}\sum_{i=1}^{q} c_i L_{i,j}, ..., \sum_{j=7\mathrm{n}+1}^{7\mathrm{n}+s}\sum_{i=1}^{q} c_i L_{i,j}]^T + sI^s{}_t), \sigma_e) \quad (18)$$

The minimization of Equation (18) results in estimates for the parameters c_i. The choice of the weighting function W is somewhat more complex here than it was in the multi-scale temporal-flow. The weighting function can be designed using the current flow estimates computed by the model (U, V). This weighting leads to different weights within the region according to the magnitude of the flow, so that points where the flow estimate is low at later time instants are more dominant, while the larger flow estimates will determine the earlier time instants. Alternatively, W can be designed using the parameters of the model $a_i(s)$ (i.e., $W(\bar{a}, s, \sigma_w)$ where $\bar{a}$ is the set of model parameters). The former leads to a computation based on a weighted combination of spatio-temporal derivatives while the latter leads to a weighted combination of parametric models. Once a choice for the weighting function has been made, the computation of the parameters of the model follows the approach proposed in [4].

In the examples in this paper we adopt the weighted combination of parametric models. In the following examples the estimation of flow within a region is motivated by computing a particular motion of the region, therefore W is designed to be most sensitive to a particular subset of these parameters. For example, if the translation of the region is of most interest then the parameters a_0 and a_3 can be substituted as $||(a_0, a_3)||$ for $||(u, v)||$ in Equation (8).

The above treatment of polynomial flow is also applicable to the orthogonal-basis modeling of spatial flow recently proposed in [6]. The coefficients used in the linear combination replace the parameters a_i in the above equations.

5 A Rigid Motion Example

The use of a temporally parameterized motion model that explicitly accounts for image acceleration and is computed directly from image intensity variations simultaneously was discussed in [10]. Here, we demonstrate how a learned temporal-flow model can capture image acceleration by watching a book "falling" in an image sequence.

The learning of the temporal-flow model is performed as follows,

- The area corresponding to the book is manually segmented in the first frame in the sequence.
- The image motion parameters of this area are estimated for 40 frames assuming a planar model (flow estimation is carried out between consecutive images only).
- A basis set for the temporal-flow parameters is computed by taking four groups of 10 consecutive instantaneous flow vectors.
- The basis set is used to compute the coefficients using Equation (18) for the whole sequence (100 frames).

In this experiment the first eigenvalue captured 99.9% of the variation among the 4 data-sets as one might expect for such a uniform motion. Therefore, a single eigenvector is used in the motion estimation stage.

Figure 2 shows the results of tracking the book using the temporal-flow model. The graphs in the middle row show the value of $a_0(s)$ and $a_3(s)$ (for $s = 1...10$) of the eigenvector used in estimation. While $a_0(s)$ is a nearly zero (corresponding to little horizontal motion), the vertical motion component $a_3(s)$ is linear. The lower graph shows the estimated coefficient c_0 throughout the long image sequence. This coefficient grows linearly, which is what one would expect since the motion is second order (i.e., a constant acceleration model).

The learned spatio-temporal models can be applied to other objects performing similar motions. The temporal-flow basis-vector learned for the book is used to estimate the falling of a different object, a cardboard box. Figure 3 shows the images, the tracking results and the coefficient c_0 that is also recovered throughout the falling. Notice that despite the accurate translational tracking some counterclockwise rotation is recovered. This is not surprising since the motion of the book included a rotational component, while the box fell without rotation. We elaborate on the implications of this in a later section.

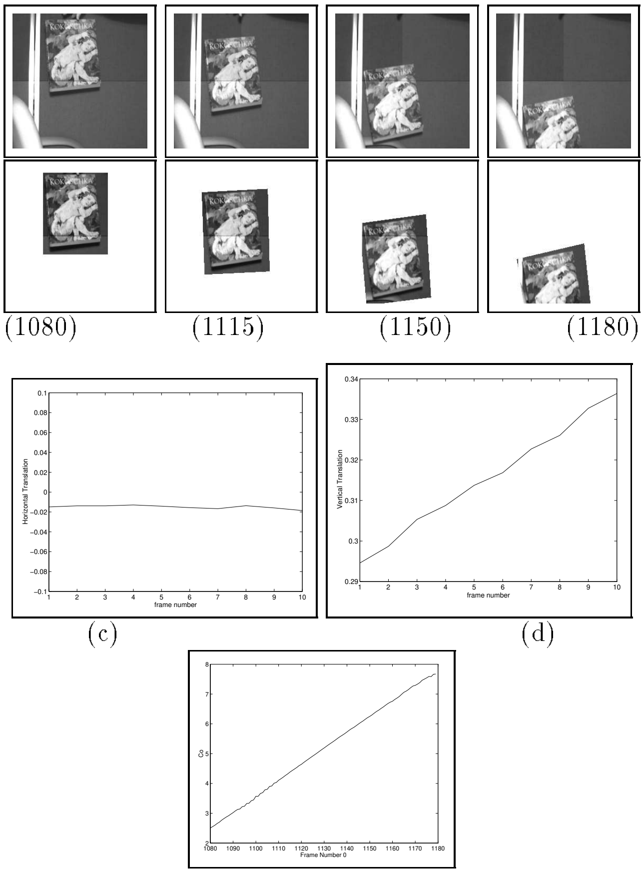
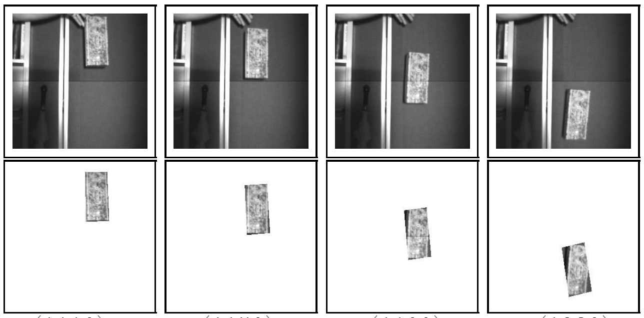


Figure 2: Four frames of a falling book tracked by a temporal-flow model (top rows), the horizontal and vertical velocities components of the learned basis-vector (third row) and the recovered expansion coefficient throughtout the sequence (bottom row).

It is worth noting that the motion trajectory of the box creates a line parallel (see Figure 3 bottom row) to the falling book's trajectory. Equation (18) minimizes the error within a subspace (of a single basis vector, in this case) in which the linear combinations of one line lead to parallel lines.

6 Learned Models of Articulated Human Motion

The *cardboard* [8] model for tracking five-part human movement (arm, torso, thigh, calf and foot) involves recovering 40 motion parameters; this requires substantial computation. Furthermore, due to the chain-like structure of the tracking, any error in the computation in an early part (in the chain structure) propagates to the succeeding parts. In the following, we use a camera with resolution 256×256 at 99Hz; this temporal sampling rate is high enough for us to effectively employ differential flow estimation. In most sequences the full



(1110) (1150) (1190) (1230)

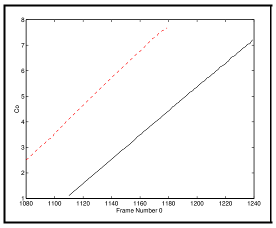


Figure 3: A sequence of falling box (top row), the tracked box (middle row) and the recovered temporal-flow coefficient throughout the sequence (solid line) and for comparison the temporal-flow coefficient for the falling book (dashed line)

human body is observed performing an activity; therefore, the image support for each body part is usually limited to a fairly small number of pixels.

Learning models of articulated motion can lead to much simpler representations in which redundancies are removed and motion couplings learned. A set of samples of the temporal-flow values of the parts of articulated object covering one entire period of an activity are modeled using principal component analysis. Applying this model to a new sequence of the articulated object motion requires temporally "registering" the model to the observation at the initial time t_0[3]. In the experiments presented in this section, we time-register sequences manually.

Similar to the accelerating book example, we assume initially that:

- The body is manually segmented into five parts in the first frame.
- People are moving at a similar viewing angle to the camera during the training and testing phases.

[3]Actually, both static and temporal information could be used to automate this initialization. Static information includes identifying specific configurations of body parts that unambiguously indicate the temporal stage of the activity (e.g., for walking, two feet on the ground, a straight leg just landed on the ground, etc.). Dynamic information includes exploiting knowledge about the motion of body parts during activity performance (e.g., the feet are not moving, etc.).

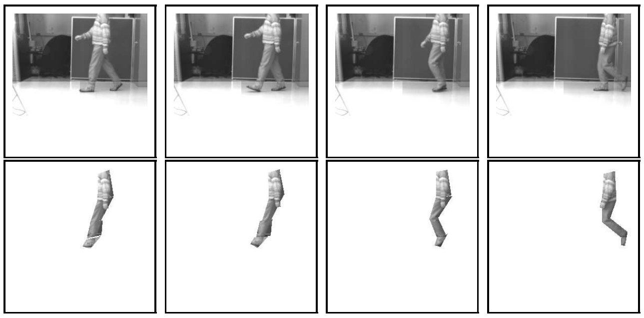


Figure 4: A few frames from a long image sequence of a subject walking with the cardboard tracking [8].

- A single activity, such as "walking," is learned and tracked. The stage of the "walking" cycle of the first frame of the sequence is known.

Learning of a the "walking" cycle temporal-flow model is performed by first employing the algorithm of Ju et al. [8] to compute each region's instantaneous motion parameters during the observed cycle of the activity. Then, the motion paramters of the activity cycles of several people are used to derive the basis-set of temporal-flows of the activity. It is worth noting that although the basis-vectors are computed for a whole cycle of "walking" the instantaneous motion recovery is conducted using a small computation temporal window (typically 6-10 frames). The five parts are tracked using Equation (18), the body parts are considered as a single object with individual motion parameters for each part coordinated through the principal components model.

Figure 4 displays a few frames of a walking sequence from the training set of one subject with the five-part body tracking as in [8]. Notice that the tracking accumulates errors, some of which also appear in the temporal-flow tracking. In learning the model from ten people's gait[4], the first basis vector accounts for about 67% of the variations and reflects very clearly the "walking" cycle. The next 4 basis vectors capture about 23% of the variations and capture imaging, individual variations and some differencies in image acquisition conditions.

Figure 5 shows the results of tracking a new instance of walking of a subject using only the first basis-vector of the temporal-flow. It also shows the coefficient, c_0, recovered throughout the sequence ($n = 8$). Low image contrast leads to accumulation of tracking errors.

The learned temporal-flow models remain effective in tracking articulated motion even when distance from the camera and the viewpoint vary from the train-

[4]The distance and viewing direction in the training data was constant. The viewing direction was approximately fronto-parallel

ing set. The variation in distance introduces practical problems of optical flow estimation since the model was learned for a "distant" object from the camera, and the tracking is conducted at a closer distance; here, the non-rigid motion of clothing and stronger perspective effects are visible. Varying the viewpoint poses a more fundamental problem since the appearance of the activity changes as we move farther from the learned viewpoint. In the following figures we provide results in which the viewing angle is about 20 degrees off the fronto-parallel plane. In experiments, not shown here, in which the viewing angle was close to 45 degrees off the fronto-parallel plane, we observed that the calf and foot are not tracked well while the torso and thigh tracking was satisfactory. Moreover, the estimation process was observed to rely heavily on the correctly tracked torso and thigh, while the other parts were found to be nonconforming with respect to the temporal-flow model of walking.

Figures 6 and 7 show the tracking of walking over a long sequence, where the distance and viewing angle are different from those used in learning. Also, in Figure 7, a subject not part of the training set is performing the activity. This example shows tracking errors, especially at the body extremities, (note that most of these errors are due to learning errors from the original data-for example the enlargement of the foot area).

Learned temporal-flow of activities can also support tracking of partially occluded parts. We demonstrate the performance of our approach on sequences of two activities, walking and marching. Since in these activities the movement of the legs and arms is symmetric, once a model for the visible parts is learned it can be applied to the occluded ones. Specifically, we assume that the phase difference between the visible and occluded parts to be one half cycle. We also assume that the difference in distance between the legs and the camera are insignificant relative to the distance of the body from the camera. In the first frame we initialize the regions for ten body parts (when parts are occluded we simply hypothesize their locations). Then, we minimize Equation (18), where all ten regions are regarded as a single object with multiple motion parameters. Only the un-occluded pixels of each region are used in the motion recovery. Each activity model was learned separately from a single example of its performance.

The results of the tracking of a marching activity in a long sequence is shown in Figure 8. The two legs are tracked well despite some inaccuracies that are due to the learned model inaccuracies.

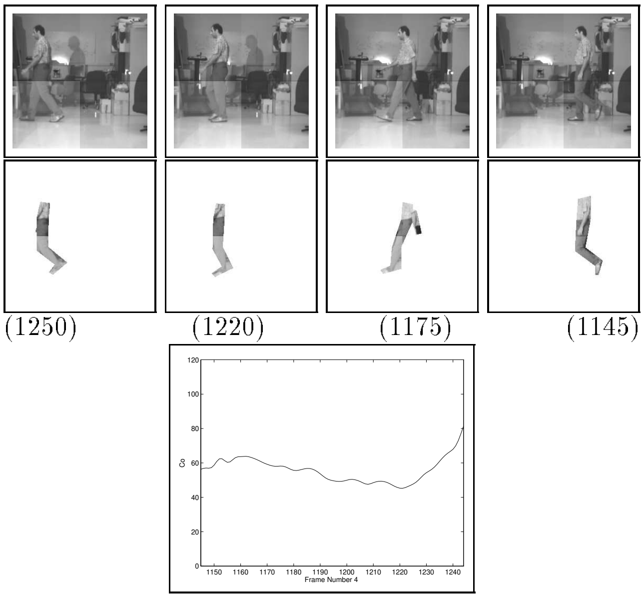


Figure 5: A few frames from a long image sequence of a subject walking with the temporal-flow tracking of a new subject's walk and the recovered coefficient.

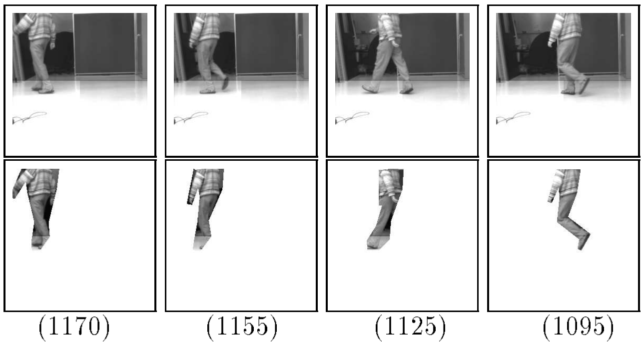


Figure 6: A few frames from a long image sequence of a subject walking as seen from a different viewing direction with the computed temporal-flow tracking.

7 Discussion

In this paper we proposed a new approach for image motion estimation from multiple frames that uses learned models of temporal-flow. Demonstration of the performance of the algorithm on both rigid and articulated motions were provided. An activity learned from one specific viewpoint was used to estimate the motion of a novel subject performing the same activity from a different but similar viewpoint. Also, it was demonstrated that the tracking of the occluded body parts is possible when a temporal model of one side of the body have been learned.

Learning plays a critical role in the accuracy of flow estimation. In our experiments on articulated motion, we observed that the inaccuracies of the cardboard model from [8] used to generate the training set for

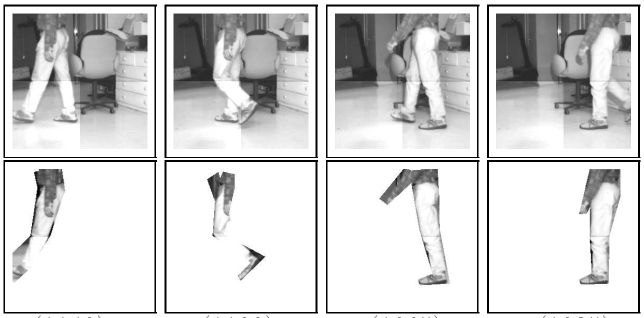


Figure 7: A few frames from a long image sequence of a subject walking with the computed temporal-flow tracking.

the learning algorithm lead to similar inaccuracies in the temporal-flow estimation. The tracking of the foot has been particularly problematic since in most image sequences it occupies a region of only about 30-100 pixels.

The learning of temporal-flow models of activities was performed independently for each activity considered (e.g., a separate model for each of walking, marching etc.). Subsequent body motion tracking would simultaneously employ all models to estimate the image motion, with the "best" model selected. How to reduce the number of temporal flow models, while minimizing the likelyhood that activities combined into one model do not result in tracking errors. (since only certain linear combinations of the basis functions would correspond to actual activities) is an open problem.

The temporal-flow estimation performs successfully even when the motions of some parts do not conform to the model, as long as a the majority of parts do conform. For example, in the case of walking, overall tracking is not disrupted if the arm is not moving in a manner consistent with its motion during the learning stage. Of course, the arm tracking fails, but the overall body tracking remains accurate. The current strong coupling between motions of body parts will be relaxed in future research to allow weaker motion couplings for certain parts.

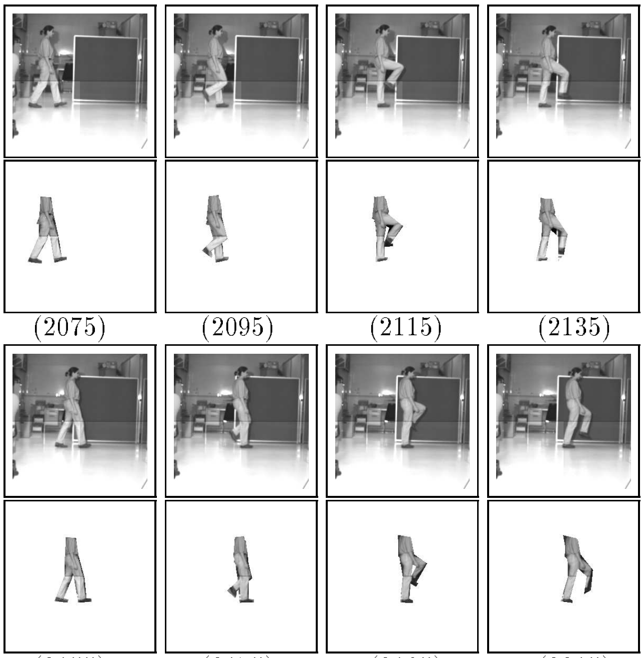


Figure 8: A few frames from a long image sequence of a subject marching with the temporal-flow tracking of a subject's marching for both the visible and occluded parts.

References

[1] Adiv G. *Determining three-dimensional motion and structure from optical flow generated by several moving objects.* IEEE PAMI, Vol. 7(4), 1985, 384-401.

[2] J.R. Bergen, P. Anandan, K.J. Hanna and R. Hingorani. *Hierarchical model-based motion estimation.* In G. Sandini, editor, ECCV-92, Vol. 588 of LNCS-Series, Springer-Verlag, 1992, 237-252.

[3] C. Bregler and S. Omohundro, Nonlinear Manifold Learning for Visual Speech Recognition, *ICCV 95*, 494-499.

[4] M. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1), 1996, 75–104.

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

[6] M. Black, Y. Yacoob, A. Jepson and D. Fleet, Learning Parameterized Models of Image Motion, *IEEE CVPR,* 1997, 561-567.

[7] A. Bobick and J. Davis. An appearance-based representation of action. *ICPR*, 1996, 307-312.

[8] S. X. Ju, M. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated image motion. in *Proc. Int. Conference on Face and Gesture*, Vermont, 1996, 561-567.

[9] D. Reynard, A. Wildenberg, A. Blake and J. Marchant, Learning Dynamics of Complex Motions from Image Sequences. *ECCV 96*, 357-368.

[10] Y. Yacoob and L. Davis, Temporal Multi-scale Models for Flow and Acceleration. In *IEEE CVPR 97*, 921-927.