Estimation of Composite Object and Camera Image Motion

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

Abstract

An approach for estimating composite independent object and camera image motions is proposed. The approach employs spatio-temporal flow models learned through observing typical movements of the object, to decompose image motion into independent object and camera motions. The spatio-temporal flow models of the object motion are represented as a set of orthogonal flow bases that are learned using principal component analysis of instantaneous flow measurements from a stationary camera. These models are then employed in scenes with a moving camera to extract motion trajectories relative to those learned. The performance of the algorithm is demonstrated on several image sequences of rigid and articulated bodies in motion.

1 Introduction

In recent years there has been increased interest in tracking and estimation of rigid and non-rigid object motion. Most approaches focused on tracking dynamic objects viewed from a stationary camera. In this paper we address the tracking and estimation of object motion while the camera itself is also moving. We propose an approach that employs a model for the composite motion of the object and camera to recover the original motion components.

Although it may be possible, in principal, to compute camera rigid motion first and then factor it out during object motion estimation (e.g., see a related example [10]), a recovery of the structure of both the scene and the object are necessary to decompose the flow over the object region into the object and camera motion components (this was not dealt with in [10]). This structure recovery is, itself, a very challenging problem. Furthermore, there exists situations where rigidity of scene structure does not hold (such as open textureless space, scene consisting entirely of multiple motions). In these situations this approach cannot be employed.

Composite object and self motion can be resolved by the human visual system equally well in textured or textureless static environments (e.g., ball catching indoors or in open-air while looking upward). This motivates us to explore the estimation of composite motion based only on the observed motion of the object region alone, and disregarding the (possibly unavailable) motion field due to the static environment.

We note that certain object or camera motions may lead to unresolvable ambiguities in composite motion estimates. For example, when one views a vehicle from a second moving vehicle (disregarding the static environment cues) it is ambiguous whether or not the observed vehicle is moving and with what direction or speed (i.e., the well known "train motion illusion").

Our approach makes extensive use of 2D region-based parameterized flow models which have been employed to recover rigid [3], deformable [4] and articulated [8, 12] object motions. We show how to extend these models to estimate composite object and camera image motions.

We make the following simplifying assumptions,

1. The decomposition of camera-object motion will be pursued without exploiting peripheral scene information. Therefore, we analyze the image motion over the object region only.
2. The moving object is observed "off-line" from a stationary camera while it performs its typical movements. This allows us to construct a *view-based* representation of these types of movements.
3. A 2D image motion estimation framework is used to describe both the object and the camera motions. As a result, the motion trajectory model of the object is view-point dependent. Therefore, only camera motions that do not significantly alter the appearance of the independent object motion can be recovered. (e.g., if the object is free falling, the camera cannot rotate by 90 degrees so that the object appears to move horizontally). We also assume that the appearance of the object does not change dramatically through the sequence due to the motion of the camera. For example, we assume that the camera motion is not so large as to make an initial frontal view of a person's walk become a parallel view.

4. The image region corresponding to the independently moving object is identified in the first frame of the image sequence, perhaps by an algorithm such as [7]. This region will be used for estimation of object and camera motion.

In summary, while these assumptions are somewhat restrictive, we propose and demonstrate a first step in addressing the estimation of composite motion. We demonstrate the performance of the approach on rigid and articulated bodies in motion.

2 Modeling Composite Motion

Let $P = (X, Y, Z)$ be an object point and $p = (x, y)$ be its projection on the image plane of the camera. Object motion leads to flow (u^o, v^o) at p. The motion of p is also affected by camera self motion. Let the flow resulting from the camera motion be (u^c, v^c); For the composite motion we have a brightness constancy

$$I(x, y, t) = I(x + u^c + u^o, y + v^c + v^o, t + 1). \qquad (1)$$

The estimation of u_c, u_o, v_c and v_o is underconstrained (one equation with four variables) and an infinite number of solutions exists unless constraints on object and camera motions are given. Employing a neighborhood-region flow constancy, as is typically done, does not allow us to separate the flow into its camera and object components.

Let $I(x, y, t), ..., I(x, y, t + \mathrm{n})$ be a sequence of $\mathrm{n} + 1$ images. The brightness constancy assumption for any time instant $s, 1 \leq s \leq \mathrm{n}$, result in

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

$$y + \sum_{j=1}^{s} v^o(j) + \sum_{j=1}^{s} v^c(j), t + s) \qquad \forall s, s = 1, ..., \mathrm{n} \quad (2)$$

where $[\sum_{j=1}^{s} u^o(j), \sum_{j=1}^{s} v^o(j)]$, $[\sum_{j=1}^{s} u^c(j), \sum_{j=1}^{s} v^c(j)]$ are the *cumulative* image motion in the horizontal and vertical directions between time instant t and $t + s$ for point p due to object and camera motions, respectively. The two, 2n long vectors constructed by concatenating the horizontal and vertical flows at each time instant $\forall j, j = 1, ..., \mathrm{n}$

$$\vec{O} = [u^o(j), v^o(j)]_{j=1}^{n} \qquad , \qquad \vec{C} = [u^c(j), v^c(j)]_{j=1}^{n}$$

will be referred to as the *motion temporal trajectories* of point p due to object and camera motions, respectively. The vectors $\vec{C}$ and $\vec{O}$ define two points in $\mathcal{R}^{2\mathrm{n}}$. Consider the separability of the sum $\vec{C} + \vec{O}$ with respect to the angle between the vectors as expressed by the normalized scalar product $cos(\gamma) = \frac{\vec{C} \cdot \vec{O}}{||\vec{C}|| * ||\vec{O}||}$:

- If $cos(\gamma) = 1$ then the vectors are parallel and there are infinite decompositions of the sum into two vectors $\vec{C}$ and $\vec{O}$. This occurs, for example, in the case of the train motion illusion.

- If $cos(\gamma) = 0$ then the vectors are orthogonal and thus separable. If we have a model for the class from which the vector $\vec{C}$ is constructed we can accurately divide the sum into its correct components.

- If $0 < cos(\gamma) < 1$ then the vectors are separable only in their orthogonal components. Specifically, the projection of $\vec{C}$ onto $\vec{O}$ and a hyperplane perpendicular to $\vec{O}$ results in one component that is parallel to $\vec{O}$ that may not be recoverable, and a second component that is orthogonal to $\vec{O}$ and can be fully recovered if we know the model that $\vec{C}$ is drawn from. It is worth noticing that if there exists a *structural* relationship between these two projected components (e.g., they are of equal length) then a full separation may again become possible. Furthermore, if the majority of the points of the vector belong to the perpendicular component then we will show that we can recover the correct decomposition.

In the rest of this section we will select the representations used for $\vec{C}$ and $\vec{O}$ and discuss how these choices impact the estimation of the two motion components. We distinguish between two models of image motion: general models [1, 3, 11] and learned models [5, 12]. The choices of models for use in composite motion estimation are given in Table 1. Using general models for both camera and object motions leads to an underconstrained problem as reflected by Equation (1). The use of learned models of camera motion and general models for object motion has potential only for rigid objects moving in simple ways but the extension to deformable, articulated objects or complex rigid motion trajectories is challenging since these motions are difficult to represent analytically. The case of both learned object and camera motions is a simplification, as will be discussed later in this paper, of the general camera motion and learned object motion models addressed below.

2.1 Camera Motion Model

We employ the standard conventions [9] for representing the spatio-temporal variation of the optical flow as the camera moves in a static scene. Assume a camera moving in a static scene with instantaneous 3D translational velocity (T_x, T_y, T_z) and rotational velocity $(\Omega_x, \Omega_y, \Omega_z)$ relative to an external coordinate system fixed with respect to the camera. A textured element P in the scene with instantaneous coordinates (X, Y, Z) will create an optical flow vector (u^c, v^c) where u^c and

	Learned Models of Object Motion	General Models of Object Motion
Learned Models of Camera Motion	Future work	Limited to simple object motions
General Models of Camera Motion	Developed in this paper	Underconstrained

Table 1: Estimation strategies for composite object and camera motions

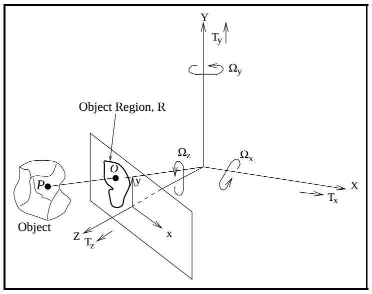

Figure 1: The motion and geometry of the camera.

v^c are the horizontal and vertical instantaneous velocities

$$u^c = \Omega_x xy - \Omega_y(1+x^2) + \Omega_z y - (T_x - T_z x)/Z$$
$$v^c = \Omega_x(1+y^2) - \Omega_y xy - \Omega_z x - (T_y - T_z y)/Z \quad (3)$$

Here, (x, y) are the image coordinates of (X, Y, Z) relative to a coordinate system in which the positive Z is aligned with the line of sight of the camera (see Figure 1). Consider an image region R that corresponds to a stationary object represented by a set of points $P_i, i = 1, ..., \mathrm{M}$ and instantaneous optical flow vectors (u^c, v^c). Assume that the object points are approximately at a constant distance from the camera, Z_0. In this case it is well known that the flow measured over the region R can be modeled by an eight parameter model,

$$\begin{aligned} u^c(x,y) &= a_0 + a_1 x + a_2 y + a_6 x^2 + a_7 xy \\ v^c(x,y) &= a_3 + a_4 x + a_5 y + a_6 xy + a_7 y^2 \end{aligned} \quad (4)$$

$$\begin{array}{ll} a_0 = -\Omega_y - T_x/Z_0 & a_1 = T_z/Z_0 \\ a_2 = \Omega_z & a_3 = \Omega_x - T_y/Z_0 \\ a_4 = -\Omega_z & a_5 = T_z/Z_0 \\ a_6 = -\Omega_y & a_7 = \Omega_x \end{array}$$

These eight parameters are estimated by pooling the motion of many points in R into an overconstrained system that can be effectively solved.

We allow general camera motion but do assume that the camera motion, and so the camera-induced flow, is constant* over the temporal window of computation (i.e., $s = 1, ..., \mathrm{n}$). This can be simply expressed as

$$u^c(s) = u^c(1) \quad v^c(s) = v^c(1) \qquad \forall s, s = 1, ..., \mathrm{n}.$$

2.2 Learned Object Motion Models

Let $(u^o(s), v^o(s))$ be the horizontal and vertical instantaneous image velocity of the object point (x, y) between frames $(t+s-1)$ and $(t+s)$. The temporal-flow vector created by the concatenation $[u^o(s), v^o(s)]_{s=1}^n$ includes 2n elements and is called a flow trajectory. Temporal-flow models can be constructed by applying principal component analysis to exemplar flow trajectories. So, the values of $[u^o(s), v^o(s)]_{s=1}^n$ are approximated by a linear combination of a *temporal-flow* basis-set of $1 \times 2 * \mathrm{n}$ vectors, U_l. The flow vector $\bar{e} = [u^o(s), v^o(s)]_{s=1}^n$ can be reconstructed using

$$\bar{e} = \sum_{l=1}^{q} c_l U_l \quad (5)$$

where c_l is the expansion coefficient of the U_l-th temporal-flow basis vector; q is the number of vectors used as the basis set (see [12]).

2.3 A Composite Motion Model

Expanding Equation (2) using a Taylor series approximation (assuming smooth spatial and temporal intensity variations) and dropping terms results in

$$0 = {I^s}_x \sum_{j=1}^{s}(u^o(j)+u^c(j)) + {I^s}_y \sum_{j=1}^{s}(v^o(j)+v^c(j)) + s{I^s}_t \quad (6)$$

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.

Equation (6) is ordinarily solved using an error minimization procedure with a robust error norm [3], $\rho(\mathbf{x}, \sigma_e)$ (σ_e is a scale parameter) of the flow over a very small neighborhood, R, of (x, y), We have n equations of the form of Equation (6), one for each time instant. The *time-generalized* error is defined as

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

Using explicitly the camera and object motion models Equation (7) can now be rewritten as:

$$E_D(u,v) = \sum_{s=1}^{n} \sum_{(x,y)\in R} \rho([{I^s}_x \quad {I^s}_y][\sum_{j=1}^{s} \sum_{m=1}^{q} c_m U_{m,j} + \sum_{j=1}^{s} u^c(j), \sum_{j=n+1}^{n+s} \sum_{m=1}^{q} c_m U_{m,j} + \sum_{j=1}^{s} v^c(j)]^T + s{I^s}_t, \sigma_e) \quad (8)$$

*A constant acceleration model can easily be substituted [11].

where $[\,]^T$ is the transpose of the temporal-flow vector.

It should be noted that in the case of an active camera with known motion, it would be possible to compute a camera temporal-flow basis set and use it instead of the general camera model used here. In this case, as long as the camera and object basis vectors are linearly independent we can unambiguously decompose the two motion components. Ambiguity arises when any of the camera bases is linearly dependent on the object bases (or vice versa). In the remainder of this paper we employ the general camera model.

3 Parameterized Composite Motion

Recall that the flow constraint given in Equation (7) 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 is an approximation to the flow generated by a plane moving in 3-D under perspective projection,

$$[UV]^T = \mathbf{X}\mathbf{P}^T \tag{9}$$

$$\begin{aligned} \mathbf{X}(x,y) &= \begin{bmatrix} 1 & x & y & 0 & 0 & 0 & x^2 & xy \\ 0 & 0 & 0 & 1 & x & y & xy & y^2 \end{bmatrix}, \\ \mathbf{P} &= \begin{bmatrix} a_0 & a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 \end{bmatrix} \end{aligned}$$

where a_i's are constants and (U, V) is the instantaneous velocity vector. To exploit the economy of parameterized models, we re-formulate the flow-bases models to learn the temporal evolution of the parameters of the planar model instead of the flow values. Specifically, consider the parameters a_i to be functions of s, so that

$$\mathbf{P}(s) = [a_i(s)]_{i=0}^{7} \tag{10}$$

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

Equation (8) can be rewritten as

$$E_D(u,v) = \sum_{s=1}^{n} \sum_{(x,y)\in R} \rho([{I^s}_x {I^s}_y](\mathbf{X}[\sum_{j=1}^{s} \mathbf{P}(j)]^T + [\sum_{j=1}^{s} u^c(j) \quad \sum_{j=1}^{s} v^c(j)]^T) + s{I^s}_t, \sigma_e) \tag{11}$$

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. $\mathbf{P}(s)$, $s = 1, ..., \mathrm{n}$, 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 $8 * n$ since it generates the eight parameters for each time instant s. We can write $[\mathbf{P}(s)]_{s=1}^{n}$ as the following sum

$$\bar{e} = [e(j)]_{j=1,...,8*n} = [\sum_{i=1}^{q} c_i L_{i,j}]_{j=1,...,8*n} \tag{12}$$

where c_i is the expansion coefficient of the L_i temporal-parameter basis vector. Equations (11) and (8) can now be rewritten as

$$E_D(u,v) = \sum_{s=1}^{n} \sum_{(x,y)\in R} \rho([{I^s}_x {I^s}_y](\mathbf{X}[\sum_{j=1}^{s} \sum_{i=1}^{q} c_i L_{i,j}, ..., \sum_{j=7n+1}^{7n+s} \sum_{i=1}^{q} c_i L_{i,j}]^T + [\sum_{j=1}^{s} u^c(j) \quad \sum_{j=1}^{s} v^c(j)]^T) + s{I^s}_t, \sigma_e) \tag{13}$$

4 Computation Algorithm

Object and camera motions can be uniquely decomposed based on Equation (13) only when the spatio-temporal motion trajectories of the camera and object are *separable* (i.e., the trajectories of the motion models are linearly independent). First it is worth exploring how well we can recover the coefficients from the sum of the flows. Let us consider the simplified case of a single basis vector $\vec{O}$ that represents the object motion (this is a $1 \times 8 * \mathrm{n}$ for the case of a single planar region in motion). Let $\beta\vec{O}$ denote the actual flow of the region due to independent motion, and let $\vec{C}$ be the unknown camera motion. Consider the problem of estimating the coefficient α that reflects the amount of independent motion in the image sequence that has a combined motion $\beta\vec{O} + \vec{C}$. Estimation of α can be posed as minimizing,

$$E = ||\alpha\vec{O} - (\beta\vec{O} + \vec{C})||^2 \tag{14}$$

The solution to Equation (14) is given by

$$\alpha = \beta + \frac{||\vec{C}|| \cos(\gamma)}{||\vec{O}||} \tag{15}$$

where γ is the angle between $\vec{C}$ and $\vec{O}$. Recall that the eigenvectors $\vec{O}$ are orthonormal, therefore $||\vec{O}|| = 1$. Equation (15) simply states that we can recover α with an error equal to the *projected* component of the camera motion onto the object motion (the term $||\vec{C}|| \cos(\gamma)$). This may look discouraging since $\vec{C}$ and $\vec{O}$ will typically not be orthogonal. However, the incorporation of a robust error norm instead of least squares allows us to relax the orthogonality requirement. Specifically, consider a robust formulation of Equation (14) as follows

$$E = \sum_{j=1}^{8*\mathrm{n}} \rho(\alpha\vec{O}_j - (\beta\vec{O}_j + \vec{C}_j), \sigma_e) \tag{16}$$

Furthermore, consider the two components of $\vec{C}$, $\vec{C}^{\perp}$ orthogonal to $\vec{O}$ and $\vec{C}^{\|}$ parallel to $\vec{O}$. Consider the first case in which the *majority* (in a robust estimation sense) of points in the vector $\vec{C}$ belong to $\vec{C}^{\perp}$. In this case, the estimate of α is accurate since the majority of the points in $\vec{C}$ are orthogonal to $\vec{O}$. As a by-product, the $\vec{C}^{\parallel}$ can be determined from α. In the second case the "majority" of points in the vector $\vec{C}$ belong to $\vec{C}^{\parallel}$; in this case the recovered α is the summation of two linearly dependent motions and therefore the motions are *inseparable*. Since robust estimators are able to overcome about 35% of the points being outliers, we can tolerate linear-dependence of up to 35% of the points and expect accurate recovery.

The computation procedure for Equation (13) is accompanied by a spatial coarse-to-fine strategy (for more information see [2]). Minimizing Equation (13) can either be done simultaneously for all parameters (i.e., $c_1, ..., c_q$ and $a_0, ..., a_7$) or, alternatively, computing $c_1, ..., c_q$ first, then warping the image sequence accordingly before computing $a_0, ..., a_7$. Since the camera model may be able, in some cases of planar objects, to account for object motion with the "assistance" of the robust error norm (e.g., a planar region moving with low acceleration) we chose a modified version of the latter alternative. Specifically, the minimization is initially started at the coarsest level of the pyramid without a camera motion model so that a linear combination of trajectories (in the multi-dimensional space of basis flow vectors) relative to the learned object motion is recovered. Then, the residual image motion in the sequence (after compensating for object motion by spatio-temporally warping the image regions throughout the sequence) is fit with the general camera model by minimizing the residual error. At subsequently finer levels of the pyramid, a refinement of these estimates is carried out similarly, after spatio-temporal warping based on the estimates from the coarse level, by first accounting for object motion and then camera-motion.

The bias of the algorithm towards accounting for object motion is motivated by our assumption that the object motion is more "constrained" than the camera motion and therefore it provides a better starting point for the minimization. The minimization steps for the object and camera motion parameters employ the Graduated-non-Convexity and a gradient descent (simultaneous-over-relaxation) algorithm as described in [3, 6].

5 Rigid Motion Experiment

We demonstrate learning the flow trajectory model of a book falling in an image sequence and the recovery of object and camera motions of different instances of book falling in new sequences. Learning the temporal-flow model is performed as follows:

- The area corresponding to the book is manually segmented in the first frame of 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 on the four non-overlapping groups of 10 consecutive instantaneous flow vectors. The resulting flow bases describe the flow trajectories for 10 frames at a time.

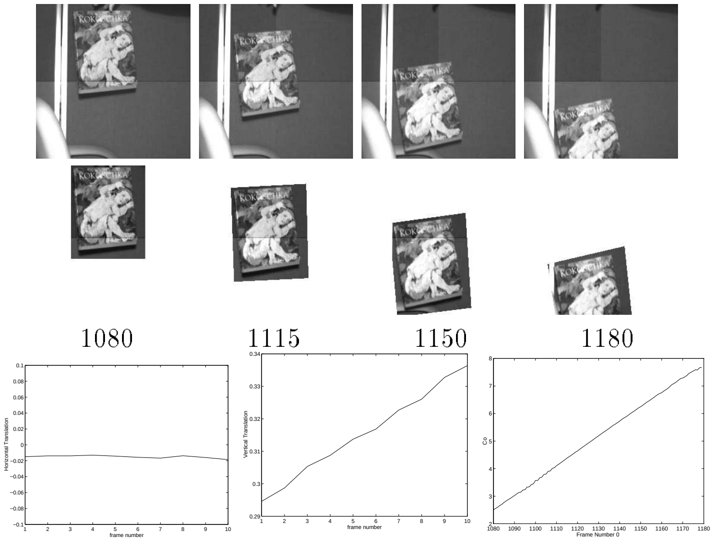

Figure 2: Four frames of a falling book tracked by a temporal-flow model (top rows), the horizontal and vertical velocities components of the learned basis-vector (left and center, bottom row) and the recovered expansion coefficient, c_0, through out the sequence (right, bottom row).

In this experiment the first eigenvalue captured 99.9% of the information 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.

The basis vector is used to compute the coefficient using Equation (13) for the whole sequence. The temporal computation window is 8 frames and could be as much as 10 frames using this flow basis set. Figure 2 shows the results of tracking the book with a stationary camera using the temporal-flow model. The graphs in the left and middle show the value of $a_0(s)$ and $a_3(s)$ (for $s = 1...10$) of the flow-vector used in estimation. While $a_0(s)$ is nearly zero (corresponding to little horizontal motion), the vertical motion component $a_3(s)$ is linear. The right side 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 has constant acceleration.

Figure 3 shows the results of composite motion estimation of a book fall while the camera is translating to

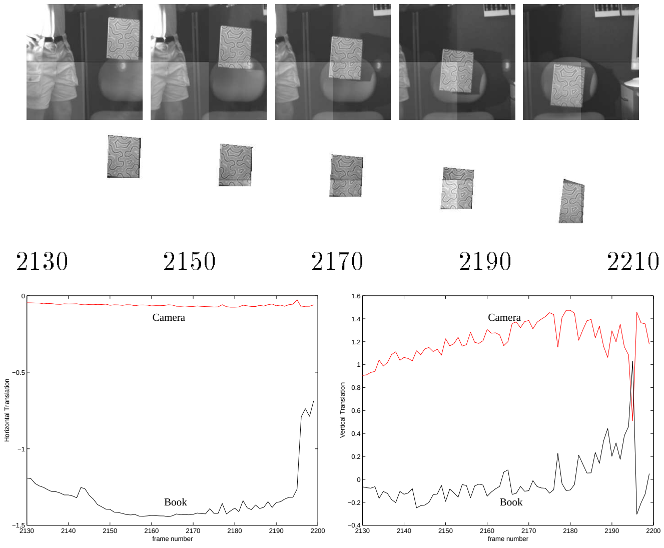

Figure 3: Frames from an image sequence of a book falling while the camera is moving horizontally and the tracked book region (top and middle rows). The horizontal and vertical translations of the book and the camera are shown in the bottom row (Red and Black plots are for camera and book, respectively).

the right. The bottom left graph shows the recovered horizontal velocities of the book and camera. As expected the book falling leads to zero horizontal speed, while the camera moves at a constant speed of about 1.4 pixels per frame. The bottom right graph shows that the camera's vertical motion is very close to zero while the book's speed increases linearly due to gravity. Towards the end of the sequence the accumulation of errors decreases the accuracy of the estimates.

Figure 4 shows the results for another book fall in which the camera is moving *away* from the book. The graphs in the third row show the recovered horizontal and vertical velocities of the book and camera. The book velocities are close to what is expected while the camera has some horizontal velocity component. The bottom row graphs are for the divergence and deformation components. Clearly the book is shrinking in size at a linear rate (then accelerated rate) as the negative divergence indicates. Moreover, since the falling book is rotating slightly away from the camera, there is a measureable deformation in the horizontal direction.

6 Articulated Human Motion

A set of samples of the spatio-temporal flow values of a human spanning one entire period of "walking" are modeled. The motion parameters of all the body parts are concatenated into one vector that captures at once the motion of five continuously visible body parts. Applying this model to a new sequence of the

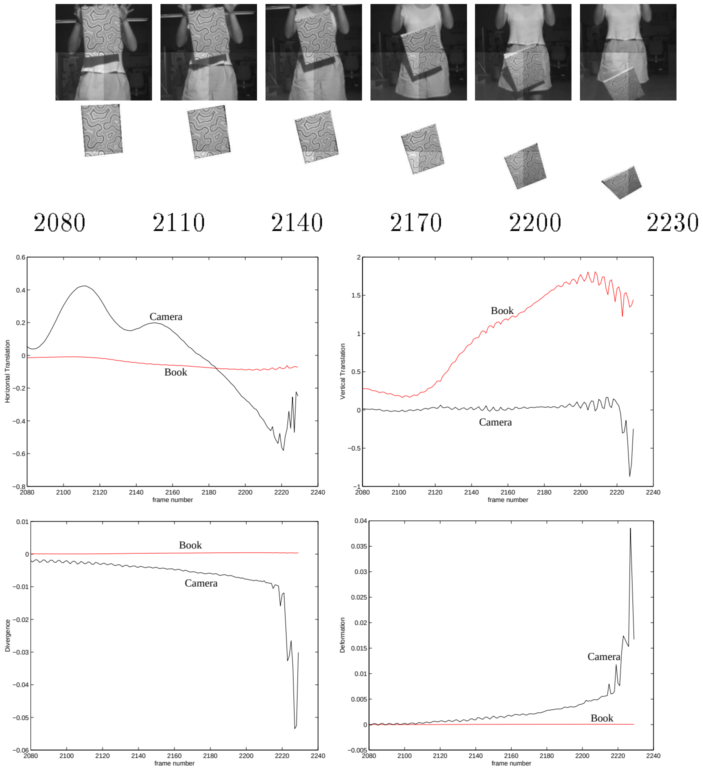

Figure 4: A few frames from a long image sequence of a book falling while the camera is moving away in depth and the tracked book region (top and second rows). The horizontal and vertical translations (third row) and the deformation and divergence parameters of the book and the camera are shown in the bottom row (Red and Black plots are for camera and book, respectively).

articulated object motion requires temporally "registering" the model to the observation at the initial time t_0. In the experiments presented in this section, we time-register sequences manually by starting the estimation at the beginning of a walking cycle.

Similar to the rigid examples, we assume that:

- The body is manually segmented into five parts in the first frame.
- People are moving at a similar viewing angle with respect to the camera during the training and testing.
- A single activity, such as "walking," is learned and tracked. The stage of the "walking" cycle of the first frame of the sequence is manually determined.
- Walking speed is approximately the same during learning and execution.

Learning the "walking" cycle temporal-flow model is performed by first employing the algorithm of Ju et al. [8] to compute each region's motion parameters during the observed cycle of the activity. Then, the motion parameters 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 temporal window (typically 6-10 frames). The five parts are tracked using Equation (13); the body parts are treated as a single object with individual motion parameters for each part.

Figure 5 displays a few frames of a walking sequence from the training set of one subject with the five-part body tracking as in [8]. Also, three illustrative flow parameters are shown, namely the horizontal, vertical and image rotation of the five body parts. In learning the model from ten people's gait, 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 individual variations and some differencies in image acquisition conditions.

Figure 6 shows the results of composite motion estimation for a new instance of walking of a subject using only the first basis-vector of the spatio-temporal flow while the camera is translating vertically. The bottom row shows the horizontal, vertical translations and the curl of the five body parts and the camera. As recovered, the camera has zero horizontal velocity and an initial downward vertical translation due to upward camera motion (frames 2045-2090) after which the opposite occurs. No camera rotation was measured. Notice the close similarity between the measurement of the five body parts relative to the graphs in Figure 5. Figure 7 shows the results of composite motion estimation for a new instance of walking of a subject while the camera is rotating clockwise around an axis off its center. Since the rotation angles are small they are often substituted by horizontal and vertical translations. Otherwise, the performance is similar to that shown in Figure 6.

7 Summary

The approach for decomposing camera and object image motions advances current research on tracking and estimation of object motion. Image motion decomposition is pursued in a direct manner without employing secondary motion clues. Specifically, progressive solution by first estimating camera motion (e.g., as the dominant motion [10]) and then object motion is replaced by direct association of image motion in the object region to object typical-motion trajectories and camera model.

Pre-learned object-typical motions are used to separate the sources of image motion. The problem is transformed into finding the motion parameters in the subspace of object motions and the motion parameters of the camera.

The separability of camera and object motions is most challenging when these motions are linearly dependent in a *subspace* $\mathcal{R}^{\mathrm{w}}$ of $\mathcal{R}^{2\mathrm{n}}$. Our robust formulation of the error minimization leads to the observation that we can recover the correct components as long as the orthogonal subspace (i.e., $\mathcal{R}^{2\mathrm{n}-\mathrm{w}}$) is the "majority" component (in a robust estimation sense). The reason is that the orthogonal component can be recovered and will itself determine the linearly dependent components by the implicit exploitation of their couplings through the basis vectors. In cases where the linearly dependent subspace is too large, recovery is not possible using our current formulation. It remains open whether other constraints can be employed in this case.

References

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

[2] J.R. Bergen, P. Anandan, K.J. Hanna and R. Hingorani. Hierarchical model-based motion estimation. *ECCV-92,* 1992, 237-252.

[3] 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.

[4] M. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motions. *IJCV*, 25(1), 1997, 23-48.

[5] M. Black, Y. Yacoob, A. Jepson and D. Fleet, Learning parameterized models of image motion. *IEEE CVPR,* 1997, 561-567.

[6] A. Blake and A. Zisserman. *Visual Reconstruction* MIT Press, 1987.

[7] C. Fermuller and Y. Aloimonos. Qualitative egomotion. *IJCV*, 15, 1995, 7-29.

[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] H.C. Longuet-Higgins and K. Prazdny, The interpretation of a moving retinal image. *Proc. Royal Society of London, B*, 208, 1980, 385-397.

[10] T.Y. Tian and M. Shah. Recovering 3D motion of multiple objects using adaptive Hough transform. *IEEE PAMI,* Vol. 19(10), 1997, 1178-1183.

[11] Authors, Temporal multi-scale models for flow and acceleration. *CVPR 97*, 921-927.

[12] Authors, Learned temporal models of image motion. *ICCV-98.* Mumbai, India, 1998, 446-453.

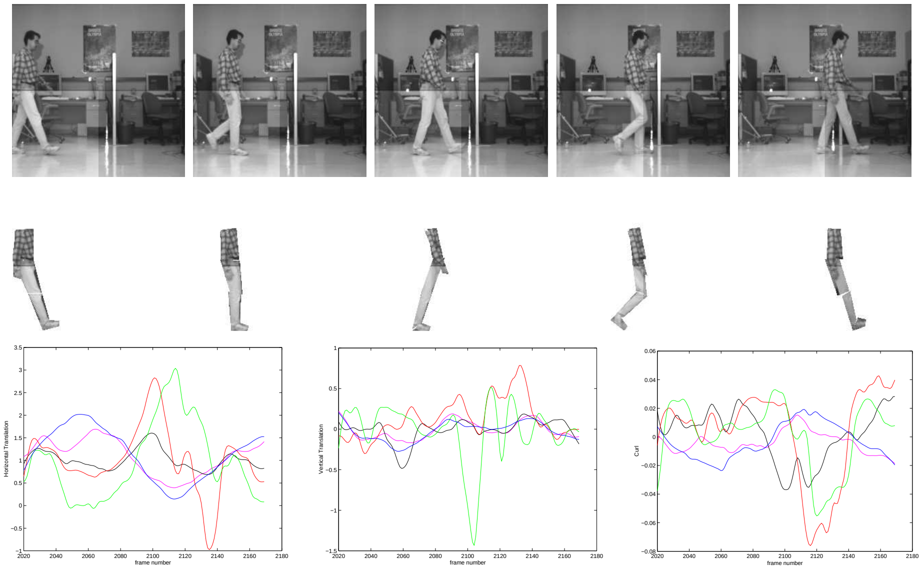

Figure 5: Tracking subject walking using the cardboard tracking [8]. The computed horizontal, vertical and image rotation of the five body parts, (torso (magenta), thigh (black), calf (red), foot (green) and arm (blue)).

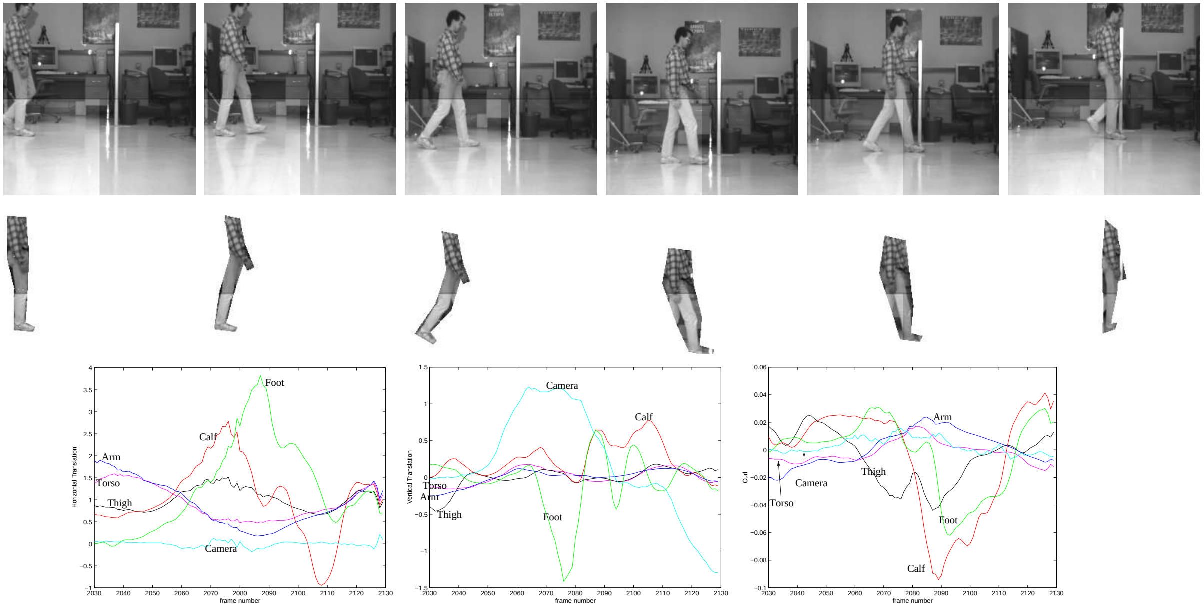

Figure 6: Tracking subject walking with vertical camera translation, recovered horizontal, vertical and image rotation of the five body parts (colors as above+camera (in cayan).

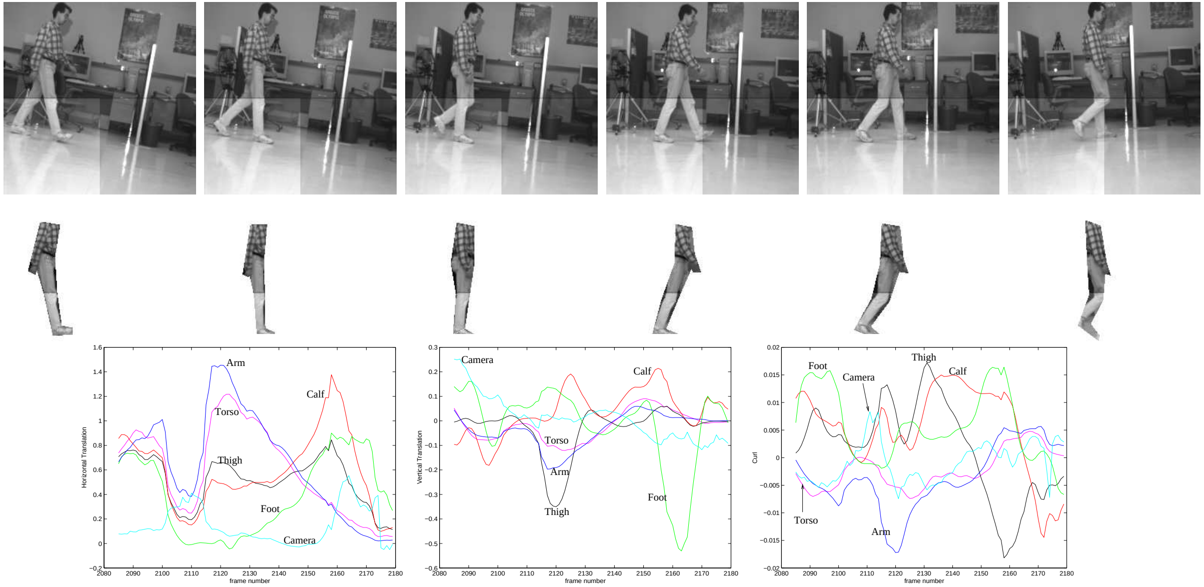

Figure 7: Tracking subject walking with camera rotation.