IJCV, ??, 1–37 (??)

Learned Models for Estimation of Rigid and Articulated Human Motion from Stationary or Moving Camera

YASER YACOOB AND LARRY S. DAVIS

yaser/lsd@umiacs.umd.edu

Computer Vision Laboratory Center for Automation Research, University of Maryland, College Park, MD 20742, USA

Received ??; Revised ??

Editors: ??

Abstract.

We propose an approach for modeling, measurement and tracking of rigid and articulated motion as viewed from a stationary or moving camera. We first propose an approach for learning temporal-flow models from exemplar image sequences. 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 then incorporated to model the movement of regions of rigid or articulated objects. These spatio-temporal flow models are subsequently used as the basis for simultaneous measurement and tracking of brightness motion in image sequences. Then we address the problem of estimating composite independent object and camera image motions. We employ the spatio-temporal flow models learned through observing typical movements of the object from a stationary camera to decompose image motion into independent object and camera motions. The performance of the algorithms is demonstrated on several long image sequences of rigid and articulated bodies in motion.

keywords: Tracking, Optical Flow, Camera Motion, Non-rigid Motion, Motion Learning.

1. Introduction

Tracking the motion of a human body in action is an exceptionally challenging computer vision problem. Even ignoring the fine structure of the hands, a human body is composed of fourteen basic parts, several 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. Particularly, monocular viewing and clothing present significant challenges to a 3D physically-based approach. In the rest of this paper, we consider an alternative 2D approach to modeling and measuring human motion.

While appearance-based (intensity) representations have been demonstrated for modeling and recognition of faces and textured 3-D objects (Murase and Nayar, 1995, Turk and Pentland, 1991), this approach does not lend itself directly to the diverse and unconstrained appearance of humans in motion sequences. The main challenges to appearance-based methods are viewpoint dependence, dealing with appearance variability (due to changes in clothing, shadowing, body size and proportions between individuals), self-occlusion, etc. An alternative approach is to develop appearance-based models for the *flow trajectories* of humans (called *motion appearances*), and to use these models to constrain the measurement and tracking of human motion.

In this paper we show how low-dimensional motion appearance models of articulated human movement can be reconstructed from observations of exemplar movements and how these models can be used to measure and track other humans performing similar movements. We present experimental evidence that suggests that the number of viewpoint-dependent motion appearance models that one would need to model a given movement is not too large (see also the discussion in (Yacoob and Black, 1998)), and also show how these models can be employed 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 motion appearance models are created by applying a standard principal components analysis to time sequences of parametric models of body part motion. These motion parameters of the exemplar movement observations are obtained using the "cardboard" body model introduced in (Ju et al., 1996), which employs the simple constraint that the instantaneous motion of body parts must agree at the joints where those parts meet. 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. Much of the analysis is carried out in a multi-temporal optical flow framework described in (Yacoob and Davis, 1999), 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 subpixel to many pixels per frame.

The measurement of human motion is further complicated when the camera itself in motion. In this case, the motion measured at each point on the human body is composed of two independent sources, body-part motion and camera motion. The aim is to recover human motion relative to the static environment after compensating for the estimated motion of the camera.

Although it may be possible, in principal, to compute camera motion first and then factor it out during object motion estimation (e.g., see a related example (Tian and Shah, 1997)), 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 (Tian and Shah, 1997)). This structure recovery is itself a very challenging problem due to the effective instantaneous change of scene structure as a result of the composite motion. Furthermore, such techniques generally depend on the availability of a "rigid" background for camera motion estimation. However, humans are able to recognize typical human movements from a moving platform even when no such rigid background is available- i.e., in situations where the motion of every pixel is a combination of camera and independent motion. Therefore we seek to determine conditions under which the object and camera image motion can be separated.

The simultaneous occurrence of object and ego-motion is typical for the human visual system. In many routine activities, humans easily identify independently moving objects and analyze their motions while they themselves are in motion (for example, all ball games involve some type of human-object interaction during simultaneous independent motion). Also, human interactions often occur during simultaneous motion; e.g., normal walking in a crowd involves estimating independent human motion; in dance, composite motion estimation is critical to performance.

Composite object and self motion can be resolved by the human visual system equally in a 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 object regions alone, 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 the observed vehicle is moving and in what direction or with what speed (i.e., the well known "train motion illusion").

Based on these observations we will propose a model-based approach for estimating the composite 2D motion of object and camera. We will demonstrate the performance of the approach on rigid and articulated bodies in motion. We make the following simplifying assumptions,

1. The independently moving object is observed "off-line" from a stationary camera while it performs its typical movements. This allows us to construct a representation of these types of movements (Yacoob and Davis, 1998).
2. 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). This will be made more precise in the body of the paper.
3. The image region corresponding to the independently moving object is identified in the first frame of the image sequence, perhaps by algorithms such as (Fejes and Davis, 1998, Fermuller and Aloimonos, 1995, Irani and Anandan, 1996). This region will be the basis for estimation of the simultaneous motion of the object and camera.

In Section 2 we discuss related research. Section 3 develops the learning and measurement of temporal models for image motion. In Section 4 we develop the spatio-temporal flow equations for parameterized regions. In Section 5 an illustration of the modeling and estimation of spatio-temporal flow for rigid and articulated motions is shown. Section 6 develops the modeling and measurement of composite motions. In Section 7 several examples for composite motion estimation are shown for rigid and articulated motions. Finally, Section 8 summarizes the paper and discusses some open problems.

2. Relevant Research

2.1. Human motion measurement from a stationary camera

Approaches to tracking the movement of humans have focused on detecting and tracking the body silhouette edges of the figure in the images to avoid the interference of the non-rigid motion of clothing. Gavrila and Davis (Gavrila and Davis, 1996) proposed a model-based approach for tracking and recovering the 3D body structure from image sequences taken from multiple cameras. The rendering of the edges of the 3D body model are matched to the edge images in each camera at each time instant to recover the degrees of freedom of each body part using an elaborate parameter search procedure. A somewhat similar approach involving a single camera has been proposed by (Goncalves et al., 1995), where a Kalman filter was used to estimate a reduced set of the degrees of freedom of a moving arm from a set of points sampled from the image based on rendering the 3D arm structure (thus, requiring prior knowledge of seven parameters of the arm). Yamamoto et al. (Yamamoto et al., 1998) proposed an approach that tracks a 3D model of a human body as seen from multiple cameras. A 3D model is initialized over the regions of calibrated cameras. Then, a direct estimation of motion parameters of 12 articulated parts of the body is performed. Baumberg and Hogg (Baumberg and Hogg, 1994) proposed an approach for tracking the outline of a moving body using an active shape model. Modal-based flexible shape models were used to capture the considerable outline variations in the human silhouette during movement. Rohr (Rohr, 1994) described a model-based approach for tracking and recognizing human movement from a single camera. Several movement states of human figures pre-captured as straight line images were used to best-fit the area of detected change in the image sequence by maximizing a similarity measure.

Pentland and Horowitz (Pentland and Horowitz,1991) describe the fitting of a 3D physically-based articulated model to optical flow data. Parts of a person are described as superquadrics with constraints on the articulated motion of the parts.

In contrast to these approaches which require various 3D models of the body, Ju et al. (Ju et al., 1996) proposed an approach for tracking humans in motion assuming that a person can be represented by a set of connected planar patches. To track this articulated motion, recovery of the relative motion of each of the limbs was performed iteratively. This is done by first estimating the motion of the torso and removing it from the image sequence using warping. The relative motions of the thighs, upper arms, and head can be then estimated relative to the torso-stabilized images. Finally, the image sequence can be stabilized with respect to these regions and the relative motions of the calf and lower arm regions can be estimated. Furthermore, the planar model is augmented to model articulated motion by constraining the motion of connected patches to be the same at the point of articulation.

Bregler and Malik (Bregler and Malik, 1998) recently proposed a 3-D approach for tracking human parts using a kinematic chain model. Each part motion is represented by a six-parameter model that encodes the relative scale and twist motion between consecutive frames (where twist motion represents the parts motion as a rotation around a 3D axis and a translation along this axis). The representation is linearized assuming an orthographic projection.

Most existing work on human motion tracking assumes that the region of the human figure has been initially detected and the body part regions localized. However, this remains a challenging goal despite some encouraging results reported in (Haritaoglu et al.,1998). Upon detection of human silhouettes (using foreground/background detection), Haritaoglu et al. (Haritaoglu et al.,1998) used a rule-based system to label the human body parts allowing for occurrence of four canonical postures.

2.2. *Human motion estimation from a moving camera*

In recent years there has been increased interest in independent object motion detection and tracking. The detection of independently moving objects has generally been posed as the problem of detecting regions in the image that are moving non-rigidly (see (Fejes and Davis, 1998, Fermuller and Aloimonos, 1995, Irani and Anandan, 1996, Tian and Shah, 1997)). Qualitative (Fejes and Davis, 1998, Fermuller and Aloimonos, 1995) and quantitative (Irani and Anandan, 1996, Tian and Shah, 1997) information derived from the image flow field is used to infer camera motion and to segment the image into independently moving patches. In (Irani and Anandan, 1996, MacLean et al., 1994, Tian and Shah, 1997), assumptions on the structure of the scene (Irani and Anandan, 1996) and camera motion (MacLean et al., 1994, Tian and Shah, 1997) were employed to segment the image into stationary and moving object regions. A limitation of these approaches is their assumption that the image motion is predominantly rigid, and that moving objects occupy a relatively small region in the image.

Some related work on detection and estimation of multiple motions have been reported for the 3D case (Boult and Brown, 1991, Costeira and Kanade, 1995, Tian and Shah, 1997) (camera motion was involved in (Tian and Shah, 1997), while a stationary camera was used in (Boult and Brown, 1991, Costeira and Kanade, 1995)). Displacements of sparse features were used to segment point motions (Boult and Brown, 1991, Tian and Shah, 1997) or recover different structures (Costeira and Kanade, 1995) that reflect motions in non-overlapping regions. In these approaches an orthographic projection was assumed. Although demonstration of performance for two independent motions of rigid objects was shown (Costeira and Kanade, 1995), it is not clear that these algorithms remain effective when a larger collection of independently moving objects are present in the scene such as in the case of human motion in front of a moving camera. The composite camera and object motion estimation problem differs from these multiple motion problems because of the confounding of camera and object motion over the object region. A motion decomposition is needed here in contrast to a spatial segmentation (Boult and Brown, 1991, Costeira and Kanade, 1995, Tian and Shah, 1997).

3. A Temporal Model for Image Motion

In this section we extend the traditionally instantaneous formulation of image motion in the time dimension. As a result, the motion vector (u, v) of a point (x, y) is extended in time by defining a *motion trajectory*, $(u(s), v(s))$ $(s = 1, ..., \mathrm{n})$ where $(u(s), v(s))$ is the image motion of point (x, y) between time instants $s - 1$ and s. This expansion of a point in 2D to a trajectory increases the dimensionality from $\mathcal{R}^2$ to $\mathcal{R}^3$ since the trajectory is equivalent to a set of points $(s, u(s), v(s))$ in (s, u, v) space.

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 $s, 1 \le s \le \mathrm{n}$, is given by

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

where $(u(s), v(s))$ is the horizontal and vertical instantaneous image velocity of the point (x, y) between frames $(t + s - 1)$ and $(t + s)$ and $[\sum_{j=1}^{s} u(j), \sum_{j=1}^{s} v(j)]$ is the *cumulative* image motion in the horizontal and vertical directions between time instant t and $t + s$. The special cases where $(u(s), v(s))$ are constant or satisfy a constant acceleration model relative to t were considered in (Yacoob and Davis, 1999):

$$u(s) = b_0 + b_1 s$$
$$v(s) = b_2 + b_3 s$$

$(b_0, b_1, b_2, b_3$ are the constant and linear order parameters of the model). Let the range of time over which temporal-flow (sequences of instantaneous flow) is estimated be $s = 1, .., \mathrm{n}$. Expanding Equation (1) using a Taylor series approximation (assuming smooth spatial and temporal intensity variations) and dropping terms results in

$$0 = {I^s}_x(x, y, t) \sum_{j=1}^{s} u(j) + {I^s}_y(x, y, t) \sum_{j=1}^{s} v(j) + s{I^s}_t(x, y, t) \quad \forall s, s = 1, ..., \mathrm{n} \quad (2)$$

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. It is important to limit the range of s so that the respective derivatives can be accurately computed. Clearly, if a large n is chosen so that large motions can occur between image I^1 and I^n the differential representation does not hold. In the context of human motion we use a high frame rate camera (85 and 99 Hz) to reduce the per frame motion to a couple of pixels, so that by using a pyramid estimation process we can compute the derivatives for 3-5 frames at a time.

Since Equation (2) is underconstrained for the recovery of $(u(s), v(s))$, the estimation of $(u(s), v(s))$ can be ordinarily posed as an error minimization over a small region R using a robust error norm, $\rho(\mathbf{x}, \sigma_e)$, that is a function of a scale parameter σ_e. The error of the flow over R is,

$$E(u, v, s) = \sum_{(x,y)\in R} \rho({I^s}_x(x, y, t) \sum_{j=1}^{s} u(j) + {I^s}_y(x, y, t) \sum_{j=1}^{s} v(j) + s{I^s}_t(x, y, t), \sigma_e) \quad (3)$$

assuming points in R conform to the same motion trajectory. We have n equations of the form of Equation (3), one for each time instant. The *time-generalized* error is defined as

$$E_D(u, v) = \sum_{s=1}^{\mathrm{n}} \sum_{(x,y)\in R} \rho({I^s}_x(x, y, t) \sum_{j=1}^{s} u(j) + {I^s}_y(x, y, t) \sum_{j=1}^{s} v(j) + s{I^s}_t(x, y, t), \sigma_e) \quad (4)$$

4. Learned Parameteric Image Motion

In subsection 4.1 we show how the space of flow trajectories can be efficiently encoded using a linear representation so that a parametric model of trajectories is created. Subsection 4.2 reformulates the parametric model of trajectories to exploit spatially parameterized optical flow models. Finally, in subsection 4.3 we describe the computational aspect of the algorithm.

4.1. Learning Temporal-Flow Models

As defined, motion trajectories $(u(s), v(s))$ $(s = 1, ..., \mathrm{n})$ require the computation of the trajectory of a single point which involves estimating the value of 2n parameters. In reality, however, physical processes constrain the space of actual motion trajectories of points. Physical considerations include static and dynamic properties of real object motions. Notice that these processes do not apply at the instantaneous level since a point can move with any velocity (u, v) and that the addition of the temporal dimension implicitly introduces physical (e.g., Newtonian) constraints projected onto the camera plane. In this subsection we propose an approach for learning a model of the space of feasible trajectories.

Purely spatial constraints on image motions were recently proposed by Black et al. (Black et al., 1997). 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 here 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.

Temporal-flow models are constructed by applying principal component analysis to exemplar flow sequences. So, the functions $(u(s), v(s))$ 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 $\bar{e} = [(u(s), v(s))]_{s=1}^{\mathrm{n}}$ can be reconstructed using

$$\bar{e} = [e(j)]_{j=1,\ldots,2*\mathrm{n}} = [\sum_{i=1}^{q} c_i U_{i,j}]_{j=1}^{2\mathrm{n}} \qquad (5)$$

where $\bar{e}$, the temporal-flow vector, denotes the concatenation of $u(s)$ and $v(s)$ 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.

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 (Yacoob and Davis, 1999) 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 $2 * \mathrm{n}$ rows by N columns. Matrix A can be decomposed using Singular Value Decomposition (SVD) as

$$A = U \Sigma V^T \qquad (6)$$

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

$$\bar{e}^* = \sum_{l=1}^{q} c_l U_l \qquad (7)$$

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.

Using the temporal-flow basis set Equation (4) can also be expressed as:

$$E_D(u, v) = \sum_{s=1}^{\mathrm{n}} \sum_{(x,y)\in R} \rho([I^s{}_x \;\; I^s{}_y][\sum_{j=1}^{s} \sum_{i=1}^{q} c_i U_{i,j}, \sum_{j=\mathrm{n}+1}^{\mathrm{n}+s} \sum_{i=1}^{q} c_i U_{i,j}]^T + s I^s{}_t, \sigma_e) \qquad (8)$$

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 (8) essentially describes how image motion of a point (x, y) changes over time under the constraint of a temporal-flow basis-set.

4.2. Parameterized Spatio-Temporal Image-Motion

Recently, it has been demonstrated that spatially parameterized flow models are a powerful tool for modeling instantaneous image motion ((Black and Yacoob, 1997, Black et al., 1997, Ju et al., 1996)). 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 polynomially parameterized image motion models.

Recall that the traditional flow constraint 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 (Adiv, 1985). For example, the planar motion model (Adiv, 1985) is an approximation to the flow generated by a plane moving in 3-D under perspective projection. The model is given by

$$\begin{aligned} U(x,y) &= a_0 + a_1 x + a_2 y + a_6 x^2 + a_7 xy \\ V(x,y) &= a_3 + a_4 x + a_5 y + a_6 xy + a_7 y^2 \end{aligned} \quad (9)$$

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 (9) can be written in matrix form as

$$[UV]^T = \mathbf{X}\mathbf{P}^T \quad (10)$$

where

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

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

$$\mathbf{P}(s) = [\mathbf{a}_0(s)\ \ a_1(s)\ \ a_2(s)\ \ a_3(s)\ \ a_4(s)\ \ a_5(s)\ \ a_6(s)\ \ a_7(s)]_{s=1}^{\mathrm{n}}$$

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

Equation (8) can be rewritten as

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

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

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

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

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

The minimization of Equation (13) results in estimates for the parameters c_i. The above treatment of polynomial flow is also applicable to the orthogonal-basis modeling of spatial flow recently proposed in (Black et al., 1997). The coefficients used in the linear combination replace the parameters a_i in the above equations.

4.3. Computation Details

The robust error norm and its derivative are adopted from (Geman and McClure, 1987),

$$\rho(x, \sigma_e) = \frac{x^2}{\sigma_e + x^2} \qquad \psi(x, \sigma_e) = \frac{2x\sigma_e}{(\sigma_e + x^2)^2} \quad (14)$$

The minimization of Equation (13) is carried out using a descent method, Simultaneous Over-Relaxation (SOR). The minimization of $E_D(u, v)$ with respect to c_i is achieved using an iterative update equation, so at step $q + 1$

$$c_i{}^{(q+1)} = c_i{}^{(q)} - \omega \frac{1}{T(c_i)} \frac{\partial E_D}{\partial c_i}(c_i{}^{(q)}) \qquad (15)$$

where $0 < \omega < 2$ is an overrelaxation parameter which is used to overcorrect the estimate of $c_i{}^{(q+1)}$ at stage $q + 1$. The value of ω determines the rate of convergence. The term $T(c_i)$ is an upper bound on the second partial derivative of E_D

$$T(c_i) \geq \frac{\partial^2 E_D}{\partial^2 c_i} = \sum_{j=1}^{8n} L_{i,j}{}^2 \max_j \psi' \qquad (16)$$

where $L_{i,j}{}^2$ is the square of element j of L_i and

$$\max \psi' = \max_x \frac{\partial^2}{\partial x^2} \rho(x, \sigma) = \frac{2}{\sigma^2} \qquad (17)$$

To achieve a globally optimal solution the robust error norm ρ is started with a large enough scale parameter σ_e to find a solution using the SOR technique. Then this process is iteratively repeated while decreasing σ_e and starting with the last estimate. The choice of a large enough σ_e guarantees convexity of the error function at the beginning of the process, which is followed by the use of the Graduated Non-Convexity method developed in (Blake and Zisserman, 1987). The iterated decrease in σ_e reduces the influence of the outlier measurements and thereby refines the estimates.

This implementation employs the standard spatial coarse-to-fine strategy (Bergen et al., 1992) that constructs a pyramid of the spatially filtered and sub-sampled images and computes the coefficients initially at the coarsest level and then propagates the results to finer levels.

5. Experiments with Stationary Camera

5.1. A Rigid Motion Example

The use of a temporally parameterized motion model that explicity accounts for image velocity and acceleration and is computed directly from image intensity variations was discussed in (Yacoob and Davis, 1999). Here, we demonstrate how a learned spatio-temporal flow model can capture image acceleration by observing a book free-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 using the parameterized flow algorithm of (Black and Anandan, 1996)).
- A basis set for the temporal-flow parameters is computed by performing PCA on the four non-overlapping groups of 10 consecutive instanteneous flow vectors.
- The basis set is used to compute the coefficients using Equation (13) 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.

KP: INSERT FIGURE 1 HERE

Figure 1 shows the results of tracking the book using the spatio-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 with positive slope that implicitly captures the constant acceleration of the fall. 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 linear order (i.e., a constant acceleration model).

The learned spatio-temporal models can be applied to other objects performing similar motions. The spatio-temporal flow basis-vector learned for the book is used to estimate the falling of a different object, a cardboard box. Figure 2 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 small rotational component, while the box fell without rotation. The single basis vector used encodes both the falling and rotation and therefore these cannot be segregated during estimation.

It is worth noting that the motion trajectory of the box creates a line parallel (see Figure 2 bottom row) to the falling book's trajectory. Equation (13) 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.

KP: INSERT FIGURE 2 HERE

5.2. *Learned Models of Articulated Human Motion*

The *cardboard* (Ju et al., 1996) model for tracking five-part human movement (arm, torso, thigh, calf and foot) involves recovering 40 motion parameters per frame; 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. Learning models of articulated movement can lead to much simpler representations in which redundancies are removed and motion parameter couplings learned. A set of samples of the motion parameters of the parts of articulated human covering one entire period of an activity are modeled using principal component analysis. In the following, we use video cameras with resolution 256×256 at 99Hz and 686×484 at 85Hz; this temporal sampling rate is high enough for us to effectively employ differential flow estimation over several frames. In most of the sequences used here the full 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.

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 modeling and measurement phases.
- A single activity, such as "walking," is learned and tracked.

Learning of the "walking" cycle spatio-temporal flow model is performed by first employing the algorithm of Ju et al. (Ju et al., 1996) to compute each region's instantaneous 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 spatio-temporal flows of the activity. It is worth noting that although the basis-vectors are computed for a whole cycle of "walking" the spatio-temporal motion recovery is conducted using a small computation temporal window (typically 6-10 frames) that slides along the movement. The five parts are tracked using Equation (13), the body parts are considered as a single object with individual motion parameters for each part coordinated through the principal components model.

KP: INSERT FIGURE 3 HERE

Figure 3 displays a few frames of a walking sequence from the training set of one subject with the five-part body tracking as in (Ju et al., 1996). Notice that the tracking accumulates errors, some of which also appear in the spatio-temporal flow tracking. In learning the model from ten people's gait[1], 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 differences in image acquisition conditions.

Applying this model to measure a human movement in a new sequence requires temporally "registering" the model to the observation at the initial time t_0. Determining the temporal stage of the activity plays an important role since it determines the local temporal window in the basis-vectors which is employed in the error minimization. One simple method to determine the correct stage is to match the initial frames against all stages of the performance of the activity. At each matching instance the quality of the match is measured as the ratio of the pixels explained by the basis-vectors over the overall number of pixels. The best match is that in which brightness motion can be best explained by the basis-vectors.

To demonstrate the performance of the temporal initialization we use the sequence shown in

Figure 5 and match its frames against all temporal stages of "walking" as captured by the basis set. Figure 4 shows the percentage of outliers (i.e., points not explained well by the basis vectors) as a function of the temporal stage of "walking." In this example the least number of outliers occurs in the beginning of the sequence and after about 112 frames which is the beginning of the next walking cycle.

KP: INSERT FIGURE 4 HERE

Figure 5 shows the results of tracking a new instance of walking of a subject using only the first basis-vector of the spatio-temporal flow. It also shows the coefficient, c_0, recovered throughout the sequence (n = 8). Low image contrast leads to accumulation of tracking errors. Figure 6 shows the tracking results using the first one, two and three basis vectors. A close look at the results reveals that the best tracking is achieved using the first two basis vectors. The third basis vector degraded the performance since the information captured by this vector is relatively small (about 5%) and does not provide an effective constraint on the brightness.

KP: INSERT FIGURE 5 HERE

KP: INSERT FIGURE 6 HERE

The learned spatio-temporal flow models remain effective in tracking articulated motion even when distance from the camera and the viewpoint vary from the training 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 remained 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 spatio-temporal flow model of walking.

Figures 7 and 8 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 8, 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 spatio-temporal flow of activities can also be employed for tracking partially occluded parts. We demonstrate the performance of our approach on sequences of two activities, walking and marching. These activities involve symmteric movement of the legs and arms that are half a cycle apart. Therefore, once a motion model for the visible parts is learned it can be applied to the occluded ones. We assume that the difference in distance between the legs and the camera are equal to the distance of the body from the camera. In the first frame we initialize the regions for nine body parts (when parts are occluded we simply hypothesize their locations). Then, we minimize Equation (13), where all nine regions are regarded as a single object with multiple motion parameters that are represented by the two-phase motion of the right and left side parts of the body. Only the un-occluded pixels of each region are used in the motion recovery while the occluded part is moved to reflect the movement in the activity model. Each activity model was learned separately from a single example of its performance.

The results of the tracking of the two activities in long sequences are shown in Figures 9-11. The two legs are tracked well despite some inaccuracies that are due to the learned model inaccuracies. One important property of the models we propose is that the recovered coefficients readily incorporate the interpretation of the activity seen in the image. In the case of multiple familiar activities we employ these models competitively to "account" for the brightness changes in the image. The model that best accounts for the brightness motion achieves the "recognition" of the observed activity.

KP: INSERT FIGURE 7 HERE
KP: INSERT FIGURE 8 HERE
KP: INSERT FIGURE 9 HERE

KP: INSERT FIGURE 10 HERE
KP: INSERT FIGURE 11 HERE

6. Modeling and Measuring Composite Motion

We consider next observing human movement from a moving camera. First consider the simpler case of a single rigid object and the flow that is observed due to the 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

$$I(x, y, t) = I(x + u^c + u^o, y + v^c + v^o, t + 1). \quad (18)$$

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 $s, 1 \leq s \leq \mathrm{n}$, is

$$I(x, y, t) = I(x + \sum_{j=1}^{s} u^o(j) + \sum_{j=1}^{s} u^c(j), \\ y + \sum_{j=1}^{s} v^o(j) + \sum_{j=1}^{s} v^c(j), t + s) \\ \forall s, s = 1, ..., \mathrm{n} \quad (19)$$

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 $\forall j, j = 1, ..., \mathrm{n}$

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

will be referred to as the 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 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 (Adiv, 1985, Black and Anandan, 1996, Yacoob and Davis, 1999) and learned models (Black et al., 1997, Yacoob and Davis, 1998). 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 (18). 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.

6.1. Camera Motion Model

We employ the standard conventions (Longuet-Higgins and Prazdny, 1980) for representing the spatio-temporal variation of the optical flow as the camera moves through 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 texture 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 v^c are the horizontal and vertical instantaneous velocities

$$u^c = \Omega_x xy - \Omega_y(1+x^2) + \Omega_z y - (T_x - T_z x)/Z$$
$$v^c = \Omega_x(1+y^2) - \Omega_y xy - \Omega_z x - (T_y - T_z y)/Z \quad (20)$$

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 12).

KP: INSERT FIGURE 12 HERE

Consider an image region R that corresponds to a stationary object represented by a set of points $p_i, i = 1, ...,$ 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,

$$\begin{aligned} u^c(x,y) &= a_0 + a_1 x + a_2 y + a_6 x^2 + a_7 xy \\ v^c(x,y) &= a_3 + a_4 x + a_5 y + a_6 xy + a_7 y^2 \end{aligned} \quad (21)$$

where

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

These eight parameters are estimated by pooling the motion of many points in R into an overconstrained system.

We allow general camera motion but do assume that the camera motion, and so the camera-induced flow, is time-wise constant[2] (between consecutive frames) over the temporal window of computation (i.e., $s = 1, ...,$ n),

$$\begin{aligned} u^c(x,y,s) &= u^c(x,y,1) = a_0 + a_1 x + a_2 y + a_6 x^2 + a_7 xy \\ v^c(x,y,s) &= v^c(x,y,1) = a_3 + a_4 x + a_5 y + a_6 xy + a_7 y^2 \end{aligned} \quad (22)$$

6.2. A Composite Model for Object and Camera Motion

Expanding Equation (19) using a Taylor series approximation (assuming smooth spatial and temporal intensity variations) and dropping terms results in

$$0 = I^s{}_x(x,y,t)(\sum_{j=1}^{s} u^o(j) + \sum_{j=1}^{s} u^c(j)) + I^s{}_y(x,y,t)(\sum_{j=1}^{s} v^o(j) + \sum_{j=1}^{s} v^c(j)) + sI^s{}_t(x,y,t) \qquad s = 1, ..., n \quad (23)$$

The *time-generalized* error is given by

$$E_D(u,v) = \sum_{s=1}^{n} \sum_{(x,y)\in R} \rho(I^s{}_x(x,y,t)(\sum_{j=1}^{s} u^o(j) + \sum_{j=1}^{s} u^c(j)) + I^s{}_y(x,y,t)(\sum_{j=1}^{s} v^o(j) + \sum_{j=1}^{s} v^c(j)) + sI^s{}_t(x,y,t), \sigma_e) \quad (24)$$

Substituting the object motion model from Equation (8) into Equation (23) results in

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

Table 1. Estimation strategies for composite object and camera motions

	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

$$\sum_{j=1}^{s} v^c(j)]^T + sI^s{}_t, \sigma_e) \qquad (25)$$

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 (25) essentially describes how the image motion of a point (x, y) changes over time under the constraint of a temporal-flow basis-set and general camera motions.

Using the spatially parameterized flow model of planar motion (see Section 4), Equation (25) can be rewritten as

$$E_D(u,v) = \sum_{s=1}^{n} \sum_{(x,y)\in R} \rho([I^s{}_x I^s{}_y](\mathbf{X}[\sum_{j=1}^{s} \mathbf{P}(j)]^T + [\sum_{j=1}^{s} u^c(j) \sum_{j=1}^{s} v^c(j)]^T) + sI^s{}_t, \sigma_e) \qquad (26)$$

where R denotes the region over which the planar motion model is applied. $\mathbf{P}(s)$, $s = 1, ..., \mathrm{n}$, can be represented by a linear combination of basis vectors in a manner similar to the temporal-flow representation developed in Section 4. Therefore,

$$E_D(u,v) = \sum_{s=1}^{n} \sum_{(x,y)\in R} \rho([I^s{}_x I^s{}_y](\mathbf{X}[\sum_{j=1}^{s} \sum_{i=1}^{q} c_i L_{i,j}, ..., \sum_{j=7n+1}^{7n+s} \sum_{i=1}^{q} c_i L_{i,j}]^T + [\sum_{j=1}^{s} u^c(j) \sum_{j=1}^{s} v^c(j)]^T) + sI^s{}_t, \sigma_e) \qquad (27)$$

Equation (27) measures the brightness motion trajectories assuming a composition of familiar object motion and general camera motion.

6.3. Computation Algorithm

Object and camera motions can be uniquely decomposed based on Equation (27) 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,

$$E = ||\alpha\vec{O} - (\beta\vec{O} + \vec{C})||^2 \qquad (28)$$

The solution to Equation (28) is given by

$$\alpha = \beta + \frac{||\vec{C}|| \cos(\gamma)}{||\vec{O}||} \qquad (29)$$

where γ is the angle between $\vec{C}$ and $\vec{O}$. Recall that the eigenvectors $\vec{O}$ are orthonormal, therefore $||\vec{O}|| = 1$. Equation (29) 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 (28) as follows

$$E = \sum_{j=1}^{8*\mathrm{n}} \rho(\alpha\vec{O}_j - (\beta\vec{O}_j + \vec{C}_j), \sigma_e) \qquad (30)$$

Furthermore, consider the two components of $\vec{C}$, $\vec{C}^{\perp}$ orthogonal to $\vec{O}$ and $\vec{C}^{\parallel}$ 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.

Minimizing Equation (27) 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. The bias of the algorithm towards accounting for object motion is motivated by our assumption that the human motion is more "constrained" than the camera motion and therefore it provides a better starting point for the minimization.

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 computation of the motions for a set of frames $I^t, I^{t+1}, ...I^{t+n}$ consists of the following stages of computation

1. Compute the amount of motion that can be associated with the object given an activity and the temporal stage of performance of the activity. This is done according to the estimation process described in Section 4. As a result, a coefficient vector $\vec{c}^* = c_1, ..., c_q$ is recovered.
2. Use the estimate $\vec{c}^*$ to warp the sequence to "remove" object movement. This is a spatio-temporal warp that registers the frames $I^{t+1}, ...I^{t+n}$ relative to I^t. This warping leaves out a residual motion that is due to camera motion.
3. Compute the camera motion model parameters $a_0, ..., a_7$ from the warped sequence of stage (2). The computation follows the multi-temporal parameteric motion estimation in (Yacoob and Davis, 1999).
4. Warp the sequence of (1) using the combined estimated object and camera motions represented by the parameters $\vec{c}^*$ and $a_0, ..., a_7$.
5. Repeat steps (1)-(4) using a coarse-to-fine pyramid estimation process.

The error minimization steps for the object and camera motion parameters in stages (1) and (3) employ the Graduated-non-Convexity and a gradient descent (simultaneous-over-relaxation) algorithm as described in Section 4.

7. Experiments

In this section we report on several experiments carried out to demonstrate the approach. Composite motion of both rigid and articulated objects are presented. The object motions will be modeled using parameterized motion models such as the planar model discussed earlier.

Figure 13 illustrates the interpretation of the image motion captured by the first six parameters of the parameterized planar model used to approximate flow. The translation in the horizontal and vertical directions is captured by a_0 and a_3 respectively, while the *divergence*, *deformation* and *curl* are captured by the following equations (see Figure 13)

$$\begin{aligned} \text{divergence} &= a_1 + a_5 = (U_x + V_y), \\ \text{curl} &= -a_2 + a_4 = -(U_y - V_x), \\ \text{deformation} &= a_1 - a_5 = (U_x - V_y) \end{aligned}$$

KP: INSERT FIGURE 13 HERE

7.1. Rigid Motion

We demonstrate composite motion estimation on the falling book using the model developed in Section 5. Figure 14 shows the results of composite motion estimation of a book fall while the camera is translating to 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 our estimates.

Figure 15 shows the results of composite motion estimation of the book's fall while the camera initially rotates clockwise then counter-clockwise about an axis off its center. The bottom left graph shows the recovered horizontal velocities of the book and camera. As expected the falling book has zero horizontal speed, while the camera starts with a movement rightward then leftward. The vertical speed of the book is only partly correct. The increase in speed up-to frame 2380 corresponds correctly to the falling model; then, for a few tens of frames, the velocity decreases as the camera's motion successfully accounts for the missing book falling component. This occurs since the book appears to fall horizontally at that stage of camera rotation, and therefore is not well described by the basis flow that represents vertical motion. This corresponds to the case of lack of majority of points belonging to the orthogonal component of the motion as discussed in Section 6.3. As the camera reverses its rotation (around frame 2400), the estimation recovers from this ambiguous state. The bottom right graph shows the recovered image rotation (curl) of the object and camera. As expected, the book does not rotate and the camera rotates clockwise then counter clockwise at about frame 2400.

Figure 16 shows the results of simultaneous motion estimation 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 (positive deformation).

KP: INSERT FIGURE 14 HERE

KP: INSERT FIGURE 15 HERE

KP: INSERT FIGURE 16 HERE

7.2. Articulated Human Motion

We employ the basis vectors computed in Section 5.2 with new sequences involving camera motion. Figure 17 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; 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 3. Figure 18 shows the results of composite motion estimation for a new instance of walking of a subject; here, the camera is rotating clockwise around an axis off its center. Since the rotation angles are small they are often confused with horizontal and vertical translations. Otherwise, the performance is similar to that shown in Figure 17.

Figure 19 shows the results of composite motion estimation for a new instance of walking of a subject; here, the camera is rotating clockwise around an axis off its center.

KP: INSERT FIGURE 17 HERE

KP: INSERT FIGURE 18 HERE

KP: INSERT FIGURE 19 HERE

8. Discussion

We presented a new approach for image motion estimation from multiple frames that uses learned models of spatio-temporal flow. Demonstration of the performance of the algorithm on both rigid and articulated motions was provided. An activity learned from one specific viewpoint was used to estimate the motion of a different subject performing the same activity from a similar viewpoint. Also, it was demonstrated that tracking of the occluded body parts is possible when a temporal model of one side of the body has 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 (Ju et al., 1996) used to generate the training set for the learning algorithm lead to similar inaccuracies in the spatio-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 spatio-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. It remains an open problem to develop a single representation for all activities so that a single processing of the data be sufficient

The spatio-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.

Our current temporal-models are "strict" in their interpretation of the time axis. In articulated motion, we assumed that the learned models and subsequent observations progress on an equal time scale, i.e., each time increment in the tracking leads to an equal increment in the model. This assumption is limiting since activity instances might have a significantly different "pace." In recent work (Yacoob and Black, 1998) we proposed an approach for analytically accounting for temporal variations in performance of movements. This approach explicitly computes a time-scaling parameter that can directly be used in the current spatio-temporal measurement model.

The approach for decomposing camera and object image motions is a departure from current research on multiple motion estimation. The following briefly summarizes the differences:

- Multiple motions have been generally considered as motions occurring in non-overlapping regions. Our approach considers two motion sources (object and camera) within a single region. In other words, multiple motion computation is posed as a spatial segmentation problem (Boult and Brown, 1991, Costeira and Kanade, 1995), composite object and camera motion is a spatio-temporal decomposition problem.
- 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 (Tian and Shah, 1997)) and then object motion is replaced by direct association of image motion in the object region to object typical-motion trajectories and camera model.
- The hypothesis of pre-learned object-typical motions is proposed as a means 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 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.

Acknowledgements

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

Notes

1. The distance and viewing direction in the training data was constant. The viewing direction was approximately fronto-parallel.
2. A constant acceleration model can easily be substituted, see (Yacoob and Davis, 1999).

References

Adiv, G.:1985, Determining three-dimensional motion and structure from optical flow generated by several moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(4), pp. 384-401.

Baumberg, A.M., Hogg, D.C.:1994, Computing optical flow across multiple scales: An adaptive coarse-to-fine strategy, *IEEE Workshop on Motion of Non-rigid and Articulated Objects,* 194-199.

Bergen, J.R., Anandan, P., Hanna, K.J. and Hingorani, R.:1992, Heirarchical model-based motion estimation, In G. Sandini, editor, *European Conference on Computer Vision*, Vol. 588 of LNCS-Series, Springer-Verlag, pp. 237-252.

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

Black, M.J. and Yacoob, Y.:1997, Recognizing facial expressions in image sequences using local parameterized models of image motion, *International Journal of Computer Vision*, 25(1), pp. 23-48.

Black, M.J., Yacoob, Y., Jepson, A., Fleet, D.:1997, Learning parameterized models of image motion, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 561-567.

Blake, A. and Zisserman, A.:1987, *Visual Reconstruction* The MIT Press, Cambridge, Massachusetts, 1987.

Boult, T.E., and Brown, L.G.:1991, Factorization-based segmentation of motion, *IEEE Workshop on Visual Motion*, pp. 179-186.

Bregler, C., and Malik, J.:1998, Estimating and tracking kinematic chains, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8-15.

Costeira, J., and Kanade, T.:1995, A multibody factorization method for motion analysis, *International Conference on Computer Vision*, pp. 1071-1076.

Fejes, S., and Davis, L.S.:1998, What can projections of flow fields tell us about the visual motion, *International Conference on Computer Vision*, pp. 979-986.

Fermuller, C. and Aloimonos, Y.:1995, Qualitative egomotion, *International Journal of Computer Vision*, 15, pp. 7-29.

Gavrila, D., and Davis, L.S:1996, 3D Model-based tracking of humans in action: A multi-view approach, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 73-80.

Geman, S. and McClure, D.E.:1987, Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute*, LII-4:5-21.

Goncalves, L., Di Bernardo, E., Ursella, E. and Perona, P.:1995, Monocular tracking of the human arm in 3D, *International Conference on Computer Vision*, pp. 764-770.

Haritaoglu, I., Harwood, D., and Davis, L.S.:1998, Ghost: A human body part labeling system using silhouettes *Fourteenth International Conference on Pattern Recognition*, pp. 77-82.

Horn, B.K.P., and Shunk, B.:1981, Determining optical flow, *Artificial Intelligence*, 17, pp. 185-203.

Irani, M. and Anandan, P.:1996, A unified approach to moving object detection in 2D and 3D scenes, *ARPA Image Understanding Workshop*, pp. 707-718.

Ju, S.X., Black, M. and Yacoob, Y.:1996, Cardboard people: A parameterized model of articulated image motion, *International Conference on Face and Gesture*, pp. 561-567.

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

MacLean, W.J., Jepson, A.D. and Frecker, R.C.:1994, Recovery of egomotion and segmentation of independent object motion using the EM algorithm. *British Machine Vision Conference*, pp. 13-16.

Murase, H. and Nayar, S.K.:1995, Visual learning and recognition of 3D objects from appearance, *International Journal of Computer Vision,* (14)1, pp. 5-24.

Pentland, A. and Horowitz, B.:1991, Recovery of nonrigid motion and structure. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 13(7), pp. 730-742.

Rohr, K.:1994, Towards model-based recognition of human movements in image sequences, *CVGIP: Image Understanding*, 59, pp. 94-115.

Tian, T.Y., and Shah, M.:1997, Recovering 3D motion of multiple objects using adaptive Hough transform, *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 19(10), pp. 1178-1183.

Turk, M. and Pentland, A.:1991, Eigenfaces for recognition, *Journal of Cognitive Neuroscience*, (3)1, pp. 71-86.

Yacoob, Y. and Davis, L.S.:1999, Temporal multi-scale models for flow and acceleration, *International Journal of Computer Vision*, (32)2, pp. 1-17.

Yacoob, Y. and Davis, L.S.:1998, Learned temporal models of image motion. *International Conference on Computer Vision,* pp. 446-453.

Yacoob, Y. and Black, M.:1998, Parameterized modeling and recognition of activities, *Journal of Computer Vision and Image Understanding*, 73(2), pp. 232-247.

Yamamoto, M., Sato, A., Kawada, S., Kondo, T. and Osaki, Y.:1998, Incremental tracking of human actions from multiple views, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2-7.

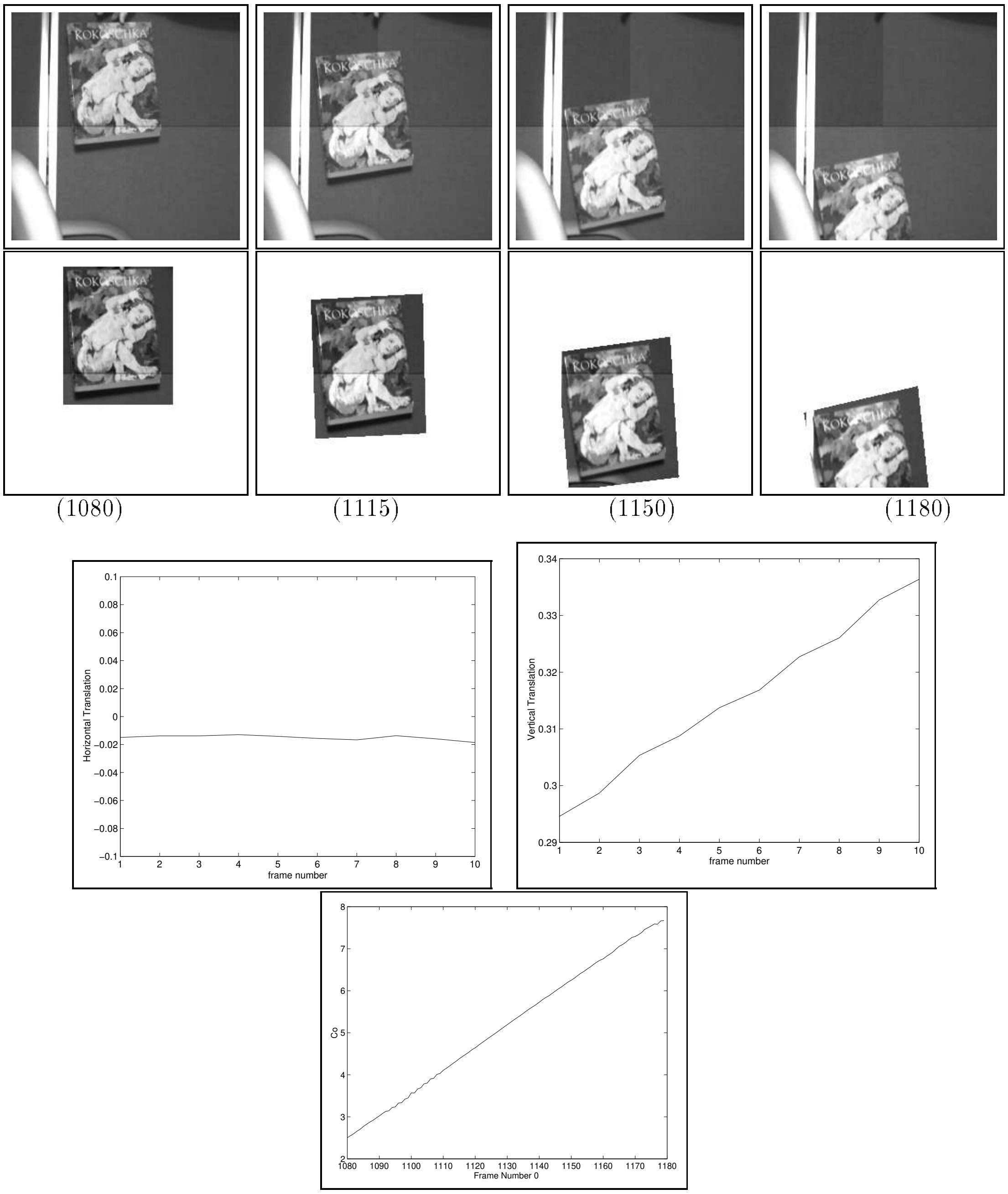


Fig. 1. Four frames of a falling book tracked by a spatio-temporal flow model (top rows), the horizontal and vertical velocities components of the learned basis-vector (third row) and the recovered expansion coefficient throughtout the sequence (bottom row).

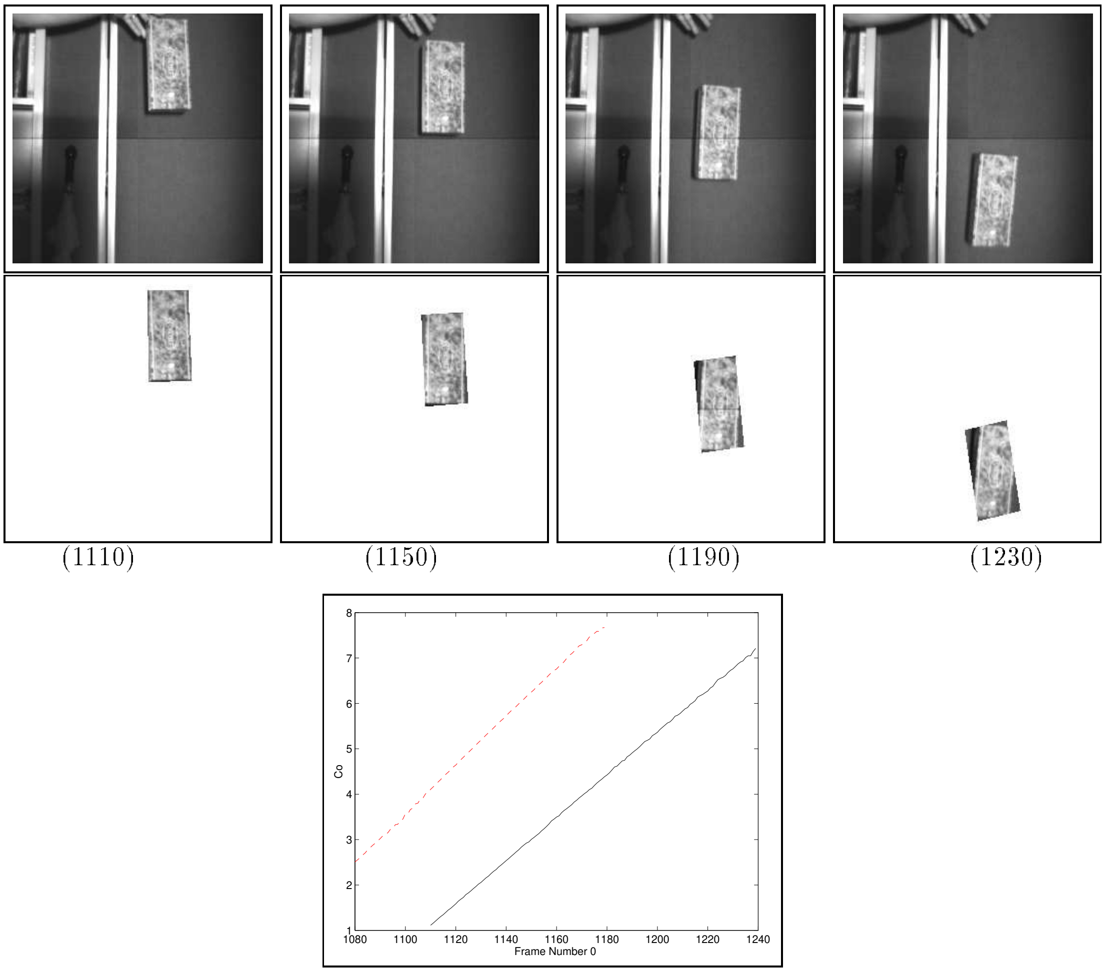


Fig. 2. A sequence of falling box (top row), the tracked box (middle row) and the recovered spatio-temporal flow coefficient throughout the sequence (solid line) and for comparison the spatio-temporal flow coefficient for the falling book (dashed line).

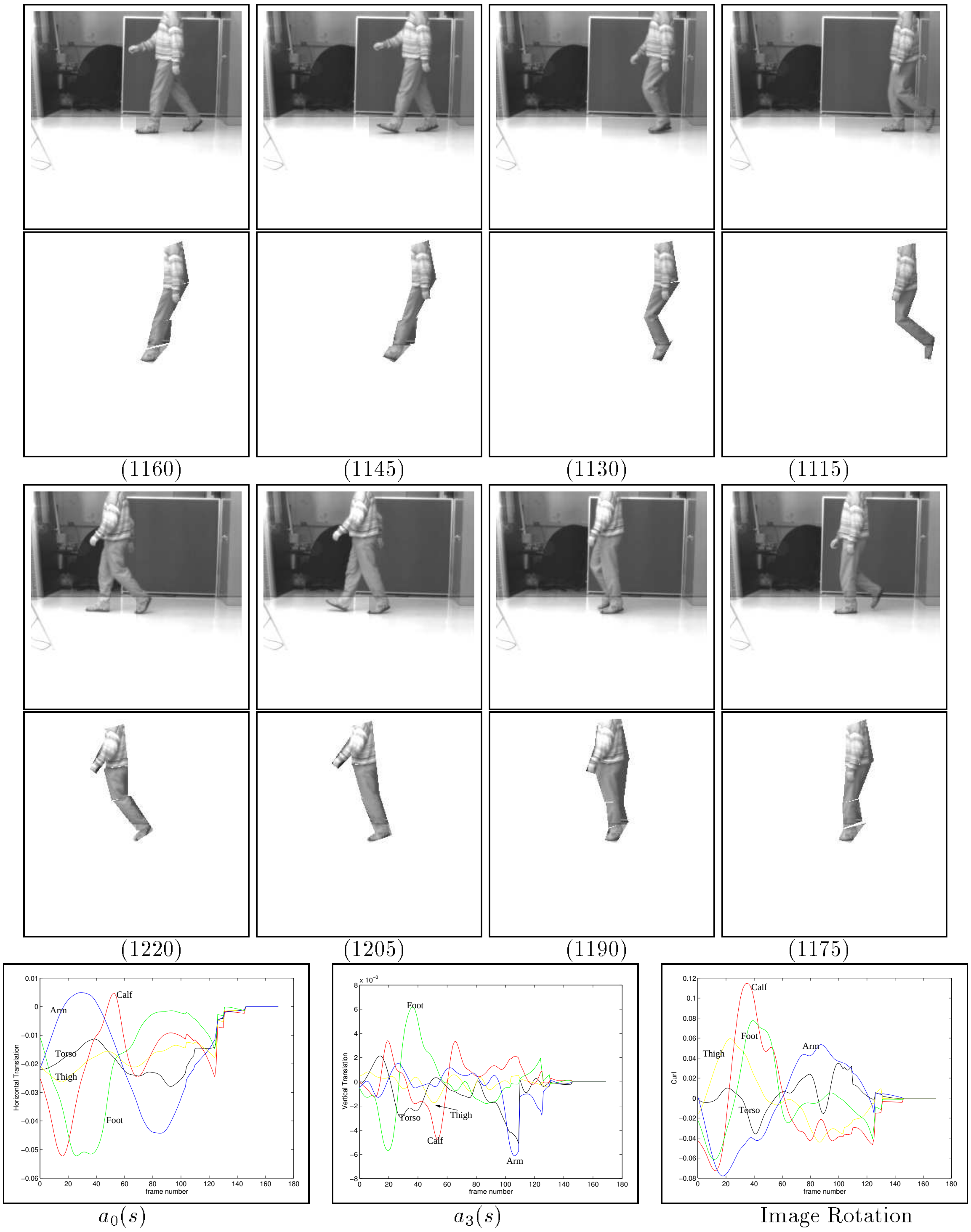


Fig. 3. A few frames from a long image sequence of a subject walking with the cardboard tracking (Ju et al., 1996). The computed horizontal, vertical and image rotation of the five body parts as modeled by the first spatio-t emporal basis vector of walking.

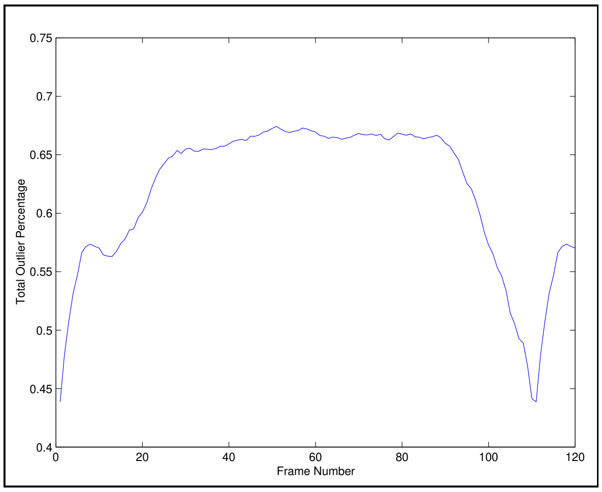


Fig. 4. Percentage of outlier points in the matching of image change to the basis set of walking for the person shown in Figure 5.

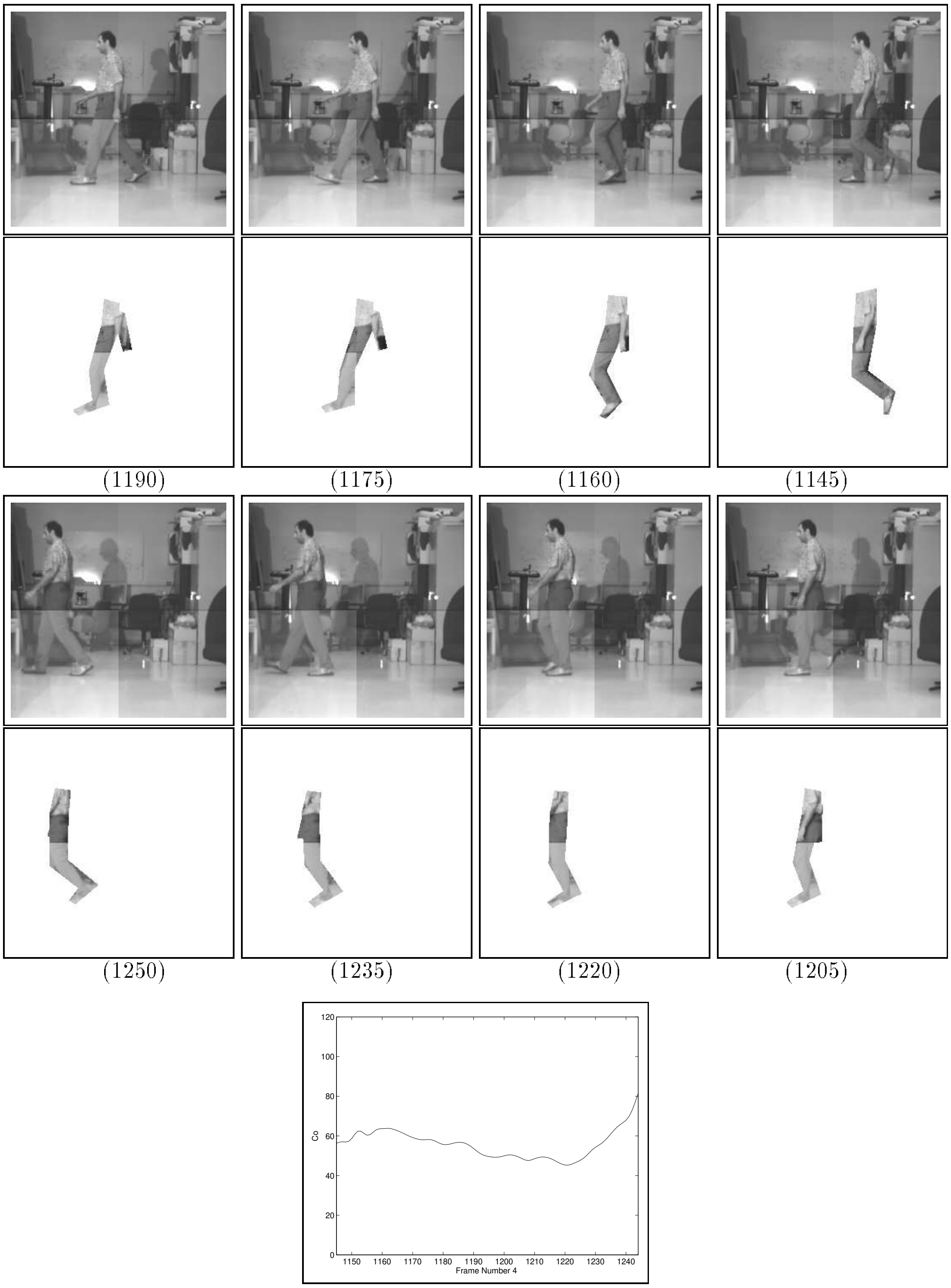


Fig. 5. A few frames from a long image sequence of a subject walking with the spatio-temporal flow tracking of a new subject's walk and the recovered coefficient.

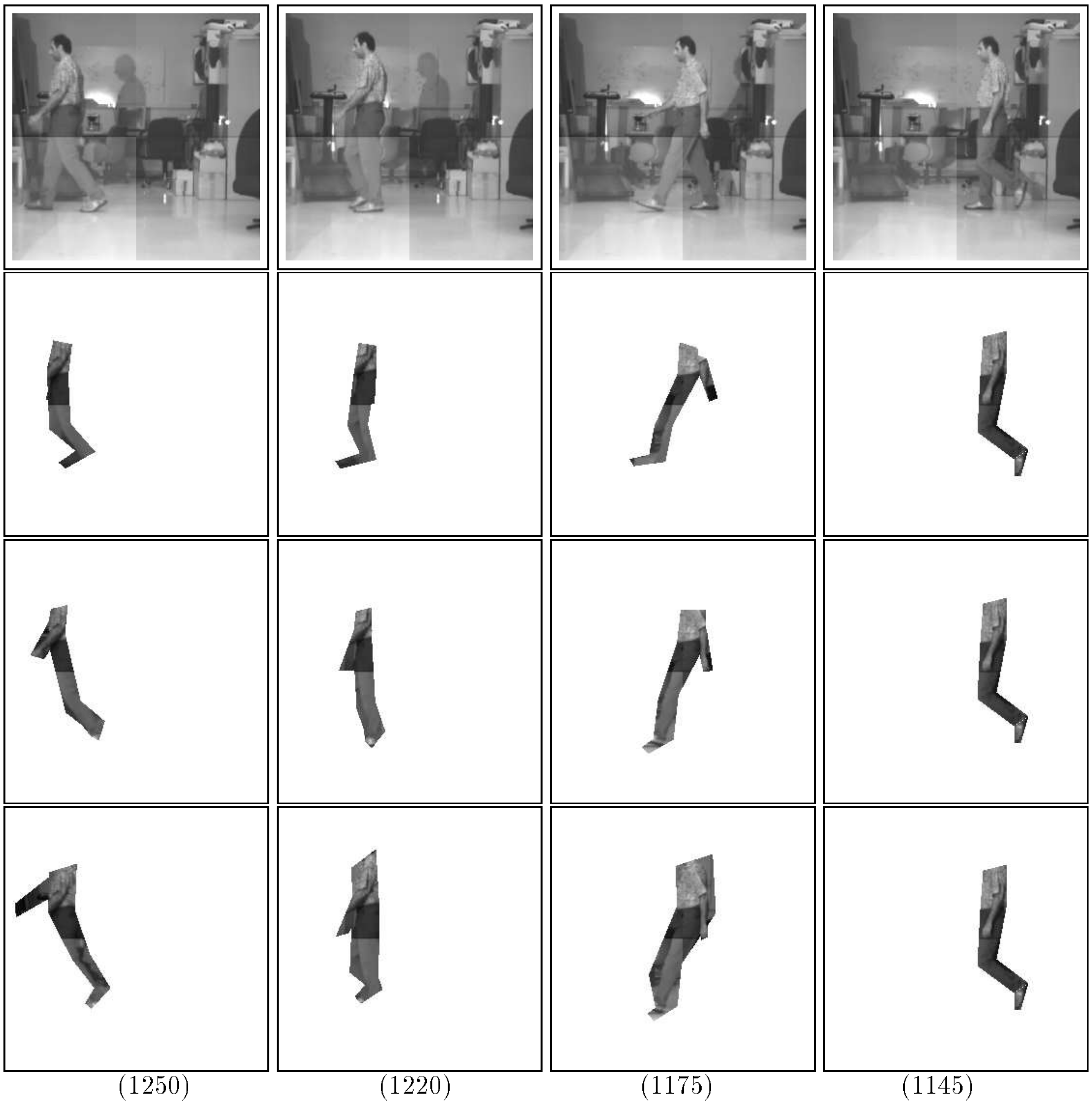
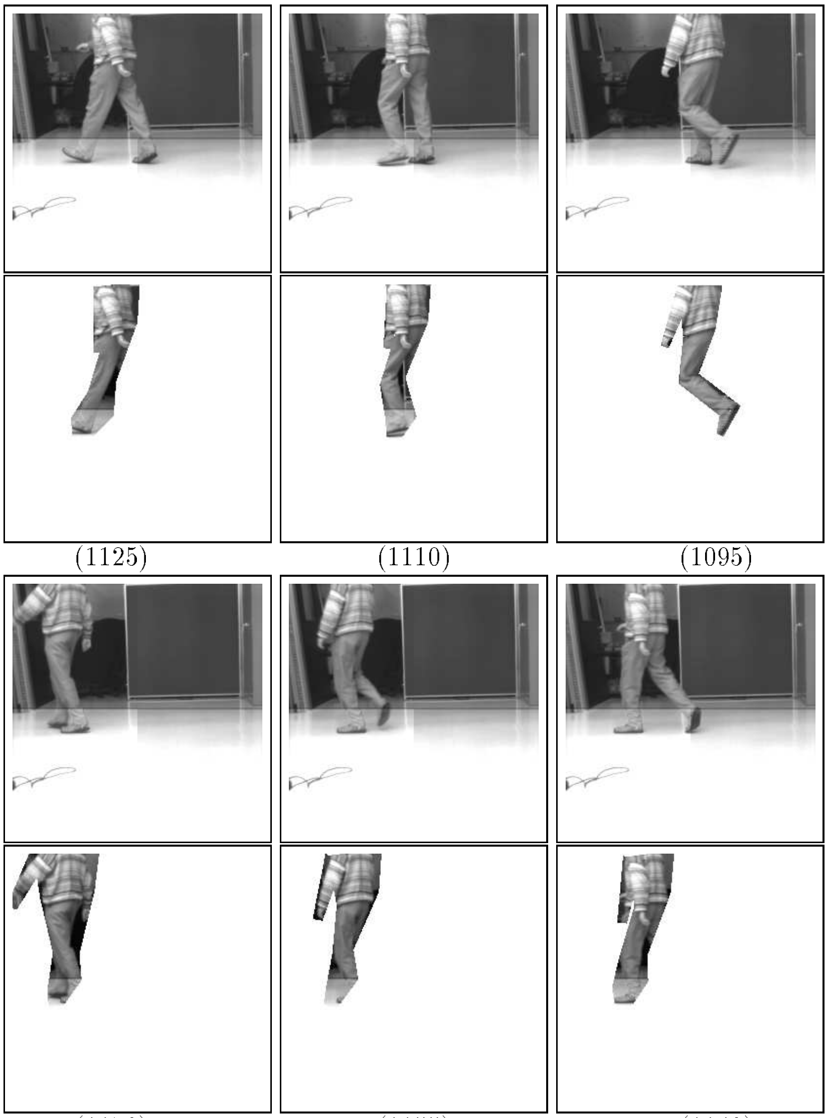


Fig. 6. A few frames from a long image sequence of a subject walking with the temporal-flow tracking of a new subject's walk employing one, two and three basis vectors (top to bottom).



(1125) (1110) (1095)

(1170) (1155) (1140)

Fig. 7. A few frames from a long image sequence of a subject walking as seen from a different viewing direction with the computed spatio-temporal flow tracking.

Fig. 8. A few frames from a long image sequence of a subject walking with the computed spatio-temporal flow tracking.

Fig. 9. A few frames fron a long image sequence of a subject walking with the spatio-temporal flow tracking of a new subject's walk for both the visible and occluded parts and the recovered coefficient.

Fig. 10. A few frames from a long image sequence of a subject marching with the spatio-temporal flow tracking of a subject's marching for both the visible and occluded parts.

Fig. 11. A few frames from a long image sequence of a subject walking with the spatio-temporal flow tracking of a new subject's walk for both the visible and occluded parts and the recovered coefficient.

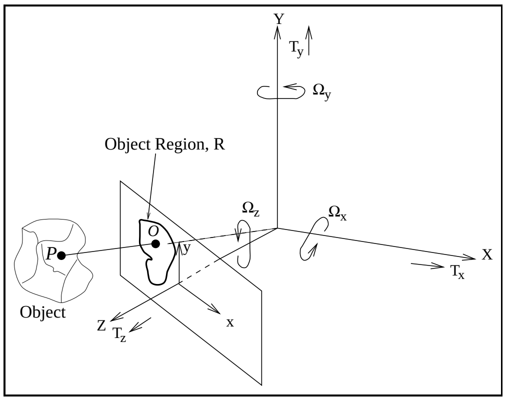


Fig. 12. The motion and geometry of the camera.

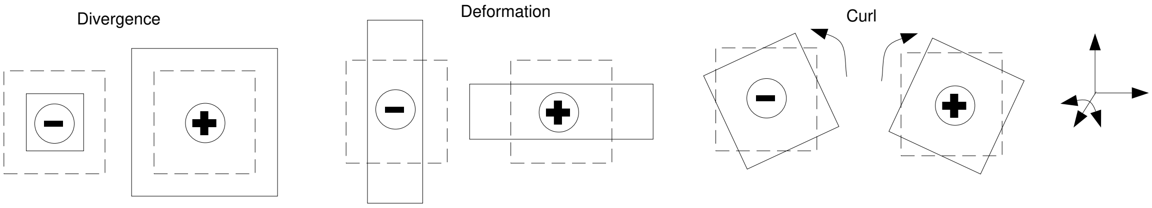


Fig. 13. The figure illustrates the motion captured by the various parameters used to represent the motion of the regions. The solid lines indicate the deformed image region and the "–" and "+" indicate the sign of the quantity.

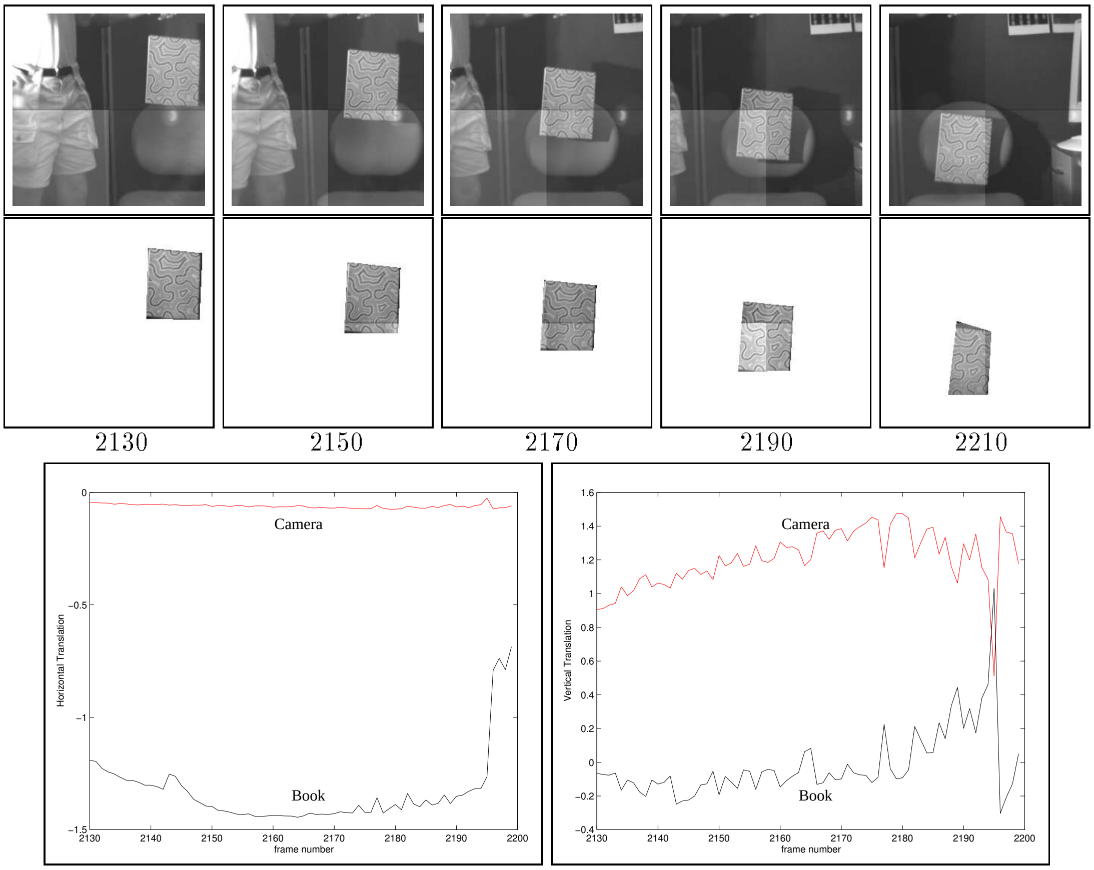


Fig. 14. A few frames from a long image sequence of a book falling while the camera is moving horizontally and the tracked book region (top and middle rows). The horizontal and vertical translations of the book and the camera are shown in the bottom row.

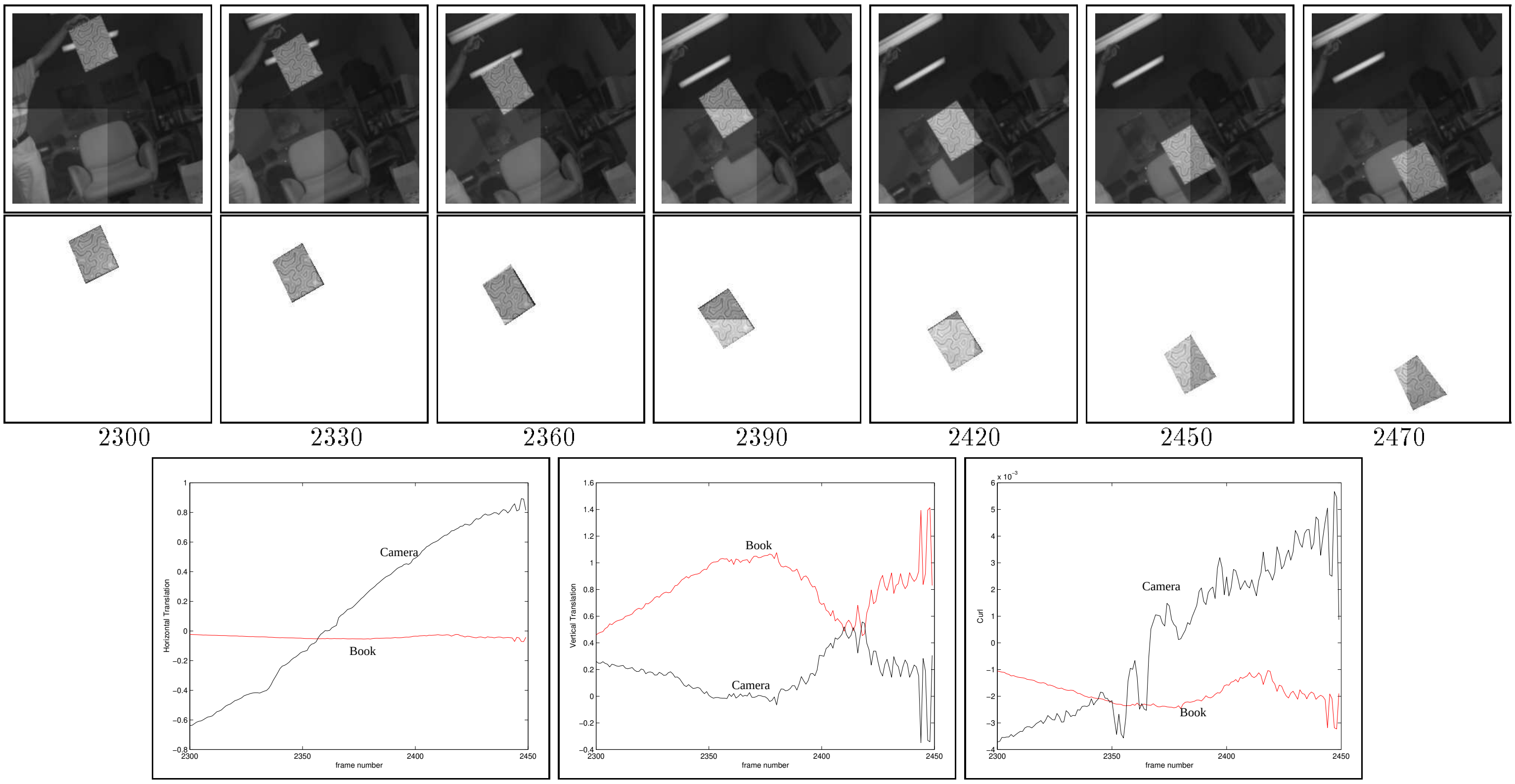


Fig. 15. A few frames from a long image sequence of a book falling while the camera is rotating clockwise and the tracked book region (top and middle rows). The horizontal and vertical translations and the rotation of the book and the camera are shown in the bottom row, left to right,respectively.

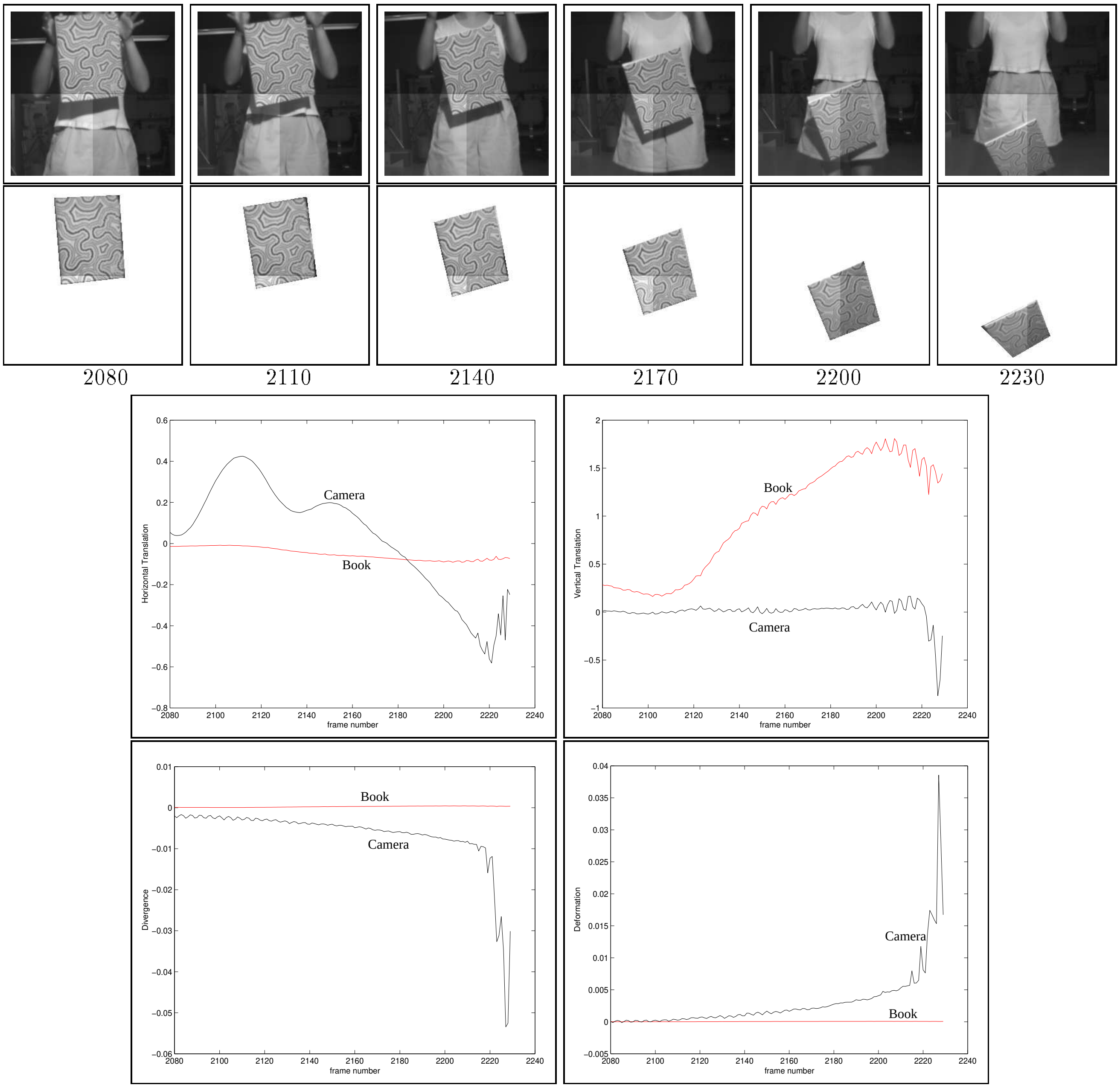


Fig. 16. A few frames from a long image sequence of a book falling while the camera is moving away in depth and the tracked book region (top and second rows). The horizontal and vertical translations (third row) and the deformation and divergence parameters of the book and the camera are shown in the bottom row.

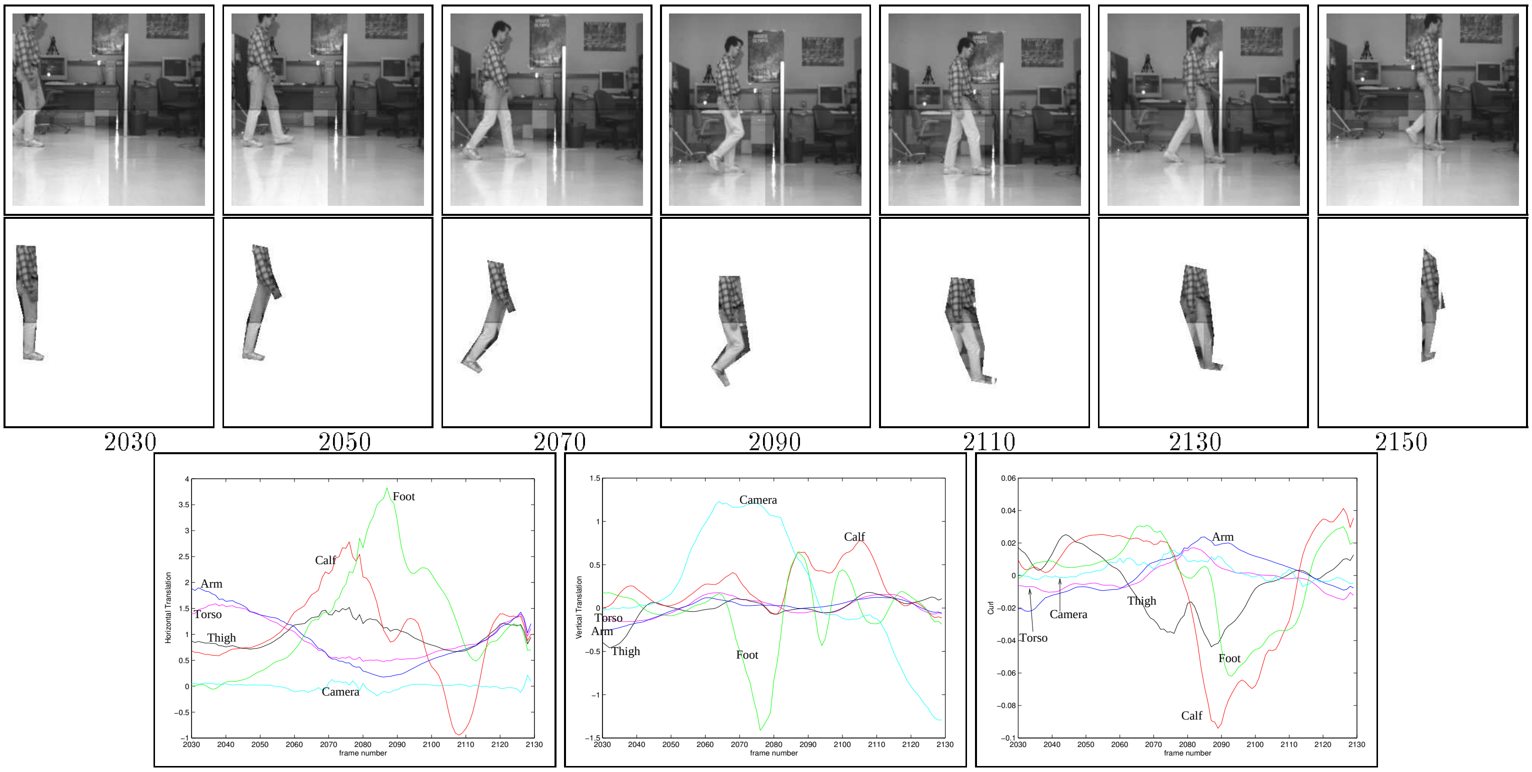


Fig. 17. A few frames from a long image sequence of a subject walking with vertical camera translation and the temporal-flow tracking of a new subject's walk and the recovered parameters for the parts and camera. The computed horizontal, vertical and image rotation of the five body parts and for the camera.

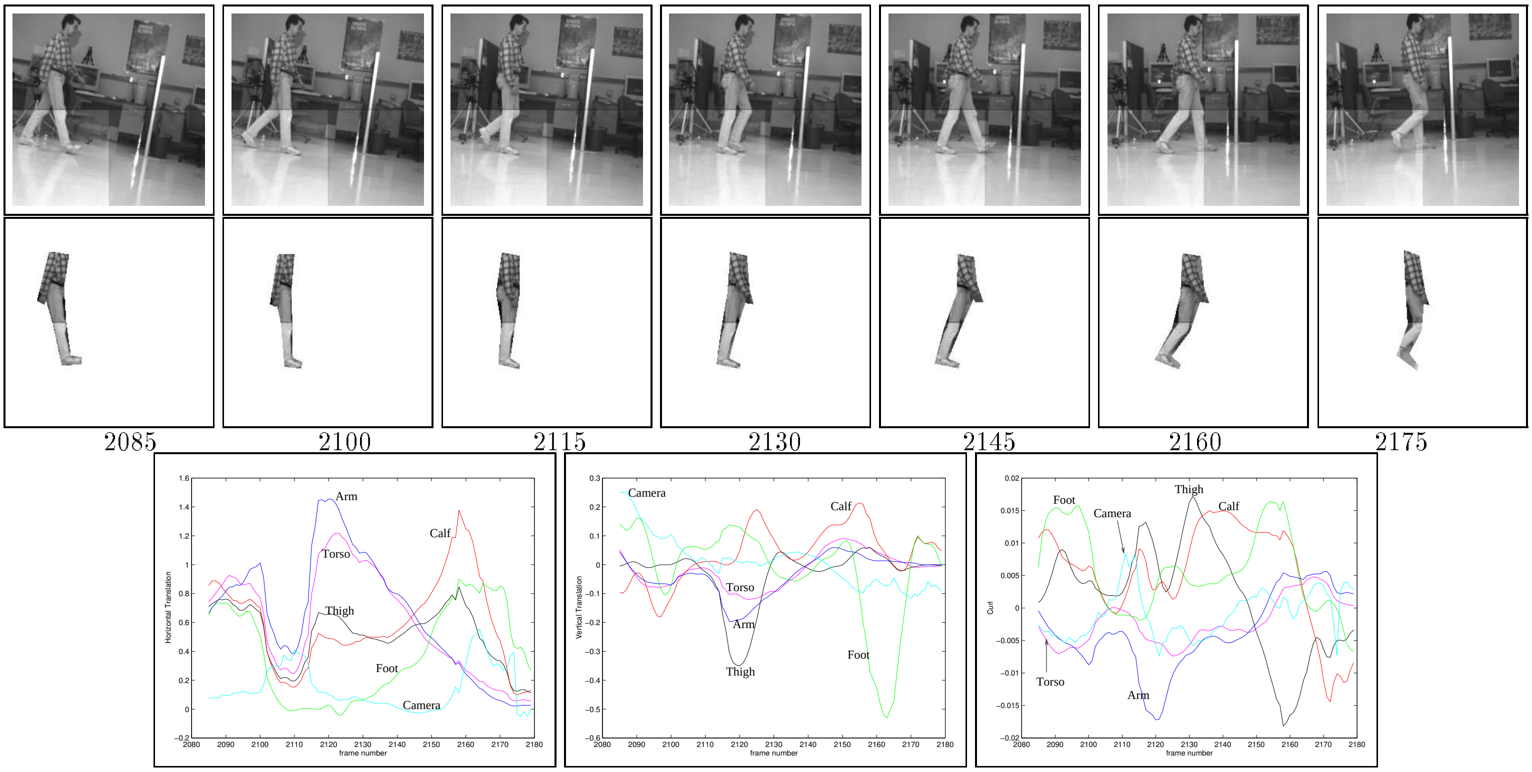


Fig. 18. A few frames from a long image sequence of a subject walking with camera rotation and the temporal-flow tracking of a new subject's walk and the recovered parameters for the parts and camera. The computed horizontal, vertical and image rotation of the five body parts, and for the camera.

Fig. 19. A few frames from a long image sequence of a subject walking with the spatio-temporal flow tracking of a subject's walk with camera rotation and the recovered parameters.