A 3D Shape Constraint on Video

Hui Ji and Cornelia Fermuller, *Member*, *IEEE*

Abstract—We propose to combine the information from multiple motion fields by enforcing a constraint on the surface normals (3D shape) of the scene in view. The fact that the shape vectors in the different views are related only by rotation can be formulated as a rank = 3 constraint. This constraint is implemented in an algorithm which solves 3D motion and structure estimation as a practical constrained minimization. Experiments demonstrate its usefulness as a tool in structure from motion providing very accurate estimates of 3D motion.

Index Terms—3-dimensional motion estimation, integration of motion fields, decoupling translation from rotation, shape and rotation.

1 Introduction

Structure from motion from single flow fields has been extensively studied and over the years many good techniques have been developed. However, the information in one motion field is not rich enough to allow for accurate estimation of 3D motion and structure. There are two issues. First, there is the ambiguity in the estimation of the motion parameters. For standard cameras with limited field of view, there is a confusion between translation and rotation [7], [10], [12], [13]. Thus, any motion algorithm, because of noise, can estimate the motion only within a range of the true solution. The second issue is the stability of structure estimation. An erroneous estimate of the motion parameters clearly will lead to errors in the estimation of structure. Furthermore, even for the correct motion parameters, the estimation of structure, because of the small displacement between the cameras (small baseline), is very unreliable. Usually, a global description of the scene, that is, the relative depth estimates of different scene patches, can be obtained rather well. However, a local description of structure, that is, shape estimates of scene patches, is very unstable.

To obtain good motion and structure estimates, we need to combine the information of consecutive motion fields. One flow field, or in the abstraction two image frames, are constrained only by the rigidity of 3D motion. The rigidity can be expressed with two constraints on the image measurements: the epipolar constraint, which says that individual flow vectors lie on a line, and the depth positivity constraint, which says that the reconstructed scene points have to be in front of the camera. Two or more motion fields are constrained in addition by the observed scene which remains constant. Existing approaches formulate this as a constraint enforcing the estimated structure, or depth of the scene, to be the same. In order to make use of this constraint, image points (or lines) over multiple frames need to be corresponded. However, automatic correspondence usually is not possible. Drifting occurs, that is, errors in correspondence accumulate till eventually the correspondence cannot be established anymore. Another problem is that, since structure in the coordinate system of one frame is related to structure in the coordinate system of another frame through both the translation and rotation, the resulting constraints are not simple and do not allow for robust estimation of structure and motion. We have the trilinear constraint resulting in 27 and the quadrilinear constraint resulting in 64 parameters whose estimations are very sensitive. [5], [9], [18], [19].

• *The authors are with the Center for Automation Research, University of Maryland, College Park, MD 20742-3275.*
E-mail: {jihui, fer}@cfar.umd.edu.

Manuscript received 3 Jan. 2005; revised 2 Oct. 2005; accepted 20 Oct. 2005; published online 13 Apr. 2006.
Recommended for acceptance by K. Daniilidis.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0004-0105.

0162-8828/06/$20.00 © 2006 IEEE Published by the IEEE Computer Society

In this paper, we propose to combine multiple motion fields, not through depth or inverse depth values, but through 3D shape. The 3D shape of a scene patch is described by the surface normal of the patch, that is, by two parameters. Consider the scene as consisting of planar patches. The surface normal of a planar patch in one view is related to the surface normal of the corresponding patch in another view only through rotation. That is, let r_1 and r_2 be the surface normals of a patch in the first and second frame, and let Ω be the rotation relating the two frames, then $r_2 = \Omega r_1$. This relationship can be formulated as a rank 3 constraint on a matrix containing the normal vectors of all the patches over all views. The advantage of this constraint is two-fold. First, we do not need to correspond image points over multiple frames, but we need to correspond image patches only, which is a much easier task. Second, the constraint can be combined easily with the estimation of motion and structure from individual flow fields. This way we can estimate structure and motion from multiple flow fields as a practical constrained minimization, where the constraint is the rank constraint on the surface normals.

The new constraint is implemented in an algorithm, which involves the following computations: We first segment planar patches in the scene and match the patches over the image sequence. The segmentation (Section 5) is based on color and motion information. Then, using as input the image gradients, the 3D motion parameters and the shape of the extracted scene patches are estimated (Section 3). Starting from motion estimates from single flow fields, we solve the constrained minimization iteratively in a two-step optimization; in one step, the surface normal are obtained, and, in the next step, the motion parameters. Section 6 presents experiments, and Section 7 discusses the role of the approach in a complete structure from motion framework.

The multiple view constraints defined on point correspondences are well understood [5], [9], [18], [19]. Nowadays, most point correspondence methods employ the technique of bundle adjustment [20] to refine 3D structure and viewing parameters. Oliensis [14], [21] proposed algorithms, which first eliminate the rotational components and then decompose the residual correspondences into structure and translation. A number of studies considered the estimation from multiple flow fields assuming continuity in the motion [2], [3], [11], [16].

Baillard and Zisserman [1] and [8] developed algorithms using line and point correspondences for the reconstruction of scenes with planar objects. The first ones to present a subspace constraint on homographies of planes in multiple views were Shashua and Avidan [17]. In the sequel, Zelnik-Manor and Irani [22] presented a subspace constraint on the relative homographies of pairs of planes across the different views. Most closely related to our work is the study of Zelnik-Manor and Irani [23], which introduced a subspace constraint on image motion. The approach assumes that differential motion between a reference frame and any other frame at the same scene points can be obtained. Clearly, this assumption limits its application to small image sequences. The improvement from our methods stems from the use of 3D shape, which makes it computationally feasible to use longer sequences.

2 Preliminaries

Let $P = (X, Y, Z)$ denote a 3D scene point and $p = (x, y)$ denote its corresponding point in the image plane $Z = f$, where $(x, y) = \frac{f}{Z}(X, Y)$. Without loss of generality, we assume $f = 1$. Then, the differential motion of the point p is expressed as:

$$\begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} = \frac{1}{Z}\begin{pmatrix} -t_x + t_z x \\ -t_y + t_z y \end{pmatrix} + \begin{pmatrix} xy\omega_x - (x^2+1)\omega_y + y\omega_z \\ -xy\omega_y + (y^2+1)\omega_x - x\omega_z \end{pmatrix},$$

where $t = (t_x, t_y, t_z)$ and $\omega = (\omega_x, \omega_y, \omega_z)$ are the translation and rotation parameters, respectively. Let P be on the world plane

$\alpha X + \beta Y + \gamma Z = 1$. The plane is described by the parameter vector $n = (\alpha, \beta, \gamma)$, which describes the depth and the surface normal. Thus, the inverse depth at P amounts $\frac{1}{Z} = \alpha x + \beta y + \gamma$. Substituting the expressions for the image motion and the plane into the well-known brightness consistency constraint:

$$\frac{\partial I}{\partial x} \cdot \frac{dx}{dt} + \frac{\partial I}{\partial y} \cdot \frac{dy}{dt} + \frac{\partial I}{\partial t} = 0,$$

we obtain the constraint

$$\begin{aligned} &[-I_x t_x - I_y t_y + (I_x x + I_y y) t_z][x\alpha + y\beta + \gamma] + \\ &(I_x xy + I_y(y^2 + 1))\omega_x - \\ &(I_y xy + I_x(x^2 + 1))\omega_y + (I_x y - I_y x)\omega_z = \\ &- I_t, \end{aligned} \tag{1}$$

which relates the image gradients to the motion and plane parameters. This equation is bilinear in the motion parameters and the plane parameters. That is, knowing t, we can solve linearly for ω and n, and vice versa. Or, similarly knowing n, we can solve linearly for t and ω. Because of the scaling ambiguity between translation and depth, we can only estimate the direction of vectors t and n.

Let $p_i = (x_i, y_i), i = 1, \cdots, N$ denote the image points on a single world plane. From (1), we obtain an equation system for this plane, which we write as:

$$\begin{aligned} &f(t, \omega, n) = 0 = \\ &[t_x A_1 + t_y A_2 + t_z A_3] \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} + B \begin{pmatrix} \omega_x \\ \omega_y \\ \omega_z \end{pmatrix} - b, \end{aligned} \tag{2}$$

where A_1, A_2, A_3, B are $N \times 3$ matrices and b is an $N \times 1$ vector, whose elements are described by the image point coordinates and the intensity gradients at points p_i. That is,

$$\begin{aligned} A_1 &= -I_{x_i}(x_i, y_i, 1)_N, \\ A_2 &= -I_{y_i}(x_i, y_i, 1)_N, \\ A_3 &= (I_{x_i} x_i + I_{y_i} y_i)(x_i, y_i, 1)_N \\ b &= (I_{t_i})_N, \\ B &= (I_{x_i} xy + I_{y_i}(y_i^2 + 1), \\ &\quad - (I_{y_i} x_i y_i + I_{x_i}(x_i^2 + 1)), \\ &\quad + (I_{x_i} y_i - I_{y_i} x_i))_N. \end{aligned}$$

2.1 Motion and Shape Estimation from Individual Flow Fields

Our method starts with 3D motion estimates from individual flow fields. In principal, many of the algorithm from the literature could be employed. We used the algorithm in [4], which is based on the equations above.

Consider a segmentation of the scene into P planar patches $V^1, V^2 \cdots V^P$. Combining equations (2) for all the patches, we obtain an overdetermined bilinear equation system of the form

$$f^p(t, \omega, n^p) = 0 \quad \text{for} \quad p = 1, \cdots, P, \tag{3}$$

whose solution provides estimates for the direction of translation, the rotation, and the structure parameters for the individual patches. The algorithm in [4] solves this minimization as a search in the space of translational directions. For each candidate translation one can solve closed-form for the rotation and the plane parameters. The translational direction minimizing (3) in the least squares sense provides the solution. Note that the algorithm does not require optical flow, but only the image gradients, which define the so-called normal flow.

2.2 Ambiguities in 3D Motion Estimation from Single Flow Fields

Several studies have addressed the noise sensitivity in structure from motion. In particular, it has been shown that for standard cameras with a small field of view imaging a shallow scene, translation, and rotation are easily confused. This can be understood by examining the differential flow equation (1). Notice that for a shallow scene with $Z(x, y)$ varying very little, a zeroth order approximation of the flow amounts to $\frac{dx}{dt} \approx \frac{-t_x}{Z} - \omega_y$ and $\frac{dy}{dt} \approx \frac{t_y}{Z} + \omega_x$. Intuitively, we can see how t_x translation along the x axis can be confused with ω_y rotation around the y axis, and t_y with ω_x for a small field of view. Thus, in the presence of noise, it is hard to distinguish these motions. The most likely estimation error is such that the projection of the translational error and the rotational error on the image are perpendicular to each other, and the estimated translation direction lies along a line passing through the true translation direction and the viewing direction [7], [10], [12], [13]. Since the estimated motion field always is noisy, we can obtain only a range of possible solutions for the motion parameters, among which the correct one lies. If we consider the 2D subspace of translational directions of this range and visualize it on the image (or on a sphere), we usually obtain an elongated region, which we refer to as the *motion valley* of solutions. Each translation direction in the motion valley, along with its best corresponding rotation and structure will agree with the observed noisy flow field. Fig. 2 shows error functions (residual of the minimization) plotted on the 2D spherical surface. The best solutions lie in the bright area of the surface. The error function makes it evident that attempting to pick a single solution in this valley is futile. Such valleys are ubiquitous, and if we pick an erroneous motion estimate, this results in the estimation of distorted structure [6].

3 Rank Constraint on 3D Shape Parameters

Let the image frames be denoted as $I_1, I_2, \cdots, I_F$ and the scene patches as V_f^p, where subscript f indicates the frame index and superscript p indicates the patch index. The translational and rotational velocities of the normal flow field at frame I_f are t_f and ω_f. One thing to be emphasized here is that the frames we use do not need to be consecutive. Actually, there is no need to combine all consecutive frames; one may combine frames far apart. The reason is that because of temporal smoothing and because of the small baseline, views close by do not provide very different information. However, note that the motion parameters t_f and ω_f are the differential velocities computed from the flow field at frame f between consecutive frames, and not the motions between the chosen frames.

The normal flow field at every frame I_f provides an equation system $\{f(t_f, \omega_f, n_f^p)\}$ (3) for the estimation of motion and depth. In order to combine the systems of the different I_f, we need to formulate some constraint which models the fact that all the frames view the same scene. We enforce the surface normals of scene patches to be the same. The surface normals in the different coordinate systems of the frames are easily related. The relative orientation of the different patches is invariant across views. The absolute orientations of a patch in the different views are related by rotation only. This invariance can be expressed as a rank constraint on a matrix containing the surface normals.

Let n_f^p be the parameter vector of the inverse depth which describes the plane with index p in frame I_f. The normal vector r_f^p of this plane is obtained by normalizing the vector n_f^p, i.e., $r_f^p = \frac{1}{\|n_f^p\|_2} n_f^p$. The normal vectors of a plane in two frames I_{f_1}, I_{f_2} are related by the rotation matrix Ω_{f_1, f_2} as

$$r_{f_1}^p = \Omega_{f_1, f_2} r_{f_2}^p \quad \text{for} \quad p = 1, \cdots, P,$$

which can be expressed in matrix form as:

$$\begin{aligned} R_{f_1} &= (r^1_{f_1}, r^2_{f_1}, \cdots, r^P_{f_1}) \\ &= (\Omega_{f_1,f_2} r^1_{f_2}, \Omega_{f_1,f_2} r^2_{f_2}, \cdots, \Omega_{f_1,f_2} r^P_{f_2}) \\ &= \Omega_{f_1,f_2}(r^1_{f_2}, r^2_{f_2}, \cdots, r^P_{f_2}) = \Omega_{f_1,f_2} R_{f_2}. \end{aligned}$$

We then combine the normal vectors in all the F frames into a $3F \times P$ matrix M of rank 3

$$M = \begin{pmatrix} R_1 \\ R_2 \\ \vdots \\ R_F \end{pmatrix} = \begin{pmatrix} I_3 R_1 \\ \Omega_{2,1} R_1 \\ \vdots \\ \Omega_{F,1} R_1 \end{pmatrix} = \begin{pmatrix} I \\ \Omega_{2,1} \\ \vdots \\ \Omega_{F,1} \end{pmatrix} R_1. \tag{4}$$

A simple, straightforward way to utilize this constraint would be to perform egomotion estimation independently on each frame I_f and subsequently perform a subspace projection on the estimated r^p_fs to regularize the estimates. Since the individually estimated r^p_fs are already very erroneous, such an approach will not lead to significant improvement. Instead, we incorporate the rank 3 constraint directly into the estimation process.

4 Motion and Shape Estimation from Multiple Flow Fields

The exposition in this chapter is as follows. We first reformulate the estimation of motion and structure on the basis of individual flow fields only. Then, we embed the 3D shape estimation into this formulation (Section 4.1) and (Section 4.2).

For every frame I_f, there is a bilinear system $\{f^p_f(t_f, \omega_f, n^p_f) = 0\}$ which we rewrite as:

$$A^p_f(t_f) n^p_f = d^p_f(\omega_f), \tag{5}$$

with

$$A^p_f(t_f) = t_{x_f} A^p_{1_f} + t_{y_f} A^p_{2_f} + t_{z_f} A^p_{3_f}$$

and

$$d^p_f(\omega_f) = -B^p_f \omega_f + b^p_f.$$

The estimation of motion and structure from the individual flow fields is formulated as a least squares optimization, that is

$$\begin{aligned} &\min_{\omega_f, t_f, n^p_f} \sum_p ||A^p_f(t_f) n^p_f - d^p_f(\omega_f)||^2 = \\ &\min_{\omega_f, t_f} \min_{n^p_f} \sum_p ||A^p_f(t_f) n^p_f - d^p_f(\omega_f)||^2. \end{aligned} \tag{6}$$

We can address this minimization in two iterative steps as follows: Given initial values for t_f and ω_f:

- Step 1: Solve for structure. Substitute the values t_f, ω_f into the system and solve the overdetermined linear system for all n^p_f using linear least squares estimation.
- Step 2: Solve for motion. Substitute the values n^p_f into the system, and solve for t_f and ω_f using least squares estimation.

Go back to Step 1 until the estimation converges.

4.1 Adding the Shape Constraint to the Estimation

The motion constraint above is defined on the vectors n^p_f, but our shape constraint is defined on the normalized surface normal vectors r^p_f. Thus, we first need to transform the nonhomogeneous system $A^p_f n^p_f = d^p_f$ for the unknown vector n^p_f into a homogeneous system $W^p_f r^p_f = 0$ with $||r^p_f|| = 1$. This is shown next.

Let H^p_f be the matrix, such that the vector d^p_f spans the null space of H^p_f, i.e., $H^p_f d^p_f = 0$. H^p_f amounts to

$$H^p_f = I - \frac{d^p_f d^{p\,t}_f}{d^{p\,t}_f d^p_f}.$$

Then, multiplying H^p_f on the both sides of the equation $A^p_f n^p_f = d^p_f$ yields

$$(H^p_f A^p_f) n^p_f = H^p_f d^p_f = 0.$$

Let $W^p_f = H^p_f A^p_f$ and $r^p_f = \frac{n^p_f}{\|n^p_f\|}$ (the normalized version of n^p_f). We then obtain the following homogeneous constraint on the normal vectors r^p_f:

$$W^p_f r^p_f = 0 \quad \text{for} \quad p = 1, \cdots, P \qquad f = 1, \cdots, F.$$

Incorporating the shape constraint, the minimization becomes:

$$\begin{aligned} &\min_{t_f, \omega_f} \min_{\|r^p_f\|=1} \sum_{p,f} ||W^p_f(t_f, \omega_f) r^p_f||^2, \\ &\text{subject to} \quad rank(M(r^p_f)) = 3. \end{aligned} \tag{7}$$

We adopt the two-step optimization for the estimation of motion and structure from multiple frames. Step 2, that is, the estimation of t_f and ω_f remains the same. Given $n^p_f = \|n^p_f\| r^p_f$, we minimize (6) using least squares. However, Step 1, the estimation of n^p_f is rather different and more difficult. We solve it by first estimating r^p_f and then estimating $\|n^p_f\|$. This is described in more detail in the next section.

4.2 Estimating the Shape Parameters

We are given estimates of t_f and ω_f and wish to estimate the r^p_f. We can pick frames which we want to combine, for example, the first, 10th, and 20th frame of the sequence. In the general case, we will not concatenate the estimates ω_f to obtain a rotation between the selected frames, but reestimate the rotation matrices.

Recall from Section 3 that $R_f = \Omega_{f,1} R_1$. Therefore,

$$R^t_f R_f - R^t_1 R_1 = R^t_1 (\Omega^t_{f,1} \Omega_{f,1} - I) R_1 = 0.$$

In other words, $R^t_f R_f$, which is the matrix encoding the relative orientations between the different patches, does not depend on the frame number. Using this, we can rewrite the minimization (7) as:

$$\begin{aligned} &\min \sum_{f,p} ||W^p_f r^p_f||^2, \quad \text{subject to} \\ &||r^p_f|| = 1, R^t_f R_f = R^t_1 R_1. \end{aligned} \tag{8}$$

This is a well-defined least squares minimization with quadratic constraints. There are many standard algorithms dealing with such types of minimization. We used the Mukai-Polak version of the Augmented Lagrangian method (ALS) ([15]) which guarantees superlinear convergence.

After having obtained the r^p_f from (8), we need to estimate $\|n^p_f\|$, the length of the n^p_f, in order to solve subsequently for translation and rotation. This is done using the first equation in $A^p_f n^p_f = d^p_f$.

4.2.1 Consecutive Frames

When combining a few consecutive frames (as opposed to significantly separated frames) $\Omega_{f,1}$ is well approximated by concatenating the differential motions ω_f. That is,

$$\Omega_{f,1} = (I + [\omega_f]_\times) \Omega_{f-1,1}.$$

For this case, the problem is much simpler. Since $\Omega_{f,1}$ is known, the minimization (7) becomes finding the least squares solution of an homogeneous system. That is,

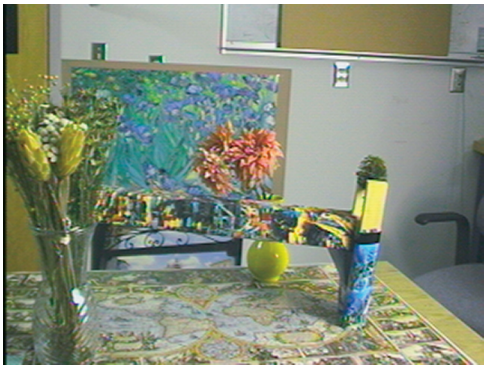


Fig. 1. Key frame for the sequence "office."

$$\min_{||r_1^p||=1, r_f^p=\Omega_{f,1} r_1^p} \sum_{p,f} ||W_f^p r_f^p||^2 =$$

$$\min_{||r_1^p||=1} \sum_p \left(\sum_f ||W_f^p \Omega_{f,1} r_1^p||^2 \right),$$

which is solved using SVD decomposition.

5 Patch Segmentation and Matching

For our purpose, the image segmentation does not need to give an "object-related" division of the scene. It only needs to locate some planar patches which are suitable for egomotion estimation and which can be matched over the image sequence. Therefore, the segmentation could be a little bit "too fine" in the sense that different patches could correspond to the same planar surface in space. However, the segmentation should not be be "too coarse" in the sense that individual image patches should not correspond to heavily curved surfaces or multiple planar surfaces.

We perform segmentation and matching using a graph-based approach, which consists of the following three components:

1. An edge enforced color segmentation in the individual frames that provides an oversegmentation.
2. Matching of the color patches in the different frames of the sequence. The geometric transformation of the patches between the frames is modeled by the transform T which includes 2D translation, rotation, and scaling. The matching is carried out by phase correlation.
3. Taking a bottom-up approach to segmentation, neighboring color patches are merged using a graph-based algorithm.

Clearly, there are more sophisticated segmentation and tracking methods in the literature. However, there is no need for these methods here. Actually, we want to make the point that our motion estimation method does not require high accuracy in tracking and segmentation.

6 Experiments

6.1 3D Motion Estimation

The color image sequence, "office" (see Fig. 1) was taken by a hand-held camera. The motion is a translation mostly along the x-axes and y-axes and a rotation. For this motion, because of the camera's small field of view, the ambiguity between translation and rotation makes the motion estimation very difficult. Fig. 2 compares the effects of different segmentations on the motion estimation from single flow fields. Although the residual value decreases (by a factor of 2.5) with better segmentation, the valleys are qualitatively very similar, demonstrating that the ambiguity in motion estimation cannot be avoided.

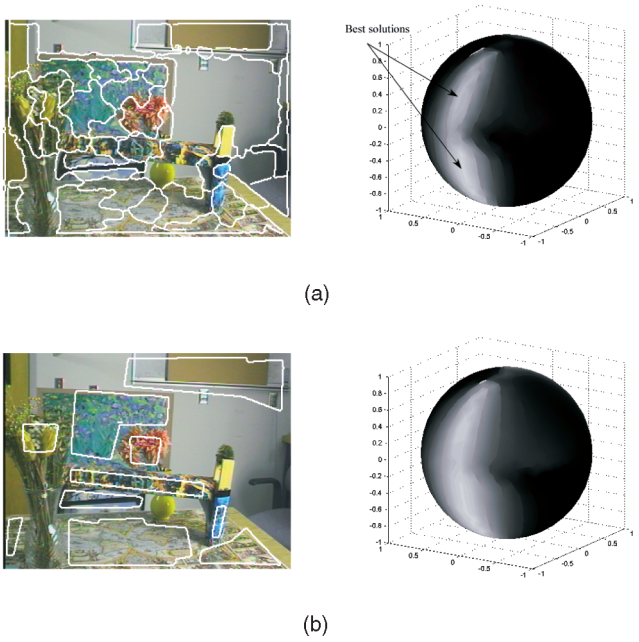


Fig. 2. Residual spheres for single frame egomotion estimation in the sequence "office." The smallest values (shown in white) denote the possible candidate solutions, which we refer to as the "motion valley." The error is found by computing for each translation the optimal rotation. (a) Automatic segmentation by the method in the paper. (b) Manual segmentation.

Next, we integrated multiple flow fields using the algorithm described in the paper. We used three frames separated by a significant baseline. Fig. 3 demonstrates a significant reduction in the ambiguity. It also shows that our segmentation performs similar to the manual one. Table 1 compares the absolute smallest values for the translational directions $(\frac{t_x}{t_z}, \frac{t_y}{t_z})$.

We compared our algorithm to the method of Zelnik-Manor and Irani [23] using the synthetic scene in Fig. 4. This method only estimates 8-parameters of the projective flow model for each patch. We added another step to obtain the 3D motion from these parameters. The results of the comparison for four different motions are presented in Fig. 5. The error in translation is measured by the angular difference between the estimated translation direction and the true direction. The error in rotation is measured by the L_2 norm between estimated and true values. Referring to Fig. 5, it can be seen that our algorithm performs significantly better on data sets 1 and 2 and a bit better on data sets 3 and 4. The reason for the decreased performance on data set 3 is the large noise in the image gradients due to the large image displacement caused by zooming. This could be remedied by introducing a hierarchical framework. All algorithms perform

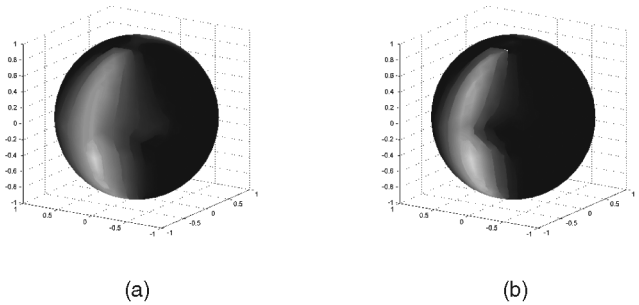


Fig. 3. Residual for multiframe motion estimation in the sequence "office." (a) Automatic segmentation. (b) Manual segmentation.

TABLE 1
Comparison of the Absolute Smallest Values for the Translational Directions $(\frac{t_x}{t_z}, \frac{t_y}{t_z})$

	Single frame	Multiple frame
Automatic Part.	(3.14, 0.02)	(2.15, -0.02)
Manual Part.	(2.30, 0.01)	(2.07, -0.01)

poorly on data set 4. The reason is the large displacement. Since there is only small translation and, thus, the flow carries very little information on structure, the combination of multiple frames cannot improve the estimation significantly.

7 The Structure from Motion Feedback Loop

We want to convey with this paper that multiple image motion fields can be combined through a constraint on 3D shape. We have implemented this constraint in a technique, and demonstrated that it provides very good 3D motion estimation. However, there may be better ways of utilizing this constraint. We consider our estimation as one module in a structure from motion framework.

We cannot obtain good models using only local measurements (image motion or correspondence) in a bottom up approach. Local image measurements do not allow for good structure estimation and localization of the discontinuities. 3D motion is not effected very much by noise, because it is globally encoded in the image, but structure is spatially local. To obtain good structure, we need processes that involve larger spatial areas. But to employ such processes, we need models of the scene. In other words, there need to be feedback loops.

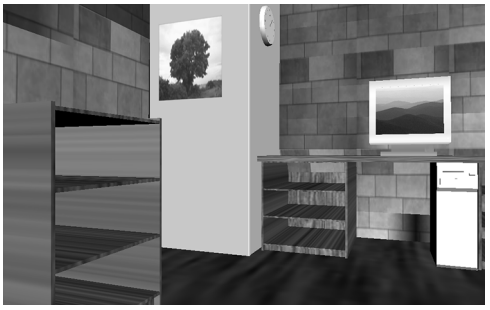
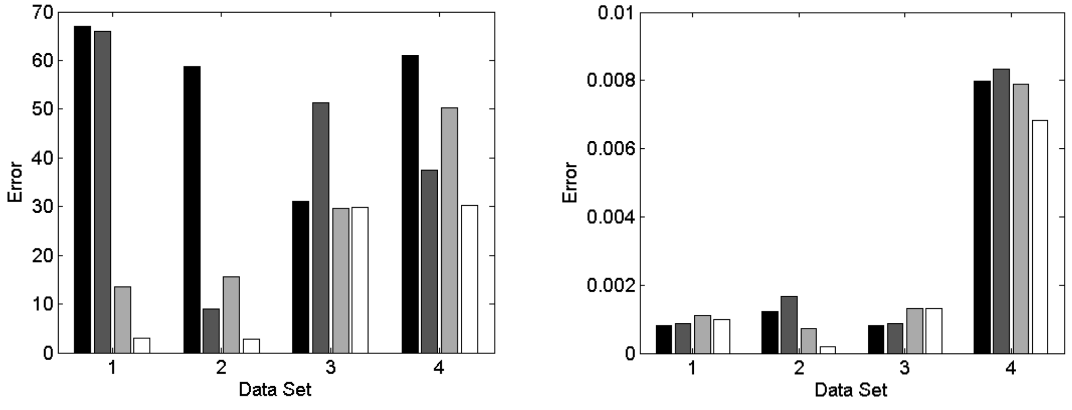


Fig. 4. Key frame of the synthesized sequence used for comparison.

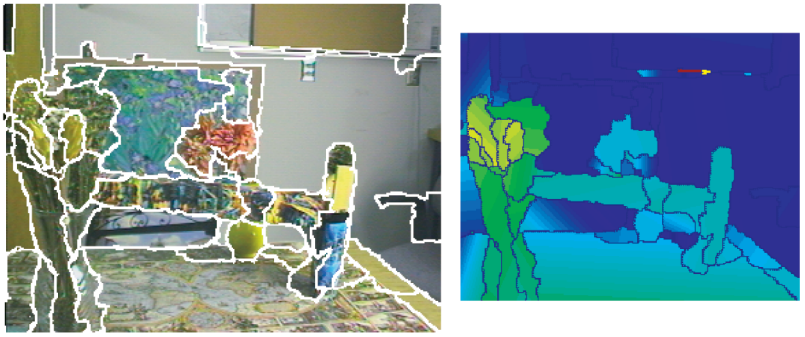
Our algorithm provides us with 3D motion estimates over multiple frames. In the sequel, we can obtain depth from image motion and we can fit shape models to the segmented patches. Using this information, we can then go back to better segment and estimate structure using images significantly separated by baseline.

Fig. 6 shows first experimental results. Using the 3D motion estimate, we rectified two significantly separated frames and computed the depth map using stereo. Then, we inserted the



(a) (b)

Fig. 5. Errors in motion estimation for different types of camera motion. The bars from black to white denote in turn: The algorithm in [23] for single image motion fields, the algorithm in [23] for multiple frames, our egomotion estimation for single flow fields, and our approach for multiple frames. The motions in the four data sets are as follows: Data set 1: translation in the $x - z$ plane and small rotation. Data set 2: translation along the y-axis and small rotation. Data set 3: dominating translation along the z-axis and small rotation. Data set 4: mostly rotation and small translation. (a) Error in translation. (b) Error in rotation.



(a) (b)

Fig. 6. Segmentation from motion and stereo and corresponding depth estimation from normal flow.

boundaries obtained from stereo into the segmentation of our algorithm and merged areas for which the flow gave continuous depth values (Fig. 6a). This gives a segmentation based on motion, stereo and color. Fig. 6b shows the depth map. On the basis of the 3D motion estimate, the parametric motion (corresponding to planes) was fit to the image intensity derivatives within the segmented areas. As can be seen, this computation, although based on flow and not disparity between faraway views, leads to a very good reconstruction.

REFERENCES

[1] C. Baillard and A. Zisserman, "Automatic Reconstruction of Piecewise Planar Models from Multiple Views," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 559-565, 1999.

[2] M.J. Black, "Combining Intensity and Motion for Incremental Segmentation and Tracking over Long Image Sequences," *Proc. European Conf. Computer Vision,* pp. 485-493, 1992.

[3] M.J. Black and P. Anandan, "Robust Dynamic Motion Estimation over Time," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 296-302, 1991.

[4] T. Brodsky, C. Fermüller, and Y. Aloimonos, "Structure from Motion: Beyond the Epipolar Constraint," *Int'l J. Computer Vision,* vol. 37, pp. 231-258, 2000.

[5] S. Carlson and D. Weinshall, "Dual Computation of Projective Shape and Camera Positions from Multiple Images," *Int'l J. Computer Vision,* vol. 27, no. 3, pp. 227-241, 1998.

[6] L. Cheong, C. Fermüller, and Y. Aloimonos, "Effects of Errors in the Viewing Geometry on Shape Estimation," *Computer Vision and Image Understanding,* vol. 71, pp. 356-372, 1998.

[7] K. Daniilidis and H.-H. Nagel, "Analytical Results on error Sensitivity of Motion Estimation from Two Views," *Image and Vision Computing,* vol. 8, pp. 297-303, 1990.

[8] A. Dick, P. Torr, and R. Cipolla, "Automatic 3D Modelling of Architecture," *Proc. British Machine Vision Conf.,* pp. 372-381, 2000.

[9] O.D. Faugeras and T. Papadopoulo, "A Nonlinear Method for Estimating the Projective Geometry of 3 Views," *Proc. Int'l Conf. Computer Vision,* pp. 477-484, 1998.

[10] C. Fermüller and Y. Aloimonos, "Observability of 3D Motion," *Int'l J. Computer Vision,* vol. 37, pp. 43-63, 2000.

[11] D. Forsyth, S. Ioffe, and J. Haddon, "Bayesian Structure from Motion," *Proc. European Conf. Computer Vision,* pp. 660-665, 1999.

[12] D.J. Heeger and A.D. Jepson, "Subspace Methods for Recovering Rigid Motion I: Algorithm and Implementation," *Int'l J. Computer Vision,* vol. 7, pp. 95-117, 1992.

[13] S.J. Maybank, "Algorithm for Analysing Optical Flow Based on the Least-Squares Method," *Image and Vision Computing,* vol. 4, pp. 38-42, 1986.

[14] J. Oliensis, "A Multi-Frame Structure-from-Motion Algorithm under Perspective Projection," *Int'l J. Computer Vision,* vol. 34, no. 2, pp. 163-192, 1999.

[15] E. Polak, *Optimization: Algorithm and Consistent Approximation.* Springer, 1996.

[16] G. Qian and R. Chellappa, "Structure from Motion Using Sequential Monte Carlo Methods," *Int'l J. Computer Vision,* vol. 59, pp. 5-31, 2004.

[17] A. Shashua and S. Avidan, "The Rank Constraint in Multiple (>= 3) View Geometry," *Proc. European Conf. Computer Vision (ECCV),* pp. 196-206, 1996.

[18] A. Shashua and M. Werman, "On the Trilinear Tensor of Three Perspective Views and Its Underlying Geometry," *Proc. Int'l Conf. Computer Vision,* 1995.

[19] M.E. Spetsakis and J. Aloimonos, "A Unified Theory of Structure from Motion," *Proc. DARPA Image Understanding Workshop,* pp. 271-283, 1990.

[20] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, "Bundle Adjustment—A Modern Synthesis," *Vision Algorithms: Theory and Practice,* B. Triggs, A. Zisserman, and R. Szeliski, eds., Springer Verlag, 2000.

[21] R. Vidal and J. Oliensis, "Structure from Planar Motions with Small Baselines," *Proc. European Conf. Computer Vision,* vol. 2, pp. 383-398, 2002.

[22] L. Zelnik-Manor and M. Irani, "Multi-View Subspace Constraints on Homographies," *Proc. Int'l Conf. Computer Vision,* 1999.

[23] L. Zelnik-Manor and M. Irani, "Multi-Frame Estimation of Planar Motion," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 10, pp. 1105-1116, Oct. 2000.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.