Simultaneous Estimation of Viewing Geometry and Structure

Tomáš Brodský, Cornelia Fermüller, and Yiannis Aloimonos

Computer Vision Laboratory, Center for Automation Research,
University of Maryland, College Park, MD 20742-3275, USA
{brodsky,fer,yiannis}@cfar.umd.edu

Abstract. Up to now, structure from motion algorithms proceeded in two well defined steps, where the first and most important step is recovering the rigid transformation between two views, and the subsequent step is using this transformation to compute the structure of the scene in view. This paper introduces a novel approach to structure from motion in which both aforementioned steps are accomplished in a synergistic manner. Existing approaches to 3D motion estimation are mostly based on the use of optic flow which however poses a problem at the locations of depth discontinuities. If we knew where depth discontinuities were, we could (using a multitude of approaches based on smoothness constraints) estimate accurately flow values for image patches corresponding to smooth scene patches; but to know the discontinuities requires solving the structure from motion problem first. In the past this dilemma has been addressed by improving the estimation of flow through sophisticated optimization techniques, whose performance often depends on the scene in view. In this paper we follow a different approach. The main idea is based on the interaction between 3D motion and shape which allows us to estimate the 3D motion while at the same time segmenting the scene. If we use a wrong 3D motion estimate to compute depth, then we obtain a distorted version of the depth function. The distortion, however, is such that the worse the motion estimate, the more likely we are to obtain depth estimates that are locally unsmooth, i.e., they vary more than the correct ones. Since local variability of depth is due either to the existence of a discontinuity or to a wrong 3D motion estimate, being able to differentiate between these two cases provides the correct motion, which yields the "smoothest" estimated depth as well as the image location of scene discontinuities. Although no optic flow values are computed, we show that our algorithm is very much related to minimizing the epipolar constraint and we present a number of experimental results with real image sequences indicating the robustness of the method.

1 Introduction and Motivation

One of the biggest challenges of contemporary computer vision is to create robust and automatic procedures for recovering the structure of a scene given multiple views. This is the well known problem of structure from motion (SFM) [4] [10].

Here the problem is treated in the differential sense, that is, assuming that a camera moving in an unrestricted rigid manner in a static environment continuously takes images. Regardless of particular approaches, the solution always proceeds in two steps: first, the rigid motion between the views is recovered and, second, the motion estimate is used to recover the scene structure.

Traditionally, the problem has been treated by first finding the correspondence or optic flow and then optimizing an error criterion based on the epipolar constraint. Although considerable progress has been made in minimizing deviation from the epipolar constraint [11] [13] [14], the approach is based on the values of flow whose estimation is an ill-posed problem.

The values of flow are obtained by applying some sort of smoothing to the locally computed image derivatives. When smoothing is done in an image patch corresponding to a smooth scene patch, accurate flow values are obtained. When, however, the patch corresponds to a scene patch containing a depth discontinuity, the smoothing leads to erroneous flow estimates there. This can only be avoided if a priori knowledge about the locations of depth discontinuities is available. Thus, flow values close to discontinuities often contain errors (and these affect the flow values elsewhere) and when the estimated 3D motion (containing errors) is used to recover depth, it is unavoidable that an erroneous scene structure will be computed. The situation presents itself as a chicken-and-egg problem. If we had information about the location of the discontinuities, then we would be able to compute accurate flow and subsequently accurate 3D motion. Accurate 3D motion implies, in turn, accurate location of the discontinuities and estimation of scene structure. Thus 3D motion and scene discontinuities are inherently related through the values of image flow and the one needs the other to be better estimated. Researchers avoid this problem by attempting to first estimate flow using sophisticated optimization procedures that could account for discontinuities, and although such techniques provide better estimates, their performance often depends on the scene in view, they are in general very slow and require a large number of resources [7] [12] [15].

In this paper, instead of attempting to estimate flow at all costs before proceeding with structure from motion, we ask a different question: Would it be possible to utilize any available local image motion information, such as normal flow for example, in order to obtain knowledge about scene discontinuities which would allow better estimation of 3D motion? Or, equivalently, would it be possible to devise a procedure that estimates scene discontinuities while at the same time estimating 3D motion? We show here that this is the case and we present a novel algorithm for 3D motion estimation. The idea behind our approach is based on the interaction between 3D motion and scene structure that only recently has been formalized [3]. If we have a 3D motion estimate which is wrong and we use it to estimate depth, then we obtain a distorted version of the depth function. Not only do incorrect estimates of motion parameters lead to incorrect depth estimates, but the distortion is such that the worse the motion estimate, the more likely we are to obtain depth estimates that locally vary much more than the correct ones. The correct motion then yields the "smoothest" estimated depth and we can define a measure whose minimization yields the correct ego-motion parameters. The measure can be computed from normal flow only, so the computation of optical flow is not needed by the algorithm. Intuitively, the proposed algorithm proceeds as follows: first, the image is divided into small patches and a search — for the 3D motion — which as explained in Sect. 3 takes place in the two-dimensional space of translations — is performed. For each candidate 3D motion, using the local normal flow measurements in each patch, the depth of the scene corresponding to the patch is computed. If the variation of depth for all patches is small, then the candidate 3D motion is close to the correct one. If, however, there is a significant variation of depth in a patch, this is either because the candidate 3D motion is inaccurate or because there is a discontinuity in the patch. The second situation is differentiated from the first if the distribution of the depth values inside the patch is bimodal with the two classes of values spatially separated. In such a case the patch is subdivided into two new ones and the process is repeated. When the depth values computed in each patch are smooth functions, the corresponding motion is the correct one and the procedure has at the same time given rise to the location of a number of discontinuities. The rest of the paper formalizes these ideas and presents a number of experimental results.

1.1 Organization of the Paper

Sect. 2 defines the imaging model and describes the equations of the motion field induced by rigid motion; it also develops certain constraints that will be of use later and makes explicit the relationship between distortion of depth and errors in 3D motion. Sect. 3 is devoted to the description of the algorithm and Sect. 4 describes a number of experimental results with real image sequences. It also formalizes the relationship of the approach to algorithms utilizing the epipolar constraint.

2 Preliminaries

We consider an observer moving rigidly in a static environment. The camera is a standard calibrated pinhole with focal length f and the coordinate system $OXYZ$ is attached to the camera, with Z being the optical axis.

Image points are represented as vectors $\mathbf{r} = [x, y, f]^{\mathrm{T}}$, where x and y are the image coordinates of the point and f is the focal length in pixels. A scene point $\mathbf{R}$ is projected onto the image point

$$\mathbf{r} = f \frac{\mathbf{R}}{\mathbf{R} \cdot \hat{\mathbf{z}}} \tag{1}$$

where $\hat{\mathbf{z}}$ is the unit vector in the direction of the Z axis.

Let the camera move in a static environment with instantaneous translation $\mathbf{t}$ and instantaneous rotation $\boldsymbol{\omega}$ (measured in the coordinate system $OXYZ$). Then a scene point $\mathbf{R}$ moves with velocity (relative to the camera)

$$\dot{\mathbf{R}} = -\mathbf{t} - \boldsymbol{\omega} \times \mathbf{R}\,. \tag{2}$$

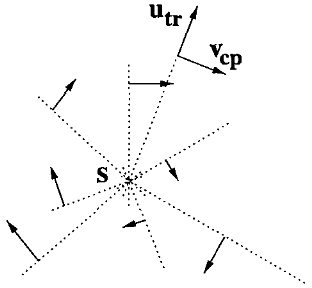

Fig. 1. The copoint directions corresponding to $\hat{\mathbf{t}}$

The image motion field is then the usual [9]

$$\dot{\mathbf{r}} = -\frac{1}{(\mathbf{R}\cdot\hat{\mathbf{z}})}(\hat{\mathbf{z}} \times (\mathbf{t} \times \mathbf{r})) + \frac{1}{f}\,\hat{\mathbf{z}} \times (\mathbf{r} \times (\boldsymbol{\omega} \times \mathbf{r})) = \frac{1}{Z}\,\mathbf{u}_{\rm tr}(\mathbf{t}) + \mathbf{u}_{\rm rot}(\boldsymbol{\omega}) \qquad (3)$$

where Z is used to denote the scene depth $(\mathbf{R}\cdot\hat{\mathbf{z}})$, and $\mathbf{u}_{\rm tr}, \mathbf{u}_{\rm rot}$ the direction of the translational and the rotational flow respectively. Due to the scaling ambiguity, only the direction of translation (focus of expansion—FOE, or focus of contraction—FOC, depending on whether the observer approaches or moves away from the scene) and the three rotational parameters can be estimated from monocular image sequences.

2.1 Projections of Motion Fields

The search for candidate 3D motions is achieved by searching for the translational component, i.e., the FOE. Given a candidate FOE, the rotation that best fits image data can be obtained by examining normal flow vectors that are perpendicular to lines passing from the FOE, since these flow values would not contain any translational part. For this we need to formalize properties of motion fields projected onto particular directions. Indeed, by projecting the motion field vectors $\dot{\mathbf{r}}$ onto certain directions, it is possible to gain some very useful insights [5] [6] [8] [1]. Of particular interest, are the *copoint* projections [5], where the flow vectors are projected onto directions perpendicular to a certain translational flow field (see Fig. 1). Let point $\hat{\mathbf{t}}$ be the FOE in the image plane of this translational field and consider the vectors emanating from $\hat{\mathbf{t}}$. Vectors perpendicular to such vectors are $\mathbf{v}_{\rm cp}(\mathbf{r}) = \hat{\mathbf{z}} \times \mathbf{u}_{\rm tr}(\hat{\mathbf{t}}) = \hat{\mathbf{z}} \times (\hat{\mathbf{z}} \times (\hat{\mathbf{t}} \times \mathbf{r}))$.

Let the camera motion be $(\mathbf{t}, \boldsymbol{\omega})$. Projection of the flow (3) onto $\mathbf{v}_{\rm cp}$ is

$$\dot{\mathbf{r}} \cdot \frac{\mathbf{v}_{\rm cp}}{\|\mathbf{v}_{\rm cp}\|} = \frac{1}{\|\mathbf{v}_{\rm cp}\|}\left(\frac{1}{Z}p_{\rm t}(\mathbf{t}, \hat{\mathbf{t}}, \mathbf{r}) + p_\omega(\boldsymbol{\omega}, \hat{\mathbf{t}}, \mathbf{r})\right) \qquad (4)$$

where the functions $p_{\rm t}$, p_ω are:

$$p_{\rm t}(\mathbf{t}, \hat{\mathbf{t}}, \mathbf{r}) = f\,(\mathbf{t} \times \hat{\mathbf{t}}) \cdot \mathbf{r}$$
$$p_\omega(\boldsymbol{\omega}, \hat{\mathbf{t}}, \mathbf{r}) = (\boldsymbol{\omega} \times \mathbf{r}) \cdot (\hat{\mathbf{t}} \times \mathbf{r})\ .$$

In particular, if we let $\hat{\mathbf{t}} = \mathbf{t}$, the translational component of the copoint projection becomes zero and (4) simplifies into

$$\dot{\mathbf{r}} \cdot \frac{\mathbf{v}_{\mathrm{cp}}}{||\mathbf{v}_{\mathrm{cp}}||} = \frac{1}{||\mathbf{v}_{\mathrm{cp}}||} p_\omega(\boldsymbol{\omega}, \hat{\mathbf{t}}, \mathbf{r}) \; . \tag{5}$$

2.2 Estimating Rotation from Known Translation

If we can estimate the translation of the camera, it is quite simple to estimate the rotational component of the camera motion. Let us for now assume that the translation $\mathbf{t}$ is known. Then the rotation can be estimated based on (5), as long as there are some normal flow measurements in the direction of the appropriate copoint vectors. Since (5) is linear in the elements of $\boldsymbol{\omega}$, we can setup a least squares minimization to estimate $\boldsymbol{\omega}$ based on the appropriate normal flow measurements.

Specifically at points with suitable normal flow direction, we have equations

$$\dot{\mathbf{r}} \cdot \frac{\mathbf{v}_{\mathrm{cp}}}{||\mathbf{v}_{\mathrm{cp}}||} = \left(\frac{1}{||\mathbf{v}_{\mathrm{cp}}||} (\mathbf{t} \times \mathbf{r})^{\mathrm{T}} \right) \boldsymbol{\omega}$$

that are linear in $\boldsymbol{\omega}$.

2.3 Depth Estimation from Motion Fields

This section introduces the novel criterion of "smoothness of depth" which is used to evaluate the consistency of the normal flow field with the estimated 3D motion and also to segment the scene at its depth boundaries. The idea is based on the interrelationship of estimated 3D motion $(\hat{\mathbf{t}}, \hat{\boldsymbol{\omega}})$ and estimated depth of the scene $\hat{Z}$.

The structure of the scene, i.e., depth computed, can be expressed as a function of the estimated translation $\hat{\mathbf{t}}$ and the estimated rotation $\hat{\boldsymbol{\omega}}$. At an image point $\mathbf{r}$ where the normal flow direction is $\mathbf{n}$, the inverse scene depth can be estimated from (3) as

$$\frac{1}{\hat{Z}} = \frac{\dot{\mathbf{r}} \cdot \mathbf{n} - \mathbf{u}_{\mathrm{rot}}(\hat{\boldsymbol{\omega}}) \cdot \mathbf{n}}{\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}}) \cdot \mathbf{n}} \tag{6}$$

where $\mathbf{u}_{\mathrm{rot}}(\hat{\boldsymbol{\omega}}), \mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})$ refer to the estimated rotational and translational flow respectively.

Substituting into (6) from (3), we obtain:

$$\frac{1}{\hat{Z}} = \frac{\frac{1}{Z}\mathbf{u}_{\mathrm{tr}}(\mathbf{t}) \cdot \mathbf{n} - \mathbf{u}_{\mathrm{rot}}(\delta\boldsymbol{\omega}) \cdot \mathbf{n}}{\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}}) \cdot \mathbf{n}} \quad \text{or}$$

$$\frac{1}{\hat{Z}} = \frac{1}{Z} \frac{\mathbf{u}_{\mathrm{tr}}(\mathbf{t}) \cdot \mathbf{n} - Z\mathbf{u}_{\mathrm{rot}}(\delta\boldsymbol{\omega}) \cdot \mathbf{n}}{\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}}) \cdot \mathbf{n}}$$

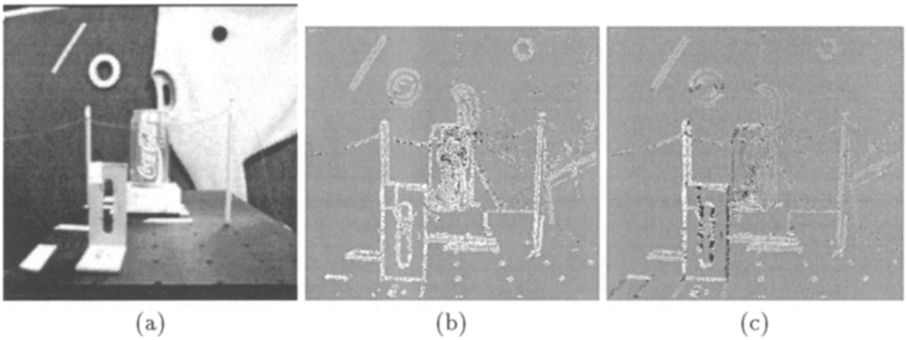

Fig. 2. (a) One frame of the NASA Coke can sequence. (b) Inverse depth estimated for the correct FOE in the middle of the image. (c) Inverse depth for an incorrect FOE 100 pixels to the left of the correct FOE (the image size was 512 × 512 pixels). The grey-level value represents inverse estimated depth with mid-level grey representing zero, white representing large positive $1/Z$ and black representing large negative $1/Z$

where $\mathbf{u}_{\text{rot}}(\delta\boldsymbol{\omega})$ is the rotational flow due to the rotational error $\delta\boldsymbol{\omega} = (\hat{\boldsymbol{\omega}} - \boldsymbol{\omega})$. To make clear the relationship between actual and estimated depth we write

$$\hat{Z} = Z \cdot D \tag{7}$$

with

$$D = \frac{\mathbf{u}_{\text{tr}}(\hat{\mathbf{t}}) \cdot \mathbf{n}}{(\mathbf{u}_{\text{tr}}(\mathbf{t}) - Z\mathbf{u}_{\text{rot}}(\delta\boldsymbol{\omega})) \cdot \mathbf{n}}$$

hereafter termed the distortion factor. Equation (7) shows how wrong depth estimates are produced due to inaccurate 3D motion values. The distortion factor for any direction $\mathbf{n}$ corresponds to the ratio of the projections of the two vectors $\mathbf{u}_{\text{tr}}(\hat{\mathbf{t}})$ and $\mathbf{u}_{\text{tr}}(\mathbf{t}) + Z\mathbf{u}_{\text{rot}}(\delta\boldsymbol{\omega})$ on $\mathbf{n}$. The larger the angle between these two vectors is the more the distortion will be spread out over the different directions. Thus, considering a patch of a smooth surface in space and assuming that normal flow measurements are taken along many directions, a rugged (i.e., unsmooth) surface will be computed on the basis of wrong 3D motion estimates. To give an example, we show the estimated depth for the NASA Coke can sequence (one frame is shown in Fig. 2 (a)). For two different translations we estimate the rotation as in Sect. 2.2 and then plot the estimated values of (6). Notice the reasonably smooth depth estimates for the correct FOE in Fig. 2 (b) and compare with sharp changes in the depth map (neighboring black and white regions) in Fig. 2 (c).

The above observation constitutes the main idea behind our algorithm. For a candidate 3D motion estimate we evaluate the smoothness of estimated depth within image patches. If the chosen image patches correspond to smooth 3D scene patches the correct 3D motion will certainly give rise to the overall smoothest image patches. To obtain such situation we attempt a segmentation of scene on the basis of estimated depth while at the same time testing the candidate 3D motions. As all computations are based on normal flow and we thus have available the full statistics of the raw data a good segmentation generally is possible.

3 Finding the Focus of Expansion

3.1 Algorithm Description

The focus of expansion is found by localizing the minimum of function $\Phi(\hat{\mathbf{t}})$, which constitutes a measure of the estimated depth variation. Let $\hat{\mathbf{t}}$ be a candidate translation, and estimate the scene depth we would obtain if $\hat{\mathbf{t}}$ were the correct translation. To obtain $\Phi(\hat{\mathbf{t}})$ we examine variation of the estimated depth over small image regions and compute a weighted sum of the local results.

We first summarize the computation of $\Phi(\hat{\mathbf{t}})$ and then explain the algorithm. To obtain $\Phi(\hat{\mathbf{t}})$

1. Estimate the camera rotation $\hat{\omega}$ from the copoint vectors corresponding to $\hat{\mathbf{t}}$.
2. Compute the depth estimates $1/\hat{Z}$ from (6) and multiply them by the average size of $\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})$ in the region.
3. Partition the image into small regions, for each region $\mathcal{R}$ check whether it needs to be segmented as described in Sect. 3.2. If the region cannot be segmented, the error measure $\Theta(\hat{\mathbf{t}}, \mathcal{R})$ is the variance of all the estimated $1/\hat{Z}$ values. Otherwise, compute the variances of $1/\hat{Z}$ in each subregion separately and let $\Theta(\hat{\mathbf{t}}, \mathcal{R})$ be the sum of the variances.
4. The depth variation measure $\Phi(\hat{\mathbf{t}})$ is the sum of $\Theta(\hat{\mathbf{t}}, \mathcal{R})$ over all image regions $\mathcal{R}$.

Several steps of the algorithm need additional explanation. The scaling of the inverted depth values in step 2 corresponds to making $\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})$ a unit vector and it is equivalent to the proper scaling of the epipolar constraint.

The smoothness measure used here corresponds to deviations of the inverse depth from a fronto-parallel plane. If more precision is required or if large image regions are used, we may fit a plane through the estimated $1/\hat{Z}$ values and replace each $1/\hat{Z}$ by its distance from the plane. Then we reduce the errors caused by small changes of the flow field within the image region.

As shown in Sect. 3.4, the smoothness measure gives provably correct results for any image patch that corresponds to a single smooth scene surface. In order to obtain only such image patches we perform a segmentation of the scene, as described in the next section. Of course, a correct segmentation at all locations cannot be guaranteed. This, however, should not influence much the solution of the 3D motion estimation, as even for patches that contain depth discontinuity, in many cases the smoothness measure is smaller for the correct 3D motion than for incorrect ones.

To find the minimum of Φ and thus the FOE/FOC, we perform a hierarchical search over the two-dimensional space of FOE positions. In practice, the function Φ is quite smooth, that is small changes in $\hat{\mathbf{t}}$ give rise to only small changes in Φ. One of the reasons for this is that for any $\hat{\mathbf{t}}$, the value of $\Phi(\hat{\mathbf{t}})$ is influenced by all the normal flow measurements and not only by a small subset.

For most motion sequences, the motion of the camera does not change abruptly. Then the FOE position does not change much between frames and a complete search only has to be performed for the first flow field. In the successive flow fields, we can search only in a smaller area centered around the previously estimated FOE.

3.2 Patch Segmentation

By segmenting a patch, we decrease its contribution to $\Phi(\hat{\mathbf{t}})$. Ideally, an image patch should be segmented if it contains two smooth scene surfaces separated by a depth discontinuity and the depth is estimated using the correct 3D motion and it should not be segmented if it corresponds to one smooth surface and the depth is estimated on the basis of incorrect 3D motion. If that is the case, we decrease $\Phi(\hat{\mathbf{t}})$ for the correct translation while keeping the measure large for incorrect translations.

We use a simple segmentation criterion. First of all, segmentation is attempted only for patches where the estimated depth is not smooth. If the histogram of depth values in the region is bimodal and the region can be divided into two spatially coherent regions that correspond to the two modes of the histogram, we divide the patch into the two subregions.

It is simple to show that this strategy can be expected to yield good patch segmentation. When the motion estimate is correct, also the depth estimates are correct and patches with large amount of depth variation contain depth discontinuities.

For incorrect motions, the distortion factor depends on the direction of normal flow. While for any patch we can split the depth estimates into two groups and decrease the variance, it is highly unlikely that the two groups of measurements define two spatially coherent separate subregions. More likely, the depth estimates from both groups are randomly distributed within the patch.

However, as the segmentation is based on local information only, we cannot expect it to perfectly distinguish the depth discontinuities from depth variation due to incorrect 3D motion. Sometimes, if a patch contains two subregions with different distributions of normal flow directions, we may split the patch even for incorrect 3D motion. An improvement, however, could be achieved by taking into account also the distribution of normal flow directions in relation to the estimated depth.

The local results are used in a global measure and occasional segmentation errors are unlikely to change the overall results. For the segmentation to cause an incorrect motion to yield the smallest $\Phi(\hat{\mathbf{t}})$, special normal flow configurations would have to occur in many patches in the image.

3.3 The Estimated Rotation

The analysis of $\Phi(\hat{\mathbf{t}})$ should start by relating the estimated rotation $\hat{\boldsymbol{\omega}}$ and the true motion parameters $(\mathbf{t}, \boldsymbol{\omega})$. When $\hat{\boldsymbol{\omega}}$ is computed, the measurements used in (5) are given by (4), i.e.,

$$\dot{\mathbf{r}} \cdot \mathbf{n} = \frac{1}{\|\mathbf{v}_{\mathrm{cp}}\|} p_\omega(\boldsymbol{\omega}, \hat{\mathbf{t}}, \mathbf{r}) + \frac{1}{\|\mathbf{v}_{\mathrm{cp}}\|} \frac{1}{Z} p_{\mathrm{t}}(\mathbf{t}, \hat{\mathbf{t}}, \mathbf{r}) \ .$$

Note that unless $\hat{\mathbf{t}}$ is the correct translation, values of the flow along the copoint vectors are influenced by the translational component of the flow field. Due to linearity, the estimated rotation is (ignoring measurement errors):

$$\hat{\boldsymbol{\omega}} = \boldsymbol{\omega} + \delta\boldsymbol{\omega} \tag{8}$$

where $\boldsymbol{\omega}$ is the true rotation and $\delta\boldsymbol{\omega}$ is the rotation of a field that best fits the translational component of the copoint projections.

The residual of the estimate is one possible source of information about the true FOE. If $\hat{\mathbf{t}}$ is the correct solution, then $\delta\boldsymbol{\omega} = \mathbf{0}$ and the residual is small regardless of the scene depth. If $\hat{\mathbf{t}}$ is an incorrect estimate, then the residual, in general, will be large; the residual could still be small, but only under special circumstances; for the residual to be zero, the depth Z has to satisfy

$$\mathbf{u}_{\mathrm{rot}}(\delta\boldsymbol{\omega}) \cdot \mathbf{v}_{\mathrm{cp}} = p_\omega(\delta\boldsymbol{\omega}, \hat{\mathbf{t}}, \mathbf{r}) = \frac{1}{Z} p_{\mathrm{t}}(\mathbf{t}, \hat{\mathbf{t}}, \mathbf{r}) \ . \tag{9}$$

Multiplying by Z^2 and substituting $Z\mathbf{r} = f\,\mathbf{R}$, we obtain a second order equation for the scene points $\mathbf{R}$

$$p_\omega(\delta\boldsymbol{\omega}, \hat{\mathbf{t}}, \mathbf{R}) - p_{\mathrm{t}}(\mathbf{t}, \hat{\mathbf{t}}, \mathbf{R}) = (\delta\boldsymbol{\omega} \times \mathbf{R}) \cdot (\hat{\mathbf{t}} \times \mathbf{R}) - f\,(\mathbf{t} \times \hat{\mathbf{t}}) \cdot \mathbf{R} = 0 \ . \tag{10}$$

Therefore, the residual is a measure of the difference between the surface in view and the quadric given by (10), but only at points where suitable normal flow measurements are available.

Assuming the scene cannot be approximated sufficiently well by a single second order surface, one might try to use the copoint residual to find the correct FOE. Such approaches can be found in [2]. In practice, however, this method may run into problems. Only a small and changing subset of measurements is used to compute the rotation, the measurements are noisy and the noise level may vary across the image. It is very difficult to weigh the residuals for different candidate translations so that they can be compared.

The main problem is that the residuals only provide negative information about the FOE. If the residual is large, we can be quite sure that the candidate translation is incorrect. However, a small residual may be caused by the lack of data and does not necessarily imply that the solution is correct. Thus, our method tests for the correct translation by going through the estimation of depth and analyzing how it is varying locally.

3.4 Algorithm Analysis

The function $\Phi(\hat{\mathbf{t}})$ depends not only on the true and the estimated motion parameters, but also on the actual distribution of normal flow directions as well as the unknown scene depth. Therefore only a statistical analysis of Φ is possible. What we study here is $\Theta(\hat{\mathbf{t}}, \mathcal{R})$, a contribution to $\Phi(\hat{\mathbf{t}})$ from a single image region $\mathcal{R}$.

We substitute both $\hat{\boldsymbol{\omega}}$ and the true motion parameters into (6) and simplify the expression

$$\begin{aligned}\frac{1}{\hat{Z}} &= \frac{(1/Z)\mathbf{u}_{\mathrm{tr}}(\mathbf{t})\cdot\mathbf{n} + \mathbf{u}_{\mathrm{rot}}(\boldsymbol{\omega})\cdot\mathbf{n} - \mathbf{u}_{\mathrm{rot}}(\boldsymbol{\omega}+\delta\boldsymbol{\omega})\cdot\mathbf{n}}{\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})\cdot\mathbf{n}} \\ &= \frac{(1/Z)\mathbf{u}_{\mathrm{tr}}(\mathbf{t})\cdot\mathbf{n} - \mathbf{u}_{\mathrm{rot}}(\delta\boldsymbol{\omega})\cdot\mathbf{n}}{\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})\cdot\mathbf{n}} \ .\end{aligned}$$

The vectors $\mathbf{u}_{\mathrm{tr}}(\mathbf{t})$, $\mathbf{u}_{\mathrm{rot}}(\delta\boldsymbol{\omega})$, and $\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})$ are polynomial functions of image position $\mathbf{r}$ and can usually be approximated by constants within a small image region.

Consider a local coordinate system where $\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})$ is parallel to $[1,0,0]^{\mathrm{T}}$. Because the computed depths are scaled in step 2 of the algorithm, we can equivalently set the length of $\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})$ to one and write $\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}}) = [1,0,0]^{\mathrm{T}}$. In the coordinate system, we denote

$$\begin{aligned}\mathbf{u}_{\mathrm{rot}}(\delta\boldsymbol{\omega}) &= u_{\mathrm{r}}\,[\cos\varphi_{\mathrm{r}}, \sin\varphi_{\mathrm{r}}, 0]^{\mathrm{T}} \\ \mathbf{u}_{\mathrm{tr}}(\mathbf{t}) &= u_{\mathrm{t}}\,[\cos\varphi_{\mathrm{t}}, \sin\varphi_{\mathrm{t}}, 0]^{\mathrm{T}} \\ \mathbf{n} &= [\cos\psi, \sin\psi, 0]^{\mathrm{T}} \ .\end{aligned} \tag{11}$$

Then $\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})\cdot\mathbf{n} = \cos\psi$ and

$$\begin{aligned}1/\hat{Z} &= (1/Z)u_{\mathrm{t}}\,(\cos\varphi_{\mathrm{t}} + \sin\varphi_{\mathrm{t}}\tan\psi) - u_{\mathrm{r}}\,(\cos\varphi_{\mathrm{r}} + \sin\varphi_{\mathrm{r}}\tan\psi) = \\ &= ((1/Z)u_{\mathrm{t}}\cos\varphi_{\mathrm{t}} - u_{\mathrm{r}}\cos\varphi_{\mathrm{r}}) + \tan\psi\,((1/Z)u_{\mathrm{t}}\sin\varphi_{\mathrm{t}} - u_{\mathrm{r}}\sin\varphi_{\mathrm{r}}) \ .\end{aligned} \tag{12}$$

As $u_{\mathrm{r}}\cos\varphi_{\mathrm{r}}$ is approximated by a constant, it does not influence the variance of $1/\hat{Z}$ and it is thus sufficient to study the variance of $1/\hat{Z}'$:

$$\frac{1}{\hat{Z}'} = (1/Z)u_{\mathrm{t}}\cos\varphi_{\mathrm{t}} + \tan\psi\,((1/Z)u_{\mathrm{t}}\sin\varphi_{\mathrm{t}} - u_{\mathrm{r}}\sin\varphi_{\mathrm{r}}) \ . \tag{13}$$

For the correct translation, we have $\delta\boldsymbol{\omega} = \mathbf{0}$, i.e., $u_{\mathrm{r}} = 0$, and $\varphi_{\mathrm{t}} = 0$, i.e., $\sin\varphi_{\mathrm{t}} = 0$. Therefore $1/\hat{Z}'$ simplifies into

$$\frac{1}{\hat{Z}'} = \frac{1}{Z}u_{\mathrm{t}} \tag{14}$$

and the variance of $1/\hat{Z}'$ is proportional to the variance of the inverse real depth $1/Z$ and is independent of $\tan\psi$.

For incorrect translation estimates, on the other hand, $\varphi_{\mathrm{t}} \neq 0$ except for points on a certain line. Unless the surface is special so that $\delta\boldsymbol{\omega} = \mathbf{0}$, also $u_{\mathrm{r}} \neq 0$

at most points in the image. In most cases $1/\hat{Z}'$ depends on $\tan\psi$ and it can become very large (or very small) when there are normal flow measurements in the region such that ψ is close to $\pm\,\pi/2$.

Obviously, large variance of $\tan\psi$ will lead to large $\Theta(\hat{\mathbf{t}}, \mathcal{R})$ only if expression $(1/Z)u_{\mathrm{t}}\sin\varphi_{\mathrm{t}} - u_{\mathrm{r}}\sin\varphi_{\mathrm{r}}$ is not small.

Consider now a typical image patch corresponding to a smooth scene patch, where the variance of $1/Z$ is small. For the correct motion estimate also the variance of $1/\hat{Z}'$ is small. For other translations, the variance of $1/\hat{Z}'$ is proportional to the variance of $\tan\psi$. It can be small if all the normal flow measurements in the patch have approximately the same direction, or if the scene patch is close to the ambiguous quadric (10).

If the depth of the scene patch varies significantly, the exact distribution of normal flow directions as well as their correlation with the scene depth will determine the 3D motion yielding the smallest depth variance. However, on average it can be expected that the depth variance for the correct 3D motion is smaller than the variances for most incorrect motions.

If the image region contain sufficiently many different directions of normal flow so that the value of $\tan\psi$ changes significantly, the variance of $1/\hat{Z}$ is dominated by

$$\tan\psi((1/Z)u_{\mathrm{t}}\sin\varphi_{\mathrm{t}} - u_{\mathrm{r}}\sin\varphi_{\mathrm{r}})\ . \tag{15}$$

Expression (15) is zero for the correct motion and non-zero for most incorrect motions. Only a scene patch close to the ambiguous quadric can give small depth variance for an incorrect 3D motion.

3.5 The Epipolar Constraint

The depth smoothness measure is closely related to the traditional epipolar constraint and we examine the relationship here. In the instantaneous form the epipolar constraint can be written as

$$(\hat{\mathbf{z}} \times \mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})) \cdot (\dot{\mathbf{r}} - \mathbf{u}_{\mathrm{rot}}(\hat{\boldsymbol{\omega}})) = 0 \tag{16}$$

or in the usual simplified form as

$$(\hat{\mathbf{t}} \times \mathbf{r}) \cdot (\dot{\mathbf{r}} + (\hat{\boldsymbol{\omega}} \times \mathbf{r})) = 0\ .$$

Usually, the distance of the flow vector $\dot{\mathbf{r}}$ from the epipolar line (determined by $\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})$ and $\mathbf{u}_{\mathrm{rot}}(\hat{\boldsymbol{\omega}})$) is computed, and the sum of the squared distances, i.e.,

$$\sum\left((\hat{\mathbf{z}} \times \mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})) \cdot (\dot{\mathbf{r}} - \mathbf{u}_{\mathrm{rot}}(\hat{\boldsymbol{\omega}}))\right)^2 \tag{17}$$

is minimized.

Methods based upon (17) suffer from bias, however, and a scaled epipolar constraint has been used to give an unbiased solution:

$$\sum \frac{\left(\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}}) \cdot (\dot{\mathbf{r}} - \mathbf{u}_{\mathrm{rot}}(\hat{\boldsymbol{\omega}}))\right)^2}{||\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})||^2}\ . \tag{18}$$

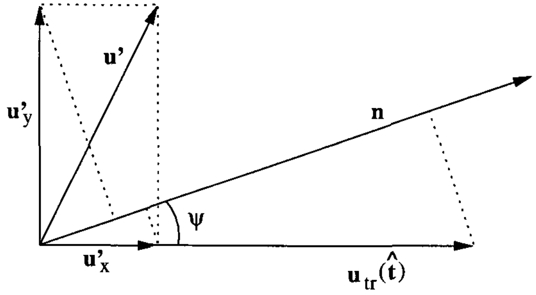

Fig. 3. The relationship between the epipolar constraint and the estimated depth

To relate the epipolar constraint and the depth smoothness measure, we rewrite (16) and substitute the true motion parameters to obtain

$$\mathbf{v}_{\mathrm{cp}} \cdot \left(\frac{1}{Z}\mathbf{u}_{\mathrm{tr}}(\mathbf{t}) - \mathbf{u}_{\mathrm{rot}}(\delta\boldsymbol{\omega})\right) = \frac{1}{Z}p_{\mathrm{t}}(\mathbf{t},\hat{\mathbf{t}},\mathbf{r}) - p_{\omega}(\delta\boldsymbol{\omega},\hat{\mathbf{t}},\mathbf{r}) = 0 \ .$$

Since $||\mathbf{v}_{\mathrm{cp}}|| = ||\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})||$, the properly scaled epipolar constraint can be written as

$$\frac{1}{||\mathbf{v}_{\mathrm{cp}}||}\left(\frac{1}{Z}p_{\mathrm{t}}(\mathbf{t},\hat{\mathbf{t}},\mathbf{r}) - p_{\omega}(\delta\boldsymbol{\omega},\hat{\mathbf{t}},\mathbf{r})\right) = 0 \ . \tag{19}$$

Consider now the local coordinate system used in Sect. 3.4. In the rotated coordinate system, $\mathbf{v}_{\mathrm{cp}} = [0,1,0]^{\mathrm{T}}$ and thus

$$(1/Z)p_{\mathrm{t}}(\mathbf{t},\hat{\mathbf{t}},\mathbf{r}) - p_{\omega}(\delta\boldsymbol{\omega},\hat{\mathbf{t}},\mathbf{r}) = (1/Z)u_{\mathrm{t}}\sin\varphi_{\mathrm{t}} - u_{\mathrm{r}}\sin\varphi_{\mathrm{r}} \ . \tag{20}$$

We see that one part of the estimated depth (12) is proportional to the epipolar distance (20).

The geometric relationship of the two quantities is illustrated in Fig. 3. We denote the derotated flow $\dot{\mathbf{r}} - \mathbf{u}_{\mathrm{rot}}(\hat{\boldsymbol{\omega}})$ by $\mathbf{u}'$ and decompose it into two components, $\mathbf{u}'_{\mathrm{x}}$ parallel to $\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})$, and $\mathbf{u}'_{\mathrm{y}}$ perpendicular to $\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})$.

In the figure, the epipolar distance is just $||\mathbf{u}'_{\mathrm{y}}||$.

The estimated depth (12) is the sum of

$$\frac{\mathbf{u}'_{\mathrm{x}} \cdot \mathbf{n}}{\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}}) \cdot \mathbf{n}} + \frac{\mathbf{u}'_{\mathrm{y}} \cdot \mathbf{n}}{\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}}) \cdot \mathbf{n}} \ .$$

Since $\mathbf{u}'_{\mathrm{x}}$ is parallel to $\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})$, the first part of the sum is $||\mathbf{u}'_{\mathrm{x}}||/||\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})||$, i.e., independent of $\mathbf{n}$. This is the first part of (12).

The angle between vectors $\mathbf{n}$ and $\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})$ is ψ, so for the second part of the sum, $\mathbf{u}'_{\mathrm{y}} \cdot \mathbf{n} = ||\mathbf{u}'_{\mathrm{y}}|| \cos(\pi/2 - \psi)$ and $\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}}) \cdot \mathbf{n} = ||\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})|| \cos(\psi) = \cos\psi$, since $\mathbf{u}_{\mathrm{tr}}(\hat{\mathbf{t}})$ is scaled to be a unit vector. Consequently, the second part of the sum is

$$||\mathbf{u}'_{\mathrm{y}}|| \tan\psi \ .$$

Consider now a constant depth patch, where $\mathbf{u}'$ is approximately constant. Then the variance of $1/\hat{Z}'$ is equal to the epipolar constraint multiplied by the variance of $\tan\psi$.

The above analysis makes clear the close relationship of the smoothness measure to the traditional epipolar constraint. Indeed, in a smooth patch, $(1/Z)u_t$ is approximately constant and the variance of $1/\hat{Z}$ can be written (using (20)) as

$$\Theta(\hat{\mathbf{t}},\mathcal{R}) = \mathrm{Var}(1/\hat{Z}) = \mathrm{Var}(\tan\psi)\left((1/Z)p_t(\mathbf{t},\hat{\mathbf{t}},\mathbf{r}) - p_\omega(\delta\boldsymbol{\omega},\hat{\mathbf{t}},\mathbf{r})\right)^2 . \qquad (21)$$

Therefore, in a smooth patch

$$\frac{\mathrm{Var}(1/\hat{Z})}{\mathrm{Var}(\tan\psi)}$$

is a good approximation of the epipolar constraint

$$\left((1/Z)p_t(\mathbf{t},\hat{\mathbf{t}},\mathbf{r}) - p_\omega(\delta\boldsymbol{\omega},\hat{\mathbf{t}},\mathbf{r})\right)^2 .$$

We thus define a scaled measure

$$\Theta'(\hat{\mathbf{t}},\mathcal{R}) = \frac{\Theta(\hat{\mathbf{t}},\mathcal{R})}{\mathrm{Var}(\tan\psi)}$$

and the corresponding global smoothness measure

$$\Phi'(\hat{\mathbf{t}}) = \sum_{\mathcal{R}} \Theta'(\hat{\mathbf{t}},\mathcal{R}) .$$

Measure $\Phi'(\hat{\mathbf{t}})$ is approximately equal to the epipolar measure if the scene patches are smooth (or successfully segmented). However, when measure $\Phi'(\hat{\mathbf{t}})$ is used as opposed to a measure evaluating the deviation from the epipolar constraint, no computation of optical flow is required. Furthermore, as discussed above, measure $\Phi'(\hat{\mathbf{t}})$ is advantageous in its ability to handle depth discontinuities as even for discontinuous patches that weren't successfully segmented, in many cases the depth smoothness measure gives the correct solution.

4 Experimental Results

Experiment 1. The depth segmentation described in Sect. 3.2 was tested for the Coke can sequence. Fig. 4 shows the image patches that were segmented for the correct FOE. Many depth discontinuities in the scene are such that computed normal flow measurements appear only on one side of the discontinuity. Consequently, such patches are not segmented and pose no problem for the motion estimation algorithm.

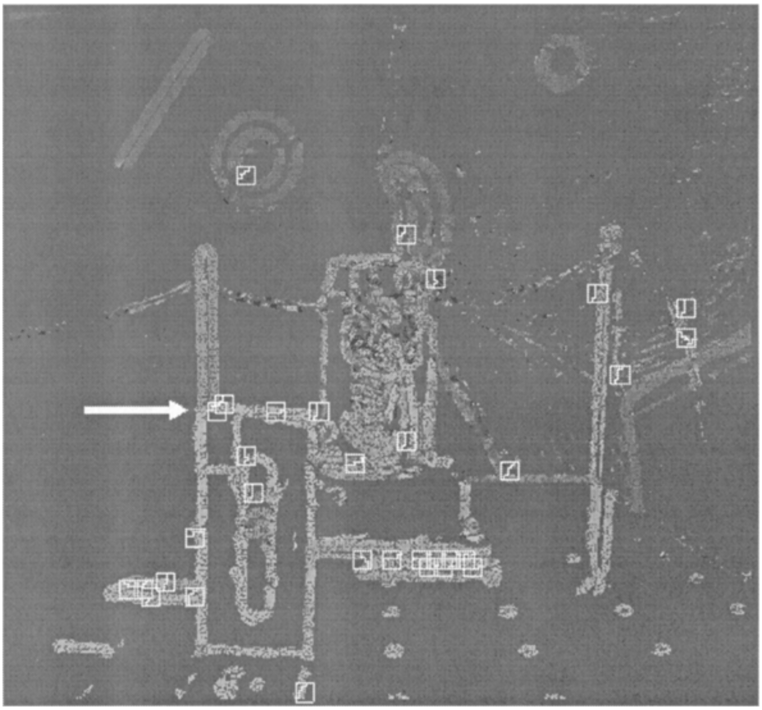

Fig. 4. Patch segmentation of the Coke can sequence overlayed on the computed inverse depth. The correct FOE (in the middle of the image) was used. The arrow points to the patch used in experiment 2

Experiment 2. The effects of the patch segmentation are best demonstrated by showing for varying $\hat{\mathbf{t}}$ function $\Theta'(\hat{\mathbf{t}}, \mathcal{R})$ for a fixed image region $\mathcal{R}$. The results are shown in Fig. 5. In particular, Fig. 5 (a) shows for all positions $\hat{\mathbf{t}}$ the smoothness measure $\Theta'(\hat{\mathbf{t}})$ for a fixed image region when segmentation was performed and Fig. 5 (b) shows the smoothness measure when no segmentation took place. Fig. 5 (c) shows all the FOE positions for which a segmentation was performed.

Experiment 3. The Yosemite fly-through sequence (one frame is shown in Fig. 6) contains independently moving clouds. Each image frame was thus clipped to contain only the mountain range.

We tested both depth smoothness measures $\Phi(\hat{\mathbf{t}})$ and $\Phi'(\hat{\mathbf{t}})$. For comparison, we estimated optical flow by assuming it locally to be constant and computed the epipolar constraint measure, using (18). The FOE voting results are shown in Fig. 7. Since the depth of the scene is changing slowly, the scaled depth smoothness measure $\Phi'(\hat{\mathbf{t}})$ gives results that are very close to the results using the epipolar constraint.

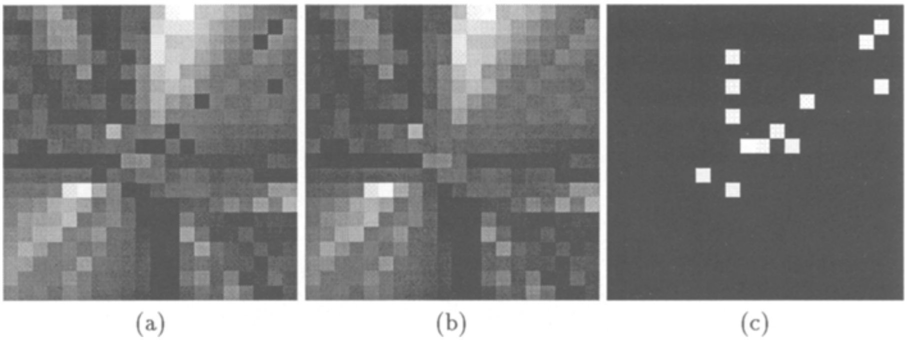

Fig. 5. Function $\Theta(\hat{\mathbf{t}}, \mathcal{R})$ for a fixed region $\mathcal{R}$ marked in Fig. 4. The area shown is three times the image size with the image in the center of the area. Black denotes the smallest and white largest values of $\Theta(\hat{\mathbf{t}}, \mathcal{R})$. (a) The patch was segmented if possible. (b) No patch segmentation was performed. (c) White blocks denote candidate FOE positions for which the patch was segmented

Fig. 6. One frame from the Yosemite fly-through sequence. Only the lower part of the image was used in the algorithm

In the coordinate system centered in the middle of the image, the correct focus of expansion was at $(-20, -100)$ in pixels.

The following FOE positions were found using hierarchical search:

- Using $\Phi(\hat{\mathbf{t}})$: $(21, -87)$
- Using $\Phi'(\hat{\mathbf{t}})$: $(-24, -104)$
- Using the epipolar constraint : $(-21, -116)$

5 Conclusions

There exists a lot of structure in the world. With regard to shape, this structure manifests itself in surface patches that are smooth, separated by abrupt discontinuities. This paper exploited this fact in order to provide an algorithm that estimates 3D motion while at the same time it recovers scene discontinuities.

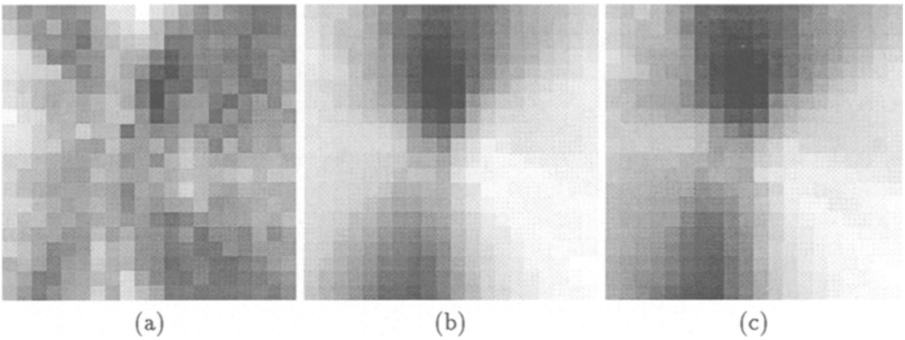

Fig. 7. The results of the FOE voting. The voting area is three times the image size with the image in the center of the area. The gray value of the bin represents the value of the depth smoothness measure with black representing the smallest value. (a) The smoothness measure $\Phi(\hat{\mathbf{t}})$. (b) The scaled smoothness measure $\Phi'(\hat{\mathbf{t}})$. (c) The epipolar constraint measure

The basis of the technique lies in the understanding of the interaction between 3D shape and motion. Wrong 3D motion estimates give rise to depth values that are locally unsmooth, i.e., they vary more than the correct ones. This was exploited in order to obtain the 3D motion that locally provides the "smoothest" depth while recovering scene discontinuities. Finally, it was shown that the technique is very much related to epipolar minimization since the function to be minimized in image areas corresponding to smooth scene patches takes the same values as deviation from the epipolar constraint. The function used here, however, is obtained from normal flow measurements. The fact that values of the flow are not needed in the approach, along with the aforementioned equivalence of our objective function to the epipolar constraint, demonstrates that the presented algorithm is an improvement over existing techniques.

References

1. T. Brodsky, C. Fermüller, and Y. Aloimonos. Directions of motion fields are hardly ever ambiguous. *International Journal of Computer Vision*, 26:5–24, 1998.
2. S. Carlsson. Sufficient image structure from motion and shape estimation. In *Proc. European Conference on Computer Vision*, pages 83–94, Stockholm, Sweden, 1994.
3. L. Cheong, C. Fermüller, and Y. Aloimonos. Effects of errors in the viewing geometry on shape estimation. *Computer Vision and Image Understanding*, 1998. In press. Earlier version available as Technical Report CAR-TR-773, June 1996.
4. O. D. Faugeras. *Three-Dimensional Computer Vision.* MIT Press, Cambridge, MA, 1992.
5. C. Fermüller. Navigational preliminaries. In Y. Aloimonos, editor, *Active Perception*, Advances in Computer Vision, pages 103–150. Lawrence Erlbaum Associates, Hillsdale, NJ, 1993.

6. C. Fermüller and Y. Aloimonos. Direct perception of three-dimensional motion from patterns of visual motion. *Science*, 270:1973–1976, 1995.
7. S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution and Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.
8. B. K. P. Horn. Motion fields are hardly ever ambiguous. *International Journal of Computer Vision*, 1:259–274, 1987.
9. B. K. P. Horn and E. J. Weldon, Jr. Direct methods for recovering motion. *International Journal of Computer Vision*, 2:51–76, 1988.
10. J. J. Koenderink and A. J. van Doorn. Affine structure from motion. *Journal of the Optical Society of America*, 8:377–385, 1991.
11. Q.-T. Luong and O. D. Faugeras. The fundamental matrix: Theory, algorithms, and stability analysis. *International Journal of Computer Vision*, 17:43–75, 1996.
12. J. Marroquin. *Probabilistic Solution of Inverse Problems.* PhD thesis, Massachusetts Institute of Technology, 1985.
13. S. J. Maybank. Algorithm for analysing optical flow based on the least-squares method. *Image and Vision Computing*, 4:38–42, 1986.
14. S. J. Maybank. *A Theoretical Study of Optical Flow.* PhD thesis, University of London, England, 1987.
15. D. Mumford and J. Shah. Boundary detection by minimizing functionals. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 22–25, 1985.