# Direct Estimation of Motion and Extended Scene Structure from a Moving Stereo Rig


Gideon P. Stein

Artificial Intelligence Laboratory
MIT
Cambridge, MA 02139
gideon@ai.mit.edu

Amnon Shashua

Institute of Computer Science
Hebrew University of Jerusalem
Jerusalem 91904, Israel
http://www.cs.huji.ac.il/∼ shashua/



### Abstract

*We investigate the relationship between the kinematics (infinitesimal motion model) of a calibrated Stereo Rig and point and line image feature measurements seen at two time instances of the rig's motion (four images in all). In particular, we are interested in the byproduct of this analysis providing a direct connection between the spatio-temporal derivatives of the images at two time instances and kinematics of the 3D motion of the Rig.*

*We establish a fundamental result showing that 3 quadruples of point-line-line-line matches (i.e., point in the reference image and lines coincident with the corresponding points in the remaining three images) are sufficient for a unique linear solution for the kinematics of the rig. In other words, the projected instantaneous motion of "one and a half" 3D lines is sufficient for recovering the kinematics of the moving rig. In particular, spatio-temporal derivatives across 3 points are sufficient for a direct estimation of the rig's motion.*

*Consequently, we describe a new direct estimation method for motion estimation and 3D reconstruction from stereo image sequences obtained by a stereo rig moving through a rigid world. Correspondences (optic flow) are not required as spatio-temporal derivative are used instead. One can then use the images from both pairs combined, to compute a dense depth map. Finally, since the basic equations are linear, we combine the contribution coming from all pixels in the image using a Least Squares approach.*

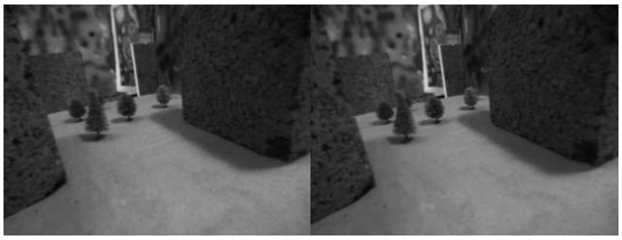

## 1 Introduction

We investigate the continuous model of a moving Stereo rig — also known as the "Stereo-Motion" problem. In a stereo-motion problem one is given a fixed (known) stereo configuration moving rigidly in the 3D world. The goal is to recover the motion (kinematics) of the rig and, by combining the stream of image pairs, to obtain a 3D reconstruction of the scene depicted by the image stream. Details on various approaches for stereo-motion cooperation can be found in [7, 4, 15, 8].

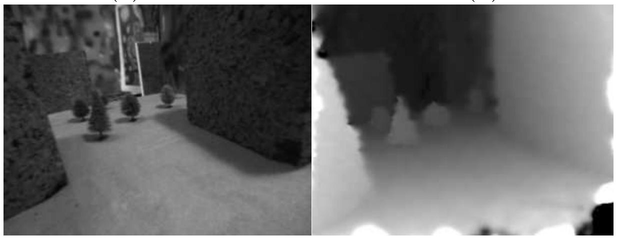
We first investigate the theoretical lower bound on the amount of image measurements required for recovering the kinematics of the rig. This problem was studied in the past by [9] who showed that the projections of a pair of 3D lines are sufficient for a Linear solution for the rig's kinematics, i.e., two matching lines across two image-pairs (four images in all) are sufficient for recovering the parameters of the continuous 3D motion of the rig between the two time instances.

We show that in fact only a "line and a half" are required for the estimation of kinematics, which means that three matching quadruples of a point and three lines, termed point-line-line-line, are sufficient. A quadruple of matching lines, for example, is equivalent to a pair of point-line-line-line matches — making three such matches be equivalent to the information content coming from a 3D line-and-half.

The point-line arrangement is especially useful when considering it as consisting of a *partial* point matching measurement. The partial measurement can be completed by introducing the spatio-temporal derivatives of image brightness, thus providing a "direct" estimation of motion and structure without the need to recover point and line matches in the process. In the case of three views, the combination of camera motion and spatio-temporal derivatives gives rise to a model-based brightness constraint known as the "tensor brightness constraint" [10, 14, 13]. We extend the model-based brightness equation for the Stereo-Motion problem and introduce a "direct" estimation procedure for recovering the kinematics of the rig and the 3D structure of the scene.



(a) (b)



(c) (d)

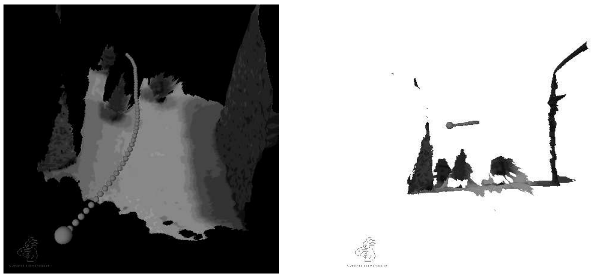
(e)

Figure 1: *The three input images (a,b,c). Image (a) and (c) are the first stereo pair in the sequence with the cameras displaced vertically. Image (b) is the image from camera (a) taken at the next time instance. (d) the estimated depth map.. (e) shows the results of a Canny edge detector on the depth map. This identifies occluding contours.*

The advantage of the new stereo-motion-brightness constraints is three fold. First, every pixel contributes linear constraints for the kinematics of the rig — thus producing a densely over-constrained system of equations for the kinematics of the rig. Second, we obtain a dense 3D reconstruction *per pixel* including those pixels whose brightness distribution cannot provide unique correspondence across images (the aperture problem). Third, the method is stable for collinear motion, for planar configurations and for singular motions like pure rotation.

## 2 Geometric Linear Constraints for the Stereo-Motion Problem

Consider a Stereo Rig producing two pairs of images at two consecutive time instances. We refer to the first pair of images as images 1,2 and the second pair as images 3,4 (see fig. 3). Let $P = (X, Y, Z, 1)$ be the



(a) (b)

(c)

Figure 2: *(a,b,c) Show 3D renderings of the scene. Very distant surfaces have been removed. The spheres show the motion of the stereo rig in the whole sequence. The large sphere is the the location of the camera when this image was taken. Note in view (b) that the ground plane is correctly found to be flat at the the two cork walls are at right angles to the it.*

homogeneous coordinates of a point in the Euclidean 3D space aligned with the coordinate system of image 1. Thus $x = \frac{X}{Z}, y = \frac{Y}{Z}$ are the non-homogeneous coordinates of the first image plane (that of image 1). The relationship between the coordinate frames of the two cameras of the rig is represented by:

$$P' = \mathbf{R}'P + t'$$

where $P'$ is the representation of object space in the coordinate system of the second camera and $\mathbf{R}', t'$ are the rotational and the translational components of the change of coordinate systems. Thus, images 1,2 are related by the motion $1 \rightarrow 2$

$$p' \cong \mathbf{R}'p + kt' \tag{1}$$

where $p = (x, y, 1), p' = (x', y', 1)$, $k = \frac{1}{Z}$ and $\cong$ denotes equality up to scale. The stereo rig makes an *infinitesimal* motion represented by the kinematic screw $(w'', t'')$ model:

$$p'' \cong p + [w'']_x p + kt'' \tag{2}$$

where $p''$ are the homogeneous coordinates of image 3, $w''$ is the instantaneous rotation of the motion $1 \rightarrow 3$

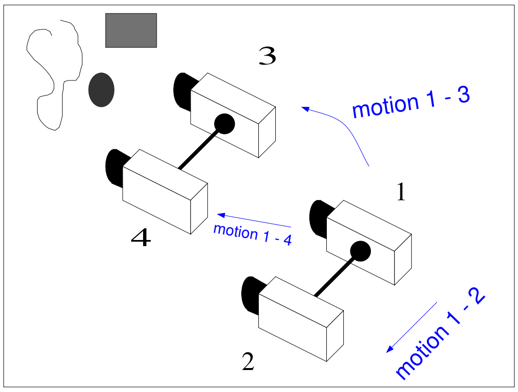



Figure 3: *Schematic diagram of a moving stereo rig.*

and $[w'']_x$ is the skew-symmetric matrix of cross products, i.e., $[w'']_x p = w'' \times p$ and $t''$ is the instantaneous translation of the motion $1 \rightarrow 3$. Likewise, the instantaneous motion $1 \rightarrow 4$ is represented by

$$p''' \cong p + [w''']_x p + kt''' \quad (3)$$

where $p'''$ are the homogeneous coordinates of image 4 and $(w''', t''')$ is the kinematic screw of the motion $1 \rightarrow 4$.

In the Stereo-Motion problem, the parameters $\mathbf{R}', t'$ of the motion $1 \rightarrow 2$ are known (and remain fixed throughout subsequent movement of the rig) and the kinematics of the motion $1 \rightarrow 3$ (and $1 \rightarrow 4$) are unknown. We prove the following claim:

**Theorem 1** *The kinematics $(w'', t'')$ of the motion $1 \rightarrow 3$ can be recovered from three quadruples of a point $p$ in image 1 and lines $s', s'', s'''$ coincident with the point matches $p', p'', p'''$, i.e., $s'^\top p' = s''^\top p'' = s'''^\top p''' = 0$.*

**Proof:** We will show that each matching point-line-line-line gives rise to two linear constraints for the unknowns $w'', t''$. Premultiply equations 1 and 2 by $s'$ and $s''$ to obtain:

$$\begin{aligned} ks'^\top t' + s'^\top \mathbf{R}' p &= 0 \quad (4) \\ ks''^\top t'' + v''^\top w'' + s''^\top p &= 0 \quad (5) \end{aligned}$$

where we have used the identity:

$$s''^\top [w'']_x p = w''^\top [p]_x s''$$

and defined: $v'' = [p]_x s'' = p \times s''$. Eliminate $k$ from equations 4 and 5 and obtain a linear constraint for $w'', t''$:

$$(s'^\top \mathbf{R}' p)(s''^\top t'') - (s'^\top t')(v''^\top w'') = (s'^\top t')(s''^\top p). \quad (6)$$

Likewise, premultiply equation 3 by $s'''$ to obtain:

$$ks'''^\top t''' + v'''^\top w''' + s'''^\top p = 0 \quad (7)$$

where $v''' = p \times s'''$. Eliminate $k$ from equations 4 and 7 and obtain a linear constraint on $w''', t'''$:

$$(s'^\top \mathbf{R}' p)(s'''^\top t''') - (s'^\top t')(v'''^\top w''') = (s'^\top t')(s'''^\top p). \quad (8)$$

Since the motion $1 \rightarrow 2$ and $3 \rightarrow 4$ are identical we can represent the motion $1 \rightarrow 4$ by concatenating the motions $1 \rightarrow 3$ and $3 \rightarrow 4$:

$$\begin{aligned} \mathbf{I} + [w''']_x &= \mathbf{R}'(\mathbf{I} + [w'']_x) \quad (9) \\ t''' &= \mathbf{R}' t'' + t' \quad (10) \end{aligned}$$

Following substitution and rearrangements, equation 8 reduces to the second linear constraint on $w'', t''$:

$$\begin{aligned} (s'^\top \mathbf{R}' p)(s'''^\top \mathbf{R}' t'') - (s'^\top t')(u''^\top w'') = \\ (s'^\top t')(s'''^\top \mathbf{R}' p) - (s'^\top \mathbf{R}' p)(s'''^\top t') \quad (11) \end{aligned}$$

where $u'' = p \times \mathbf{R}'^\top s'''$. Thus, equations 6 and 11 provide two (non-homogeneous) linear constraints for the 6 unknowns of $w'', t''$. $\square$

**To Summarize:** Each point $p$ in image 1 and lines $s'$, $s''$ and $s'''$ passing through the corresponding points in images 2,3 and 4 provide two linear equations in the unknown motion $w''$ and $t''$ (equations 6 and 11).

Given 3 point-line-line-line correspondences it is possible to recover the motion $w''$ and $t''$. Since 3 points are sufficient it is also obvious that the motion can be recovered from a planar scene. However, there still remains the problem of finding those correspondences. This is addressed in the next section in which we introduce the photometric constraint.

## 2.1 Photometric Constraints

A first order approximation of the constant brightness constraint leads to the optical flow constraint equation [5]:

$$u' I_x + v' I_y + I'_t = 0 \quad (12)$$

where $(u', v')$ are the *optical flow* values at $(x, y)$ between image 1 and image 2 (i.e. $u' = x' - x$ and $v' = y' - y$). $(I_x, I_y, I'_t)$ are the spatial and temporal derivatives at the coordinates $(x, y)$. In practice $I'_t = I_2(x, y) - I_1(x, y)$.

The optical flow constraint equation 12 can be rewritten in the form:

$$(I_x, I_y, I'_t)^\top (u, v, 1) = 0.$$

The line

$$s = (I_x, I_y, -xI_x - yI_y)^\top$$

in the projective plane passes through the point $p = (x, y, 1)^\top$ since:

$$(I_x, I_y, -xI_x - yI_y)^\top (x, y, 1) = 0.$$

Combining those two equations together:

$$\begin{pmatrix} I_x \\ I_y \\ -xI_x - yI_y \end{pmatrix}^\top \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} + \begin{pmatrix} I_x \\ I_y \\ I'_t \end{pmatrix}^\top \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = 0 \tag{13}$$

Rearranging the the terms results in:

$$\begin{pmatrix} I_x \\ I_y \\ I'_t - xI_x - yI_y \end{pmatrix}^\top \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = 0 \tag{14}$$

where we have used $x' = x + u$ and $y' = y + v$. Therefore, the line:

$$s' = (I_x, I_y, I'_t - xI_x - yI_y)^\top \tag{15}$$

passes through the point $p' = (x', y', 1)^\top$.

Thus the photometric constraints provide a matching constraint between a point $p$ and a line $s'$ passing through the corresponding point $p'$ in image 2. *The coordinates of the point $p'$ in image 2 are not required.*

**To summarize:** the lines:

$$s' = (I_x, I_y, I'_t - xI_x - yI_y)^\top \tag{16}$$

$$s'' = (I_x, I_y, I''_t - xI_x - yI_y)^\top \tag{17}$$

and

$$s''' = (I_x, I_y, I'''_t - xI_x - yI_y)^\top \tag{18}$$

are lines coincident with $p'$,$p''$ and $p'''$ respectively and parallel to the iso-brightness contour at $(x, y)$. ($I''_t$ is the temporal derivative between the image 3 and image 1. i.e. $I''_t = I_3(x, y) - I_1(x, y)$ and likewise $I'''_t = I_4(x, y) - I_1(x, y)$).

### 2.2 Model-Based Brightness Constraint for a Moving Stereo Rig

We now combine the geometric constraints (equations 6 and 11) together with the photometric constraints (equations 16, 17 and 18). This results in two equations for each point in the image. In practice it is convenient to rectify the images in the stereo pair as a preprocessing stage so that $\mathbf{R}' = \mathbf{I}$.

Since $s'^\top p = I'_t$ and $s'''^\top p = I'''_t$ the brightness constraint equations are then:

$$I'_t s''^\top t'' - s'^\top t' v''^\top w'' = s'^\top t' I''_t \tag{19}$$

and

$$I'_t s'''^\top t'' - s'^\top t' v'''^\top w'' = s'^\top t' I'''_t - I'_t s'''^\top t' \tag{20}$$

**Notes:**

- There are two such linear constraints for each point in the image.
- The constraints involve image coordinates in image 1 only and the image gradients. It does not require correspondences $(x', y', x'', y'', x''', y''')$.
- Since the constraints are based on the 'optical flow constraint equation' they are correct only for infinitesimal motion. Consequently, motion $1 \rightarrow 2$ is also assumed to be small, thus a coarse-to-fine framework is required for the implementation stage.

## 3 Implementation Details

### 3.1 Calibration

We assume the internal parameters of the first camera are known. These can be found using methods described in [11]. For true Euclidean reconstruction an accurate estimate of the focal length is required but the whole process degrades gracefully when only approximate values are provided.

Calibration of the stereo pair is performed in two stages. First we take an image of a distant scene (the plane at infinity) and find the homography between image 2 and image 1 using the method described in [6]. Since the rotation angle is small we can assume an affine model rather than a full planar projective transformation. This stage takes into account both the rotation between the two cameras and also the variation in internal camera parameters. We can now use this mapping (projective or affine) to preprocess all the images coming from camera 2.

The second stage is to find the translation between the two cameras. We move the whole stereo rig in pure translation. Under the assumption of pure translation we can rewrite equation 19 as:

$$I''_t s'^\top t' - I'_t s''^\top t'' = 0. \tag{21}$$

This equation, which appeared in [14], can be used to accurately compute both the translation of the rig $t''$ and the displacement between the two cameras $t'$ (up to an unknown scale factor).

#### 3.1.1 Lens Distortion

Since we us are using a wide field-of-view (FOV) lens there is noticeable lens distortion. This does not affect the stability of the method but the accuracy of the motion estimates is reduced and the 3D reconstruction suffers non-projective distortion. Flat surfaces and straight line appear slightly curved. A variety of methods to compute lens distortion appear in the literature (see [12]).

### 3.2 Computing Depth

To compute the depth at every point we use equations 4,(5 and 7. Information is combined from both image pairs and over a local region by minimizing:

$$\min_k \arg \sum_{x,y \in R} \sum_j \beta(x,y) |s^{j\top} t^j|^p \left( k s^{j\top} t^j + v^{j\top} w^j + s^{j\top} p \right)^2 \tag{22}$$

where $j$ denotes $'$, $''$ or $'''$.

- The windowing function $\beta(x,y)$ allows one to increase the weight of the closer points.
- The $|s^{j\top} t^j|^p$ term reduces the weight of points which have a small gradient or where the gradient is perpendicular to that camera motion since these cases are highly affected by noise. We used $p = 1$.
- During the iteration process we used a region $R$ of $7 \times 7$.
- After the last iteration, we reduced the region $R$ to $1 \times 1$ but added a very weak global smoothness term and performed multi-grid membrane interpolation. This stabilizes regions where there is no image gradient.

### 3.3 Coarse to fine processing and iterative refinement

In order to deal with image motions larger than 1 pixel we use a Gaussian pyramid for coarse to fine processing [1, 2]. For a $640 \times 480$ image we used a 5 level pyramid.

The linear solution can be thought of as a single iteration of Newton's method applied to the problem. At each level of the pyramid we iterate as follows:

1. Calculate motion (using equations 19 and 20).
2. Compute depth (using equation 22).
3. Using the depth and motion, warp images 2,3 and 4 towards image 1.
4. Compute new time derivatives $I'_t$,$I''_t$ and $I'''_t$.
5. Compute a new motion and depth estimate.

One cannot simply compute the incremental model from the previous iteration because as the iterations proceed the system of equations of the incremental model will become badly conditioned. We followed the procedure in [14]. At the finest level ($640 \times 480$) we performed 2 iterations and we recursively doubled the number of iterations at the coarser levels. We can afford to do this because the number of computations per iteration at each levels drops by a factor of 4.

After we have computed the structure and motion at the finest level we keep the motion constant and repeat the depth computation down the whole pyramid. This fixes a few 'holes' particularly near the borders of the image.

## 4 Experiments

### 4.1 Experimental details

A single camera was mounted on a 3 degree of freedom motion stage (horizontal and vertical translation and rotation around the vertical axis). A stereo pair of images was captured by translating the camera vertically by $8.4mm$ between images. This in effect means that the first stage of calibration (sec. 3) is not required and that none of the measurement error can be attributed to different internal geometric or photometric parameters between the cameras. Initial experiments (not reported here) have been performed using two cameras mounted vertically one above the other. These show qualitatively similar results to those presented but quantitative results are not available.

The camera was a high resolution BW CCD camera with an $4.8mm$ lens giving a corner to corner viewing angle of $100^o$. The images were captured at $640 \times 480$ resolution.

### 4.2 Extended Scene Reconstruction

For the 'fly through' sequence (fig. 1) the camera was mounted on a extension bar which positioned the camera's Center of Projection (COP) $400mm$ from the axis of rotation (see fig. 4). Thus a rotation of $1.5^o$ produces a translation of $10.5mm$. The camera was mounted facing $20^o$ down from the horizontal and $10^o$ inwards. The focus of expansion (FOE) is therefore inside the image towards the top right.

The extension bar position perpendicular to the translation stage will be denoted by an angle of $\theta = 0^o$. The bar was moved in $1.5^o$ decrements from a starting position of $\theta = 45^o$ to a position of $\theta = 0^o$. The camera was then translated in $10mm$ increments through a distance of $120mm$ and then rotated again $1.5^o$ decrements to $\theta = -12^o$. At each camera location vertical stage motion provided a stereo pair.

The camera motion and depth map were computed for each motion in the sequence. The motion estimates

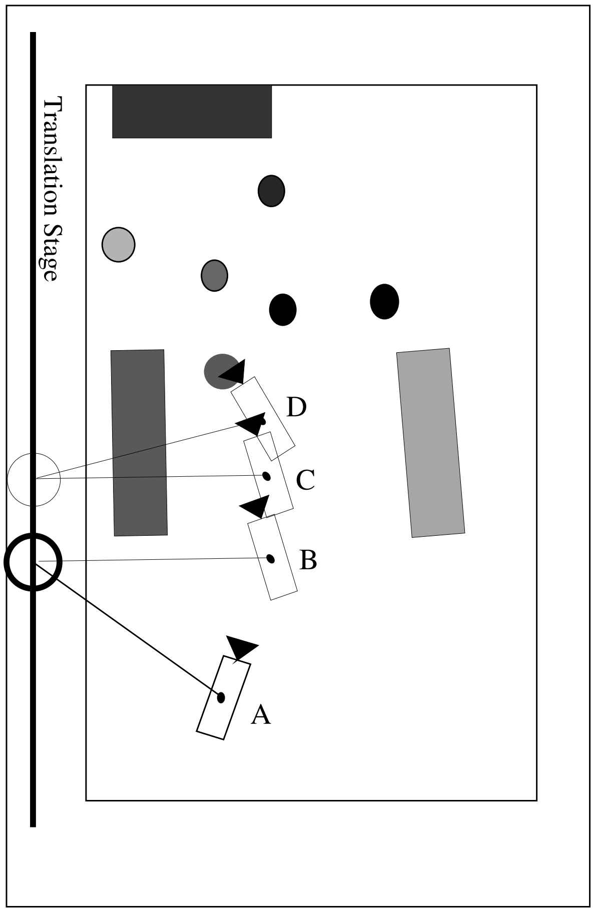



Figure 4: *Schematic overhead view of the 'fly-through' scene indicating the three cork blocks, 6 model tree (circles) and the motion stage (not to scale). The camera arm is rotated* $45^o$ *in* $1.5^o$ *steps from position A to position B. The arm then translates in twelve* 1cm *steps to position C. Then it Rotates through* $12^o$ *to position D.*

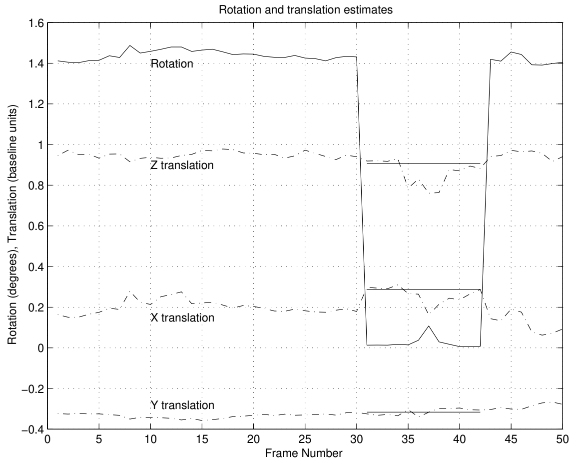
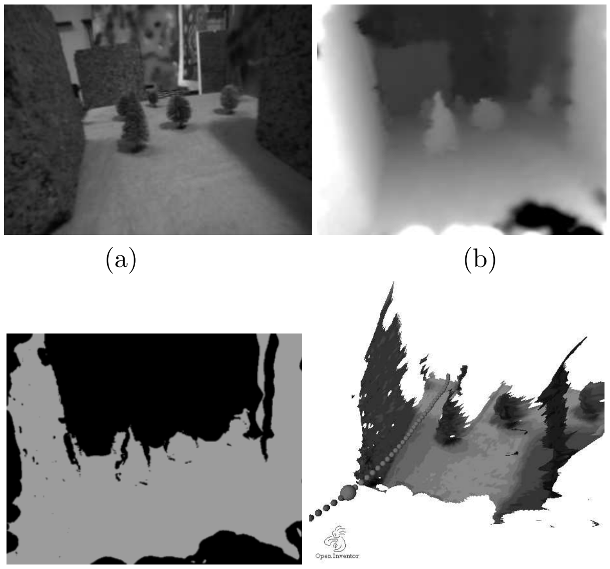



Figure 5: *Estimated Rotation and translation. The rotation angle (solid lines) should be* $1.5^o$ *and is on average* $1.45^o$*. The translation estimates (dot-dashed lines) are plotted along with the true values for the pure translation section (short solid lines).*

are plotted in fig. 5. Figs. 1,6,7 and 8 show four example reconstructions made at four points along the path. Areas which were very close or very far ($60 \times baseline$) from the camera are masked out and not rendered and so are points where the (inverse) depth map has a large gradient since these are points where depth reconstruction is inaccurate. Fig. 1e shows the results of a Canny edge detector applied to the (inverse) depth image highlighting potential occluding contours.

Fig. 9 shows the four reconstructions aligned using the estimated camera motions. The scene has more information than can be seen from any one camera.



(a) (b)

(c) (d)

Figure 6: *(a) The 14th image in the sequence. (b) The depth map. (c) The mask. We do not render surfaces that are far from the camera or where the surface is facing the camera with a sharp angle. (d) rendering of the surfaces.*

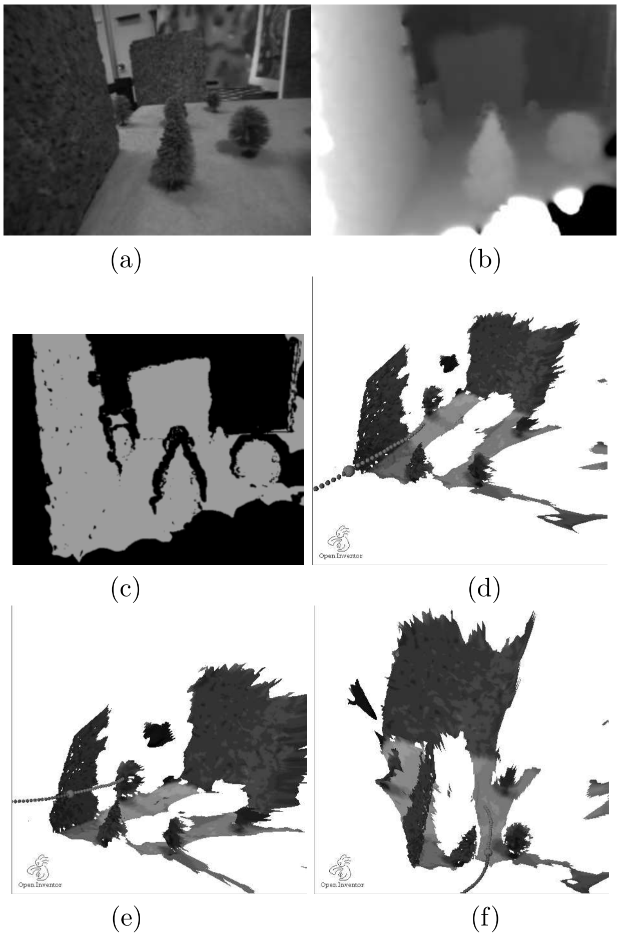



Figure 7: *(a) The 28th image in the sequence. (b) The depth map. (c) The mask. We do not render surfaces that are far from the camera or where the surface is facing the camera with a sharp angle. (d), (e),(f) rendering of the surfaces.*

## 5 Discussion and Future Work

We have presented a new method for recovering motion and structure from stereo image sequences. The method gives good motion estimates with errors less than 5% but this is not a zero mean error and there is a clear bias even when the focus of expansion (FOE) is inside the field of view (FOV). In the experiments presented here only images 1,2 and 3 were used for the motion and structure estimation and thus we did not use all the available information. Using the additional information from image triplet 1,2 and 4 will possibly result in even better motion estimates. This will be tested in the near future.

While we have shown that this method can give good motion estimates these are only incremental motion estimates. A full system could incorporate this

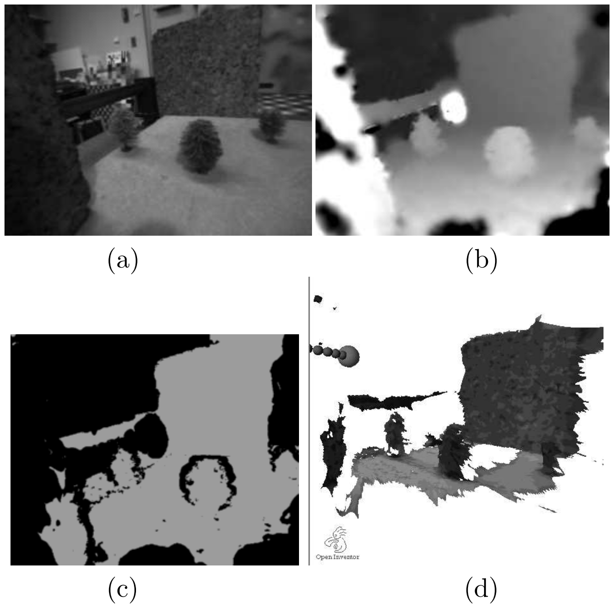



Figure 8: *(a) The 49th image in the sequence. In this view a bit of the translation stage can be seen (top left quadrant). (b) The depth map. (c) The mask. (d) 3D renderings of the surfaces.*

method as a better estimation stage inside a Kalman filter or a good initial guess in a batch global estimation framework. We have also not dealt with the issue of how to represent the final 3D reconstruction. At this point we draw multiple $2\frac{1}{2}D$ surfaces. Better methods are described in [3]. We have focused on direct methods but multiview geometric constraints can also be applied to merge 3D reconstructions from a stereo sequence using feature correspondences.

## References

[1] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *Proceedings of the European Conference on Computer Vision*, Santa Margherita Ligure, Italy, June 1992.

[2] P. J. Burt and E. H. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31:532–540, 1983.

[3] P. Fua. Reconstructing complex surfaces from multiple stereo views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1078–1085. IEEE Computer Society Press, June 1995.

[4] K. J. Hanna and N. E. Okamoto. Combining stereo and motion analysis for direct estimation of scene structure. In *Proceedings of the International Conference on Computer Vision*, Berlin, Germany, May 1993.

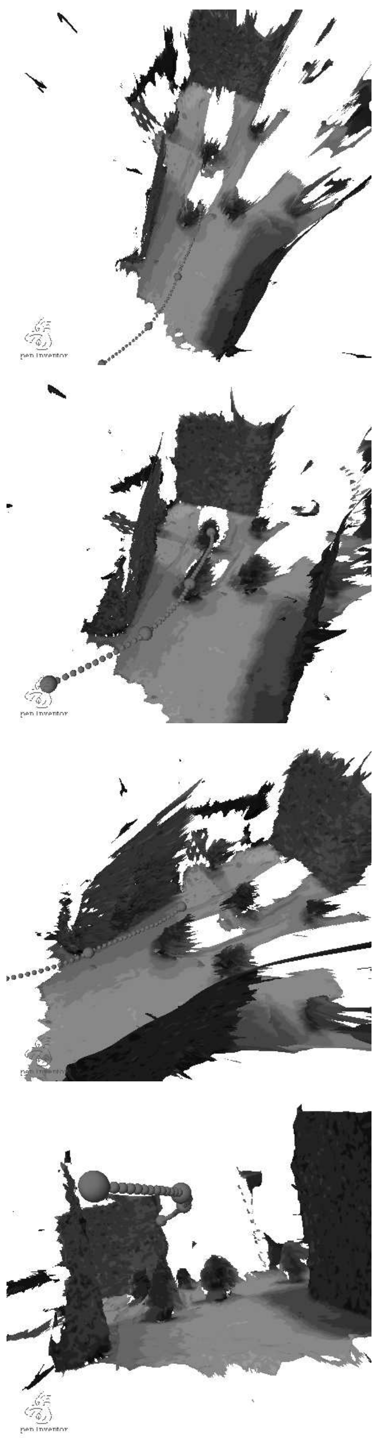


Figure 9: *Euclidean reconstruction of the extended scene created by aligning four separate reconstructions into one coordinate system using the recovered camera motion. The small spheres show the camera path. The large spheres show the camera positions of the four Euclidean reconstructions.*


[5] B. K. P. Horn and B. G. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[6] M. Irani, B. Rousso, and S. Peleg. Recovery of ego-motion using image stabilization'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 454–460, Seattle, Washington, June 1994.

[7] Y. C. Kim and J. K. Aggarwal. Determining object motion in a sequence of stereo images. *IEEE Journal of Robotics and Automation*, 3(6):599–614, December 1987.

[8] R. Koch. 3-d surface reconstruction from stereoscopic image sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–114. IEEE Computer Society Press, June 1995.

[9] N. Navab, R. Deriche, and O. D. Faugeras. Recovering 3d motion and structure from stereo and 2d token tracking cooperation. In *Proceedings of the International Conference on Computer Vision*, pages 513–516, Osaka, Japan, December 1990.

[10] A. Shashua and K. J. Hanna. The tensor brightness constraints: Direct estimation of motion revisited. Technical report, Technion, Haifa, Israel, November 1995.

[11] G. P. Stein. Accurate internal camera calibration using rotation, with analysis of sources of error. In *Proceedings of the International Conference on Computer Vision*, pages 230–236. IEEE Computer Society Press, June 1995.

[12] G. P. Stein. Lens distortion calibration using point correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Puerto Rico, June 1997.

[13] G. P. Stein. *Geometric and Photometric Constraints: Motion and Structure from Three Views.* PhD thesis, M.I.T Artificial Intelligence Laboratory, February 1998.

[14] G. P. Stein and A. Shashua. Model based brightness constraints: On direct estimation of structure and motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Puerto Rico, June 1997.

[15] Z. Zhang and O. D. Faugeras. Three dimensional motion computation and segementation in a long sequnce of stereo frames. *International Journal of Computer Vision*, 7(3):211–241, 1992.