MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo No. 1621 December, 1998

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/

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu. The pathname for this publication is: ai-publications/1500-1999/AIM-1621.ps.Z

Abstract

We describe a new method for motion estimation and 3D reconstruction from stereo image sequences obtained by a stereo rig moving through a rigid world. We show that given two stereo pairs, one can compute the motion of the stereo rig directly from the image derivatives (spatial and temporal). Correspondences are not required. One can then use the images from both pairs combined, to compute a dense depth map. The motion estimates between stereo pairs enable us to combine depth maps from all the pairs in the sequence to form an extended scene reconstruction. We show results from a real image sequence.

The motion computation is a linear least squares computation using all the pixels in the image. Areas with little or no contrast are implicitly weighted less, so one does not have to explicitly apply a confidence measure.

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for this research was provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-94-01-0994. G.S. would like acknowledge the financial support from ONR contracts N00014-94-1-0128 and DARPA contracts N00014-94-01-0994, N00014-97-0363 A.S. wishes acknowledge the financial support from US-IS Binational Science Foundation 94-00120/2, the European ACTS project AC074 "Vanguard", and from DARPA through ARL Contract DAAL01-97-0101.

1 Introduction

Stereo and motion together form a powerful combination [6, 20] with a variety of applications from ego motion estimation to extended scene reconstruction. We describe a new method for directly estimating the motion of the stereo rig thus greatly simplifying the process of combining the information from multiple image pairs. We show that given two stereo pairs one can compute the motion of the stereo rig directly from the image derivatives (spatial and temporal). Correspondences are not required.

The core of the method is an application of the 'Tensor Brightness Constraint' [15, 17], which combines geometric and photometric constraints of three views, to the case where the motion between one pair of views is known. The 'Tensor Brightness Constraint' equations are bilinear in the motion parameters but if one of the motions is known they result in a set of linear equations for the second motion parameters. There is one equation for each point in the image and 6 unknowns (the translation and rotation parameters) which results in a highly over-constrained and very stable computation.

The use of a stereo pair rather than a single image sequence (monocular structure from motion) has further advantages. The method is stable in the case of collinear motion, in the case of pure rotation and also for planar objects. There is no scale ambiguity since all motions are relative to the known baseline length [6].

After the motion has been found one can compute a dense depth map using information from the stereo pair and the motion (see fig. 1d). This can now be viewed as a multi baseline stereo system with all the advantages of such as system: it reduces aliasing, helps deal with occlusions and it extends the dynamic range of the system as detailed by [10] and others. Applying an edge detector to the depth map (fig. 1e) highlights possible areas of occlusion. In a traffic scene, for example, these are locations where a pedestrian might suddenly appear. The motion estimates between stereo pairs enable us to combine depth maps from all the pairs in the sequence to form an extended scene reconstruction.

The typical approach to reconstruction from a moving stereo rig ([11, 20, 19, 3]) has been to first compute a depth map from each pair (or 3D location of features points). One also computes a confidence measure for each depth estimate. Then one registers the stereo pair with the current 3D model to compute the motion and finally one updates the model (using a Kalman filter or a batch method).

Directly computing the motion of the stereo rig simplifies the problem of registering all the depth reconstructions into one coordinate frame. Our method gives good estimates of the camera motion between frames. There is no need to explicitly compute confidence values for the depth estimates since the causes for error are taken into account implicitly in the equations. By concatenating these motion estimates one can bring all the depth maps into one coordinate frame. Combining incremental motion in this way will accumulate errors. To get a more accurate reconstruction one can use the results as a starting point for global reconstruction schemes such as [5, 6].

1.1 Overview of the Camera Motion Estimation

The 'Tensor Brightness Constraint' [15, 17] combines geometric and photometric constraints of three views. It provides one homogeneous constraint equation for each point in the image where the unknown motion parameters are an implicit function of the spatial derivatives of Image 1 and the temporal derivatives between Image 1 and Image 2 and between Image 1 and Image 3. The motion parameters appear as 27 bilinear combinations of the camera motions. Point correspondences or optical flow are not required.

If one of the motions is known (e.g. motion from Image 1 to Image 2) then the 'Tensor Brightness Constraint' equation reduces to a non-homogeneous linear equation in the second motion parameters (i.e. the translation and rotation from image 1 to image 3). Again, one equation for each point in the image.

Let I'_i,I''_i,I'_{i+1},I''_{i+1} be the images taken by a stereo camera pair at times t_i and t_{i+1} respectively. One can apply the 'Tensor Brightness Constraint' to the images I'_i,I''_i and I'_{i+1}. Since the camera displacement from I'_i to I''_i is fixed and known (calibrated) the motion I'_i to I'_{i+1} can be found.

We have not used image I''_{i+1}. Theoretically it is not required since [4] show that there are no constraints among 4 (or more) images that are not simply 2 and 3 image constraints. But other image triplets can be used and the information combined in a least squares way.

In section (2) we briefly develop the 'Tensor Brightness Constraint' and then derive the constraint in the case where one motion is known. Section (3) describes some of the implementation details. In particular we describe a simple two step procedure to calibrate the stereo pair. Section (4.2) shows results for an extended scene reconstruction.

2 Mathematical Background

2.1 The Tensor Brightness Constraint

The 'Tensor Brightness Constraint' is developed in [17]. We briefly derive it here.

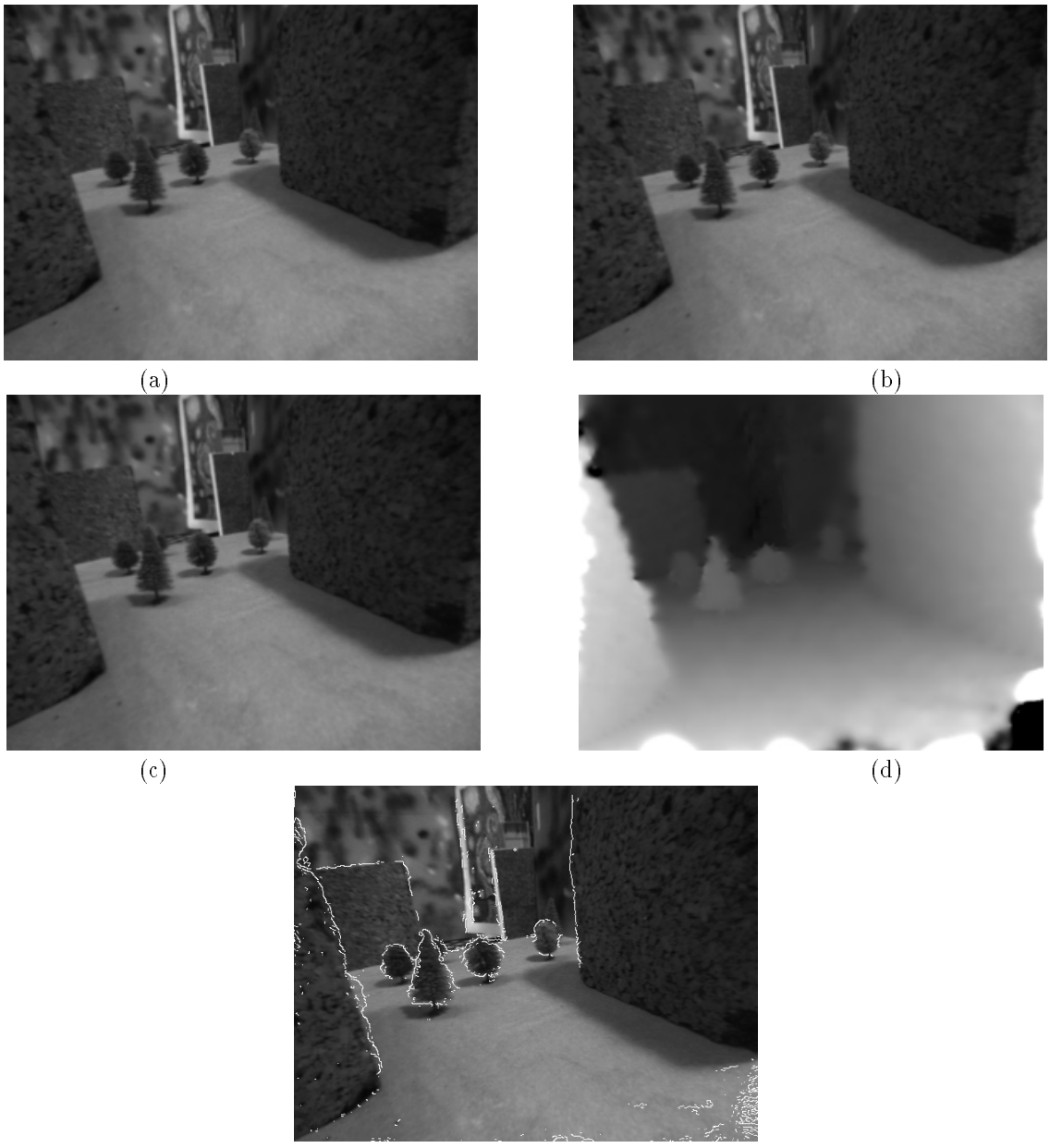
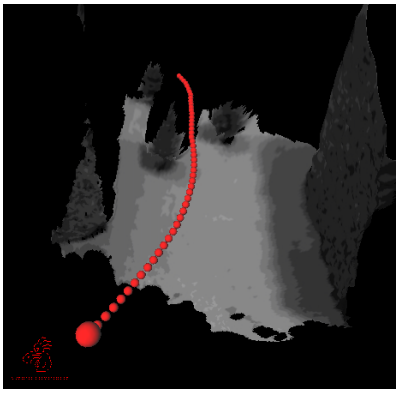
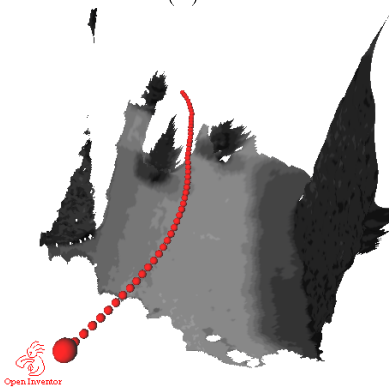


Figure 1: *The three input images (a,b,c). Image (a) and (c) are the first stereo pair in the sequence with the cameras displaced vertically. Image (b) is the image from camera (a) taken at the next time instance. (d) the estimated depth map.. (e) shows the results of a Canny edge detector on the depth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



(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. Not 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.*

2.1.1 The projective case

Let $P = (X, Y, 1, \rho)$ be a point in the projective space p^3 and it's image in three views $p = (x, y, 1)$, $p' = (x', y', 1)$, $p'' = (x'', y'', 1)$. Using homogeneous coordinates:

$$\begin{array}{lcll} p & \cong & [I;0]\,P & \\ p' & \cong & [A;t']\,P & \cong \quad Ap + \rho t' \\ p'' & \cong & [B;t'']\,P & \cong \quad Bp + \rho t'' \end{array} \tag{1}$$

Let S' and S'' be any lines passing through the image points p' and p'' respectively. Then p, S' and S'' are related through the following equation:

$$p^i S''_k S'_j \alpha_i^{jk} = 0$$

where α_i^{jk} is the $3 \times 3 \times 3$ tensor representing a bilinear function of the camera matrices: $\alpha_i^{jk} = t'^j b_i^k - t''^k a_i^j$ and first appeared in [13] and [14].

In particular, let:

$$S' = \begin{pmatrix} I_x \\ I_y \\ -x'I_x - y'I_y \end{pmatrix} \quad S'' = \begin{pmatrix} I_x \\ I_y \\ -x''I_x - y''I_y \end{pmatrix} \tag{2}$$

Now apply the 'optical flow constraint equation' [7]:

$$u'I_x + v'I_y + I'_t = 0 \tag{3}$$

where:

$$\begin{array}{lcl} u' & = & x - x' \\ v' & = & y - y' \end{array} \tag{4}$$

to get:

$$S' = \begin{pmatrix} I_x \\ I_y \\ I'_t - xI_x - yI_y \end{pmatrix} \quad S'' = \begin{pmatrix} I_x \\ I_y \\ I''_t - xI_x - yI_y \end{pmatrix} \tag{5}$$

This results in the **Tensor Brightness Constraint**:

$$\boxed{p^i s''_k s'_j \alpha_i^{jk} = 0} \tag{6}$$

where again $\alpha_i^{jk} = t'^j b_i^k - t''^k a_i^j$. The important thing to note here is that S' and S'' can both be computed without correspondences between the images.

2.1.2 The small motion model

Assuming calibrated cameras and small rotation eq. (1) can be written as:

$$\begin{array}{lcl} p' & \cong & [I + [w']_x; t']\,P \\ p'' & \cong & [I + [w'']_x; t'']\,P \end{array} \tag{7}$$

where $[\cdot]_x$ is the skew symmetric matrix of vector products. Then:

$$\begin{array}{lcl} s'^{\top}(I + [w']_x)\,p + k s'^{\top} t' & = & 0 \\ s''^{\top}(I + [w'']_x)\,p + k s''^{\top} t'' & = & 0 \end{array} \tag{8}$$

and after simplifying:

$$\begin{array}{lcl} k s'^{\top} t' + v'^{\top} w' + I'_t = & 0 \\ k s''^{\top} t'' + v''^{\top} w'' + I''_t & = & 0 \end{array} \tag{9}$$

where $v' = p \times s'$, $v'' = p \times s''$ and $k = 1/z$ has replaced ρ from eq. (1).
If we also use the Longuett-Higgins and Prazdny [12] small motion assumptions:

$$t'_z << Z \qquad \frac{w'_y x}{f} << 1 \qquad \frac{w'_x y}{f} << 1 \tag{10}$$

where f is the focal length, and w'_x, w'_y are the rotations around the X and Y axes. Thus for the first motion we get:

$$k s^{\top} t' + v^{\top} w' + I'_t \quad = \quad 0 \tag{11}$$

which first appeared in [8]. And similarly for the second motion:

$$k s^{\top} t'' + v^{\top} w'' + I''_t \quad = \quad 0 \tag{12}$$

Eliminating $k = 1/z$ from equations (11) and (12) we obtain the 15-parameter **model-based brightness constraint**:

$$\boxed{I_t'' s^\top t' - I_t' s^\top t'' + s^\top [t' w''^\top - t'' w'^\top] v = 0} \tag{13}$$

where s, v are defined below:

$$s = \begin{pmatrix} I_x \\ I_y \\ -xI_x - yI_y \end{pmatrix} \quad v = \begin{pmatrix} -I_y - y(xI_x + yI_y) \\ I_x + x(xI_x + yI_y) \\ xI_y - yI_x \end{pmatrix} \tag{14}$$

Notes:

- There is one such equation for each point in the image.
- The equation involves image gradients and image coordinates in image 0 only. It does not require correspondences (x', y', x'', y'').
- Since it is based on the 'optical flow constraint equation' it is correct only for infinitesimal motion.
- In the case of collinear motion (i.e. $t' \propto t''$) the system of equations becomes degenerate.

2.2 Moving Stereo Rig

Let us assume that t'' and w'' are known. We can then rewrite equation (13) in the form:

$$(I_t'' + w''^\top v) s^\top t' - (s^\top t'') v^\top w' = I_t' s^\top t'' \tag{15}$$

which is linear in the unknown translation and rotation, t', w' respectively. If the stereo pair has been rectified (i.e. $w'' = 0$) then equation (15) can be further simplified to:

$$I_t'' s^\top t' - (s^\top t'') v^\top w' = I_t' s^\top t''. \tag{16}$$

The above derivation was performed for the case of calibrated cameras but an equivalent derivation is possible using uncalibrated cameras for the projective case.

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 [18]. 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 [9]. 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. We then use equation (13) which gives accurate results under the assumption of pure translation [17] to compute both the translation of the rig and the displacement between the two cameras.

3.1.1 Lens Distortion

Since we us are using a wide 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 [16]).

3.2 Computing Depth

To compute the depth at every point we use equations (11). Information is combined from both image pairs and over a local region by minimizing:

$$\min_K \arg \sum_{x,y \in R} \sum_j \beta(x,y) |s^T t^j|^p \left(k s^T t^j + v^T w^j + I_t^j \right)^2 \tag{17}$$

- The windowing function $\beta(x, y)$ allows one to increase the weight of the closer points.
- The $|S^T t^j|^p$ 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 7×7.
- After the last iteration, we reduced the region R to 1×1 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 640×480 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 equation 15).
2. Compute depth (using equation 17).
3. Using the depth and motion, warp images 2 and 3 towards image 1.
4. Compute new time derivatives 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 [17]. At the finest level (640×480) 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 640×480 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 figure 3). 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 perpendicular to the translation stage will be denoted 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 was computed for each motion in the sequence. The motion estimates are plotted in figure 4. Figures (1), (6) and (7) 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. Figure (1e) shows the results of a Canny edge detector applied to the (inverse) depth image highlighting potential occluding contours.

Figure (8) shows the four reconstructions aligned using the estimated camera motions. The scene has more information than can be seen from any one camera.

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). The bias increases when the translation has a large component parallel to the image plane. This system provides 'where' information but because of the wide field of view and finite resolution of the camera it does not provide accurate 3D models of the objects in the scene ('what' information).

While we have shown that this method can give good motion estimates these are only incremental motion estimates. A full system could incorporate this method as a better estimation stage inside a Kalman filter or batch global estimation framework. We have also not dealt with the issue of how to represent the final 3D reconstruction. At this

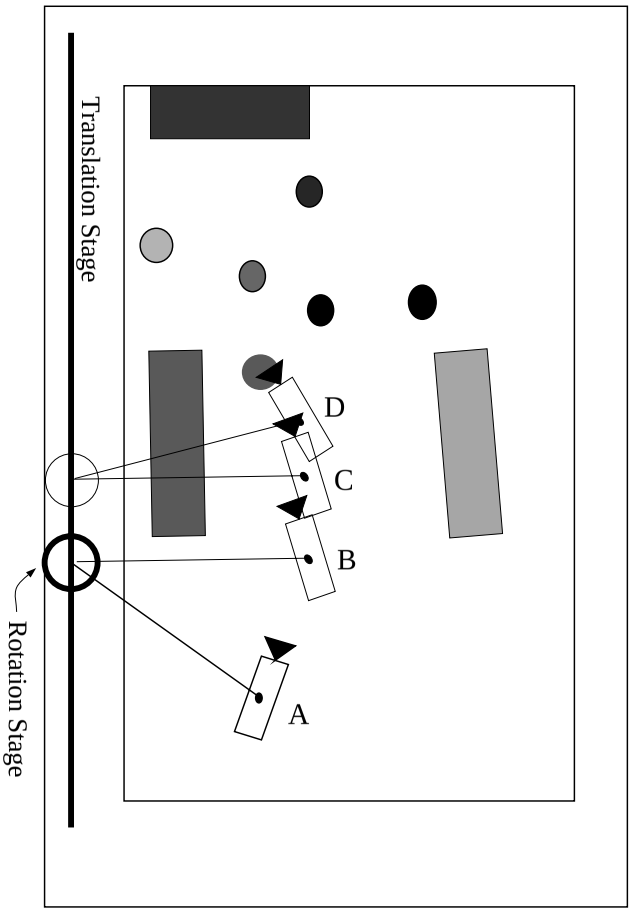


Figure 3: *Schematic overhead view of the 'fly-through' scene indicating the three cork blocks, 6 model tree (circles) and the motion stage (not to scale). The camera arm is rotated* 45^o *in* 1.5^o *steps from position A to position B. The arm then translates in twelve* $1cm$ *steps to position C. Then it Rotates through* 12^o *to position D.*

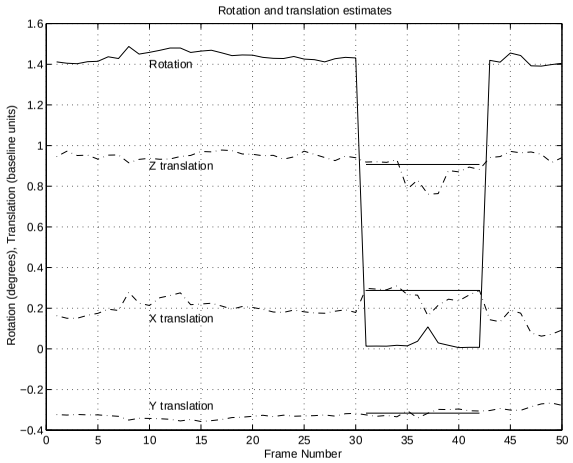


Figure 4: *Estimated Rotation and translation. The rotation angle (solid lines) should be* 1.5^o *and is on average* 1.45^o*. The translation estimates (dot-dashed lines) are plotted along with the true values for the pure translation section (short solid lines).*

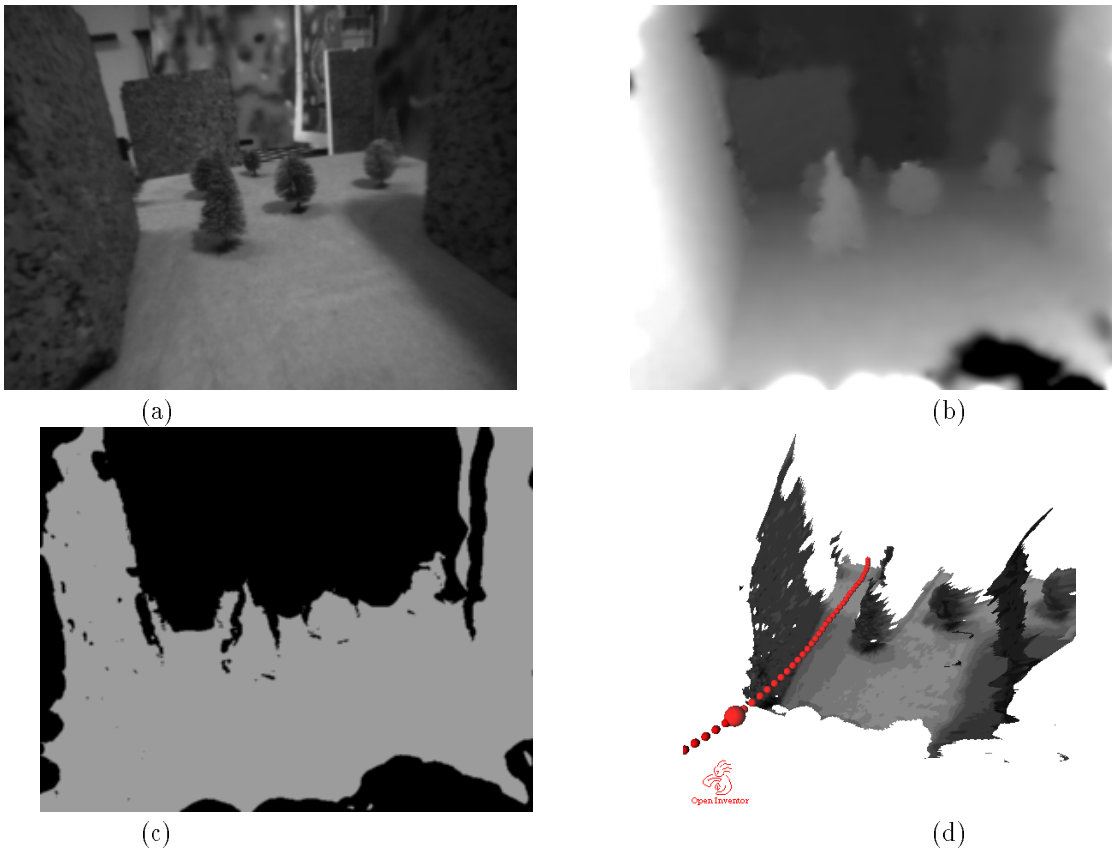


Figure 5: *(a) The 14th image in the sequence. (b) The depth map. (c) The mask. We do not render surfaces that are far from the camera or where the surface is facing the camera with a sharp angle. (d) rendering of the surfaces.*

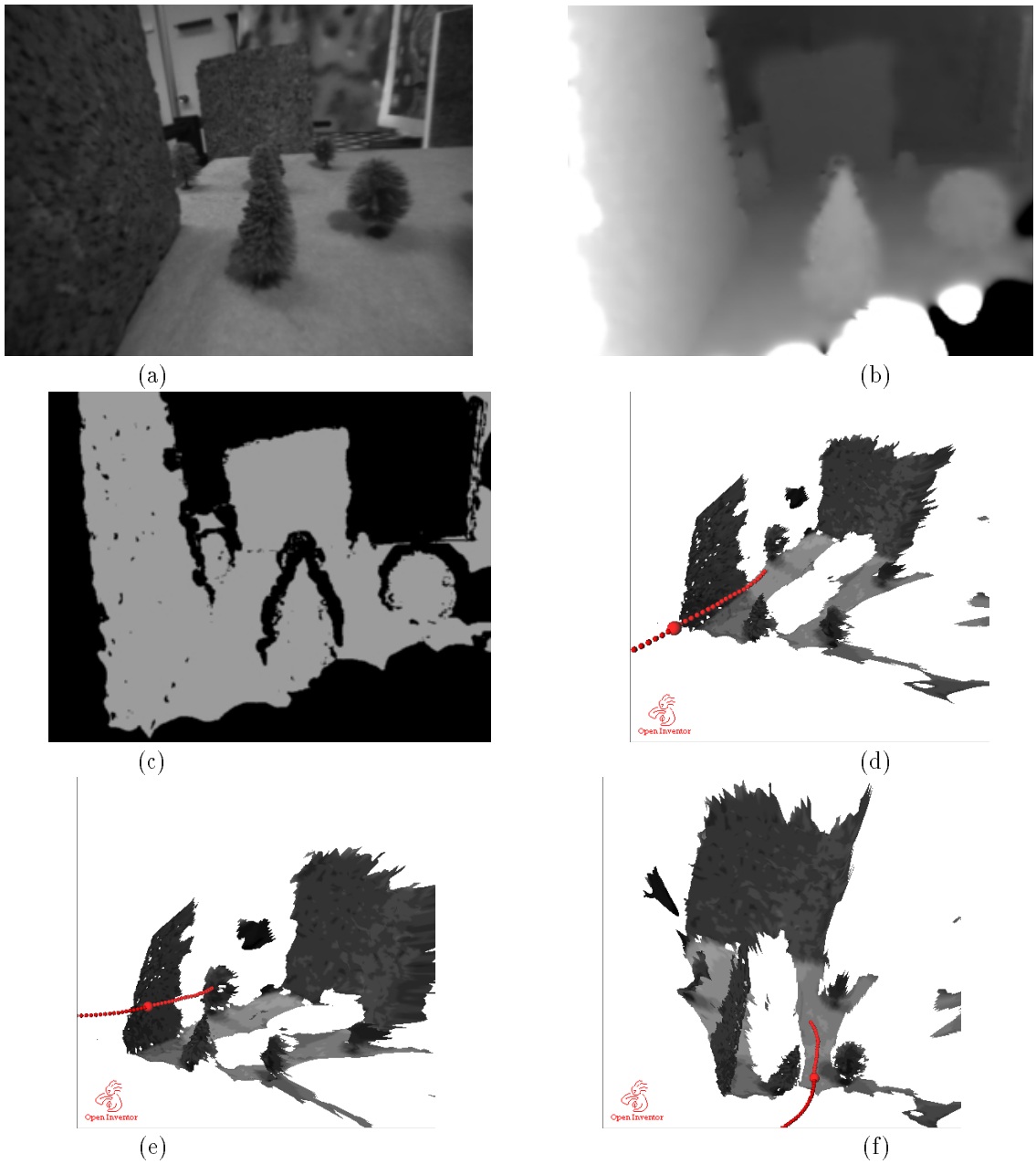


Figure 6: *(a) The 28th image in the sequence. (b) The depth map. (c) The mask. We do not render surfaces that are far from the camera or where the surface is facing the camera with a sharp angle. (d), (e),(f) rendering of the surfaces.*

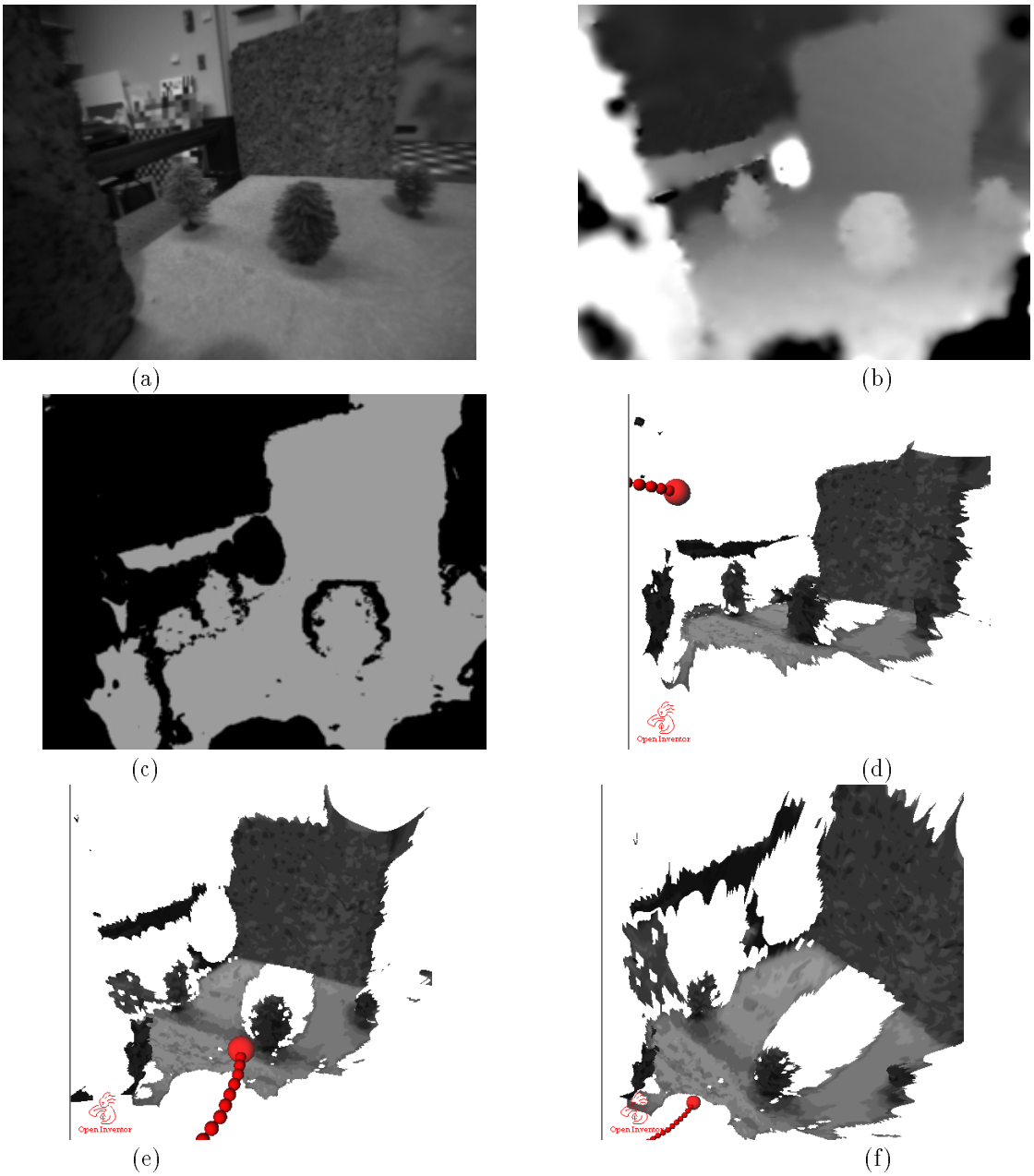


Figure 7: *(a) The 49th image in the sequence. (b) The depth map. (c) The mask. (d), (e),(f) 3D renderings of the surfaces. In this view a bit of the translation stage can be seen (top left quadrant).*

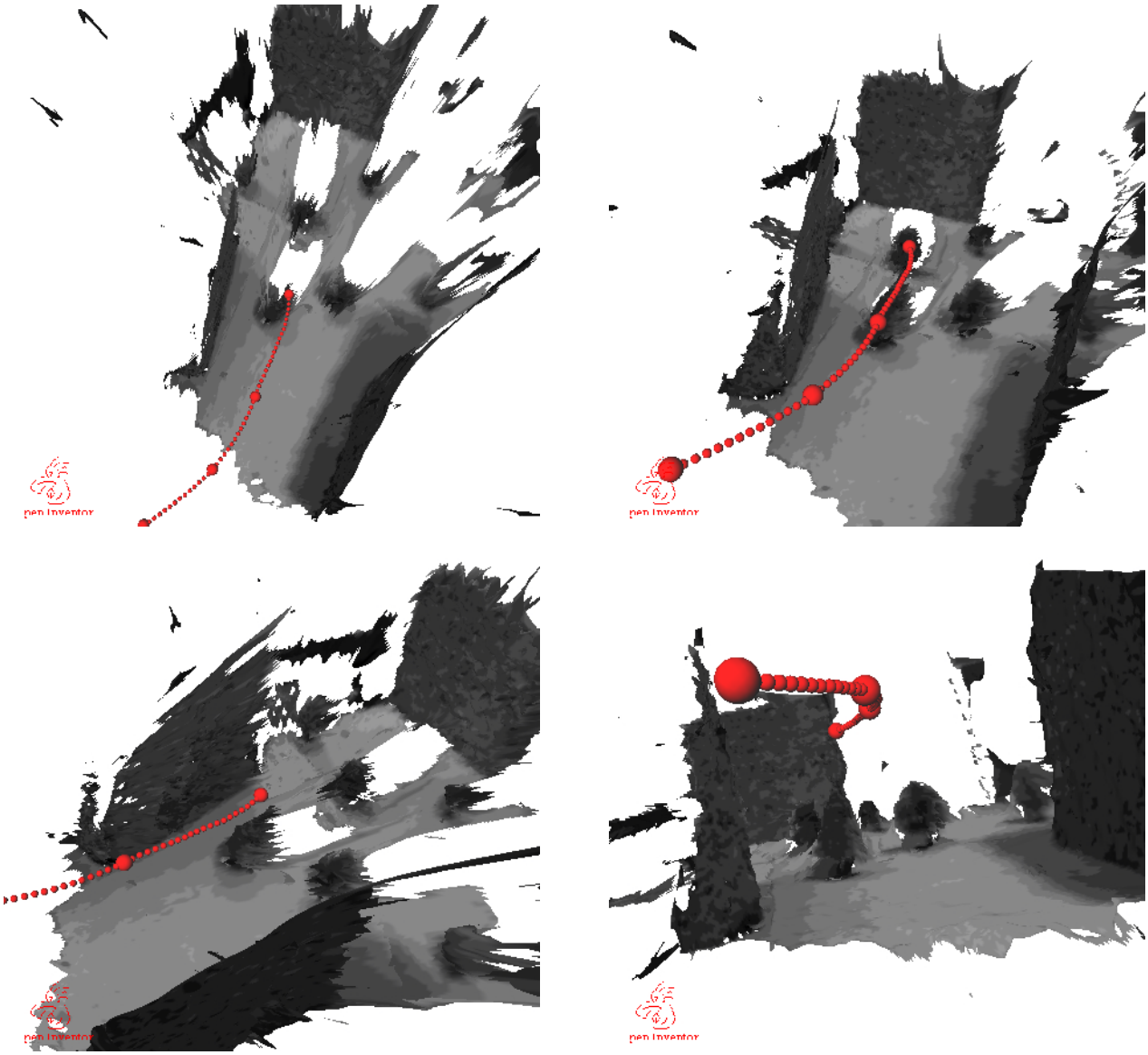


Figure 8: *Euclidean reconstruction of the extended scene created by aligning four separate reconstructions into one coordinate system using the recovered camera motion. The small spheres show the camera path. The large spheres show the camera positions of the four Euclidean reconstructions.*

point we draw multiple $2\frac{1}{2}D$ surfaces. Better methods are described in [5]. The idea of using multiview geometric constraints to merge 3D reconstruction from a stereo sequence pair can also be applied in systems 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] Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31:532–540, 1983.

[3] Olivier Faugeras. *Three Dimensional Computer Vision: a Geometric Viewpoint*. The MIT Press, Cambridge, MA, 1993.

[4] Olivier D. Faugeras and B. Mourrain. On the geometry and algebra of the point and line correspondences between N images. In *Proceedings of the International Conference on Computer Vision*, pages 951–956, Cambridge, MA, June 1995. IEEE Computer Society Press, IEEE Computer Society Press.

[5] Patrick 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.

[6] Keith J. Hanna and Neil 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.

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

[8] B.K.P. Horn and E.J. Weldon. Direct methods for recovering motion. *International Journal of Computer Vision*, 2:51–76, 1988.

[9] 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.

[10] T. Kanade, Okutomi, and Nakahara. A multiple-baseline stereo method. In *Proceedings of the ARPA Image Understanding Workshop*, pages 409–426. Morgan Kaufmann, San Mateo, CA, January 1992.

[11] Reinhard 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.

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

[13] A. Shashua. Algebraic functions for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):779–789, 1995.

[14] A. Shashua and M. Werman. Trilinearity of three perspective views and its associated tensor. In *Proceedings of the International Conference on Computer Vision*, June 1995.

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

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

[17] G. 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.

[18] Gideon 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.

[19] Juyang Weng, Paul Cohen, and Nicolas Rebibo. Motion and structure estimation from stereo image sequnces. *IEEE Transactions on Robotics and Automation*, 8(3):362–382, June 1992.

[20] Zhengyou Zhang and Olivier 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.