Estimating Rigid 3D Motion by Stereo Fixation of Vertices

D. Yang and J. Illingworth
Dept. of Electronics and Electrical Engineering,
University of Surrey, Guildford. GU2 5XH

Abstract

In this paper 3D motion estimation using an active stereo vision paradigm is addressed. The objective is to measure camera rotations of a stereo head to estimate the 3D motion of a single moving object in a scene. The method uses stereo camera fixation of junctions. The vertex of a junction is selected as the fixation point. Vertex position and junction orientations are recovered in the fixated views and used with known camera motion to estimate 3D rigid object motion. Experiments are reported for real data and promising results are obtained.

1 Introduction

Motion analysis is the process of estimating object motion parameters through analyzing temporal and spatial information contained in an image sequence. During the past decade, a great deal of effort has been devoted to motion analysis because of its use in a variety of applications such as mobile robot navigation and 3D scene interpretation. Existing methods can be divided into two categories depending upon whether they use temporal changes of inter-frame features (known as feature tracking) or temporal changes of image brightness values (known as optical flow). Much of the early work on motion estimation concentrated on using a sequence of monocular images[1, 6]. With new developments in stereo vision systems, motion estimation using a sequence of stereo images has recently been investigated[11, 13]. The use of stereo vision allows the recovery of 3D depth information which greatly reduces the complexity of the motion estimation task.

Active vision has recently received much attention. In the context of motion analysis. Active vision aims to achieve various motion tasks not only by analyzing temporal and spatial information but also by making use of controlled camera motions. Fixation, keeping the point of interest at the image center in a sequence of images, is one of the mechanisms exploitable in an active vision system. Taalebinezhaad[12] used fixation for solving the 2D motion estimation problem using a sequence of monocular images. He has shown that a constraint equation that explicitly expresses the rotational velocity as a linear function of the translational velocity can be obtained in the case of fixation. Using a pixel shifting process to obtain fixated images, he estimated 2D motion in the fixated view. Reid and Murray[10] have demonstrated that a fixation point can be obtained for a real time gaze control system. Using affine structure available from two or three views, they proposed a method of tracking corner features over time and developed a control strategy. Pahlavan et al[9] have shown impressive video-rate stereo tracking based on low-level focussing and vergence mechanisms on an active vision head.

BMVC 1993 doi:10.5244/C.7.5

Traditional approaches to estimating 3D motion using a sequence of images acquired by a passive stereo head consist of two steps. Firstly, establish the correspondences of 2D features between left and right views to recover 3D features. Secondly, match the 3D features between consecutive time instances and estimate 3D motion. However, problems may occur with either stage of these methods because in complicated dynamic scenes both stereo correspondence and temporal matching are difficult. In the method discussed in this paper some of these difficulties are alleviated by using high level features and by applying a simple motion segmentation algorithm to the line data to eliminate static scene features. Vertices formed by straight lines are used as basic features as they are easy to extract from image data yet sufficiently complicated to provide strong constraints on object pose. The motion segmentation uses the fact that if camera motion is restricted to rotation about the optical centre then the change in position of static features is predictable. Once motion segmentation has been achieved only a few linear features are left for junction finding and matching.

It has been demonstrated by several authors[10, 9] that video-rate object tracking can be achieved. However, in order to simultaneously determine object pose it is necessary to carefully analyse the changes in appearance of the object. Under general perspective transformation this is a complicated problem. The method proposed in this paper utilises the simplifying observation of Kanatani[7] that for junction features located at the image center there is no difference in appearance between orthographic and perspective projection. When the junction vertex is centred in the image then the angles between component lines which form the junction are the same under both orthographic and perspective projection. As pose estimation under orthographic projection is quite simple then the process of centering the target vertex at the image center considerably reduces the difficulty of general pose estimation.

In theory the process of dynamic tracking and vergence on a vertex will bring the vertex to the image center. However, in fast tracking there is often a delay or lag between the motion of the camera and the object, especially if the motion is not uniform. Thus there will often be a small discrepancy between the observed vertex position and the image centre. To overcome this we propose to use, in addition to actual head tracking movements, a software implemented or virtual camera rotation to bring the vertex to the exact centre of the image. The virtual rotation technique suggested by Kanatani[7] is used. Once the object is centred in the image centre then measurements on vertex features can be combined with known camera and virtual rotations to find the translation and change of pose of the target object.

The work is designed for implementation on the GETAFIX stereo camera head designed and built at Surrey, see Figure 1. This head has common neck pan and tilt, independently vergeable stereo cameras and controllable zoom, focus and aperture on each camera lens.

The paper is structured as follows: Section 2 introduces some notation and gives an overview of the proposed method. In Section 3 the problem of estimating 3D rigid motion from changes in observed junctions and known tracking movements is discussed. Experimental results on real data are given in Section 4. A brief summary in Section 5 concludes the paper.

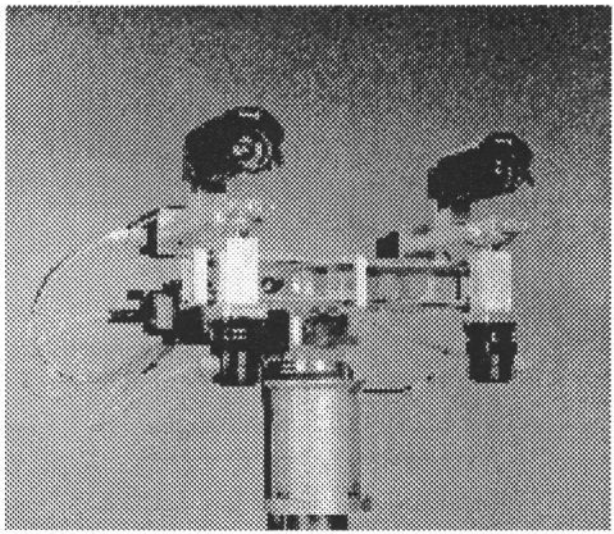

Figure 1: GETAFIX: The Surrey VSSP Group Robot Head

2 Overview

The problem which we address involves taking images with two cameras at different time instances and relating geometric transformations between various coordinate systems. The spatial relationship between left and right cameras, stereo geometry, is specified by a rotation R^{lr} followed by a translation T^{lr} from left centered camera coordinate system to right centered camera coordinate system. The left and right cameras can rotate around their optical centres independently. We use R^r and R^l to denote the rotations of left and right cameras, respectively. Sub-scripts will be added to the above notations to indicate time where this needs to be made clear, i.e., R_1^{lr} means the rotation of stereo geometry at time instance t_1 and R_1^r means the rotation of right camera around its optical center between time instance t_1 and time instance t_2. With the above notations we have that

$$R_{k+1}^{lr} = R_k^r R_k^{lr} (R_k^l)^{-1} \tag{1}$$

$$T_{k+1}^{lr} = R_k^r T_k^{lr} \tag{2}$$

which are used to update stereo geometry after cameras are rotated.

Figure 2 shows the basic stages of the proposed algorithm. Four images from two different times are input. Lines are extracted using Canny edge detection and a line finder. This set of lines could then be passed directly to a junction finder. However, the problem is simplified by rejecting those lines which belong to objects which are stationary in the scene. This pruning is achieved by restricting head motion to rotations about the camera centre. In this case the change in position of object lines which are stationary in the scene can be predicted. Lines which do not have corresponding predictions in images taken at different times are those which have undergone significant motion in 3D and it is this subset which are selected for junction finding. Consider a camera rotation about the optical center,

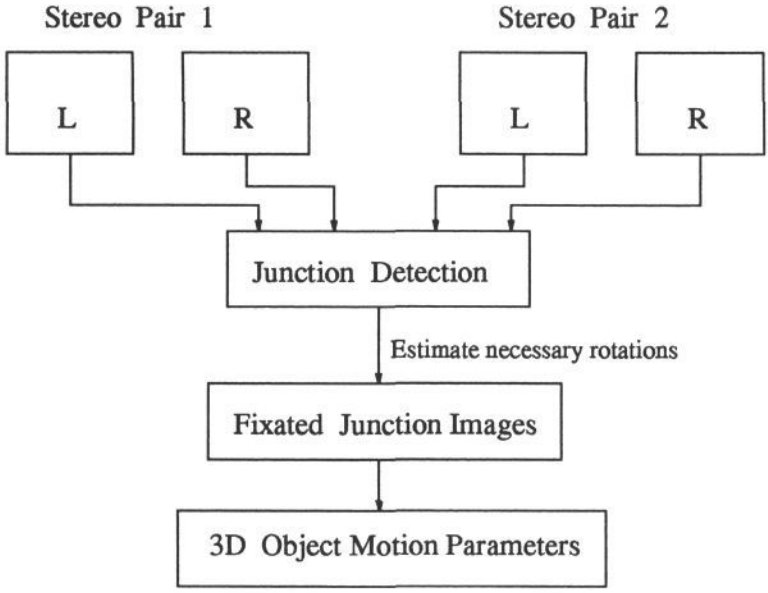

Figure 2: Summary of main steps of the method

$R = (r_{i,j}, i, j = 1 \ldots 3$. Using perspective projection the image coordinates after rotation (u_2, v_2) are related to the image coordinates before rotation (u_1, v_1) by

$$u_2 = f\frac{r_{11}u_1 + r_{21}v_1 + r_{31}f}{r_{13}u_1 + r_{23}v_1 + r_{33}f} \qquad v_2 = f\frac{r_{12}u_1 + r_{22}v_1 + r_{32}f}{r_{13}u_1 + r_{23}v_1 + r_{33}f} \tag{3}$$

where f is the focal length of the camera and the inverse of the transformation can be obtained by replacing R by R^{-1}. A line in the image plane can be parameterized as $u + a_1 v + a_2 = 0 (a_1 \neq 0)$or $v + a_2 = 0$. Without loss of generality, let us consider a line $u + a_1 v + a_2 = 0$. It can be easily verified that the line is mapped onto $u + a_1' v + a_2' = 0$ when the camera is rotated by $R = (r_{ij}), i, j = 1, 2, 3$ where

$$a_1' = \frac{r_{12}f + a_1 r_{22}f + a_2 r_{32}}{r_{11}f + a_1 r_{21}f + a_2 r_{31}} \tag{4a}$$

$$a_2' = \frac{(r_{13}f + a_1 r_{23}f + a_2 r_{33})f}{r_{11}f + a_1 r_{21}f + a_2 r_{31}} \tag{4b}$$

Therefore, if there is a line with parameters a_1' and a_2' in the image plane of the rotated camera this line is a static feature in the scene.

Stereo correspondence of junctions taken at the same time instant is established using the epipolar constraint and correspondence between times is usually easily determined as there are few vertices. The vertex of a junction which appears in both views at two consecutive time instants is chosen as the fixation point and exact fixation is achieved via virtual camera rotations. Using information about the transformation (real plus virtual) required to achieve fixation and measurements of the orientations of the vertex lines allows the calculation of the motion of the rigid object of which the junction is part.

3 Motion Estimation Using Fixation

In this section the problem of estimating 3D motion of a junction from stereo pairs of fixated images is considered. Junctions are found following edge detection, line finding and rejection of static linear scene features. Correspondence of junctions is established using simple epipolar and similarity constraints. The junctions are generally not centred in the image because of imprecision in tracking motions. It is therefore necessary to bring the vertices to the center of the image using virtual camera rotations. The virtual camera rotation defined by Kanatani [7] is used. Denote the image coordinate of the fixated vertex by (u, v). The virtual camera rotation to map (u, v) to $(0, 0)$ is given by [7]

$$R(u,v) = \begin{pmatrix} D & E & A \\ E & F & B \\ -A & -B & C \end{pmatrix} \tag{5}$$

where $A = \frac{u}{\sqrt{u^2+v^2+f^2}}$, $B = \frac{v}{\sqrt{u^2+v^2+f^2}}$, $C = \frac{c}{\sqrt{u^2+v^2+f^2}}$, $D = \frac{u^2C+v^2}{u^2+v^2}$, $E = \frac{uv(C-1)}{u^2+v^2}$ and $F = \frac{v^2C+u^2}{u^2+v^2}$.

Applying the virtual camera rotation of (5) to both left and right views, a pair of fixated images is obtained. The virtual rotations are denoted by $\bar{R}^l$ and $\bar{R}^r$ for left and right views respectively.

To this point, fixated images for left and right views at the same time instant have been constructed. In the following a method is presented for estimating 3D motion using the fixated images from two time instants. The method has two stages: (1) reconstruct 3D junction vertex and 3D line orientations of the junction (2) estimate motion parameters from the 3D junction parameters at consecutive time instances.

Firstly the 3D junction vertex position is the intersection of the z axes of the left and the right camera coordinate systems and therefore can be easily obtained. The 3D vertex of the junction at time t_1 and t_2 are denoted by F_1 and F_2, respectively. Since the junction vertex is at the image origin the orientation of the 3D line of junction can be represented in spherical coordinate as

$$\boldsymbol{l} = (\cos\Phi\sin\Theta,\ \sin\Phi\sin\Theta,\ \cos\Theta)^T \qquad (0 \le \Phi < 2\pi) \quad (0 \le \Theta \le \pi) \tag{6}$$

where Φ is related to the 2D projection line of the 3D line onto the image plane by

$$(\cos\Phi)u - (\sin\Phi)v = 0 \tag{7}$$

and Φ can be determined in the line finding process.

Consider a pair of matching lines between left and right fixated views where stereo geometry is given by $\bar{R}^{lr} = (\bar{r}_{ij}), i, j = 1, 2, 3$, we have that

$$\cos\Phi_r \sin\Theta_r - (\bar{r}_{11}\cos\Phi_l + \bar{r}_{12}\sin\Phi_l)\sin\Theta_l - \bar{r}_{13}\cos\Theta_l = 0 \tag{8a}$$
$$\sin\Phi_r \sin\Theta_r - (\bar{r}_{21}\cos\Phi_l + \bar{r}_{22}\sin\Phi_l)\sin\Theta_l - \bar{r}_{23}\cos\Theta_l = 0 \tag{8b}$$
$$\cos\Theta_r - (\bar{r}_{31}\cos\Phi_l + \bar{r}_{32}\sin\Phi_l)\sin\Theta_l - \bar{r}_{33}\cos\Theta_l = 0 \tag{8c}$$

where subscripts l and r attached to angles indicate the angles with respect to left and right camera coordinate systems, respectively. Since there are two unknowns

Θ_l and Θ_r in the three equations of (8), the solution can be obtained using any two of them provided that they are linearly independent. In most cases the equations are linearly independent and therefore we can increase the accuracy of the solution for Θ_l and Θ_r by computing the average of solutions.

It is worth noting that the proposed method of reconstructing orientation of lines is slight different from the traditional stereo method. In the traditional stereo method orientation of a 3D line is computed from the stereo images by intersection of two planes in which both rotation and translation of stereo geometry are involved. However, we can see from equation (8) that the 3D reconstruction in the fixated views is only dependent on the rotation of stereo geometry. Therefore in the proposed method errors in translation of stereo geometry do not produce errors in the orientation of estimated junction lines. In addition, it can be seen that the accuracy of orientation of junction lines can be improved by computing the average of solutions of the alternative formulates.

Having obtained the orientation of the fixated 3D junctions at time t_1 and t_2, the rotation of the moving junction between time t_1 and t_2 can be estimated. Three linearly independent basis vectors can be constructed from the lines which form the junctions at times t_1 and t_2. In the case of a Y junction, i.e., three lines l_1, l_2 and l_3 joining a common vertex, the three lines can be used directly if they are linearly independent, i.e.

$$|(l_1 \times l_2) \bullet l_3| > 0 \tag{9}$$

where $\bullet$ and $\times$ denote dot and vector cross product, respectively. Otherwise, two of the lines can be used to construct a third line using $l_3 = l_1 \times l_2$. In the case of a V junction, i.e., two non-collinear lines joining at a vertex, the third virtual line can be constructed in similar way.

Let $l_{1,i}, (i = 1, 2, 3)$ be the orientation of the three lines of the fixated junction at time t_1 and $l_{2,i}$ be the orientation of the lines of the matched fixated junction at time t_2. The rotation R_{fix} of the junction between t_1 and t_2 in the fixated view can be computed by the least-squares optimization

$$R_{fix} = \min_R \sum_{i=1}^{3} W_i \ ||l_{2,i} - R\ l_{1,i}||^2 \tag{10}$$

where W_i are nonnegative weights chosen to reflect knowledge of the precision of the data. Assuming that longer line segments have more reliable orientation the weights can be assigned as follows:

$$W_i = (w_{1,i} + w_{2,i}) / \sum_{j=1}^{2} \sum_{i=1}^{3} w_{j,i} \tag{11}$$

where $w_{j,i}$ is the length of line i $(i = 1, 2, 3)$ at time t_j $(j = 1, 2)$. The optimization problem can be converted into another form such that

$$R_{fix} = \max_R \ tr(R^T\ C) \tag{12}$$

by defining the correlation matrix C between $l_{1,i}$ and $l_{2,i}$ as $C = \sum_{i=1}^{3} W_i\ l_{1,i}\ l_{2,i}^T$ The solution can be obtained using various methods such as singular value decomposition method [2], a quaternion representation method [3] or polar decomposition [4]. Applying the singular value decomposition method an estimate of R_{fix}

is obtained as follows:

$$R_{fix} = V \ \mathrm{diag}\{1, 1, \det(V \ U^T)\} \ U^T \tag{13}$$

where V and U are the factors of the singular value decomposition of correlation matrix C such that $C = VDU^T$ where $D = \mathrm{diag}\{\sigma_1, \sigma_2, \sigma_3\}$ is a diagonal matrix and $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq 0$ are the singular values.

The 3D motion of a moving junction with respect to the world coordinate system can be obtained by combining the rotation of the junction in the fixated view R_{fix} and the depth of the fixation points F_1 and F_2 with respect to the right and left camera coordinate systems. In the fixated view

$$\bar{X}_2 - F_2 = R_{fix}(\bar{X}_1 - F_1) \tag{14}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the 3D point coordinates before and after motion in the fixated coordinate systems. Since it is known that

$$\bar{X}_1 = \bar{R}_1(X_1 - \bar{T}_1) \qquad \bar{X}_2 = \bar{R}_2(X_2 - \bar{T}_2) \tag{15}$$

where X_1 and X_2 are the 3D point coordinates before and after motion in the world coordinate system, the 3D motion of the junction with respect to the world coordinate system can be computed by combing equations (14) and (15). It is worth noting that rotation $\bar{R}_1$ and $\bar{R}_2$ in equation (15) accounts for both actual camera rotation and virtual image transform.

4 Experimental Results and Discussion

In this section results are reported for a real data example. The algorithm was implemented in C on a Sun workstation. Images were acquired using the Surrey VSSP group active stereo head (GETAFIX) (see Figure 1) which was calibrated using an implementation of Tsai's calibration routine[1]. The scene to be analysed consists of a moving box on a table. The first part of experiment is concerned with moving feature detection. Figure 3 (a) and (b) show linear features overlayed on the left camera images taken at times t_1 and t_2. Between times t_1 and t_2 the camera was rotated by 1.5 degrees around its optical center. Using this known rotation it is possible to identify and remove lines which are static in the 3D scene. This produces the images shown in Figure 3 (c) and (d). The lines which remain are moving with respect to the scene and form the basic input to the junction detection process. Most of the detected moving line segments belong to the box. Other lines which falsely remain are caused by line segments appearing in only one of the two images. This can be caused by poor performance of the Canny edge detection or the line detection processes.

The second part of the experiment shows the 3D recovery of the junctions. A different set of images is used in which the motion is larger than in the first example. Motion segmentation is once more applied and then the junction finder developed by Matas and Kittler [8] is used. In our experiment there are usually 1 to 3 Y junctions and 10 to 20 V junctions found in each view. Stereo correspondence

[1] This software was produced by British Aerospace and forms part of the LAIRD (Location And Inspection using Range Data) project. LAIRD is a SERC/IED project and involves British Aerospace, BAe/SEMA, the National Engineering Laboratory and the Universities of Edinburgh, Heriot-Watt and Surrey

of junctions is established using epipolar constraints and temporal matching is done based on junction similarity criteria. Matching Y junctions are sought first as they provide more stringent constraints for both matching and pose determination. After matching both real and virtual camera rotations are used to fixate the vertex at the center of the image. The 3D structure of the junction is recovered in the fixated images and this can then be converted back to the original views. The effectiveness of this is illustrated in Figure 4 by showing the 3D junctions backprojected to the four original input images. It can be seen that the junctions overlap the original images very well. Finally, in Figure 4 the effectiveness of the estimate of 3D motion is demonstrated by showing that the junction estimate from time t_1 superimposes on the image at time t_2 after the motion estimate has been applied. It can be seen that the calculated motion is very close to the observed image changes.

5 Conclusions

In this paper a method of using binocular stereo vergence to recover the motion of an object has been considered. The method utilises junctions and an active tracking paradigm to recover world structure. The known camera motions needed to maintain vergence and the observed appearance of image lines which form the junction provide the information to deduce the 3D rigid object motion. The method has been demonstrated on real images captured using an active stereo head.

Acknowledgement

This work was carried out as part of the ESPRIT Basic Research Action EP-7108-VAP-II "Vision as Process II".

References

[1] J.K. Aggarwal and N. Nandhakumar. On the computation of motion from sequences of images — a review. *Proceedings of the IEEE*, 76:917–935, 1988.

[2] K.S. Arun, T.S. Huang, and S.D. Blostein. Least squares fitting of two 3-d point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9:698–700, 1987.

[3] B. K. P. Horn. Closed-form solution of absolute orientation using unit quatenions. *J. Opt. Soc. Amer.*, A-4:629–642, 1987.

[4] B. K. P. Horn, H. M. Hilden, and S. Negahdaripour. Closed-form solution of absolute orientation using orthonormal matricers. *J. Opt. Soc. Amer.*, A-5:1128–1135, 1988.

[5] *Fourth International Conference on Computer Vision (German, May, 1993)*, Washington, DC., 1993. Computer Society Press.

[6] C. P. Jerian and R. Jain. Structure from motion - A critical analysis of methods. *IEEE Transactions on Robotics and Automation*, 21(3):572–588, 1991.

[7] K. Kanatani. *Group-Theoretical Methods in Image Understanding.* Springer, Berlin, 1990.

[8] G. Matas and J Kittler. Junction detection using probabilistic relaxation. *Image and Vision Computing*, 11(4):197–202, 1993.

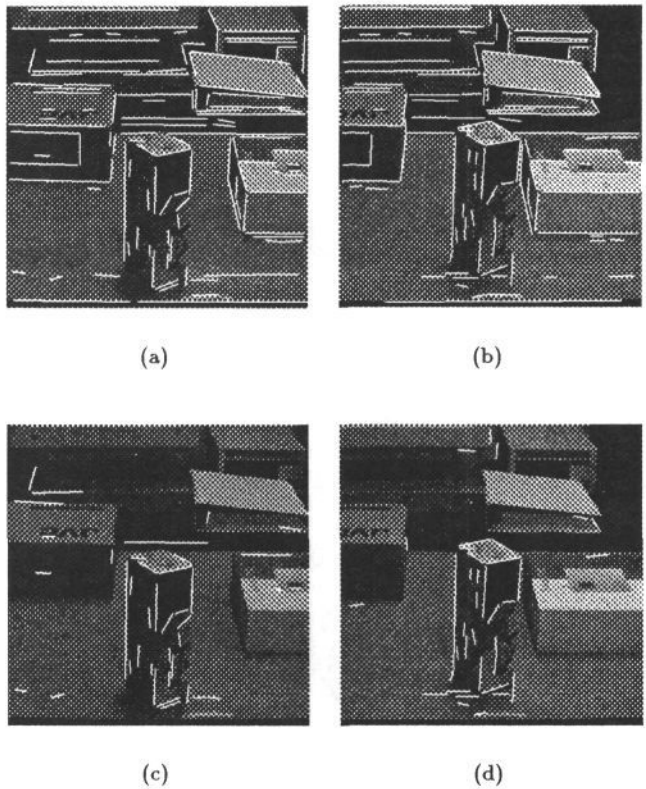

Figure 3: Moving feature detection: (a) lines in left camera image at time t_1 (b) lines in left camera image at time t_2 (c) moving lines in left camera at t_1 (d) moving lines in left camera at t_2

[9] K. Pahlavan, T. Uhlin, and J-O. Eklundh. Dynamic fixation. In Proceedings of 4th International Conference on Computer Vision, Berlin 1993, pages 412–419.

[10] I. D. Reid and D. W. Murray. Tracking foveated corner clusters using affine structure. In Proceedings of 4th International Conference on Computer Vision, Berlin 1993, pages 76–83

[11] B. Sabata and J. K. Aggarwal. Estimating of motion from a pair of range images: A review. *CVGIP: Image Understanding*, IU-54(3):309–324, 1991.

[12] M. A. Taalebinezhaad. Direct recovery of motion and shape in the general case by fixation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8):847–853, 1992.

[13] Z. Zhang and O. Faugeras. *3D Dynamic Scene Analysis: A Stereo Based Approach.* Springer, Berlin, 1992.

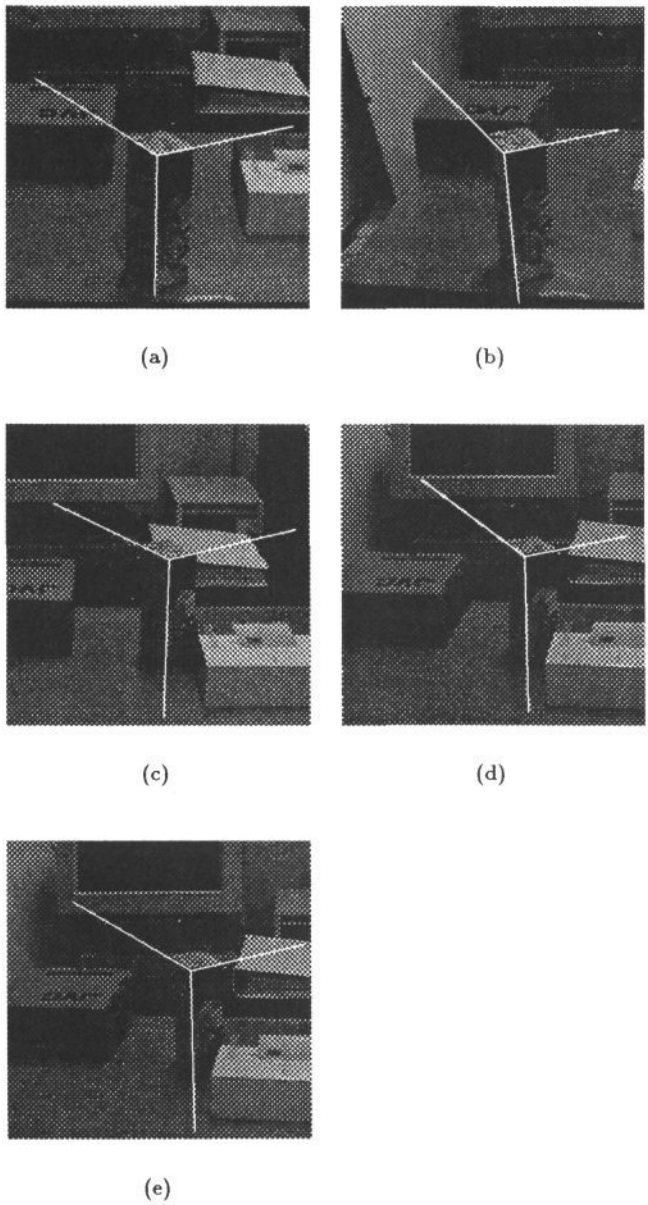

Figure 4: Back projection of 3D vertices onto the original image data: (a) left image at time t_1 (b) right image at time t_1 (c) left image at time t_2 (d) right image at time t_2 (e)Transformation of 3D vertex at time t_1 into right camera image at t_2 using motion estimate