Motion Estimation from Disparity Images

D. Demirdjian T. Darrell

Massachusetts Institute of Technology, AI Laboratory
Cambridge, MA 02139
demirdji@ai.mit.edu trevor@ai.mit.edu

Abstract

A new method for 3D rigid motion estimation from stereo is proposed in this paper. The appealing feature of this method is that it directly uses the disparity images obtained from stereo matching. We assume that the stereo rig has parallel cameras and show, in that case, the geometric and topological properties of the disparity images. Then we introduce a rigid transformation (called d-motion) that maps two disparity images of a rigidly moving object. We show how it is related to the Euclidean rigid motion and a motion estimation algorithm is derived. We show with experiments that our approach is simple and more accurate than standard approaches.

1 Introduction

The problems of estimation, detection and understanding motion from visual data are among the most challenging problems in computer vision. At a low-level, 3-D motion must be analyzed based on the 2-D features that are observable in images. At a high-level, the previously derived 2-D motion fields must be interpreted in terms of rigid or deformable objects and the associated motion parameters are estimated.

Many researchers have addressed the problem of motion/ego-motion estimation and motion segmentation for both monocular [3, 9] and stereo [7, 6, 13, 15] sensors. The approaches using a stereo rig for ego-motion estimation are appealing because they are based on 3-D information (*e.g.* Euclidean reconstructions) and less noise-sensitive than monocular approaches.

Many approaches for ego-motion and motion estimation with a stereo rig have a common structure. First, image points are matched in the image pair. A disparity image is obtained, and a scene reconstruction is performed in the Euclidean space. Then the rigid motion is estimated based on 3-D point correspondences.

Unfortunately using the 3-D Euclidean space to estimate rigid motions fails to give optimal solutions because reconstructions performed in 3-D Euclidean space have non homogeneous and non isotropic noise (as wrongly approximated in many standard SVD- or quaternion-based approaches). One should instead use methods that deal with non homogeneous and non isotropic noise [12, 13] or methods that look for an optimum solution for both structure and motion, like *bundle-adjustment* [10].

The disparity images have a well-behaved noise (theoretically isotropic for parallel stereo images) and can be used instead of the 3-D Euclidean space for motion estimation in this paper. We recall that the disparity images are related to the 3-D Euclidean reconstruction by a projective transformation. We show that two disparity images of a rigid scene are related by a transformation (called *d-motion* in the paper). We present a motion estimation algorithm based on the d-motion estimation. The approach is simple and more accurate than standard motion estimation algorithms.

2 Notations

In this paper we consider a stereo rig whose images are rectified, *i.e.* epipolar lines are parallel to the x-axis. It is not a loss of generality since it is possible to rectify the images of a stereo rig once the epipolar geometry is known [1]. We also assume that both cameras of the rectified stereo rig have similar internal parameters so that the rig can be fully described by a focal length f, a principal point (u_0, v_0) and a baseline B.

Stereo reconstruction has been studied for decades and is now standard in computer vision. Consider a rectified image pair and let (x, y) and (x', y') be two corresponding points in that image pair. Since the corresponding points must lie on epipolar lines, the relation between the two points is:

$$\begin{cases} x' = x - d \\ y' = y \end{cases}$$

where d is defined as the *disparity* of the point (x, y).

Let us define $(\bar{x}, \bar{y}) = (x - u_0, y - v_0)$ the centered image point coordinates of (x, y). Let (X, Y, Z) be the Euclidean

0-7695-1143-0/01 $10.00 (C) 2001 IEEE

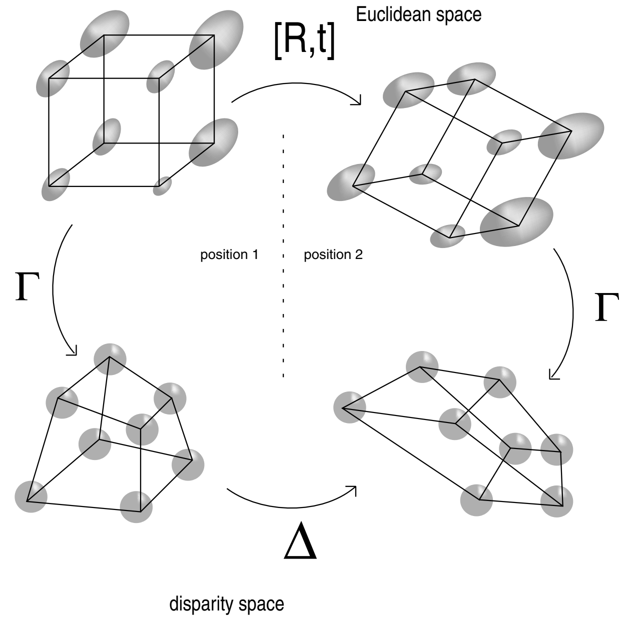

Figure 1: Euclidean reconstruction and motion of a cube *vs.* reconstruction and motion in the disparity space.

coordinates of the 3-D point $\boldsymbol{M}$ corresponding to the image point correspondence in a frame attached to the stereo rig. Then the relation between (X, Y, Z) and $(\bar{x}, \bar{y}, d)$ is:

$$\begin{cases} \bar{x} = x - u_0 = f\frac{X}{Z} \\ \bar{y} = y - v_0 = f\frac{Y}{Z} \\ d = \frac{fB}{Z} \end{cases} \tag{1}$$

Eq.(1) is widely used in the vision literature for estimating (X, Y, Z) from $(\bar{x}, \bar{y}, d)$. However $(\bar{x}, \bar{y}, d)$ happens to already be a reconstruction. We call the space of $(\bar{x}, \bar{y}, d)$ *disparity space*. This space has been used to perform such tasks as image segmentation, foreground/background detection. However, little work has been done to understand the geometric properties of the disparity space.

Organization

In this paper, we demonstrate some properties of the *disparity space*. In Section 3 we recall that the *disparity space* is a projective space and we discuss the form of the noise in this space in the case of parallel camera stereo rigs. Section 4 introduces the rigid transformations (d-motions) associated to that space. We show how d-motions are related to the Euclidean rigid transformations. An algorithm to calculate the Euclidean motion from d-motion is given. Section 5 shows experiments carried out with both synthetical and real data. Finally, our experiments are discussed Section 6.

3 Properties of the disparity image

In this section we argue the use of disparity images for spatial representation of stereo data. We claim that this representation has nice geometric and topological properties that makes it ideal for optimal motion estimation.

We show that (i) using homogeneous coordinates, the disparity image is a particular projective reconstruction of the 3-D observed scene; therefore, the disparity space is a projective space, and (ii) for parallel camera stereo rigs, the noise in the disparity space is isotropic.

3.1 Geometric feature : the disparity space is a projective space

Throughout the paper, homogeneous coordinates are used to represent image and space point coordinates and ”$\simeq$” denotes the equality up to a scale factor.

Let us consider a rectified stereo rig. Let f be the focal length, (u_0, v_0) the principal point coordinates associated with the stereo rig and B be the baseline of the stereo rig.

Let $\boldsymbol{M} = (X\ Y\ Z)$ be the 3-D coordinates of a point observed by a stereo rig. Let d be the disparity of the associated image point $\boldsymbol{m} = (x\ y)$ and $(\bar{x}\ \bar{y}) = (x - u_0\ y - v_0)$ the centered image point coordinates.

Using homogeneous coordinates and multiplying each term of the equations (1) by Z, we have:

$$\begin{aligned} Z \begin{pmatrix} \bar{x} \\ \bar{y} \\ d \\ 1 \end{pmatrix} = \begin{pmatrix} fX \\ fY \\ fB \\ Z \end{pmatrix} &= \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & fB & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \\ 1 \\ Z \end{pmatrix} \\ &= \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 0 & fB \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \end{aligned} \tag{2}$$

Let ω a vector such that $\omega = \begin{pmatrix} \bar{x} \\ \bar{y} \\ d \end{pmatrix}$. Then we have:

$$\boxed{\begin{pmatrix} \omega \\ 1 \end{pmatrix} \simeq \mathbf{\Gamma} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} = \mathbf{\Gamma} \begin{pmatrix} \boldsymbol{M} \\ 1 \end{pmatrix}} \tag{3}$$

where $\mathbf{\Gamma}$ is a 4×4 matrix such that:

$$\mathbf{\Gamma} = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 0 & fB \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

The Eq.(3) demonstrates that there is a projective transformation $\mathbf{\Gamma}$ between the homogeneous coordinates

$(X\ Y\ Z\ 1)$ in the 3-D Euclidean space and the homogeneous coordinates $(\bar{x}\ \bar{y}\ d\ 1)$. Therefore, as also shown in [2], $(\bar{x}\ \bar{y}\ d\ 1)$ is a projective reconstruction of the scene. The disparity space is then a projective space.

3.2 Topological feature

An important feature of the disparity space is that the noise associated with $(\bar{x}\ \bar{y}\ d)$ is known:

- The noise associated with $\bar{x}$ and $\bar{y}$ is due to the *image discretization*. Without any *a priori* information, the variances $\sigma_{\bar{x}}$ and $\sigma_{\bar{y}}$ of this noise is the same for *all* image points. We can write $\sigma_{\bar{x}} = \sigma_{\bar{y}} = \sigma$ where σ is the pixel detection accuracy (typically $\sigma = 1$ *pix.*).

- The noise associated with d corresponds to the *stereo matching uncertainty* and is related to the intensity variations in the image. The variance σ_d of this noise can be estimated from the stereo matching process.

From the discussion above, it is clear that the noises associated with $\bar{x}$, $\bar{y}$ and d are independent. Therefore the covariance matrix Λ_i associated with any reconstruction $(\bar{x}_i\ \bar{y}_i\ d_i)$ in the disparity space can be written as 3 × 3 diagonal matrix Λ_i such that:

$$\Lambda_i = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma_{d_i}^2 \end{pmatrix}$$

If the disparity reconstruction is restricted to the image points which have been stereo-matched with enough accuracy [8], *i.e.* when the estimated disparity is about $\sigma_{d_i} = \sigma = 1$ *pix.*, then the covariance Λ_i of the noise can be fairly approximated by $\Lambda_i = \sigma^2 \mathbf{I}$. The noise is then considered isotropic and homogeneous.

4 Rigid transformations in the disparity space

In this section, we introduce the transformation that maps two reconstructions of a rigid scene in the disparity space. We call this transformation *d-motion* and show how it is related to the rigid motion in the Euclidean space.

Let us consider a fixed stereo rig observing a moving point. Let $\boldsymbol{M} = (X\ Y\ Z)$ and $\boldsymbol{M'} = (X'\ Y'\ Z')$ be the respective 3-D Euclidean coordinates of this point before and after the rigid motion. Let $\mathbf{R}$ and $\boldsymbol{t}$ denote the rotation and translation of the rigid motion. Using homogeneous coordinates we have:

$$\begin{pmatrix} \mathbf{M}' \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R} & \boldsymbol{t} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{M} \\ 1 \end{pmatrix}$$

Replacing $\begin{pmatrix} \mathbf{M} \\ 1 \end{pmatrix}$ and $\begin{pmatrix} \mathbf{M}' \\ 1 \end{pmatrix}$ using Eq.(2) we have:

$$\Gamma^{-1} \begin{pmatrix} \omega' \\ 1 \end{pmatrix} \simeq \begin{pmatrix} \mathbf{R} & \boldsymbol{t} \\ 0 & 1 \end{pmatrix} \Gamma^{-1} \begin{pmatrix} \omega \\ 1 \end{pmatrix}$$

Let $\mathbf{H}_d = \Gamma \begin{pmatrix} \mathbf{R} & \boldsymbol{t} \\ 0 & 1 \end{pmatrix} \Gamma^{-1}$. Then we have:

$$\begin{pmatrix} \omega' \\ 1 \end{pmatrix} \simeq \mathbf{H}_d \begin{pmatrix} \omega \\ 1 \end{pmatrix} \tag{4}$$

Let $\bar{\mathbf{R}}$ be a 2 × 2 matrix, $\boldsymbol{r}$, $\boldsymbol{s}$ and $\bar{\boldsymbol{t}}$, 2-vectors and λ and μ scalars such that:

$$\mathbf{R} = \begin{pmatrix} \bar{\mathbf{R}} & \boldsymbol{r} \\ \boldsymbol{s}^T & \lambda \end{pmatrix} \qquad \boldsymbol{t} = \begin{pmatrix} \bar{\boldsymbol{t}} \\ \mu \end{pmatrix}$$

Then $\mathbf{H}_d$ can be expressed as follow:

$$\mathbf{H}_d = \begin{pmatrix} \bar{\mathbf{R}} & \frac{1}{B}\bar{\boldsymbol{t}} & f\boldsymbol{r} \\ 0 & 1 & 0 \\ \frac{1}{f}\boldsymbol{s}^T & \frac{\mu}{fB} & \lambda \end{pmatrix} \tag{5}$$

Using standard coordinates, Eq.(4) becomes:

$$\boxed{\omega' = \Delta(\omega) = \frac{1}{(\omega\ 1)^T \gamma} \left(\mathbf{A}\omega + \boldsymbol{b} \right)} \tag{6}$$

where

$$\mathbf{A} = \begin{pmatrix} \bar{\mathbf{R}} & \frac{1}{B}\bar{\boldsymbol{t}} \\ 0 & 1 \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} f\boldsymbol{r} \\ 0 \end{pmatrix} \quad \gamma = \frac{1}{f} \begin{pmatrix} \boldsymbol{s} \\ \frac{\mu}{B} \\ f\lambda \end{pmatrix}$$

We can also notice that from Eq.(6):

$$(\omega\ 1)^T \gamma = \frac{d}{d'} \tag{7}$$

The transformation $\omega' = \Delta(\omega)$ is called *d-motion*. In homogeneous coordinates, it can be defined by the matrix $\mathbf{H}_d$. In standard coordinates, it can be defined by $\mathbf{A}$, $\boldsymbol{b}$ and γ.

4.1 Motion estimation with d-motion

Let $\omega_i \rightarrow \omega_i'$ be a list of point correspondences. As we have to deal with dense disparity images, such technics as optical flow are required to estimate dense point correspondences. This paper only focuses on the 3-D motion estimation and the challenging problem of dense point correspondences will be tackled in a forthcoming paper.

The problem of estimating the rigid motion between the points ω_i and ω_i' amounts to minimizing over Δ the following error:

$$E^2 = \sum_i \varepsilon_i^2 \tag{8}$$

where $\varepsilon_i^2 = (\omega_i' - \Delta(\omega_i))^T \Lambda_i^{-1} (\omega_i' - \Delta(\omega_i))$

As demonstrated previously, if the focal length f and the baseline B of the stereo rig are known, Δ can be parameterized by $\mathbf{R}$ and $\boldsymbol{t}$. The error E^2 can therefore be minimized over $\mathbf{R}$ and $\boldsymbol{t}$.

4.2 Generalization to the uncalibrated case

The d-motion can be generalized in the uncalibrated case (f and B unknown). In that case, $\mathbf{A}$, $\boldsymbol{b}$ and $\boldsymbol{\gamma}$ can be represented by 12 general parameters such that:

$$\mathbf{A} = \begin{pmatrix} \star & \star & \star \\ \star & \star & \star \\ 0 & 0 & 1 \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} \star \\ \star \\ 0 \end{pmatrix} \quad \boldsymbol{\gamma} = \begin{pmatrix} \star \\ \star \\ \star \\ \star \end{pmatrix}$$

In this case, the d-motion can be considered as a particular case of projective transformation [11] and has the same structure as an affine transformation [5]. $\mathbf{R}$ and $\boldsymbol{t}$ cannot be recovered, but the d-motion Δ can still be estimated and used for tasks requiring no Euclidean estimation, such as motion segmentation or motion detection.

4.3 Minimization of E^2

In this paper, we are interested in the estimation of small motion; Therefore, the rotation $\mathbf{R}$ can be parameterized by:

$$\mathbf{R} = \mathbf{I} + \begin{pmatrix} 0 & -w_c & w_b \\ w_c & 0 & -w_a \\ -w_b & w_a & 0 \end{pmatrix} \qquad (9)$$

The error ε_i^2 can be expressed in a quasi-linear way:

$$\varepsilon_i^2 = || \frac{1}{(\omega_i \ 1)^T \boldsymbol{\gamma}} \mathbf{P}_i \boldsymbol{u} + \boldsymbol{v_i} - \omega_i' ||^2$$

where $\boldsymbol{u} = (\bar{\boldsymbol{t}}^T \ w_c)^T$, $\boldsymbol{v_i}$ is a 3-vector function of w_a, w_b, μ, ω_i and ω_i' and $\mathbf{P}_i$ is a matrix that depends on ω_i.

This form enables to perform the minimization of E^2 alternatively over $\boldsymbol{u}$ (w_a, w_b, μ fixed) using a linear method such as SVD, and then over w_a, w_b and μ ($\boldsymbol{u}$ fixed) using an iterative non-linear algorithm, such as Levenberg-Marquardt. In the case of larger motions, a global non-linear minimization must be performed over the 6 motion parameters.

5 Experiments

5.1 Simulated data

Experiments with simulated data are carried out in order to compare the quality of the results. A synthetic 3-D scene of

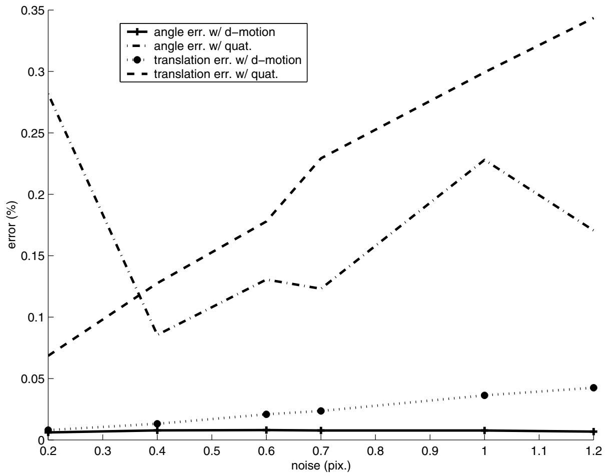

Figure 2: Estimation of the relative angle and translation errors for small motions

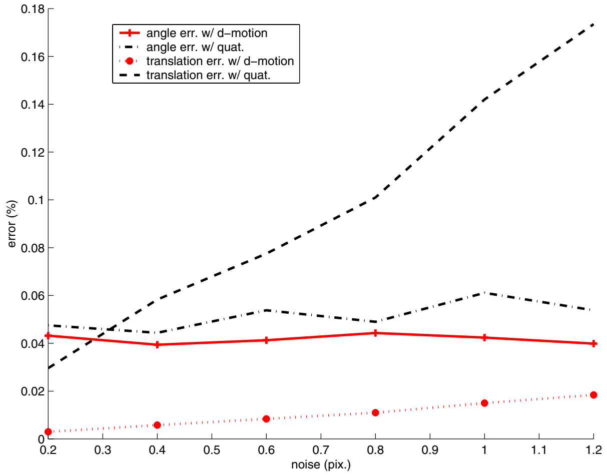

Figure 3: Estimation of the relative angle and translation errors for standard motions

0-7695-1143-0/01 $10.00 (C) 2001 IEEE

100 points is generated. A random rigid motion is generated as well. The 3-D points of each position (before and after the rigid motion) are projected onto the cameras of a virtual stereo rig, and Gaussian noise with varying standard deviation (0.0 to 1.2 *pix*) is added to the image point locations. Two different methods are applied : (i) the method based on d-motion and (ii) the quaternion-based algorithm [4]. In order to compare the results, some errors are estimated between the expected motion and the estimated ones. The criterion errors are : the relation translation error, the relative rotation angle error and the angle between the expected and estimated rotation axis.

This process is iterated 500 times. The mean of each criterion error is shown Figures 2 and 3. Figure 2 shows the estimation of the relative angle and translation errors for small motions (rotation angle smaller than 0.1 rad.). Figure 3 shows the estimation of the relative angle and translation errors for ”standard” motions (rotation angle greater than 0.1rad.. Both figures show that the method gives accurate results even for high image noise (greater than 1.0 pix.).

5.2 Real data

Experiments with real data were conducted in order to justify the accuracy and applicability of the approach. An image sequence of 100 image pairs was gathered by a parallel camera stereo rig. In that sequence, a person is moving her head in many directions

The images were stereo-processed, and the disparity images were estimated.

The optical flow field [14] is estimated between two consecutive left images of the stereo pair in order to find point correspondences. The algorithm of d-motion estimation is applied using consecutive matched disparity reconstructions and the corresponding rigid motion is derived.

Figures 4, 5, 6 and 7 show the evolution of the estimated rotation angle and translation in the $xz-$plane. This shows that the estimated motion is consistent with the observed sequence.

6 Discussion

In this paper, we have described a method to estimate the rigid motion of an object observed by a parallel camera stereo rig. We studied the reconstructions in the disparity space and demonstrated its geometric and topological features. We introduced the rigid transformations associated with this space and show how they are related to the Euclidean rigid transformation. A motion estimation algorithm has been derived and its efficiency has been proved by comparing it with a standard algorithm using both synthetical and real data.

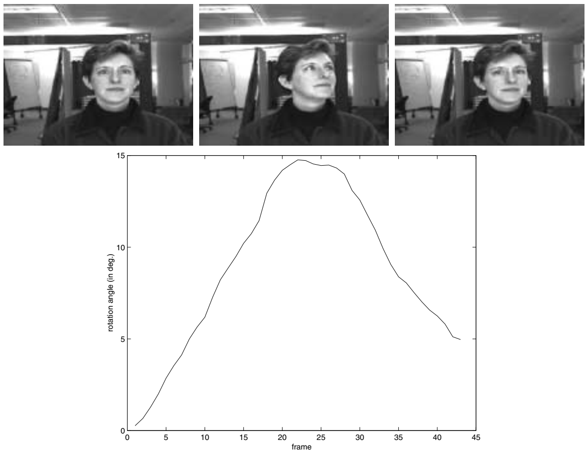

Figure 4: Estimation of the rotation angle *vs.* frame

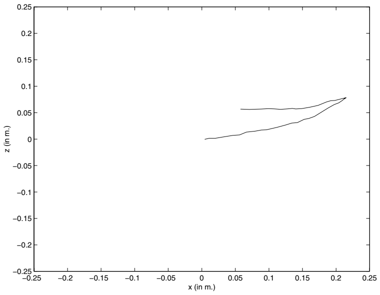

Figure 5: Trajectory of the face center in the $xz-$plane

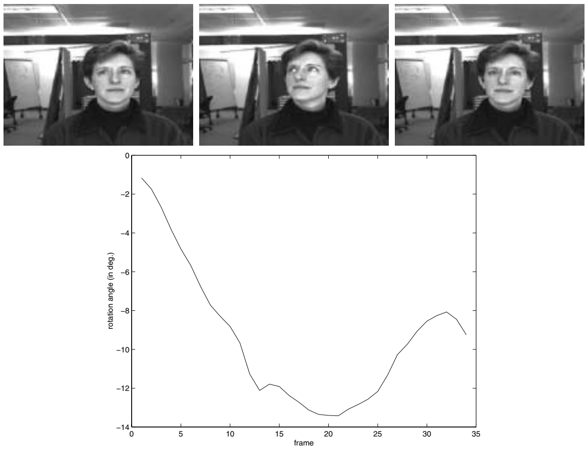

Figure 6: Estimation of the rotation angle *vs.* frame

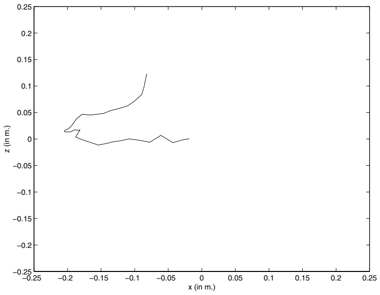

Figure 7: Trajectory of the face center in the $xz-$plane

There are many theoretical advantages of estimating the motion from the disparity space and d-motion. Minimizing E^2 gives an accurate estimation of $\mathbf{R}$ and $\boldsymbol{t}$ because, for parallel camera stereo rigs, the noise of points in the disparity space is isotropic (and nearly homogeneous when the reconstruction is restricted to well textured points). Therefore minimizing E^2 gives a (statistically) quasi-optimal estimation.

For non-parallel camera stereo rigs, the images have to be rectified and the noise is not exactly isotropic anymore (the effect depending on the vergence angle of the cameras). However, the noise in the disparity space is still far more isotropic than in the position space.

Our approach is straightforward and does not have the drawbacks of the traditional ”Euclidean” scheme which consists in first reconstruction 3-D data from disparity images and then estimating the motion from Euclidean data. Indeed the ”Euclidean” scheme implies that (i) image noise has to be propagated to 3-D reconstruction (actually approximated at the first order) and (ii) methods [12, 13] have to be designed to deal with the heteroscedastic nature of the noise (non-homogeneous, non-identical).

Finally, our approach could easily be extended in the uncalibrated case when the internal parameters of the stereo rig are unknown (see section 4.2). The minimization of E^2 should then be performed over 12 parameters instead of 6.

A topic of ongoing and future work is the use of multimodal noise distribution to model disparity uncertainties. Introducing that model in a stereo matching algorithm would give multiple-hypothesis disparity images, where each image pixel could have one or multiple disparities. A robust algorithm should be able to estimate the d-motion from two multiple-hypothesis disparity images corresponding to a rigid moving object.

References

[1] M. Pollefeys, R. Koch, and L. VanGool, “A Simple and Efficient Rectification Method for General Motion,“ *ICCV’99*, pp. 496-501, 1999.

[2] F. Devernay and O. Faugeras. From projective to Euclidean reconstruction. In *Proceedings Computer Vision and Pattern Recognition Conference*, pages 264–269, San Francisco, CA., June 1996.

[3] M. Irani and P. Anandan. A unified approach to moving object detection in 2d and 3d scenes. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 20(6):577–589, June 1998.

[4] B.K.P. Horn, H.M. Hilden, and S. Negahdaripour. Closed-form solution of absolute orientation using orthonormal matrices. *Journal of the Optical Society of America*, 5(7): 1127–1135, July 1988.

[5] J. Koenderink and A. van Doorn. Affine structure from motion. *Journal of the Optical Society of America A*, 8(2):377–385, 1991.

[6] W. MacLean, A.D. Jepson, and R.C. Frecker. Recovery of egomotion and segmentation of independent object motion using the em algorithm. In E. Hancock, editor, *Proceedings of the fifth British Machine Vision Conference, York, England*, pages 175–184. BMVA Press, 1994.

[7] T.Y. Tian and M. Shah. Recovering 3d motion of multiple objects using adaptative hough transform. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 19(10):1178–1183, October 1997.

[8] J. Shi and C. Tomasi. Good features to track. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Seattle, Washington, USA*, pages 593–600, 1994.

[9] P.H.S. Torr and D.W. Murray. Outlier detection and motion segmentation. In P.S. Schenker, editor, *Sensor Fusion VI*, pages 432–442, Boston, 1993. SPIE volume 2059.

[10] D. Brown The bundle adjustment - progress and prospect. In *XIII Congress of the ISPRS*, Helsinki, 1976.

[11] G. Csurka, D. Demirdjian, and R. Horaud. Finding the collineation between two projective reconstructions. *Computer Vision and Image Understanding*, 75(3): 260–268, September 1999.

[12] N. Ohta and K. Kanatani. Optimal estimation of three-dimensional rotation and reliability evaluation. In *Proceedings of the 5th European Conference on Computer Vision*, Freiburg, Germany, pages 175–187, 1998.

[13] B. Matei and P. Meer. Optimal Rigid Motion Estimation and Performance Evaluation with Bootstrap. In *Proceedings Computer Vision and Pattern Recognition Conference*, pages 339–345, San Francisco, CA., 1999.

[14] B.K.P. Horn and B.G. Schunck. Determining Optical Flow. *Artificial Intelligence*, 17:185–203, 1981.

[15] D. Demirdjian and R. Horaud. Motion-Egomotion Discrimination and Motion Segmentation from Image-pair Streams. *Computer Vision and Image Understanding*, volume 78, number 1, April 2000, pages 53–68.

0-7695-1143-0/01 $10.00 (C) 2001 IEEE