Local Motion Estimation from Stereo Image Sequences

N. Pérez de la Blanca[1], J. M. Fuertes[2], M. Lucena[2,] and A. Garrido[1]

[1] Department of Computer Science and Artificial Intelligence
ETSII, University of Granada, 18071 Granada, Spain
nicolas@ugr.es
agarrido@decsai.ugr.es
[2] Departmento de Informática, Escuela Politécnica Superior
Universidad de Jaén, Avenida de Madrid 35, 23071 Jaén, Spain
{jmf,mlucena}@ujaen.es

Abstract. This paper proposes a method for representing local temporal deformations of a 3D flexible surface in an orthogonal space from a sequence of stereo images. The approach uses a disparity space as the main space in order to represent all the 3D information. The local motions are estimated removing the rigid motions from the global motion in the disparity space. A robust algorithm based on the RANSAC approach is used to estimate the rigid motions through the image sequence. An incremental SVD algorithm is used to estimate the representation space of the local motions as data is received. The approach presented in this paper is valid for any type of camera.

1 Introduction

This paper proposes a technique for representing the instantaneous motions of a 3D deformable surface in a linear space from its projections in a sequence of stereo images. To date, many efforts have been made to study the problem of static surface reconstructions from features extracted from monocular or stereo images [10,12,13,15]. However, little attention has been paid to the case of deformable surfaces [1,6,11]. The main inconvenience when deformable surfaces are studied is the impossibility of predicting the location of their projection on the images in the absence of a geometrical restriction on the 3D motion of the points. In order to constrain and regularize the motion parameter estimation problem in this situation, different approaches have been proposed in the case of monocular image sequences.

The most successful approach to date is to fit a 3D template into the projected data and move this template according to the tracked motions of a set of landmark points on the image [2,6,9,14]. Although effective in many situations, the shortcoming of this approach is that it requires a 3D template of the surface. Another more general approach, based only on the coordinates of a set of image points tracked through the image sequence, defines an object model as a linear combination of the deformation axes. The estimation process of these axes and the shape coefficients associated to each instant only needed the knowledge of the point coordinates [3,5]. Interesting

F.J. Perales et al. (Eds.): IbPRIA 2003, LNCS 2652, pp. 740-747, 2003.
© Springer-Verlag Berlin Heidelberg 2003

results have been found in the case of parallel projection cameras, although the general analytical solution has been shown to be extremely complex [3]. To the best of our knowledge, the solution for general perspective cameras still remains open.

Since we are interested in studying the 3D deformations of objects near the camera, we use the general perspective camera model in order to analyze our images. An important instance of this situation appears in the 3D videoconferencing system, where the 3D shape of the head and face of each participant must be refreshed in each instant of time, and the usual short distance between cameras and surfaces introduces strong perspective effects. The problem of iteratively estimating an orthogonal linear space, which characterizes all deformations is also of great interest, since once we have learned this space, the new observations could be auto-coded by the coefficients of their projection in this space.

In this paper, the disparity space from stereo images is used to solve the instantaneous reconstruction problem from deformable surfaces, and at the same to estimate 3D rigid motions. In order to define a reconstruction linear space and to adapt it to each new observation, we propose that an incremental SVD algorithm be used which will allow us to adapt the base of the space when the new observations appear. In Section 2, we introduce the geometrical concepts of the disparity space. In Section 3, we study the rigid motion estimation in the disparity space. In Section 4, we approach the local deformation estimation by proposing the use of an incremental algorithm for adapting the base of the reconstruction space. In Section 5, simulation experiments carried out on synthetic data are shown. Finally, in Section 6, discussions and conclusions are presented.

2 Stereo Images

Let us consider a calibrated rectified stereo rig, *i.e.* the 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 [10]. We also assume that both cameras of the rectified stereo rig have known and similar internal parameters.

Stereo reconstruction has been studied for years, and is now a standard topic in computer vision. Let us 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 the epipolar line, the relation between the two points is

$$\begin{aligned} x' &= x - d \\ y' &= y \end{aligned} \tag{1}$$

where d is defined as the disparity of the point (x,y). From rectified stereo images, we can define representation spaces based on the projected coordinates that are equivalent to a 3D reconstruction of the points up to a homography of the 3D space [8]. These spaces are known as *disparity spaces*. The equations relating the 3D coordinates (X,Y,Z) with the disparity coordinates in the case of oriented and rectified cameras are [13]:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \frac{B}{\overline{x} - \overline{x}'} \begin{pmatrix} \overline{x} \\ \overline{y} \\ 1 \end{pmatrix} \qquad \overline{x} = \frac{x - x_0}{\alpha}, \quad \overline{y} = \frac{y - y_0}{\alpha}, \quad \overline{x}' = \frac{x' - x'_0}{\alpha'} \tag{2}$$

where x_0, y_0, x'_0 are the principal point coordinates of the left and right image, respectively, α and α' are the focal distance of the left and right cameras, respectively and B is the baseline of the stereo-rig. All image coordinates in are expressed in terms of pixels.

In this paper, we use the disparity space defined by the triple (x,y,d). From expression (2), taking $\alpha=\alpha'$ the homographic relationship between the 3D coordinates of a point $\mathbf{X}=(X,Y,Z)^T$ and its associated disparity vector $(\overline{x}, \overline{y}, d)^T$ can be expressed as,

$$Z\begin{bmatrix} \overline{x} \\ \overline{y} \\ d \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha & 0 & 0 & 0 \\ 0 & \alpha & 0 & 0 \\ 0 & 0 & 0 & \alpha\, B \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{3}$$

or in a shorter way as

$$\begin{pmatrix} \boldsymbol{\tau} \\ 1 \end{pmatrix} \cong \mathbf{H}_B \begin{pmatrix} \mathbf{X} \\ 1 \end{pmatrix} \qquad , \quad \boldsymbol{\tau} = (\overline{x}, \overline{y}, d)^T \tag{4}$$

From equation (3), it is clear that in the case of non-calibrated cameras each pair of rectified stereo images provides us with the reconstruction of the surface being imaged up to projectivity. From the intrinsic parameters of the stereo rig, the projective reconstruction can be upgraded to metric.

3 Rigid Motions in the Disparity Space

Let us apply a rigid motion on the 3D data. If $\mathbf{X}$ and $\mathbf{X}'$ represent the 3D coordinates of a point before and after the motion, we have

$$\begin{pmatrix} \mathbf{X}' \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R} & \mathbf{T} \\ \mathbf{0}^T & 1 \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ 1 \end{pmatrix} \tag{5}$$

From expressions (4) and (5) we obtain

$$\lambda \begin{pmatrix} \boldsymbol{\tau}' \\ 1 \end{pmatrix} = \mathbf{H}_B \begin{pmatrix} \mathbf{R} & \mathbf{T} \\ \mathbf{0}^T & 1 \end{pmatrix} \mathbf{H}_B^{-1} \begin{pmatrix} \boldsymbol{\tau} \\ 1 \end{pmatrix} = \boldsymbol{\Gamma} \begin{pmatrix} \boldsymbol{\tau} \\ 1 \end{pmatrix} \tag{6}$$

Equation (6) describes the 3D homography $\boldsymbol{\Gamma}$ relating the disparity homogeneous coordinates of a point before and after the motion.

1.1 Noise on the Data

An important feature of the disparity space is that the noise associated to the data vectors $(\bar{x}, \bar{y}, d)^T$ under some assumptions can be considered isotropic and homogeneous. The $\bar{x}, \bar{y}$ disparity coordinates are affected by the noise produced by the discretization effect and without additional information can be assumed equal for all pixels. The noise on *d* is associated to the change in the gray level of the pixels in the stereo matching process and could be estimated from this process. So, we can assume that the noises associated to $\bar{x}, \bar{y}$ and *d* are independent. If we assume that the variance of *d* is of the same magnitude as the variance of the discretization error the covariance matrix of the noise on each point of our disparity space is $\mathbf{\Omega}=\sigma^2 \mathbf{I}_{3x3}$.
In our case, apart from the above measurement errors, we also assume that in our scene there are points in motion. All the correspondences associated with these moving points are therefore potentially erroneous. In order to select point correspondences not affected by the moving points, we use the RANSAC algorithm to select subset of point correspondences that are free of this contamination.

1.2 Rigid Motion Estimation

Let $(\boldsymbol{\tau}_i, \boldsymbol{\tau}'_i)$ be a set of point correspondences. The problem of estimating the rigid motion parameters (**R,T**) from the set of points $(\boldsymbol{\tau}_i, \boldsymbol{\tau}'_i)$ amounts to minimizing the error

$$E^2 = \sum_i d(\boldsymbol{\tau}'_i, \boldsymbol{\Gamma}\boldsymbol{\tau}_i)^2 , \quad d(\boldsymbol{\tau}'_i, \boldsymbol{\Gamma}\boldsymbol{\tau}_i)^2 = (\boldsymbol{\tau}'_i - \boldsymbol{\tau}'^{\Gamma}_i)^T \mathbf{\Omega}^{-1} (\boldsymbol{\tau}'_i - \boldsymbol{\tau}'^{\Gamma}_i) \tag{7}$$

where $\boldsymbol{\tau}'^{\Gamma}_i = (\boldsymbol{\tau}'^{\Gamma}_{i1}/\boldsymbol{\tau}'^{\Gamma}_{i4} \quad \boldsymbol{\tau}'^{\Gamma}_{i2}/\boldsymbol{\tau}'^{\Gamma}_{i4} \quad \boldsymbol{\tau}'^{\Gamma}_{i3}/\boldsymbol{\tau}'^{\Gamma}_{i4})$ is the estimated Euclidean coordinate vector for $\boldsymbol{\tau}'_i$ from (6), and $\mathbf{\Omega}$ is the covariance matrix of the disparity vectors. Here we assume an i.i.d noise model. Equation (6) shows that this error function is not linear in the parameters for (**R,T**), so a non-linear method has been used to estimate the vector of six unknowns by parameterizing the rigid motion. Here we are interested in the case of small rotations (< 5 degree), so the rotation matrix can be expressed as $\mathbf{R}=\mathbf{I}+[\boldsymbol{\omega}]_x$, where I is the identity matrix and $[\boldsymbol{\omega}]_x$ represents the skew-symmetric matrix associated to the vector $\boldsymbol{\omega}$. In order to estimate the solution vector $(\boldsymbol{\omega},T)^T$ a quasi-linear iterative algorithm has been used on the normalized image coordinated [7]. An initial solution for the vector $(\boldsymbol{\omega},T)^T$ can be calculated from equation (6), solving the linear system that appear considering the equations associated to Euclidean coordinates of the all points τ and τ' and assuming all λ=1. In the next iteration we recalculate the value of λ from the above solution and solve again the equation (6) for a new solution. We iterate until the convergence of the vector $(\boldsymbol{\omega},T)^T$. In our experience three of our iterations are enough.

4 Local Deformations

Now we are interested in learning an orthogonal linear space for representing all local motions on the 3D surface. This space is estimated from the residual vectors of the points observed in the disparity space after the rigid motion component has been removed. This residual shall of course be contaminated with noise associated mainly to the matching point and rigid motion registration processes.

Let $\hat{\mathbf{R}}_i$ and $\hat{\mathbf{t}}_i$ denote the rotation and translation estimation from the (i-1)-th to the i-th images. Let $\mathbf{X}_{in}$ denote the 3D coordinates of the i-th scene point in the instant n>0. The 3D local deformation vector associated to this point is therefore calculated by

$$\hat{\mathbf{D}}_{in} = \hat{\mathbf{R}}_1^T \times \cdots \times \hat{\mathbf{R}}_n^T \left(\mathbf{X}_{in} - \hat{\mathbf{T}}_i\right) - \mathbf{X}_{i0} \ , \quad \hat{\mathbf{T}}_i = \hat{\mathbf{R}}_i \hat{\mathbf{T}}_{i-1} + \hat{\mathbf{t}}_i \ , \quad \hat{\mathbf{T}}_0 = \mathbf{0} \tag{8}$$

From (5) and (8), there is a homography $\mathbf{\Gamma}_n$ mapping the disparity vectors of the n-th image to those of the initial image plus the present local deformation

$$\begin{pmatrix} \tau_{in} \\ 1 \end{pmatrix} \cong \Gamma_n \begin{pmatrix} \tau_{i0} + \tau_{D_{in}} \\ 1 \end{pmatrix}, \qquad \Gamma_n = \mathbf{H}_B \begin{pmatrix} \hat{\mathbf{R}}_n \cdots \hat{\mathbf{R}}_1 & \hat{\mathbf{T}}_n \\ \mathbf{0}^T & 1 \end{pmatrix} \mathbf{H}_B^{-1} \tag{9}$$

Since the vectors $\boldsymbol{\tau}_{i0}$ are fixed and the matrix $\mathbf{\Gamma}_n$ can be calculated from the rigid motion estimate, we can estimate the vectors $\tau_{D_{in}}$ iteratively from (9) once the corresponding parameters ($\mathbf{R}_n$,$\mathbf{T}_n$) have been estimated. Assuming all 3D points are present in all images, for each stereo-image j we concatenate all its local deformations disparity vectors $\tau_{D_{ij}}$ in a single column vector τ_{D_j}. We also assume, without loss of generality, that the local deformations are mainly defined by cyclic patterns. Then the set of vectors $\boldsymbol{\tau}_\mathbf{D}$=$\{ \tau_{D_j}$, j=1,…,n$\}$ can be embedded, for n sufficiently large, in a orthogonal linear space using a Singular Value Decomposition (SVD) algorithm. In order to estimate this orthogonal space and update it iteratively, we applied an incremental SVD algorithm to the sequence of disparity vectors $\{ \tau_{D_j}$, j=1,…,n$\}$ [4]. If $\boldsymbol{\tau}_\mathbf{D}$ =UDV^T denote the SVD of $\boldsymbol{\tau}_\mathbf{D}$, the orthogonal matrix $\mathbf{U}$ is considered the basis for our local deformation space.

5 Simulation Experiments

Experiments with simulated data were carried out in order to compare the quality of the approach. A 3D scene defined by 243 points on two orthogonal planes, was used; see Figure 1(a). This scene was projected onto the stereo images, before and after a rigid motion of the scene, using a virtual stereo rig. We assume that the intrinsic parameters of the stereo rig (B=3, α'= α=549) are known. The distance between the scene and the stereo rig was four times the height of the scene. Gaussian noise with variable standard deviation (σ=0 to 1.4 pixel) was added to the image coordinates. In

our case, a fixed percentage, of scene points, is free to move its 3D location during the global rigid motion.

1.1 Rigid Motion Estimation

A set of point correspondences was used to estimate the rigid motion. In order to avoid point correspondences associated to moving points, a RANSAC algorithm was used from 50 random samples of two disparity vector correspondences, since two vectors provided us with six restrictions to estimate the six parameters. The accuracy of the estimate was measured in terms of the relative error of the quadratic norm of the residual. Figure 1(b) shows the average evolution of 500 simulations for increasing values of noise in the case of small motions (rotation angle < 5 degree and a maximum translation of 0.5 units on each axe).

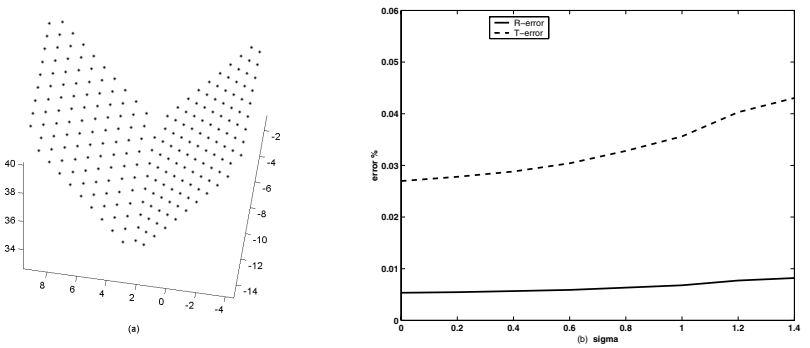

Fig. 1. a) 3D scene used in the simulation experiments, b) Estimation of the relative rotation and translation errors for small motions

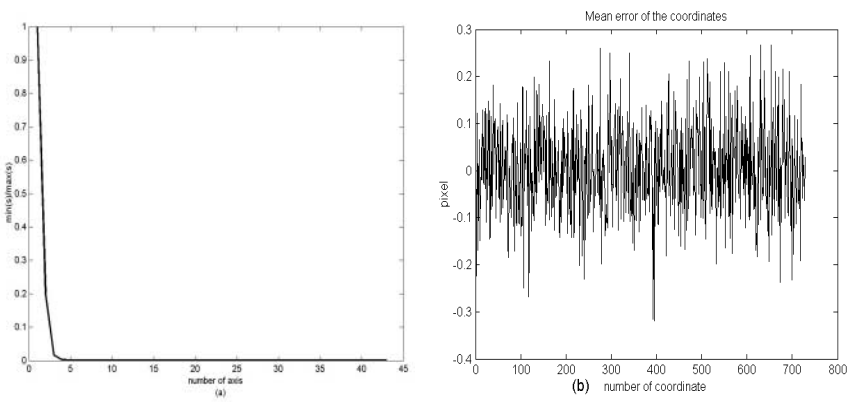

Fig. 2. a) Average curve of the learning curves from 100 sequences of 100 images with 40% of moving points in the scene, b) Average vector of the orthogonal components to the subspace spanned by **U** from 200 simulations with sigma=0.5

1.2 Linear Space Estimation

To analyze the evolution of the dimensionality of the space with the complexity of the local motions, we have applied random rigid motions to our scene and we have calculated the associated sequence of vectors $\boldsymbol{\tau}_{\mathbf{D}}$. The 3D moving points follow a periodic affine motion on each plane in our scene. The translation and scale parameters of the affine transformation were modulated with a periodic function along the sequence. As criteria for stopping the updating process we fixed a superior threshold ($< 10^{-4}$) on the ratio between the minimum and maximum singular values.

2 Discussion and Conclusions

Figure 1(b) shows how the rotation and translation estimate between two images is very accurate even for high level of noise. The algorithm we use is fast enough since no more than 3-4 linear iterations are needed for convergence. This result allows us to use this method to remove the rigid motion component from the disparity vector using equation (9). Of course, for large sequences the accumulated error could come to be very large. In order to avoid this situation, more than two images must be used in the rigid motion estimation process. In the learning process, the proposed stopping criterion behaves adequately since for low noise and global non-rigid motions the number of axes is as low as expected. However, when we have a very large number of local motions, the number of axes increases considerably. Figure 2(a) shows the average evolution of the number of axes calculated from 100 sequences of 100 images of our moving scene. Figure 2(b) shows the average graph of the vectors $(\mathbf{I}-\mathbf{U}\mathbf{U}^{T})\boldsymbol{\tau}_{\mathbf{D}}$, that is the components of $\boldsymbol{\tau}_{\mathbf{D}}$ orthogonal to the subspace spanned by $\mathbf{U}$. From the experimental simulations can be established that the estimated space $\mathbf{U}$ represents quite well the full sequence of vectors $\boldsymbol{\tau}_{\mathbf{D}}$ even with high level of noise on the observations.

This paper has presented a technique to estimate 3D local motions as vector in orthogonal space. The rigid motion fitting on the disparity space has been the key point, since in this space the errors can be considered isotropic and homogeneous. The use of an incremental SVD algorithm has allows us to update the representation space at the time that the new observations appear.

Acknowledgments

This work, has been financed by Grant IT-2001-3316 from the Spanish Ministry of Science and Technology

References

[1] Bascle, B., and Blake, A.: Separability of pose and expression in facial tracking and animation. In Proc. Int. Conf. Computer Vision. 1998.

[2] Basu, S., Oliver, N., Pentland A.: 3D modelling and tracking of human lip motions, ICCV'98, 1998.

[3] Brand, M. : 3D Morphable models from video. In Proc. CVPR'01. 2001.

[4] Brand, M.: Incremental singular value decomposition of uncertain data with missing value. In Proc. Europ. Conf. Computer Vision, ECCV-2002, LNCS-2350, 707-720.

[5] Bregler, C., Hertzmann, A., Biermann, H.: Recovering non-rigid 3D shape from image streams. In Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'00). 2000.

[6] Decarlo,D., and Metaxas, D.: Optical flow constraints on deformable models with application to face tracking. International Journal of Computer Vision, 38(2),99-127, 2000.

[7] Demirdjian, D., and Darell, T.,: Motion estimation from disparity images, In Proc. ICCV01, Vancouver Canada, 2001, vol-II, 628-635.

[8] Devernay,F. and Faugeras.O.: From projective to Euclidean reconstruction. In Proceedings Computer Vision and Pattern Recognition, 264-269,1996.

[9] Fua, P,: Regularized bundle-adjustment to models heads from image sequences without calibration data, International Journal of Computer Vision, 38(2), 2000.

[10] Hartley, R. Zisserman A.,: Multiple View geometry in computer vision. CUP, 2002.

[11] Lanitis, A., Taylor, C.J., Cootes, T.F. and Ahmed, T.: Automatic interpretation of human faces and hand gestures using flexible models. In Int. Workshop on Autom. Face-and-Gesture Recognition, 1995.

[12] Pollefeys, M., Van Gool, L., Zisserman, A., and Fitzgibbon, A: 3D Structure from images – SMILE 2000, Lecture Notes in Computer Science 2018, Springer, 2000.

[13] Tarel, J.P.,: Global 3D Planar Reconstruction with Uncalibrated Cameras a Rectified Stereo Geometry, Machine Graphics & Vision Journal, vol-6, 4, 1997, 393-418.

[14] Valente.S., and Dugelay, J.L.: A visual analysis/synthesis feedback loop for accurate face tracking, Signal Processing Image Communications, 16, 2001, 585-608.

[15] Zhang, Z., Faugeras, O.: 3D Dynamic Scene Analysis,: A stereo based approach. Springer series in Information Science, 27, Springer-Verlag, 1992.