3D Rigid Facial Motion Estimation from Disparity Maps

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

[1]Department of Computer Science and Artificial Intelligence
ETSII. University of Granada, 18071 Granada, Spain
nicolas@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 an approach to estimate 3D rigid facial motions through a stereo image sequence. The approach uses a disparity space as the main space in order to represent all the 3D information. A robust algorithm based on the RANSAC approach is used to estimate the rigid motions through the image sequence. The disparity map is shown to be a robust feature against local motions of the surface and is therefore a very good alternative to the traditional use of the set of interest points.

1 Introduction

To date, many efforts have been made to study the problem of camera motion from features extracted from monocular or stereo images [6,9,13]. The main approach estimates the motion by establishing correspondences between interest points on each image. There are two main shortcomings of such an approach: firstly, it requires the set of interest points on each image to lie on static 3D surfaces of the scene; and secondly, the surfaces of the scene must be textured enough to allow interest points to be estimated. When we approach the problem of estimating 3D rigid facial motions from images, we find that the problem of estimating the rigid motion of a 3D surface with many instantaneous local deformations is usually due to local facial motions [1,5,8]. Furthermore, it is well known that the surface of the face is not textured enough. Therefore, alternatives to the traditional use of the set of interest points must be considered. In this paper, a homography between disparity spaces is used to estimate 3D rigid motions. Dense disparity maps are used as a feature from which the homography parameters can be estimated.

Since we are interested in studying 3D object motions near the camera, we use the general perspective camera model in order to analyze our images. An important instance of this situation appears in 3D videoconferencing systems, 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 [12].

A. Sanfeliu and J. Ruiz-Shulcloper (Eds.): CIARP 2003, LNCS 2905, pp. 54–61, 2003.
© Springer-Verlag Berlin Heidelberg 2003

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, disparity map estimation is discussed. In Section 5, experiments carried out on image 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. There is no loss of generality since it is possible to rectify the images of a stereo rig once the epipolar geometry is known [6]. We also assume that both cameras of the rectified stereo rig have internal parameters which are similar and known.

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 [4]. 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]:

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

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 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 $(\bar{x}, \bar{y}, d)^T$ can be expressed as

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

or in a shorter way as

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

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, then

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

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 Γ relating the disparity homogeneous coordinates of a point before and after the motion.

3.1 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. We can therefore 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 $\Omega = \sigma^2 \mathbf{I}_{3x3}$. 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 which are unaffected by the moving points, we use the RANSAC algorithm to select the subset of point correspondences that are free of this contamination.

3.2 Rigid Motion Estimation

Let (τ_i, τ'_i) be a set of point correspondences. The problem of estimating the rigid motion parameters (**R,T**) from the set of points (τ_i, τ'_i) amounts to minimizing the error

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

where $\tau'^{\Gamma}_i = \left(\tau'^{\Gamma}_{i1}/\tau'^{\Gamma}_{i4} \quad \tau'^{\Gamma}_{i2}/\tau'^{\Gamma}_{i4} \quad \tau'^{\Gamma}_{i3}/\tau'^{\Gamma}_{i4}\right)$ is the estimated Euclidean coordinate vector for τ'_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 $\mathbf{R}=\mathbf{I}+[\omega]_X$, where I is the identity matrix and $[\omega]_X$ represents the skew-symmetric matrix associated to the vector ω. In order to estimate the solution vector $(\omega,T)^T$ a quasi-linear iterative algorithm has been used on the normalized image coordinated [3]. An initial solution for the vector $(\omega,T)^T$ can be calculated from equation (6), solving the linear system that appears by considering the equations associated to Euclidean coordinates of all the points τ and τ' and assuming all λ=1. In the next iteration we recalculate the value of λ from the above solution and again solve equation (6) for a new solution. We iterate until convergence of the vector $(\omega,T)^T$. In our experience, three or four iterations are enough.

Nevertheless, the presence of outliers in the correspondences between the disparity maps degrades the estimation considerably. In order to circumvent this problem a RANSAC based algorithm is proposed in Table 1. This algorithm makes a robust iterative linear estimation as a first approach, but because of the noise in the disparity estimation, a non-linear optimization step from the pixel color values is necessary.

4 Disparity Map Estimation

In this paper two different dense disparity maps are used. Firstly, we estimate the disparity map for each stereo image, and from this we estimate a region of interest by applying a binary thresholding operator on it. Secondly, we estimate the dense motion vector map associated to every two consecutive left and right images, respectively. In this case we assume that the region of interest is the region of moving pixels nearest the camera.

Table 1.

```
Iterative robust algorithmI. To estimate and normalize the set of
disparity vectors
II. Repeat N iterations
     To choose n>=2 disparity vectors randomly
     i. For each vector calculate λi, Ai and biii. Solve λAX=b for
   X
     iii. Count the number of inliers.III. To take the solution
with higher number of inliers as the best linear solution.
IV. To minimize the pixel color differences between images by
applying the Levenberg-Mardquart algorithm from the linear
solution.
```

Figure 1 shows how we estimate our region of interest on each stereo image. In short, we segment the subset of moving points of the scene to a distance of the camera, which is less than a fixed threshold. In our case, the planar motion is calculated in pixel units. In order to remove isolated small regions we apply a size filter. All the pictures shown correspond to the left image of the stereo pair.

(a) (b) (c) (d)

Fig. 1. This example corresponds to rotation left-right of the head. Picture (a) represents the estimated stereo disparity map, picture (b) represents the x-motion dense map, picture (c) represents the y-motion dense map, and picture (d) represents the result of the union of picture (b) and picture (c) intersection with picture (a).

Dense disparity maps from two images is a very active field of research [10]. Very recently, new energy minimization algorithms based on cut graphs was proposed [2][7]. These algorithms achieved a very good compromise between temporal efficiency and accuracy of the estimation [7]. Since the implementation of these algorithms only depends on a free parameter, $\lambda>0$, associated to the scale of the estimation [10], very different estimations can be achieved by varying the λ value. Low values of λ provide us with more accurate estimations but a larger number of points will be undefined. A scale combination scheme therefore provides us with a better estimation. In our case, four different scales (λ=3,5,10,30) have been considered in order to estimate the disparity maps. The combination scheme defined the disparity value on each pixel as the value of the lowest scale in which the disparity is defined. For motion estimation only the lowest scale has been used, since the other

scales do not contribute much information. In order to obtain as accurate a segmentation as possible, there has been some loss in computational efficiency.

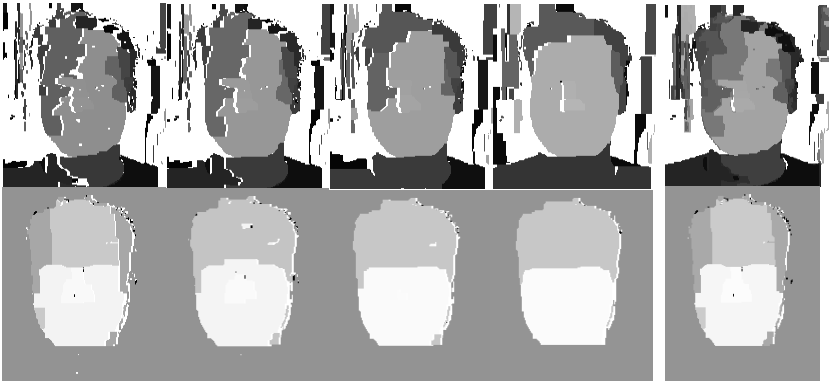

Fig. 2. The first four columns of each row show the stereo disparity map from a stereo image, and the x-motion estimation from two consecutive stereo images, respectively, for different λ values. The last column shows the resulting estimation from combining the different scales. All these images correspond to the left image of the stereo pair.

The first row of Figure 2 shows the stereo disparity map estimation from a stereo image for different values of λ joined to the final estimation obtained by combining the different scales. The second row shows the x-motion map estimation from two consecutive stereo images. It is possible to appreciate how the use of multiple scales does not greatly improve the first scale estimation in the case of motion estimation. However, the combination of different scales proves to be very useful when the stereo disparity map is estimated.

5 Experimental Results

Experiments to estimate 3D rigid facial motion have been carried out from different stereo image sequences captured by a Pointgrey stereo camera (Bumblebee) watching an actor moving his face freely. A fixed window inside the captured images fixed the sub-images of interest. Our algorithm was applied to the image sequence defined by the sub-images. The proposed algorithm was applied on every two consecutive stereo images in the sequence. In order to assess the goodness of the estimation process we synthesized a new sequence of images by interpolating from the estimated motions and the original sequence.

Figure 3 shows six sampled images to a distance of ten samples, each, of a stereo sequence of our examples. It can be seen how the strength and unpredictability of local facial motions makes it difficult to use interest points in the estimation process. Figure 4 shows how accurate the estimated motion for a particular sequence is. We

compare the norm of the difference between two consecutive images, with the norm of the residual calculated by the difference between an original image and its corresponding synthetic. The large decrease of the norm of the difference image from the first case to the second case, shows that the estimated motion is right and precise enough. We should point out that it is difficult to visualize the accuracy of the parameter estimation from this type of graph, but we prefer this type because it is much more difficult to appreciate small residual motions by comparing eye static pictures.

Fig. 3. These pictures show local motions present in a standard stereo sequence.

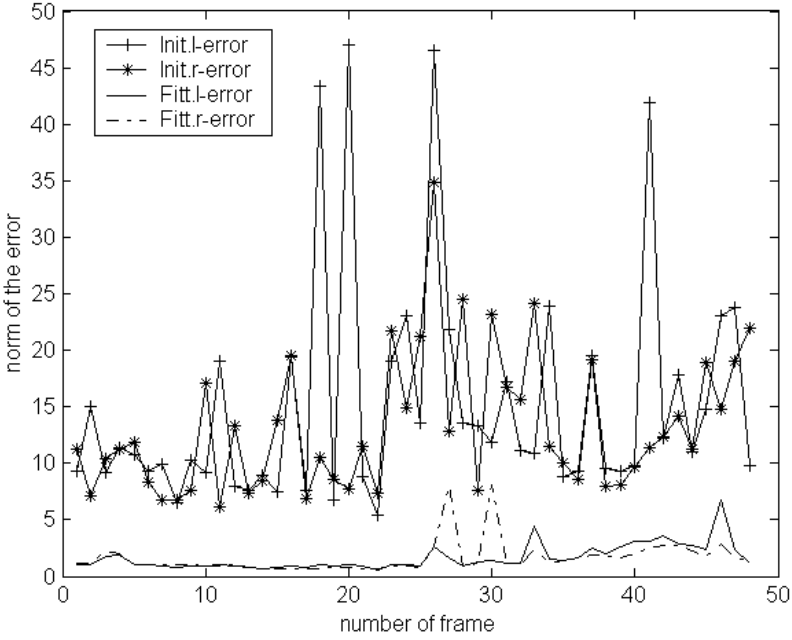

Fig. 4. This figure shows four graphs each of which is the norm of the gray level difference pixel-by-pixel from two images. The graphs Init.l-error and Init.r-error represent the case in which the images are two original consecutive right images and left images, respectively, of the sequence. The graphs Fitt_l.error and Fitt_r.error represent the case in which the two images are the original and synthesized one, using the proposed algorithm, for the right and left images, respectively.

6 Discussion and Conclusions

In this paper a new approach to estimate the 3D rigid motion of a deformable surface is proposed. The algorithm we propose is accurate and fast enough since no more than 3-4 linear iterations plus 2 non-linear iterations are needed for convergence. The use of stereo images allows us to estimate the motion without the need for external information. This result will allow us to use this approach to remove the rigid motion component from the disparity vector to estimate local deformations. Of course, in this latter case and for large image sequences, the accumulated error might get very large. In order to avoid this situation, the present accuracy of the estimated motion based on two images must be improved. An alternative in order to improve estimation would be the joint use of all images in a bundle algorithm, but this approach is inapplicable in time efficiency demanding applications

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. Boykov,Y.,Veksler,O.,and Zabih,R.,: Fast approximate energy minimization via grpah cuts, IEEE Trans. PAMI vol-23,11,1222–1239,2001.
3. Demirdjian, D., and Darell, T.,: Motion estimation from disparity images, In Proc. ICCV01, Vancouver Canada,,2001, vol-II, 628–635.
4. Devernay, F. and Faugeras, O.: From projective to Euclidean reconstruction. In Proceedings Computer Vision and Pattern Recognition,264–269,1996.
5. Fua, P.: Regularized bundle-adjustment to models heads from image sequences without calibration data, International Journal of Computer Vision, 38(2), 2000.
6. Hartley, R. and Zisserman, A.,: Multiple View geometry in computer vision. CUP, 2002
7. Kolmogorov,V., and Zabih, R.,: Visual correspondences with occlusions using graph cuts, In ECCV'02, Lecture Notes in Computer Science 2352, 82–96,2002
8. Lanitis, A., Taylor, C.J., Cootes, T.F. and Ahmed, T.: Automatic interpretation of human faces and hand gestures using flexible models. In International Workshop on Automatic Face-and-Gesture Recognition, 1995.
9. Pollefeys, M., Van Gool, L., Zisserman, A., and Fitzgibbon, A: 3D Structure from images – SMILE 2000, Lecture Notes in Computer Science 2018, Springer, 2000.
10. Scharstein,D., and Szeliski,R.,: A Taxonomy and evaluation of dense two-frame stereo correspondence algorithms, IJCV, 47(1):7–42, 2002..
11. Tarel, J.P.: Global 3D Planar Reconstruction with Uncalibrated Cameras, A Rectified Stereo Geometry, Machine Graphics & Vision Journal, vol-6, 4, 1997, 393–418.
12. Valente.S. and Dugelay, J.L.: A visual analysis/synthesis feedback loop for accurate face tracking, Signal Processing Image Communications, 16, 2001, 585–608.
13. Zhang,Z., Faugeras,O.: 3D Dynamic Scene Analysis,: A stereo based approach. Springer series in Information Science, 27, Springer-Verlag, 1992.