A COMPARATIVE STUDY ON 3D MOTION ESTIMATION UNDER ORTHOGRAPHY

Yiannis Xirouhakis and Anastasios Delopoulos

Image, Video and Multimedia Systems Lab., Dept. of Electrical and Computer Eng.,
National Technical University of Athens, GREECE, jxiro@image.ntua.gr

ABSTRACT

In the present work, the algorithm proposed in [8,10] is tested against existing approaches on 3D motion and structure estimation of rigid objects under orthography. The theoretical relation between the proposed approach and the well-known factorization and epipolar methods is discussed. At the same time, comparative simulated experiments are given, illustrating the performance of the three algorithms (the factorization, the epipolar and the proposed one).

The proposed algorithm seems to be more generic than the existing approaches, and provides superior estimates of 3D motion in most cases.

1. INTRODUCTION

Much work has been done recently for determining three dimensional motion and structure of moving rigid objects viewed at different time points and/or by multiple cameras. The extraction of motion and shape parameters of a moving rigid 3D object from a 2D image sequence (often named as the Structure From Motion problem) has been tackled by several authors. Various approaches have been proposed, which differ in the projection model assumed, the feature correspondences and the input measurements employed, and the adopted data-processing technique [7].

As far as the 2D features are concerned, line, curve and point correspondences have been utilized, with the latter being the most popular. Two well-known projection models mainly considered in the literature are the perspective and the orthographic, with the latter assumed when the object is far away from the camera. More precisely, orthographic approximation yields acceptable experimental results when the range in depth values of the 3D points in the original 3D scene is smaller than 10 percent of the average depth value.

Ullman in his classical work [0] proved that four point correspondences over three frames are sufficient to yield a unique solution to motion and structure up to a reflection. In this direction, Huang and Lee in [1] proposed a linear algorithm to obtain the 3D motion and structure parameters of a rigid object, introducing implicitly the epipolar equation. Based on the epipolar equation, a number of relative approaches, deemed as epipolar methods, have been presented in the literature, including [3,4,9] among others. Shapiro *et al* [3] rely on the affine epipolar lines' properties and solve the affine epipolar line equation. A next step determines all unknown camera motion parameters. In the same manner, Xu and Sugimoto [9] solve the epipolar equation and determine the three rotation angles (Eulerean angles) in a second step. Ostuni and Dunn [4] utilize the epipolar equation as well, along with a different parametrization for the rotation matrix.

A somehow different approach under orthography was presented in [2], which is widely known as the factorization method. Tomasi and Kanade's solution in [2] is based on a camera-centered problem representation, which may incorporate an arbitrary number of point correspondences and frame transitions to achieve robustness in the presence of noise. The solution relies on decomposing the matrix containing all measurements into camera motion and object shape. In the same context, the factorization method has been extended to include paraperspective projection [6] and sequential processing over a sequence of images [5].

A later approach, proposed by Xirouhakis and Delopoulos in [8], relies on the eigendecomposition of a matrix formed on the basis of 2x2 matrices, modeling in turn the projected motion of planar patches. As indicated in [10], appropriate choice of the planar patches (equivalently point triplets) greatly enhances the performance of the algorithm.

In this work, the theoretical relation between the three non-approximate methods, i.e. the proposed approach [8,10], the factorization [2] and the epipolar methods [3,9] is discussed. At the same time, comparative simulated experiments are given, illustrating the performance of the three algorithms.

2. BACKGROUND

The 3D motion and shape estimation problem under orthography can be posed in the following manner: assuming that three views/projections of a rigid 3D object are available, containing at least four 2D points (x_i, y_i) that their correspondence between frames is known $(x_i, y_i) \rightarrow (x_i', y_i') \rightarrow (x_i'', y_i'')$, compute the motion parameters for the two transitions, as well as the depth z_i of all points. For rigid motion, the motion parameters include the 3x3 rotation matrix **R** and the 3x1 translation vector **T**, so that

$$\begin{bmatrix} x_i' \\ y_i' \\ z_i' \end{bmatrix} = \mathbf{R} \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} + \mathbf{T}, \qquad (1)$$

for transition $(x_i, y_i) \rightarrow (x_i', y_i')$. A similar movement equation is written for transition $(x_i, y_i) \rightarrow (x_i'', y_i'')$ for

distinct **R** and **T**. Due to the nature of the orthographic projection, (x_i, y_i, z_i) projects onto (x_i, y_i).
Supposing **R**=[r_{ij}], the epipolar equation for a transition is written implicitly in [1] for **T**=0, as

$$r_{23}x'_i - r_{13}y'_i + r_{32}x_i + r_{31}y_i = 0\,. \qquad (2)$$

The affine epipolar constraint equation

$$ax'_i - by'_i + cx_i + dy_i + e = 0\,, \qquad (3)$$

(see [3] for details, p.154) constitutes an extension of (2) to the weak perspective case, which in turn differs from the orthographic one, only in the sense that it permits a scale change between different views $(x_i, y_i) \rightarrow (f\,x'_i,\, f\,y'_i)$. Epipolar methods, in fact, solve the homogeneous equation resulting by subtraction of eq. (3) for the centroid of the point set from the respective eq. (3) for any available point correspondence; the latter is achieved through the minimization of a cost function involving the 'scatter matrix' [3]. The motion parameters, as well as point depths, are estimated in a second step [4,9].

A somehow different approach is given in the factorization method [2]. A respective 'measurement matrix' is formed (see [2], p.138), which in turn is 'registered' by subtraction of the centroid of the point set. However, in this case no intermediate quantities (such as the 'epipolar geometry') are estimated. The 'registered measurement matrix' $\tilde{\mathbf{W}}$ is proved to decompose, in terms of singular-value-decomposition into

$$\tilde{\mathbf{W}} = \boldsymbol{R} \cdot \boldsymbol{S}\,, \qquad (4)$$

where the rows of $\boldsymbol{R}$ represent the orientations of the camera reference axes throughout the stream and the columns of $\boldsymbol{S}$ the 3D coordinates of the employed points.

In [8], matrix **K** models the projected motion of a 3D plane (defined by a point triplet) for a transition between two frames. The theoretical definition of **K** for a triplet of point correspondences is

$$\mathbf{K} = \mathbf{R}_{2x3} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ p & q \end{bmatrix}, \qquad (5)$$

for $p = -a_x / a_z$, $q = -a_y / a_z$ where $\mathbf{a} = [a_x \;\; a_y \;\; a_z]^{\mathrm{T}}$ represents the unit vector perpendicular to the plane, and scalars *p*, *q* contain the plane orientation information. $\mathbf{R}_{2x3}$ is the 2x3 matrix that contains the first two rows of the rotation matrix **R**. As the available point set contains more than 3 point correspondences, a 2x2 symmetric matrix **Y** is obtained on the basis of all **K**s (see [8,11] for details). From the latter, 3D motion parameters and 3D shape are obtained in a second step, just after the eigen-decomposition of **Y**.

3. DISCUSSION

Since the proposed, the factorization and the epipolar methods provide solutions for the same problem, they are expected to have some mathematical relation. As already stated, the type of the projection is not considerable problem since all methods account for the orthographic case (the epipolar for *f*=1); or even for the weak perspective case by considering [6] for the factorization method. It should be pointed out here that, for the proposed algorithm in the weak perspective case, the unknown scale factors for both transitions disappear in the final expressions yielding the rotation parameters. In fact, it can be shown that the proposed method is, in some sense, related to both others, having at the same time one more convenient property for robust 3D motion estimation in the presence of noise.

In fact, matrix **K** in (5) incorporates the epipolar equation for the particular triplet of point correspondences. As a immediate result of (5),

$$\mathrm{adj}(\mathbf{K}) \begin{bmatrix} r_{13} \\ r_{23} \end{bmatrix} + \begin{bmatrix} r_{31} \\ r_{32} \end{bmatrix} = 0\,, \qquad (6)$$

which in turn is identical to two (of the three) differential epipolar equations that can be formed from three point correspondences. The latter can be verified by the estimation procedure of **K**, given three point correspondences [8]. In this context, matrix **K** not only contains the epipolar equation, but also is formed in terms of differential epipolar equations, similarly to the 'scatter matrix' of [3]. However, the point centroid is not involved in every subtraction; on the contrary, points are divided into point triplets in order to estimate **K**.

Regarding the proposed approach, in relation to the factorization method, it can be seen that similar expressions hold for the estimation of 3D motion. In [2], matrix $\boldsymbol{R}$ of eq. (4) is defined as containing the horizontal and vertical camera reference axes throughout the stream. In addition, the camera reference axes $\mathbf{i}_0, \mathbf{j}_0$ are initialized (see [2], p.139) to be aligned with the world reference (e.g. $\mathbf{i}_0 = [1\;0\;0]^{\mathrm{T}}$ and $\mathbf{j}_0 = [0\;1\;0]^{\mathrm{T}}$). On the other hand, when the object rotates w.r.t. a 3x3 rotation matrix **R**, the camera equivalently rotates w.r.t. $\mathbf{R}^{\mathrm{T}}$. In this sense, matrix $\boldsymbol{R}$ incorporates the first two rows of the rotation matrix for each transition. The latter is closely related to matrix **K**, as introduced in equation (5) of the present work; i.e. matrix **K** can be 'factorized' into $\mathbf{R}_{2x3}$ (containing the first two rows of the rotation matrix) and a matrix containing shape parameters (*p*, *q*). As in the factorization method, a block matrix could be similarly derived by enlarging **K**, that is by adding **K**s column-wise for more triplets and row-wise for more frames.

With respect to the above, the proposed method is closely mathematically related to both other ap-

proaches. The strategy followed in the factorization method appears convenient for the incorporation of more than three frames in the estimation of 3D structure, leading though to a too large 'measurement matrix'. On the other hand, the convenient formulation of appropriate point triplets is in this way absent. Similarly, the epipolar methods by implicitly considering the difference between the epipolar equation of each point and the centroid, also lack this property.

In general, the three methods (the factorization, the epipolar and the proposed one) lead to the solution of an over-determined homogeneous system for some of the rotation parameters as a first step. The following steps depend on the formulation of the homogeneous system. In addition, all methods take advantage of the properties of eigen-decomposition or singular-value-decomposition to eliminate the variance of i.i.d. noise in motion vector estimates. The proposed method is superior in this sense, as it manages to increase SNR in differential motion field by appropriate formulation of the point triplets [10].

This is because additive noise terms in matrix quantities are in the form of $s_j^2\mathbf{I}$ where $\mathbf{I}$ is the 2x2 identity matrix and scalar s_j^2 equals

$$s_j^2 = \frac{\left\|\mathbf{r}_3^j - \mathbf{r}_1^j\right\|^2 + \left\|\mathbf{r}_2^j - \mathbf{r}_1^j\right\|^2 + \left\|\mathbf{r}_3^j - \mathbf{r}_2^j\right\|^2}{\left|\det\left[\mathbf{r}_3^j - \mathbf{r}_1^j \;\; \mathbf{r}_2^j - \mathbf{r}_1^j\right]\right|^2}, \quad (6)$$

for the j-th triplet of 2D points $\mathbf{r}_1^j$, $\mathbf{r}_2^j$ and $\mathbf{r}_3^j$ (see [10]). In fact, the nominator of this ratio equals the sum of the squared lengths of the triangle formed by the three points, while the denominator is the squared area of the corresponding rectangle. In this context, minimization of quantities $s_j^2\mathbf{I}$, and thus noise terms, relies on appropriate choice of the employed triplets so that the ratio in (6) is minimized.

4. SIMULATION RESULTS

A number of simulated experiments were carried out in error-prone environments in order to test the proposed algorithm's performance against other existing methods. For this purpose, the factorization, the epipolar and the proposed method were implemented along the lines of [2], [9,3] and [8,11] respectively.

In Figure 1, a computer generated model of a 3D smooth surface is depicted. The model was subjected to rigid movements in the 3D space and then projected, in order to derive a number of noise-free motion fields. The latter were artificially disturbed by i.i.d. noise. In Figure 2, the (noisy) reconstructed, by the proposed method, surface is depicted for SNR -10dB in the differential motion fields.

The proposed approach seemed to be superior in nearly all simulated experiments held. In Figures 3, 4 and 5, estimates of the rotation angle for varying SNR level for a transition of the 3D surface are depicted. Mean (solid line) and standard deviation (dash-dotted line) estimates of the rotation angle using the factorization, the epipolar and the proposed method respectively, are illustrated. The estimates were obtained through 50 Monte Carlo runs for each SNR level. In each run, the same set of noise-contaminated point correspondences was fed to all three algorithms. The true value for the particular rotation angle was $f_R = 17^o$. The proposed approach performed better than both the factorization and the epipolar method, in particular in the presence of noise resulting to low SNRs. It can be though pointed out that both the factorization and the epipolar method illustrate smaller standard deviation estimates compared to the proposed method for higher SNR levels. This is possibly due to the strategy adopted in the formulation of point triplets.

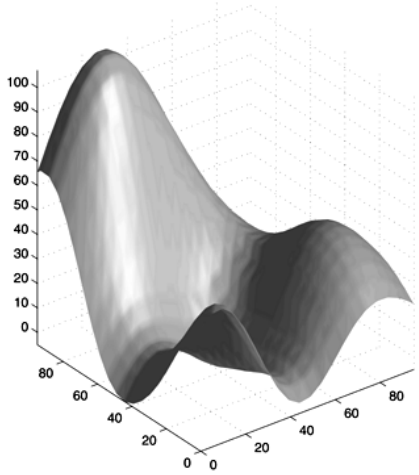

Figure 1. Original smooth 3D surface

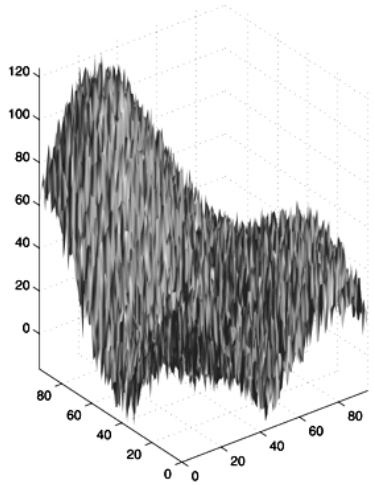

Figure 2. Reconstructed (noisy) 3D surface

5. CONCLUSIONS

In this work, the method proposed in [8,10] for 3D motion and shape estimation of rigid objects from orthographic projections is revisited and tested against existing approaches.

The proposed algorithm seems to be more efficient, compared to others, providing improved estimates of 3D motion in most cases. Superior estimates are obtained in the presence of particular noise for very low SNRs, where the rest algorithms fail.

Although the proposed algorithm performs generally better than the others, it has been shown that there is a strong mathematical relation with each one of them. Nevertheless, both the factorization and the epipolar methods lack the comfortable property of minimizing the effect of the noise terms, by appropriately selecting the point triplets fed to the algorithm.

6. REFERENCES

[0] S. Ullman, *The Interpretation of Visual Motion*. Cambridge, MA, MIT Press, 1979.

[1] T. S. Huang and C. H. Lee, "Motion and structure from orthographic projections," *IEEE Trans. PAMI*, vol.11, pp.536-540, May 1989.

[2] C. Tomasi and T. Kanade, "Shape and Motion from Image Streams under Orthography: a Factorization Method," *Int'l J. of Computer Vision*, vol. 9, no. 2, pp. 137-154, 1992.

[3] L. S. Shapiro, A. Zisserman and M. Brady, "3D Motion Recovery via Affine Epipolar Geometry," *Int'l J. of Computer Vision*, vol. 16, pp. 147-182, 1995.

[4] J. Ostuni and S. Dunn, "Motion from Three Weak Perspective Images Using Image Rotation," *IEEE Trans. PAMI*, vol. 18, no. 1, pp. 64-69, Jan. 1996.

[5] T. Morita and T. Kanade, "A Sequential Factorization Method for Recovering Shape and Motion from Image Streams," *IEEE Trans. PAMI*, vol. 19, no. 8, pp. 858-867, Aug. 1997.

[6] C. J. Poelman and T. Kanade, "A Paraperspective Factorization Method for Shape and Motion Recovery," *IEEE Trans. PAMI*, vol. 19, no. 3, pp. 206-218, March 1997.

[7] S. Soatto and P. Perona, "Reducing 'Structure From Motion': A General Framework for Dynamic Vision, Part 2: Implementation and Experimental Assessment," *IEEE Trans. PAMI*, vol. 20, no. 9, pp. 943-960, September 1998.

[8] A. Delopoulos and Y. Xirouhakis, "Robust Estimation of Motion and Shape based on Orthographic Projections of Rigid Objects," *IEEE Image and Multidimensional Digital Signal Processing Workshop '98*, Alpbach, Austria, July 1998.

[9] G. Xu and N. Sugimoto, "A Linear Algorithm for Motion from Three Weak Perspective Images Using Euler Angles," *IEEE Trans. PAMI*, vol. 21, no. 1, pp. 54-57, Jan. 1999.

[10] Y. Xirouhakis, G. Tsechpenakis and A. Delopoulos, "User Choices for Efficient 3D Motion and Shape Extraction from Orthographic Projections," *IEEE Int'l Conf. on Electronics, Circuits and Systems*, Paphos, Cyprus, Sept. 1999.

[11] Y. Xirouhakis and A. Delopoulos, "Least Squares Estimation of 3D Shape and Motion of Rigid Objects from their Orthographic Projections," to appear in *IEEE Trans. PAMI*.

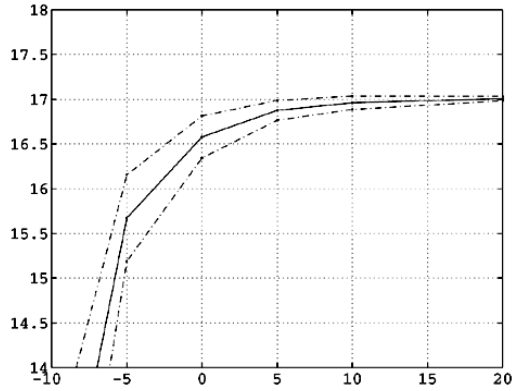

Figure 3. Estimated angle (factorization method)

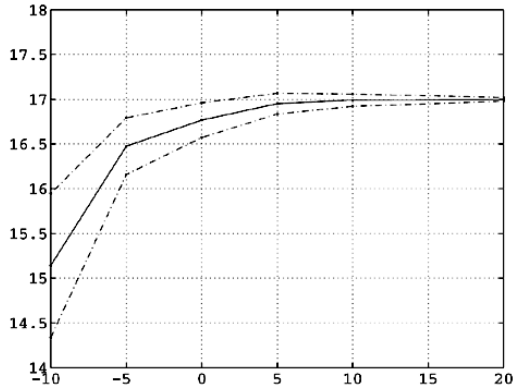

Figure 4. Estimated angle (epipolar method)

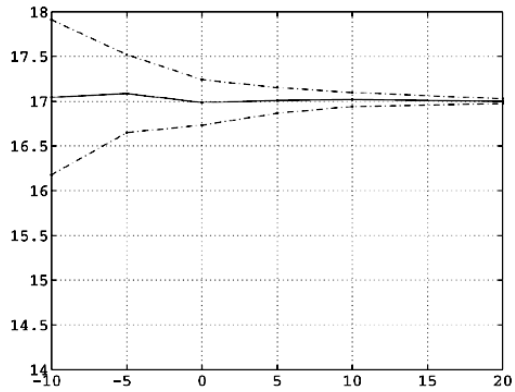

Figure 5. Estimated angle (proposed method)