Recursive 3D Reconstruction under Orthography using Kalman Filtering

Alexia Briassouli, Yiannis Xirouhakis and Anastasios Delopoulos

Computer Science Div., Dept. of Electrical and Computer Eng.,
National Technical University of Athens,
9 Iroon Polytechniou str., Athens GR-15773, GREECE
Email:jxiro@image.ntua.gr

ABSTRACT

In modern technological applications, machines that operate in three-dimensional environments have become extremely popular. Sophisticated systems, with the ability to extract the structure and motion of objects on the basis of certain $3D$ and/or $2D$ features, have resulted to a substantially new set of tasks; including, among others, $3D$ reconstruction and modeling.

The recovery of $3D$ motion and structure, or the Structure From Motion problem (SFM), has been tackled by several authors, however the obtained results are reported to suffer from noise. In this work, we investigate the improvement of depth estimates on the basis of multiple frames using a Kalman filter under orthography. Simulation results exhibit the efficiency of the proposed approach.

1. INTRODUCTION

The Structure From Motion problem emerges in several modern applications, as well as research fields, such as $3D$ modeling, video coding and compression. The problem has been tackled by several authors on the basis of different $2D$ features (input measurements), including lines, curves or points, with the latter being the most popular.

In both the case of orthographic and perspective projection, exact theoretical solutions to the SFM problem have been proposed for example in [1] and [2] respectively. Solutions in the presence of noisy point correspondences have been proposed as well, for both cases [3, 4], yielding relatively accurate $3D$ motion estimates. However, due to the inevitable noise in motion vectors, $3D$ structure estimates are of rather low quality. For the improvement of $3D$ structure, a common approach is the utilization of as many as possible frames from the available sequence. For the orthographic case, a solution is proposed in [5], based on the singular value decomposition of a large matrix, containing all employed point correspondences over the employed frames.

As reported in [6], the Kalman filter and the extended Kalman filter (EKF) theory has been employed for the estimation of both motion and structure of rigid objects from multiple frames. Such a solution can be found in [7], where an iterative extended Kalman filter (IEKF) is introduced for the perspective case. In this work, an appropriate Kalman filter is constructed for the estimation of $3D$ structure under orthography given the $3D$ motion parameters extracted by the algorithm presented in [3]. For improved estimation of motion parameters, the guidelines presented in [8] were adopted.

The employed Kalman filter is proved to yield particular improved results for increasing number of available frames (and $2D$ motion estimates). Simulation results will be included to verify the accuracy of the obtained estimates.

2. BACKGROUND

2.1. 3D motion model

The extraction of $3D$ motion under orthography is performed using the algorithm presented in [8], which is proved to yield relatively accurate estimates of the rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{T}$ for each transition between frames. In this sense, we will hereon assume that $\mathbf{R}$ is known and $\mathbf{T}$ is also available within a noise component.

Yet, the three dimensional scene's structure can't be recovered within a satisfactory degree of accuracy, despite knowledge of rotation parameters. In fact, the latter is rather expected, since structure depends not only on rotation parameters, but also on the available noisy $2D$ point correspondences, as it can be seen in the following equation when point (x, y, z) moves to

(x', y', z') in $3D$ space,

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \mathbf{R} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \mathbf{T} \ . \qquad (1)$$

2.2. Kalman filter formalism

A discrete time Kalman filter is expressed by a simple iteration, if the system's state equations have the proper formulation [9]. In particular, let $\mathbf{x}(k+1)$ denote the state variable vector at time $k+1$. Then, its dependency on its previous value $\mathbf{x}(k)$ and the random noise vector $\mathbf{w}(k)$ is supposed to be given by:

$$\mathbf{x}(k+1) = \mathbf{F}(k)\mathbf{x}(k) + \mathbf{G}(k)\mathbf{w}(k) + \mathbf{\Gamma}(k)\mathbf{u}(k) \quad (2)$$

while the measurements vector $\mathbf{q}(k)$ is expressed as

$$\mathbf{q}(k) = \mathbf{H}^T(k)\mathbf{x}(k) + \mathbf{v}(k) \ , \qquad (3)$$

where $\mathbf{v}(k)$ denotes the additive noise vector to $\mathbf{q}(k)$, $\mathbf{u}(\mathbf{k})$ is a known input sequence and $\mathbf{F}(k)$, $\mathbf{G}(k)$, $\mathbf{\Gamma}(k)$ and $\mathbf{H}(k)$ are appropriately defined matrices w.r.t. the model.

Given these matrices, $\mathbf{x}(k+1)$ is obtained in terms of the well-known Kalman filter equations [9]. Naturally, an initialization is necessary for the state's first estimate $\mathbf{x}(k_o)$, and since this initial value cannot be exact, for its mean value $E_o = E(\mathbf{x}(k_o))$ and variance $P_o = P(\mathbf{x}(k_o))$ as well. In addition, the measurement noise $\mathbf{v}(k)$ is considered as zero mean, while the variance matrices for all the random quantities, i.e. the noise $\mathbf{w}(k)$ and $\mathbf{v}(k)$, as well as the inexact initial value $\mathbf{x}_o$, are considered to be known.

In fact, these quantities are treated as tuning parameters, to be set by the filter's designer for optimal results. In the proposed model, as it will be shown in Section 4, the obtained estimates appeared considerably robust to changes in the initial values.

3. DERIVING EQUATIONS FOR 3D STRUCTURE RECOVERY

In the $3D$ reconstruction problem, the unknown state variable to be estimated is the $3D$ position of all given points on the rigid object. The latter is estimated w.r.t. to a reference frame (corresponding to a reference scene), for example the first available frame. It can be then seen, that depth in all available frames is estimated from equation (1).

Using equation (1) for all given transitions, two equations corresponding to (2) and (3) are obtained. In the proposed algorithm, we choose vector $\mathbf{z}_k$, containing the $3D$ depth positions of a given set of points in the k-th scene, to be the state vector. Hereon, we will consider for simplicity the $3D$ position of one single point z_k, without loss of generality. Let $\mathbf{R}(k)$, $\mathbf{T}(k)$ denote the rotation and translation matrices respectively for the transition from frame 0 to frame k (transition $0 \rightarrow k$). Let also (x_k, y_k) denote the $2D$ position in the k-th frame of the particular point ((x_0, y_0) in the reference 0-th frame). We define the following decomposition of the rotation matrix $\mathbf{R} = \begin{bmatrix} \mathbf{R}_{22} & \mathbf{r}_1 \\ \mathbf{r}_2^T & r_{33} \end{bmatrix}$ where $\mathbf{R}_{22} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}$, $\mathbf{r}_1 = \begin{bmatrix} r_{13} \\ r_{23} \end{bmatrix}$ and $\mathbf{r}_2 = \begin{bmatrix} r_{31} \\ r_{32} \end{bmatrix}$. In addition, let $\mathbf{T}_{12}$ denote the vector containing the first two components of $\mathbf{T}$ and T_3 the third.

Then, matrix equation (1) for transition $(0 \rightarrow k)$ can be analyzed into:

$$\begin{bmatrix} x_k \\ y_k \end{bmatrix} = \mathbf{R}_{22}(k) \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \mathbf{r}_1(k)\, z_0 + \mathbf{T}_{12}(k) \quad (4)$$

and

$$z_k = \mathbf{r}_2^T(k) \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + r_{33}(k)\, z_0 + T_3(k) \ . \qquad (5)$$

On the basis of (4), (5) and another two similar equations for transition $(0 \rightarrow k+1)$, after some manipulations, we obtain the Kalman filter model, eqs (2) and (3), for

$$\begin{aligned} x(k) &\equiv z_k, \quad H(k) \equiv \frac{\mathbf{r}_1^T(k)}{r_{33}(k)} \\ F(k) &\equiv \frac{r_{33}(k+1)}{r_{33}(k)}, \quad G(k) \equiv 1, \quad \Gamma(k) \equiv 1 \\ u(k) &\equiv (\mathbf{r}_2^T(k+1) - F(k)\mathbf{r}_2^T(k)) \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \\ &\quad T_3(k+1) - F(k)T_3(k) \\ q(k) &\equiv \begin{bmatrix} x_k \\ y_k \end{bmatrix} + (H^T(k)\mathbf{r}_2^T(k) - \mathbf{R}_{22}(k)) \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \\ &\quad H^T(k)T_3(k) - \mathbf{T}_{12}(k) \end{aligned}$$

Using the proposed model, a solution for z_k is provided by the Kalman filter equations. In addition, an estimate for depth z_0 in the reference scene is obtained through equation (5). Depth for all points in the k-th and the reference scene is obtained by direct expansion of (4) and (5), in the sense that the state vector $\mathbf{z}_k$ is of length equal to the number of points.

An initial value for each point's depth is obtained from transitions $(0 \rightarrow 1)$ and $(1 \rightarrow 2)$, on the basis of the estimated motion parameters

(using [8]) and the corresponding noisy $2D$ motion vectors. Since this initial value is not exact, it bears a degree of uncertainty that is expressed by variance P_o, which in turn is set by the filter's designer. The random noise $\mathbf{w}(k)$ of eq. (2) and the measurement noise $\mathbf{v}(k)$ of eq. (3) are supposed to have known statistical properties, i.e. mean value and covariance matrix. Their covariance matrices can be defined by the filter's designer to be as close to the actual ones as possible, since then the Kalman filter will be more efficient.Based on the above system equations, initial values and noise properties, the Kalman filter performs a recursive estimation process of depth z_0 for each point.

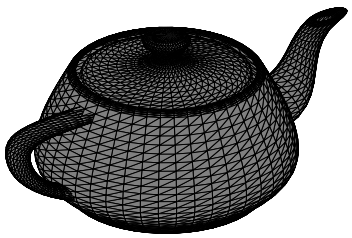

Figure 1: Synthetic model of moving teapot

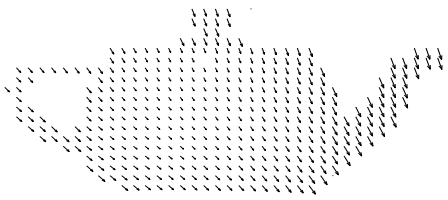

Figure 2: Noise-free motion field

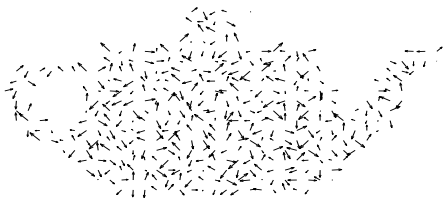

Figure 3: Noise-contaminated motion field

4. SIMULATIONS

The performance of the proposed filter was tested over a large number of synthetic models, such as the moving teapot depicted in Figure 1. The employed synthetic models were subjected to subsequent rotations and translations in order to produce arbitrary successive $3D$ scenes and respective $2D$ frames. The corresponding motion fields were artificially noise-contaminated with zero-mean i.i.d. noise of various SNR levels. The obtained noise-contaminated motion fields were next given as input to the $3D$ motion estimation algorithm in order to obtain motion parameters for each transition. Motion parameters along with the noisy motion fields were in turn fed to the Kalman filter and improvement in depth estimates was verified. In Figures 2 and 3, a noise-free motion field and its noise-contaminated counterpart respectively are depicted for the teapot model.

Figure 4 depicts the improvement in $3D$ depth for all employed points on the teapot for increasing number of frames. Mean estimates of the mean squared error (MSE), between estimated and true depth for all points were calculated on the basis of 50 Monte Carlo runs. It must be noticed, that since absolute depth cannot be determined under orthography, the object's barycentre was set to coincide with the world origin before calculating MSE factors. In Figures 5, 6 and 7, the obtained depth in the reference scene (0-scene) is depicted for all points for 5, 20 and 40 frames respectively. All three figures should be compared to the true 'visible' portion of the teapot utilized (Figure 8).

The number of frames required for a (visually) satisfactory reconstruction varied along with the induced noice, the error in initial conditions (noise mean value and covariance) and the particular model. In all cases examined, 20-40 frames were sufficient. In fact, the proposed approach proved to be robust to errors in initial conditions, since the performance of the filter was little affected by even large arbitrary errors.

5. CONCLUSIONS

In this work, a Kalman filter is designed for improved $3D$ reconstruction results on the basis of multiple frames. Kalman filtering has been widely utilized in $3D$ reconstruction in the case of perspective projections attempting to immediately estimate motion and structure. The proposed approach employs an existing algorithm for the estimation of motion parameters under orthography and proposes a filter for the improvement of depth estimates. The algorithm appears to perform well in a large number of simulated experiments. The prospect of incorporating the filter in a motion and structure estimation system from natural sequences is currently under consideration.

6. REFERENCES

[1] S. Ullman, "The Interpretation of Visual Motion," Cambridge, MA, MIT Press,

1979.

[2] R.Y. Tsai and T.S. Huang, "Uniqueness and Estimation of Three-Dimensional Motion Parameters of Rigid Objects with Curved Surfaces," *IEEE Trans. PAMI,* 6(1):13–27, 1984.

[3] A. Delopoulos and Y. Xirouhakis, "Robust Estimation of Motion and Shape based on Orthographic Projections of Rigid Objects," Proc. of *Image and Multidimensional Digital Signal Processing Workshop (IEEE IMDSP98),* pp. 151-154, Alpbach Austria, 1998.

[4] J. Weng, N. Ahuja and T.S. Huang, "Optimal Motion and Structure Estimation," *IEEE Trans. PAMI,* 15(9):864–884, 1993.

[5] C. Tomasi and T. Kanade, "Shape and Motion from Image Streams under Orthography: a Factorization Method," *Intl. J. of Computer Vision,* 9(2):137–154, 1992.

[6] S. Soatto and P. Perona, "Reducing "Structure from Motion: A General Framework for Dynamic Vision. Part 2: Implementation and Experimental Assessment," *IEEE Trans. PAMI,* 20(9):943–960, 1998.

[7] T.J. Broida, S. Chandrashekhar, and R. Chellapa, "Recursive 3-D Motion Estimation from a Monocular Image Sequence," *IEEE Trans. Aerospace and Electronic Systems,* 26(4):639–656, 1990.

[8] Y. Xirouhakis, G. Tsechpenakis and A. Delopoulos, "User Choices for Efficient 3D Motion and Shape Extraction from Orthographic Projections," accepted in the *Intl. Conf. on Electronics, Circuits and Systems (IEEE ICECS99)*, Pafos Cyprus, 1999.

[9] Brian D.O. Anderson and John B. Moore, "Optimal Filtering," Prentice-Hall, 1979.

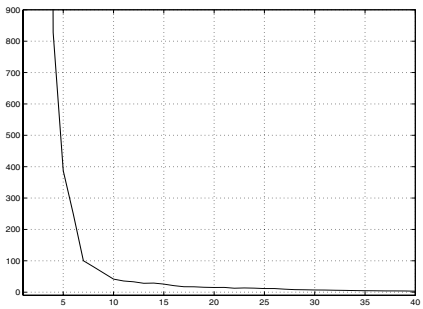

Figure 4: Improvement of depth (MSE factors) versus employed frames

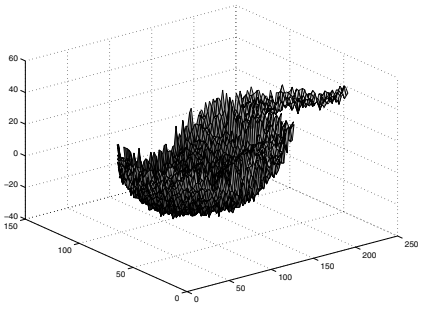

Figure 5: Estimated depth using 5 frames

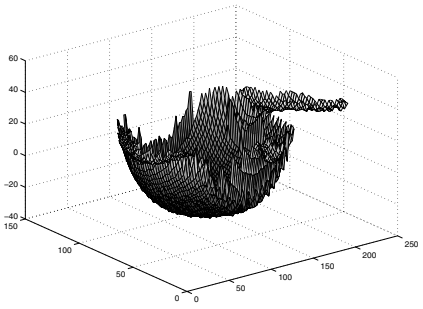

Figure 6: Estimated depth using 20 frames

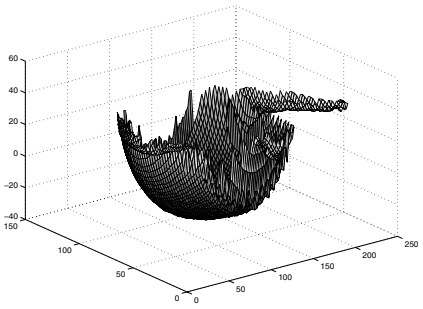

Figure 7: Estimated depth using 40 frames

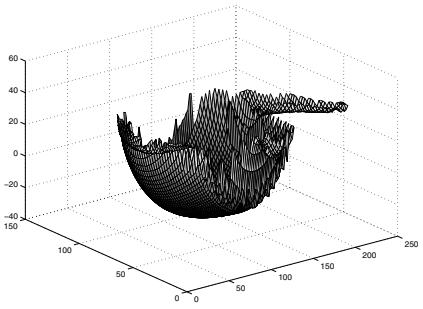

Figure 8: Visible (true) respective surface portion