STRUCTURE FROM MOTION OF PLANAR SURFACES (ANALYSIS OF AN EKF BASED APPROACH)

Jacob Ström

Linköping University
Department of Electrical Engineering, Image Coding Group
581 83 LINKÖPING

ABSTRACT

This paper analyzes the Extended Kalman filter approach to the Structure from Motion problem for the case when all points of the object belongs to a plane. It is shown that, even though the Kalman filter is underconstrained in each time step, it converges to the correct solution as well as with a non-planar 3D object. This is in contrast to fundamental matrix or trilinear tensor approaches, which are well known to fail for planar surfaces. The problem arises for instance in realtime face tracking of faces, where the points in the face that are tracked can be close to coplanar.

1. INTRODUCTION

The Structure from Motion (SfM) problem, i.e., the problem of recovering 3D structure and camera motion from a number of 2D projections of a scene, is a well known and thoroughly investigated branch of computer vision [3, 4, 5]. Some constellations of 3D points, called critical surfaces, create problems for the SfM algorithms [3, 9, 6, 7]. One such critical situation is when all 3D points belong to the same plane. Whereas it is well known that methods based on multilinear constraints (such as the fundamental matrix or the trilinear tensor) fail for planar objects [9, 6, 7], it is less well investigated how Azarbayejani and Pentland's Extended Kalman filter (EKF) approach [1] behaves in this respect. The goal of this paper is thus to investigate the convergence of the EKF approach for planar surfaces. Section 2 provides a short description of the EKF based SfM algorithm, Section 3 goes through the analysis and explains the way in which the filter converges in the planar case. A Monte Carlo experiment is performed to show that the planar case converges as well as the 3D case. To show that this problem is not only of theoretical interest, Section 4 also gives an example of a real-time head tracking system where the points of the face can be very close to being coplanar.

2. EKF-BASED SFM

The EKF based SfM algorithm in this paper comes from Azarbayejani and Pentland [1]. They reformulate the SfM problem into a stable recursive estimation problem that has been shown to converge reliably [1], [2]. By parameterizing each 3D point with one scalar only (the depth α), the once under-constrained problem becomes over-constrained and uniquely solvable in each time step. The explanation of the method below is limited due to space restrictions. The reader should consult [1] for details.

2.1. Representation for Recursive Estimation

The goal of the SfM problem is to calculate 3D structure, motion and internal camera parameters. These unknowns constitute the "state vector" $\mathbf{x}$ which the Kalman filter recovers by measurements of the 2D projections, $\mathbf{y}$. The camera is parameterized with focal length only, i.e., camera skew, aspect ratio and the principal point of the camera are assumed to be known. The focal length is assumed to be unknown, but constant throughout the sequence. It is not estimated directly, instead we estimate $\beta = \frac{1}{f}$. The 3D structure of a point is parameterized with the (known) image position (u, v) in the first image plus the unknown depth α. The cartesian coordinates (X, Y, Z) of the point can be obtained by

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} (1+\alpha\beta)u \\ (1+\alpha\beta)v \\ \alpha \end{bmatrix}. \tag{1}$$

Translation is defined relative to the first image and is parameterized as $(t_X, t_Y, t_Z\beta)$. Rotation is defined by a quaternion relative to the first image. However, between images the incremental Euler angles $(\omega_X, \omega_Y, \omega_Z)$ are estimated and used to update the quaternion.

The state vector $\mathbf{x}$ thus consists of 6 motion parameters, one internal camera parameter (β) and one scalar depth value (α) for each point. This amounts to $7+N$ parameters for $\mathbf{x}$ where N is the number of points.

$$\mathbf{x} = (t_X, t_Y, t_Z\beta, \omega_X, \omega_Y, \omega_Z, \beta, \alpha_0, \alpha_2, \ldots, \alpha_{N-1}) \quad (2)$$

This should be compared to the measurement vector $\mathbf{y}$ obtained in every time step

$$\mathbf{y} = (X_1, Y_1, X_2, Y_2, \ldots, X_N, Y_N) \quad (3)$$

which is of size $2N$ and where (X_i, Y_i) are the 2D coordinates measured from the images. Thus, the Kalman filter is over-determined in every time step when $2N \geq 7+N$ or when $N \geq 7$. Hence, a well constrained, unique solution can be found if 7 or more points are being measured in 2D. To resolve the scale ambiguity, α_0 is set to 1.0 and can thus be removed from $\mathbf{x}$.

3. ANALYSIS

The objective of the Extended Kalman filter at image k is to minimize the (scalar) error function

$$J_k(\mathbf{x}) = (\mathbf{y}_k - \mathbf{h}_k(\mathbf{x}))^T(\mathbf{y}_k - \mathbf{h}_k(\mathbf{x})) \quad (4)$$

where $\mathbf{h}_k(\mathbf{x})$ consists of the Kalman filter's estimate of the projected points $(\hat{X}_1, \hat{Y}_1, \hat{X}_2, \hat{Y}_2, \ldots, \hat{X}_n, \hat{Y}_n)$. If we know the correct solution $\mathbf{x} = \mathbf{x}^*$, $J_k(\mathbf{x})$ can be Taylor expanded around $\mathbf{x}^*$ according to

$$\begin{aligned} J_k(\mathbf{x}) \approx J_k(\mathbf{x}^*) + \frac{\partial J_k(\mathbf{x}^*)}{\partial \mathbf{x}}(\mathbf{x}-\mathbf{x}^*) + \\ + \frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^T \frac{\partial^2 J_k(\mathbf{x}^*)}{\partial \mathbf{x}^2}(\mathbf{x}-\mathbf{x}^*). \end{aligned} \quad (5)$$

If the Hessian $\frac{\partial^2 J_k(\mathbf{x}^*)}{\partial \mathbf{x}^2}$ is positive definite, $\mathbf{x}^*$ will be a local minimum to $J_k(\mathbf{x}^*)$, which means that, locally, there is no $\mathbf{x}$ that minimizes $J_k(\mathbf{x})$ as well as $\mathbf{x}^*$ does. The goal of our analysis is to calculate the Hessian at $\mathbf{x}^*$ for a number of point constellations, and analyze if it is positive definite or not. A positive definite Hessian indicates that the function has a unique optimum (at least locally), whereas a Hessian which has one or more eigenvalues of zero will allow an entire manifold of solutions $\mathbf{x}$ that will minimize the cost function $J_k(\mathbf{x})$ just as well as the known optimum, $\mathbf{x}^*$, will. Note that a positive definite Hessian in $\mathbf{x}^*$ only shows the uniqueness of the solution locally for a particular point constellation, motion and focal length. Nevertheless, these particular cases can help to understand how it works generally.

By differentiating Equation (4), we get

$$\frac{\partial^2 J_k}{\partial \mathbf{x}^2} = 2\frac{\partial h_k}{\partial \mathbf{x}}^T \frac{\partial h_k}{\partial \mathbf{x}} + 2\sum_k \frac{\partial^2 h_k}{\partial \mathbf{x}^2}(h_k(\mathbf{x}) - \mathbf{y}_k). \quad (6)$$

Note that if the measurements are perfect, $h_k(\mathbf{x}^*) = \mathbf{y}_k$ and the second term of Equation (6) is zero. Hence, in $\mathbf{x}^*$, $\frac{\partial^2 J_k}{\partial \mathbf{x}^2}$ can be conveniently calculated as $2\frac{\partial h_k}{\partial \mathbf{x}}^T \frac{\partial h_k}{\partial \mathbf{x}}$. Since $J_k(x) \geq 0$, negative eigenvalues are impossible, and the Hessian is thus positive definite unless it has a zero valued eigenvalue. Hence it suffices to investigate if the Hessian has any singular values that are zero.

3.1. Two views of a 3D object

A random 3D object containing 16 points has been rotated and translated randomly to produce 2D measurements that are fed into a SfM EKF. After the solution has converged (after about 100 images), the filter is stopped, the Hessian $\frac{\partial^2 J_k(\mathbf{x}^*)}{\partial \mathbf{x}^2}$ is calculated and decomposed with Singular Value Decomposition. The singular values are plotted in the left part of Figure 1. As we

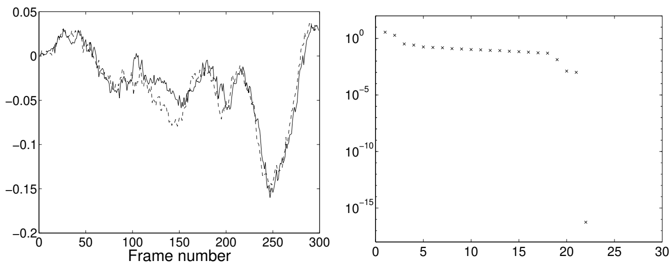

Fig. 1. Singular values for the Hessian of $J_k(\mathbf{x})$ in $\mathbf{x}^*$. Left: a 3D object. Right: an object where all points lie in a plane.

can see the smallest singular value is at around 10^{-5}, so it is fair to say that the Hessian is positive definite.

3.2. Two views of a planar object

The right diagram in Figure 1 shows the result from repeating the experiment with points that lie in a plane. The smallest singular value is now 10^{-17} — several orders of magnitude smaller than the others — practically zero. Thus, locally, there exists a curve $x_k(t)$ for which $J_k(x_k(t)) = 0$, each point on which representing a false solution to the SfM problem. An example of a false solution is shown in Figure 2, obtained by letting the filter converge and then deliberately go in the direction corresponding to the zero valued singular value. The two leftmost images are the ones that are part of the optimization function $J_k(\mathbf{x})$, namely the first image and the kth image. The crosses (estimated points) fit nicely into the circles (true points). It is not until we see a third view, (the rightmost image in Figure 2), that it becomes evident that this solution is false.

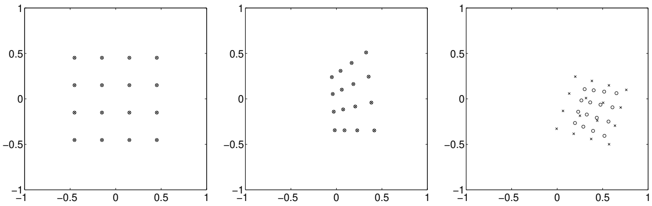

Fig. 2. A false solution. The image of the real object is marked with circles, and the reconstructed object is marked with crosses. In the 1st and the kth image (left, middle), they match ($J_k(\mathbf{x}) = 0$), but from another view (right) it is obvious that the solution is false.

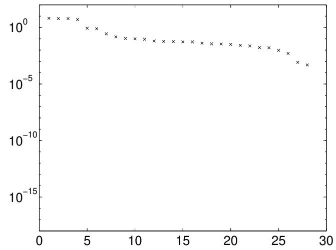

Fig. 3. Singular values for the Hessian in $\mathbf{x}^{e*}$ of the extended error function $J_k^e(\mathbf{x})$.

3.3. Three views of a planar object

As seen above, two views of a planar object with unknown focal length is not sufficient for a unique solution of the SfM problem. An interesting question is if there exists a unique solution for M views of a planar object. To investigate the case $M = 3$, a modified Kalman filter is constructed that considers not only the current image, but also the previous one. The state vector in Equation (2) is thus extended with another six degrees of freedom $(t_X^2, t_Y^2, t_Z^2\beta, \omega_X^2, \omega_Y^2, \omega_Z^2)$ for the camera motion of the previous image:

$$\begin{aligned}\mathbf{x}^e = (&t_X^2, t_Y^2, t_Z^2\beta, \omega_X^2, \omega_Y^2, \omega_Z^2, \\ &t_X, t_Y, t_Z\beta, \omega_X, \omega_Y, \omega_Z, \beta, \alpha_1, \alpha_2, \ldots, \alpha_N).\end{aligned} \tag{7}$$

The measurement vector is the concatenation of the two latest measurements, $\mathbf{y}_k^e = [\mathbf{y}_k|\mathbf{y}_{k-1}]$. Analogously with Equation (4), we get $J_k^e(\mathbf{x}) = (\mathbf{y}_k^e - \mathbf{h}_k^e(\mathbf{x}^e))^T(\mathbf{y}_k^e - \mathbf{h}_k^e(\mathbf{x}^e))$ and $\frac{\partial^2 J_k^e(\mathbf{x}^{e*})}{\partial \mathbf{x}^{e2}} = 2\frac{\partial h_k^e}{\partial \mathbf{x}^e}^T \frac{\partial h_k^e}{\partial \mathbf{x}^e}$.

An experiment with a planar object using this modified Kalman filter has been carried out. Rather big steps in the camera motion are taken in order to get reasonable different views. The Hessian is calculated and decomposed using SVD. In Figure 3 the singular values for such an experiment is shown. Note that the smallest singular value is about 10^{-4}, far greater than the 10^{-17} obtained in the two view case, and the Hessian thus positive definite. This means that, locally, there is no other constellation of points and camera positions that will generate zero error. This is in contrast to the trilinear tensor approach where, as reported by Shashua [6] and Hartley [7], there is no unique solution to the problem of three images of a planar surface.

3.4. Several views of a planar object

The Extended Kalman filter approach to the SfM problem does not rely on only two or three images, but instead it takes an entire sequence of images into account. Whether the filter will converge or not for a planar object is not obvious: On the one hand, the previous section indicates that three images are enough for the optimum to be (locally) unique. On the other hand, according to Section 3.2, the Kalman filter will be under-determined in each step and a manifold of false structures will be possible. To resolve this a Monte Carlo experiment is conducted. Random points are selected from a plane, which is randomly oriented. The filter runs for about 400 frames and the convergence is investigated. The experiment is repeated 100 times using two types of motion: Rotation around a random vector in the $z = 0$ plane (to avoid the degenerate rotation around the z-axis) and "Brownian motion", i.e., small incremental random steps in translation and rotation. The filter is then initialized with $\beta = 2.0$, (the true value is $\beta = 1.3$). Two types of initialization procedures are tried for the depths $\alpha_{1\ldots N-1}$: In the first type, called "no prior", the αs are set to random values in the interval ± 0.5 around the depth α_0 which is fixed to 1.0. In the second type, called "prior", the filter is initialized with the true α-values plus random noise in the interval ± 0.5. The entire experiment is then repeated for a random 3D object. This time, the "no prior" procedure will mean that all the αs are initialized to the same value 1.0, whereas the "prior" will mean that the αs are initialized to the the correct value plus random noise in the interval ± 0.5. The result is shown in the table below.

	no prior		prior	
	rot	Brown	rot	Brown
2D	0.66	0.79	0.98	0.96
3D	0.77	0.70	0.98	1.0

We can see that, without prior, the filter will fail to find the correct solution in about 30% of the cases, whereas in the "prior" case it will converge almost always. This should be compared to the 3D case, shown in the bottom row of the same table. As can be seen there is no big difference in convergence. The reason why the filter sometimes fails is that the random initialization is too close to the inverted solution. With any reasonable prior this should not be a problem.

A way to understand why the filter converges for planar views is the following. Each new image k gives a new constraint $J_k(\mathbf{x}) = 0$, which will be satisfied by an entire curve of false solutions $x_k(t)$. If this curve is projected down to only the structure components of x, we get a new curve $s_k(t)$, where $\mathbf{s} = (\beta, \alpha_1, \alpha_2 \dots \alpha_N)$. Since the solution is unique for three views, $s_k(t)$ and $s_{k+1}(t)$ cannot be the same curve, and must meet in the point $\mathbf{s}^*$ that represents the correct structure. This is illustrated in Figure 4. The estimate $\hat{s}$ will move

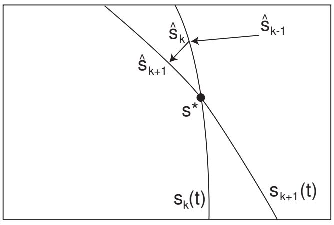

Fig. 4. Structure Convergence

towards the curves, but for each new image the curve has moved and $\hat{s}$ will continue moving until it has found the optimum. This will be a slower convergence than in the over-determined case, which was also observed during the experiment.

4. TRACKING

The above-mentioned SfM algorithm has been used for model-based tracking of a face. Figure 4 shows how the system automatically chooses a number of feature points in the face. These may be approximate coplanar. For more information, see Ström et al. [8].

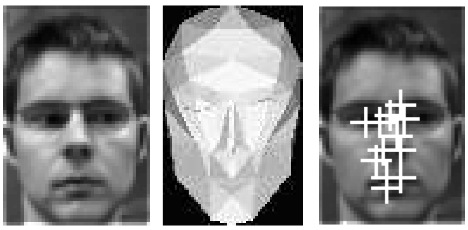

Fig. 5. From left: In-image, 3D model, selected points.

5. CONCLUSIONS

This paper has provided an analysis of how the EKF based SfM algorithm from [1] works for objects where all points lie in a plane. The problem is investigated by analyzing the singular values to the Hessian in the known optimum point. A Monte Carlo experiment is performed that shows that planar objects converge as well as 3D objects, though not as fast. Especially interesting is that the filter converges for the "Brownian motion" case, when there are very small differences between frames. In this case algorithms based on multilinear constraints will fail two-fold, since they are degenerate both for planar surfaces and for small baseline motion. Finally, an application of model-based face tracking is mentioned, where it is important to have stable behavior in the planar case.

6. REFERENCES

[1] A. Azarbayejani and A. Pentland, "Recursive Estimation of Motion, Structure, and Focal Length," *IEEE Pattern Analysis and Machine Intelligence*, vol. 17, no. 6 pp. 562-575, June 1995.

[2] T. Jebara, A. Azarbayejani and A. Pentland, "3D Structure from 2D motion," *IEEE Signal Processing Magazine,*, pp. 66-84, May 1999.

[3] O. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*, MIT Press, 1993

[4] O. Faugeras, "A nonlinear method for estimating the projective geometry of 3 views", *Sixth International Conf. on Computer Vision*, pages 477-484, Jan 1998.

[5] A. Shashua and M. Werman, "Trilinearity of Three Perspective Views and its Associated Tensor", in International Conf. on Computer Vision, Cambridge, MA, 1995, pp. 920-925

[6] G. P. Stein and A. Shashua, "On Degeneracy of Linear Reconstruction from Three Views: Linear Line Complex and Applications", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(3):244-251, 1999.

[7] R. I. Hartley, "Ambigous Configurations for 3-View Projective Reconstruction", *Sixth European Conference on Computer Vision*, Dublin, Ireland, June/July 2000 pp 922-935.

[8] J. Ström, T. Jebara, S. Basu and A. Pentland, "Real-Time Tracking and Modeling of Faces: An EKF-based Analysis by Synthesis Approach", Proceedings of the Modelling People Workshop at ICCV, August 1999.

[9] P. Torr, W. Fitzgibbon and A. Zisserman, "Maintaining multiple motion model hypotheses over many views to recover matching and structure," in *Proc. of Sixth Int. Conf. on Computer Vision*, Bombay, India, pp. 485-491, Jan 1998.