Proc. of the 2002 International Conference on Pattern Recognition, ICPR, Quebec, Canada, Aug. 2002.

Maximum Likelihood Structure and Motion Estimation Integrated over Time

Marco Zucchelli[†] José Santos-Victor[‡] Henrik I. Christensen[†]
†CVAP & CAS, Royal Institute of Technology, Stockholm, Sweden S100 44
‡Vislab, Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisboa, Portugal
zucch@nada.kth.se — jasv@isr.ist.utl.pt — hic@nada.kth.se

Abstract

Least squares minimization of the differential epipolar constraint is a fast and efficient technique to estimate structure and motion for pair of views. Previous work in this area showed how unbiased and consistent estimates could be obtained minimizing the squared errors. However, it implicitly assumes that the errors along the x and y directions are identical and uncorrelated. This is rarely the case for real data, due to the aperture problem. Instead, one should minimize the covariance weighted squared error. Moreover, when dense sequences are acquired, further robustness can be achieved by integrating the reconstruction of structure over time. This paper has two main contributions: (i) *we show that the minimization of the weighted squared errors (i.e. Maximum-Likelihood estimate) outperforms the more traditional approach of un-weighted least squares,* (ii) *we show how structure estimation can be integrated over time in a multi-view approach that drastically improves estimates.*

1. Introduction

Optical flow can be effectively used to estimate structure and motion. In the last 20 years, a number of different solutions to the problem of structure from motion in the differential setting has been proposed. Linear techniques are fast and can be expressed in closed form, but the estimation of motion and structure is biased. Zhang and Tomasi [1] recently showed that the bias is due to the incorrect choice of the objective function and that unbiased and consistent estimates can be obtained by direct minimization of the differential epipolar constraint in the least squares sense. However, that approach assumes that errors on the x and y directions are identical and uncorrelated. Whenever this is not true, severe errors and bias can be produced during the minimization process. Instead, we minimized the *mahalanobis* distance (the re-weighted squared error), which takes into account the spatial structure of the error: this is the *Maximum Likelihood* formulation of the problem.

If more than two images are available, more information can be used for structure and motion estimation. One possible approach consists in blending the various depth estimates arising from pair-wise application of structure and motion estimation methods. Alternatively we formulate a single estimation problem, where all the information is used simultaneously to determine structure and motion.

In summary, we extend previous work in two fundamental ways: (i) by considering the covariance of the noise in the estimation problem and (ii) by proposing a multi-view approach that increases statistical precision by relying on a reduced number of parameters.

2. Problem Formulation

In this section we review the basic motion model and the structure and motion estimation algorithm proposed in [1].

The relationship between the image plane motion field $\mathbf{u}(\mathbf{x})$ and the motion of the camera is given by:

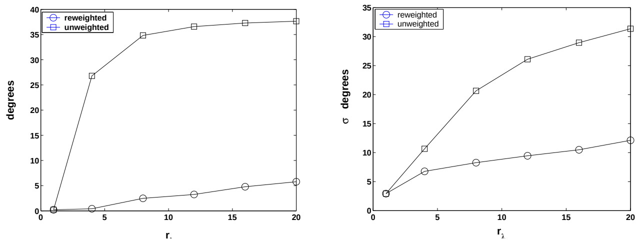
$$\mathbf{u}(\mathbf{x}) = \frac{1}{Z}A(\mathbf{x})v + B(\mathbf{x})\omega + \mathbf{n}(\mathbf{x}) \tag{1}$$

where (v, ω) are the camera linear and angular velocities and $\mathbf{n}(\mathbf{x}) \sim N(0, \Sigma)$ is zero-mean gaussian additive noise. The matrices $A(\mathbf{x})$ and $B(\mathbf{x})$ are functions of image coordinates defined as follows:

$$A = \begin{bmatrix} 1 & 0 & -x \\ 0 & 1 & -y \end{bmatrix}; B = \begin{bmatrix} -xy & (1+x^2) & -y \\ -(1+y^2) & xy & x \end{bmatrix}$$

2.1 Two-frames non linear estimation of structure and motion

Given to views of the same scene, the instantaneous motion model of Eq.(1) is valid when the camera rotation is small and the forward translation is small relative to the depth. If this condition is met optical flow between the two frames can be computed and depths and velocities can be estimated. Consider M frames $\mathcal{I}_j$ $j \in \{1 \dots M\}$ and let $\mathcal{I}_0$ be the reference view. We further assume that all the image pairs $\{\mathcal{I}_0, \mathcal{I}_j\}$ satisfy the small motion approximation. The residual for the i_{th} feature relative to a pair of



a

b

Figure 1. (a) Angle between true and estimated linear velocities for the re-weighted and un-weighted algorithms with constant error ellipse orientation. (b) Standard deviation of the estimated linear velocities for the re-weighted and un-weighted algorithms with random error ellipse orientation.

frames$\{\mathcal{I}_0, \mathcal{I}_j\}$ is defined as:

$$\mathbf{r}_i = \mathbf{u}_i - \frac{1}{Z_i} A(\mathbf{x}_i) v - B(\mathbf{x}_i)\omega \qquad (2)$$

$\mathbf{u}_i$ is the optical flow of the i_{th} feature calculated from the frames $\{\mathcal{I}_0, \mathcal{I}_j\}$ and $\mathbf{x}_i$ denotes the feature's position in the reference frame. Stacking the residuals $\mathbf{r}_i$ in the $2N \times 1$ vector $\rho = [\mathbf{r}_1, \ldots, \mathbf{r}_N]$ the motion and structure can be estimated by solving the least squares problem:

$$(\hat{v}, \hat{\omega}, \hat{\mathbf{Z}}) = arg \min_{(v,\omega,\mathbf{Z})} \|\rho\|^2 \qquad (3)$$

where $\mathbf{Z} = (Z_1, \ldots, Z_N)$. Note that the Z_i are estimated with respect to the reference frame $\mathcal{I}_0$ for each j.

The problem in Equation (3) is a non linear least squares estimation and has to be solved by an iterative technique. We used Gauss-Newton in the form:

$$J_k \Delta[v, \omega, \mathbf{Z}]_k = -\rho_k \qquad (4)$$

where J is the Jacobian of ρ and k is the iteration index. In general, J is rank deficient, due the fact that the residual function is invariant under the transformation $(v, \omega, \mathbf{Z}) \mapsto (\alpha v, \omega, \alpha \mathbf{Z})$. The rank deficient linear system (4) can be solved in the least square sense by using the $pseudoinverse$ of J. Alternatively, the constant α can be fixed by imposing the constraint $\|v\| = 1$. Such constraint can be differentiated, i.e. $v_k \Delta v_k = 0$, and this equation added as the last line of the linear system in Eq. (4). The resulting system of equations is full rank and can be solved with techniques for full rank least squares problems that are about twice as fast as the $pseudoinverse$ [2].

Iterative techniques for non linear optimization problems are locally convergent and a good initialization is needed in order to find the global minimum. In our problem initialization is easier due to the *separability* of the differential epipolar constraint equation. Defining:

$$\mathbf{e} = [e_1, e_2] = \frac{A(\mathbf{x})v}{\|A(\mathbf{x})v\|} \qquad (5)$$

the vector $\tilde{\mathbf{e}} = [e_1, -e_2]^T$ is normal to the component of the optical flow generated by the linear velocity and can be used to eliminate this from the residual in Eq. (2):

$$\tilde{\mathbf{e}} \cdot \frac{1}{Z} A(\mathbf{x}) = 0 \Rightarrow \tilde{\mathbf{e}} \cdot (\mathbf{u} - B(\mathbf{x})\omega) = 0 \qquad (6)$$

from which ω can be estimated by least squares minimization when v is known. When v and ω are known we can estimate the ratio $\frac{1}{Z}$ from:

$$\frac{1}{Z} = \frac{\mathbf{e}^T(\mathbf{u} - B(\mathbf{x})\omega)}{\|A(\mathbf{x})v\|} \qquad (7)$$

To generate the initial value for $(v, \omega, \mathbf{Z})$, it is sufficient to initialize the vector v on the half sphere of ray 1 and then estimate the corresponding ω and $\mathbf{Z}$ using equations (6) and (7).

3. Re-weighted Multi-View Formulation

In this section we re-formulate the *maximum-likelihood* and time integrated version of the algorithm described in the previous section.

3.1 Re-weighted Formulation

The algorithm described previously gives a consistent and unbiased solution to the problem when errors are isotropic and all equals. However, due to the aperture problem, the flow estimates in the direction of the image gradient are much more precise than those in the normal direction. Hence, errors are usually elliptic and correlated along the directions x and y. An estimate of the covariance matrix Σ for the computed flow vectors is given by the hessian of the images gray levels around the considered feature point [3]:

$$\Sigma^{-1} = \begin{pmatrix} I_{xx} & I_{xy} \\ I_{yx} & I_{yy} \end{pmatrix} \qquad (8)$$

where $I(x, y)$ is the image brightness. Assuming that there is no correlation between the noise relative to different features, Equation (3) can be rewritten as:

$$(\hat{v}, \hat{\omega}, \hat{\mathbf{Z}}) = arg \min_{(v,\omega,\mathbf{Z})} \|W^{\frac{1}{2}}\rho\|^2 \qquad (9)$$

where W is the block diagonal matrix whose blocks are the matrices Σ_i^{-1}: this is the *maximum-likelihood* estimator. The Gauss-Newton iterations associated to Eq. (4) become:

$$W^{\frac{1}{2}} J_k \Delta[v, \omega, \mathbf{Z}]_k = -W^{\frac{1}{2}} \rho_k \qquad (10)$$

Again the constraint $\|v\| = 1$, expressed as $v_k \Delta v_k = 0$ is added as the last line of the linear system in Eq. (10).

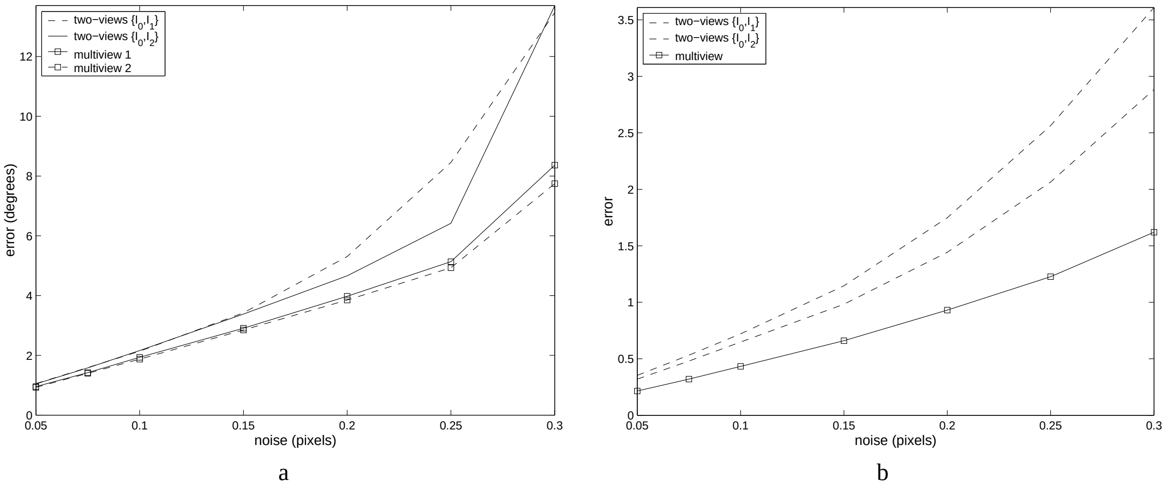


Figure 2. Structure and motion error using the two views algorithms over the pairs $\{\mathcal{I}_0, \mathcal{I}_1\}$ and $\{\mathcal{I}_0, \mathcal{I}_2\}$ and using the multi-frame algorithm over the 3 frames simultaneously. (a) Linear velocities (v_1, v_2). (b) Structure

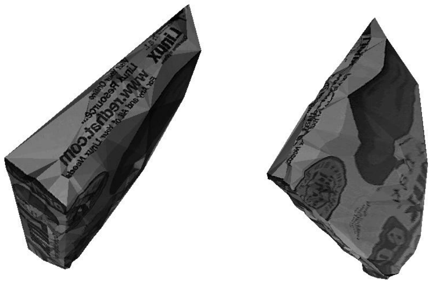


Figure 3. Multi-frame reconstruction over 3 frames. 271 features were used. Average optical flow is one pixel per frame.

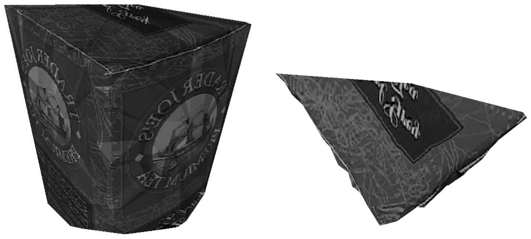


Figure 4. Multi-frame reconstruction over 3 frames. 245 features were used. Average optical flow is one pixel per frame.

3.2 Multi-view Structure and Motion Estimation

The algorithm described above can be applied to all image pairs, $\{\mathcal{I}_0, \mathcal{I}_j\}$, satisfying the small motion approximation, yielding different and independent estimates of the same parameters $\mathbf{Z}$.

Since the parameters $\mathbf{Z}$ are shared by all the minimizations of the type of Eq.(3), it is possible to minimize all the two frames residuals simultaneously, in a single non linear least square problem. Stacking the linear and angular velocities (v_j, ω_j) between pair of frames $\{\mathcal{I}_0, \mathcal{I}_j\}$ in $3M \times 1$ vectors $\vec{v} = [v_1 \dots v_M]$ and $\vec{\omega} = [\omega_1 \dots \omega_M]$ we can formulate the *multi-view* minimization as :

$$(\hat{\vec{v}}, \hat{\vec{\omega}}, \hat{\mathbf{Z}}) = arg \min_{(\vec{v}, \vec{\omega}, \mathbf{Z})} \|\mathbf{W}^{\frac{1}{2}} \vec{\rho}\|^2 \tag{11}$$

where $\vec{\rho}_{2NM \times 1} = [\rho_1, \dots, \rho_M]$ is obtained by stacking the two-frames residual vectors, ρ_j, and $\mathbf{W}_{2NM \times 2NM}$ is the block diagonal weight matrix whose diagonal blocks are the two-frames weight matrices W_j, $j \in \{1 \dots M\}$. For the minimization we used again Gauss-Newton in the form:

$$\mathbf{W}^{\frac{1}{2}} \mathbf{J}_k \cdot \Delta[\vec{v}, \vec{\omega}, \mathbf{Z}] = -\mathbf{W}^{\frac{1}{2}} \vec{\rho} \tag{12}$$

where $\mathbf{J}_{2NM, N+6M}$ is the jacobian of $\vec{\rho}$

The advantage of the multi-frame minimization is that the number of fitted parameters is significantly reduced, hence improving the statistical precision of the estimate. Assuming that $\mathbf{Z}$ is estimated M times independently from the two-frames algorithms, the precision of the estimate is about $\epsilon_s \approxeq 1/\sqrt{M(2N - p)}$, where p denotes the number of estimated parameters, in our problem $p = N + 6M$. For the multi-frame estimation we get $\epsilon_m \approxeq 1/\sqrt{M2N - p}$. Convergence properties for the two-frame and multi-frame minimization are considered in the experiments section.

In the multi-frame setting it is more convenient to handle the scale ambiguity by fixing the norm of $\mathbf{Z}$, which automatically fixes the norms of the different v_j.

3.3 Experiments

We extensively tested the algorithm using synthetic flow fields. For homogeneity and simplicity we used the same experimental conditions and benchmarks as in [4]. The focal length was set to 1 and the focal plane dimensions to 512×512 pixels. The field of view is 90^o. Random clouds of 100 points are generated in a depth range of 2-8 focal lengths. The motion is a combination of rotations and translations. The rotational speed magnitude was constant and chosen to be 0.23 degrees per frame. The magnitude of the linear velocity was chosen to fixate the point at the center of the random cloud. With this setting the average optical flow is about 1 pixel per frame, very similar to real working conditions. Zero-mean gaussian noise was added to the components of the velocity with different degrees of ellipticity and orientation. The shape of the elliptical uncertainty was varied changing the value of the parameter $r_\lambda = \sqrt{\lambda_{max}/\lambda_{min}}$ where λ_{max} and λ_{min} are the largest and smallest eigenvalues of the covariance matrix Σ.

Simulations: The two frames re-weighted algorithm was tested for different ellipticity in the range $0 \leq r_\lambda \leq 20$. We performed two different set of tests. In the first, the errors were elliptical and the orientation of the error ellipses was kept constant. Figure 1(a) shows the bias in the estimation of the linear velocity. The un-weighted algorithm fails almost systematically to find the correct camera velocity. In the second test, the ellipses orientation was random. Figure 1 (b) shows that both the un-weighted and re-weighted algorithms lead to an unbiased translational velocity, but the re-weighted version has globally a lower error, up to 3 times smaller for ellipticity $r_\lambda = 20$.

In the case of the multi-view minimization we used 3 views, of which one is fixed as the reference view. We estimated motion and structure parameters $(v_1, \omega_1, v_2, \omega_2, \mathbf{Z})$ for different noise levels using the two-views algorithm with the image pairs $\{\mathcal{I}_0, \mathcal{I}_1\}$ and $\{\mathcal{I}_0, \mathcal{I}_2\}$ and using the multi-view algorithm with the 3 views simultaneously. Figure 2 clearly shows that the multi-view algorithm outperforms the single-view.

Real Images: Figures (3) and (4) show two examples of the multi-frame reconstruction using a total of 3 frames. Features were tracked using the method in [3]. Sequences are acquired with a hand held commercial camcorder at 25 Hz. The average feature motion is about 1 pixel per frame. Due to the unavailability of the ground truth, we assessed the efficiency of our method by measuring the planarity of the 3 planar surfaces of the box-like shapes. This was done by fitting 3 planes to the 3D reconstruction and measuring the average residual of the fit. We found that the re-weighting improves the planarity of about 10% and the multi-frame integration of about 30% for both the sequences.

Convergence: Both the two-view and multi-view algorithms converge within 4-5 iterations to a minimum. The global minimum can be found starting the algorithm for different random initializations and checking the values of the residuals at the end of the minimizations. The global minimum is found essentially all the times starting with 15-20 random initializations. Initialization is made easier by variable separability described in Section 2.

4 Conclusions

We described a *Maximum Likelihood* estimation of structure and motion from optical flow, using the differential epipolar constraint. The main contributions consist in (i) considering the full directional uncertainty of the observations and (ii) formulating a multi-view approach that uses all the available image data in a single estimation problem.

Appendix A

Sensitivity in the 3D reconstruction is measured aligning the ground-truth and the reconstructed model and taking the average of the distance between estimated and true features positions, $\hat{\mathbf{X}}$ and $\mathbf{X}$, divided by the average object size. A similar approach is taken to assess the error in the reconstructed velocities:

$$\sigma_r = E[\frac{\|\mathbf{X}_i - \hat{\mathbf{X}}_i\|}{obj.\ size}]; \;\; \sigma_v = \sqrt{\frac{1}{L-1}\sum_{l=1}^{L}[\cos^{-1}(\bar{v}\hat{v}_l)]^2} \tag{13}$$

where i runs over the features and $\bar{v}$ is the average of the reconstructed velocities, that minimizes $\sum_{l=1}^{L} \cos^{-1} \hat{v}_l \bar{v}$ subject to $\|\bar{v}\| = 1$. L is the number of trials.

The average rotation matrix $\bar{R}$ is computed from the estimates, $\hat{R}_l$. The rotation sensitivity is computed as the standard deviation of the difference angle, ϕ, between $\bar{R}$ and the estimates, $\hat{R}$, for each trial sample:

$$\sigma_\phi = \sqrt{\frac{1}{L-1}\sum_{l=1}^{M}\phi_l^2} \;\; ; \;\; \phi_l = \cos^{-1}[\frac{Tr(\hat{R}_l\bar{R}) - 1}{2}] \tag{14}$$

References

[1] Zhang T. and Tomasi C. Fast, robust, and consistent camera motion estimation. In *CVPR*, volume 1, pages 164–170, 1999.

[2] Golub G.H and Van Loan C.F. *Matrix Computations*. Johns Hopkins University Press, 1996.

[3] Shi J. and Tomasi C. Good features to track. In *CVPR*, pages 593–600, 1994.

[4] Tomasi C. and Heeger D. J. Comparison of approaches to egomotion computation. In *CVPR*, pages 315–320, 1994.

[5] Bouget J. www.vision.caltech.edu/bouguetj/calib_doc/index.html.