

Structure from Motion: Combining Features Correspondences and Optical Flow

Adel Fakh, John Zelek
University of Waterloo
200 University Ave. West Waterloo ON, Canada
{afakh, jzelek}@engmail.uwaterloo.ca

Abstract

This paper suggests using discrete feature displacements and optical flow simultaneously to determine the camera motion and its velocity. This is advantageous when the number of feature correspondences is low or when the feature correspondences are noisy. The reason is that usually the available optical flow data largely outnumbers the available feature correspondences data. It is also advantageous from the perspective of the instantaneous motion estimation because it gives better estimates for the camera velocity than those obtained from optical flow by itself. We propose a probabilistic framework capitalizing on this idea. Monte-Carlo filtering is employed due to the non-linearities involved in the problem and to the non-Gaussianity of the measurements' probability distributions.

1 Introduction

Two main lines of work have marked the research in Structure from Motion (SFM): (1) a discrete-time SFM using feature correspondences when the baseline (displacement of the camera centers) is large; and (2) an instantaneous (or differential) SFM which is the limit of the discrete one when the baseline is infinitesimally small.

Two major differences between feature correspondences and optical flow are: (1) feature correspondences have a higher signal to noise ratio; and (2) the number of reliable feature correspondences is less than the number of optical flow values. The consequences are that using feature correspondences leads to more accurate estimates, and that's why they are used in most SFM approaches which have real-life applications potential [13, 8, 12]. But, although many methods have been developed to reach the motion yielding minimum reprojection error [5] (i.e., the motion best fitting the data), due to noise and to limited number of features, this latter motion is not guaranteed to be close to the true motion. From another perspective, the instantaneous motion

has many desirable properties such as providing the camera velocity (useful in many applications) and offering the possibility of obtaining a dense depth. Therefore, by combining optical flow and feature-correspondences in a simultaneous estimation of discrete and instantaneous motion, a two-fold benefit can be reaped: (1) the large number of optical flow data can help in further constraining the discrete motion thus providing a remedy to cases where an erroneous motion fits the discrete data better than the true one; and (2) More accurate instantaneous motion estimates can be obtained.

To validate the advantages of such approach we present a probabilistic formulation to efficiently estimate the motion from both optical flow and feature correspondences. Monte-Carlo sampling is employed because of the non-linearities involved and because of the similarities it bears to the principle of hypothesize-and-test [4, 14, 15, 12] widely used in motion estimation. We show that the results are better than those obtained using each approach by itself.

2 Preliminaries and Setup

In a perspective projection model, the projection of a 3-D point $\vec{X} = (X, Y, Z)^T$ on the image plane of a camera located in the origin and facing the Z axis is:

$$\vec{x} = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \frac{X}{Z} \\ f \frac{Y}{Z} \end{bmatrix}. \quad (1)$$

f is the focal "length". A calibrated Euclidean framework is assumed and without loss of generality f is taken as 1.

If the camera undergoes between two time instants t_1 and t_2 a rotation R and a translation \vec{T} , the location of the 3D point \vec{X}_1 with respect to the camera will be \vec{X}_2 defined as:

$$\vec{X}_2 = -R\vec{X}_1 - \vec{T} \quad (2)$$

which leads to the following relation:

$$\vec{x}_2 E_s \vec{x}_1 = 0; \quad (3)$$

$E_s = [\vec{T}]_{\times} R$, where $[\vec{T}]_{\times}$ is the skew symmetric matrix of \vec{T} , is called the essential matrix.

If the camera moves with a translational velocity $\vec{V} = (V_x, V_y, V_z)^T$ and a rotational velocity $\vec{\omega} = (\omega_x, \omega_y, \omega_z)^T$, the motion of \vec{X} with respect to the camera will be:

$$\left(\frac{dX}{dt}, \frac{dY}{dt}, \frac{dZ}{dt}\right)^t = -(\vec{\omega} \times \vec{X} + \vec{V}). \quad (4)$$

Substituting the time derivatives of Eq 1 in Eq 4:

$$\vec{\dot{x}}(\vec{x}) = -\frac{A(\vec{x})\vec{V}}{Z} - B(\vec{x})\vec{\omega}, \quad (5)$$

where $A(\vec{x}) = \begin{bmatrix} 1 & 0 & -x \\ 0 & 1 & -y \end{bmatrix}$ and $B(\vec{x}) = \begin{bmatrix} -xy & 1+x^2 & -y \\ -1-y^2 & xy & x \end{bmatrix}$. $\vec{\dot{x}}$ is the image velocity (i.e., optical flow) at the pixel \vec{x} .

The discrete and instantaneous motions have always been estimated separately. However, the discrete motion (R, \vec{T}) , the instantaneous motion $(\vec{\omega}, \vec{V})$, the optical flow and the feature correspondences are all inter-dependant because they all depend on the 3D locations of the features. Triggs [16] and Heyden [7] provided algebraic constraints that relate all these quantities. However, they did not fully exploit these constraints to determine both motions from both data simultaneously. They used them only to predict a discrete motion from a previous one using differential data.

Assume we have n pairs of corresponding features \vec{x}_1, \vec{x}_2 and the image velocities $\vec{\dot{x}}_2$ of these features at t_2 . Assume also that we have the image velocities of k extra points at t_2 with $k \gg n$. We use the term D_d to refer to the discrete data (feature correspondences) and D_f to refer to the instantaneous data. We seek to determine the probability distribution $p(M|D_d, D_f)$ of the discrete and instantaneous motions $M = (R, \vec{T}, \vec{\omega}, \vec{V})$ given both discrete and differential measurements. This allows us to determine the Maximum A Posteriori (MAP) or the Minimum Mean Square Error (MMSE) estimates of the motion. To achieve this we proceed as follows: (1) We determine the probability distribution $p(M|D_d)$ of the motion given the discrete data. Then (2) we determine the likelihood $p(D_f|M)$ of the motion given the instantaneous data. (3) We combine the two probabilities using Bayes rule. The obtained distribution can be considered also as the automatic initialization of a recursive filter where the subsequent steps use the same type of bayesian inference.

3 Motion Distribution Given the Feature Correspondences

If the noise in the feature correspondences is Gaussian, the covariance propagation techniques [18] can be used to

obtain the Gaussian distribution of the motion. When that noise is not Gaussian, one solution is to determine a sample-based representation of the motion distribution. For example, Chang [2] determines a sampled representation of the distribution of features correspondences and then obtains a motion sample from each features correspondences sample. We take a different approach which exhibits some resemblance to the hypothesize-and-test framework. We start by generating a set of samples $M^{(j)}$ with corresponding weights $W_d^{(j)}$. A sample is generated as follows:

1. Randomly draw five feature pairs from the available feature correspondences set and determine the corresponding essential matrix using Nister's five points algorithm [11]. Up to ten matrices might be generated.
2. For each of the generated matrices, compute the corresponding R and T using *SVD* decomposition [6].

To determine the weight of the samples:

1. For each motion sample $M^{(j)}$, compute its reprojection error $e_d^{(j)2}$ using the approximation of Torr [14].
2. Take the weight of $M^{(j)}$ as $W_d^{(j)} = \exp\left(\frac{-e_d^{(j)2}}{2\sigma_d^2}\right)$. σ_d controls the shape of the distribution. A small σ_d makes the probability function concentrated around the minima; a large σ_d allows it to be more spread out.
3. Samples with weights below a threshold are discarded.

Fair samples are obtained by drawing (with replacement) from the weighted samples according to their weights.

4 Motion likelihood given the optical flow

To determine the likelihood $p(D_f|M^{(j)})$ of the motion sample $M^{(j)}$ given the optical flow, the first step is to get the instantaneous motion $(\vec{\omega}^{(j)}, \vec{V}^{(j)})$ from the discrete motion $(R^{(j)}, \vec{T}^{(j)})$. Heyden [7] derived the following constraint on both the instantaneous and discrete motion.

$$\text{rank} \begin{bmatrix} \vec{\dot{x}} & \vec{x}_2 & 0 & \vec{T} \\ \hat{\vec{\omega}}^T \vec{x} & \vec{\dot{x}}_2 & \vec{x}_2 & \vec{V} \end{bmatrix} < 4. \quad (6)$$

The above matrix is 6×4 where $\vec{\dot{x}} = (\dot{x}, \dot{y}, \dot{z})^T = R.(x_1, y_1, 1)^T$ and $\hat{\vec{\omega}}$ is the skew symmetric matrix corresponding to $\vec{\omega}$. Being of rank (< 4) means that all the minors of order four are equal to zero. The useful minors (nine) are the ones containing two rows of the first three and two of the last three because these minors give equations relating $R, \vec{T}, \vec{\omega}$ and \vec{V} . Only two independent equations

can be obtained. We use the following two minors:

$$\det \begin{bmatrix} \tilde{x} & x_2 & 0 & T_x \\ \tilde{y} & y_2 & 0 & T_y \\ \omega_y \tilde{z} - \omega_z \tilde{y} & \dot{x}_2 & x_2 & V_x \\ \omega_z \tilde{x} - \omega_x \tilde{z} & \dot{y}_2 & y_2 & V_y \end{bmatrix} = 0 \quad (7a)$$

and

$$\det \begin{bmatrix} \tilde{y} & y_2 & 0 & T_y \\ \tilde{z} & 1 & 0 & T_z \\ \omega_y \tilde{z} - \omega_z \tilde{y} & \dot{x}_2 & x_2 & V_x \\ \omega_x \tilde{y} - \omega_w \tilde{x} & 0 & 1 & V_z \end{bmatrix} = 0 \quad (7b)$$

Expanding the above two determinants gives two linear equations in $\vec{\omega}$ and \vec{V} for each point. Hence, at least 3 points are needed to determine $\vec{\omega}$ and \vec{V} . Given the discrete motion sample $(R^{(j)}, \vec{T}^{(j)})$ generated as in Section 3, we use the system of linear equations formed by concatenating Eq 7 for each of the five features used to generate the motion sample. Solving this linear system gives the corresponding instantaneous motion sample. The likelihood of this sample given the optical flow is determined using the unweighted re-projected error function of Chiuso [3]:

$$e_f^{(j)2} = \sum_{i=1}^{n+k} \left[\tau(\vec{x}^i, \vec{V}^{(j)}, 1)^T (\dot{\vec{x}}^i - B(\vec{x}^i) \vec{\omega}^{(j)}) \right]^2, \quad (8)$$

where $\tau(\vec{x}, \vec{V}, 1)$ is a unit vector perpendicular to the translational component of the optical flow. The likelihood is taken as:

$$p(D_f|M^{(j)}) = \exp\left(\frac{-e_f^{(j)2}}{2\sigma_f^2}\right) \quad (9)$$

σ_f controls the shape of the distribution.

5 Probabilistic Integration

Resorting to Bayes rule, the distribution $p(M|D_d, D_f)$ can be written as:

$$p(M|D_d, D_f) = \frac{p(M|D_d) \times p(D_f|M)}{\int p(M|D_d) \times p(D_f|M) dM}. \quad (10)$$

To make use of this relation we rely on the *Importance Sampling* principle. Taking $p(M|D_d)$ as proposal distribution -(the samples that we already have are generated from this distribution)- then a sampled representation of $p(M|D_d, D_f)$ can be obtained by weighting the samples in the following way:

$$W(M^{(j)}) = \frac{p(M^{(j)}|D_d, D_f)}{p(M^{(j)}|D_d)} \propto \frac{p(D_f|M^{(j)})}{\sum_n p(M^{(n)}|D_d) \times p(D_f|M^{(n)})}. \quad (11)$$

The MAP estimate is defined as:

$$\hat{M}_{MAP} = \underset{M}{\operatorname{argmax}}(p(M|D_d, D_f)), \quad (12)$$

so \hat{M}_{MAP} is the estimate with the highest weight. The MMSE estimate is defined as

$$\hat{M}_{MMSE} = E(M|D_d, D_f). \quad (13)$$

Determining the expectation $E(M|D_d, D_f)$ is not trivial because it requires an averaging over motion samples. A suitable parameterization of the motion is required. We adopt a parameterization based on five points, similar to the one used for optimization purposes in [15], and we extend it to represent the instantaneous motion also. The motion is encoded by the corresponding essential matrix $E_s = [\vec{T}]_{\times} R$ which is then represented by five points (x_1, y_1) and (x_2, y_2) . Fixing the x_1, y_1 and x_2 coordinates of these five correspondences, the space of E_s is parameterized by the five y_2 coordinates. Similarly, taking the image velocities (\dot{x}_2, \dot{y}_2) of three of these points and fixing the \dot{x}_2 , the space of $\vec{\omega}$ and \vec{V} is parameterized by the three \dot{y}_2 . Now to obtain $E(M|D_d, D_f)$, we determine \hat{y}_2 as the expectation of the y_2 of all the samples. The expected values of R and \vec{T} can then be taken as the ones fitting the 5 points (x_1, y_1) and (x_2, \hat{y}_2) . Similarly $\hat{\dot{y}}_2$ is determined as the expectation of the \dot{y}_2 of all the samples. Then using $x_1, y_1, x_2, \hat{y}_2, \dot{x}_2, \hat{\dot{y}}_2$ and the expected values of R and \vec{T} in Eqs 7, the expected value of the instantaneous motion is determined.

The obtained samples can be used also as an automatic initialization of a recursive filter which works as follows; the samples can be used to predict a new set of samples:

$$\begin{aligned} \vec{T}_{t+\delta t}^j &= \vec{T}_t^j + \vec{V}_t^j \\ R_{t+\delta t}^j &= R_t^j \exp(\hat{\vec{\omega}}_t) \end{aligned} \quad (14)$$

The predicted samples can be updated using the new optical field at $t + \delta t$ as in Eq.11. A resampling step could be used to avoid sample impoverishment.

6 Experimental Results

A camera (focal length = 492.467 pixels, center=(329.944, 243.616)) fixed on a controllable pan-tilt unit is used. The motion is a combination of rotation and translation. While the translation direction is unknown, the rotation is determined from the pan-tilt unit. Lucas-Kanade [10] detector is used for feature detection and SIFT descriptors [9] are employed for feature matching. The optical flow is computed following a pyramidal implementation of Lucas-Kanade as in [1]. Figure 1

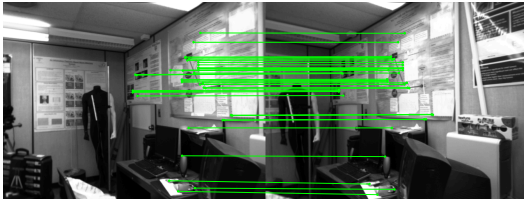


Figure 1. Features between two frames.

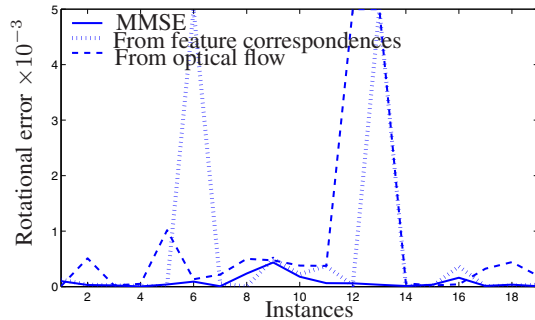


Figure 2. Discrete rotation error.

shows an instance of this sequence with the corresponding matches. We compare rotations using the distance between corresponding unit quaternions and rotational velocities using the distance between rotational velocity vectors $\vec{\omega}$. In Figure 2 the errors of the following discrete estimates are shown: (1) MMSE estimates as described in Section 5; (2) estimates minimizing the reprojected features correspondences error; and (3) estimates minimizing the reprojected optical flow error. The error in the MMSE estimates is always low while the other estimates have very high errors in some cases. In Figure 3 the errors of the following estimates are shown: (1) MMSE estimates; (2) instantaneous estimates determined from the discrete estimates as in Section 4; and (3) estimates obtained from optical flow using Zhang and Tomasi's optimal approach [17]. The MMSE estimates are much more accurate.

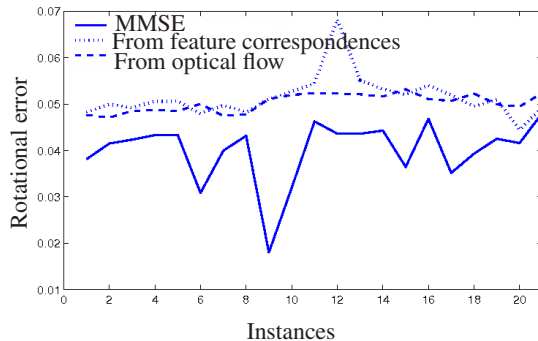


Figure 3. Instantaneous rotation error.

References

- [1] J. Bouquet. Pyramidal implementation of the lucas kanade feature tracker description of the algorithm. *Intel Corporation, Microprocessor Research Labs*, 2000.
- [2] P. Chang. *Robust Tracking and Structure from Motion with Sampling Method*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, February 2003.
- [3] A. Chiuso, R. Brockett, and S. Soatto. Optimal structure from motion: Local ambiguities and global estimates. *International Journal of Computer Vision*, 39(3):195–228, 2000.
- [4] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.
- [5] R. Hartley and F. Kahl. Global optimization through searching rotation space and optimal estimation of the essential matrix. pages 1–8, 2007.
- [6] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.
- [7] A. Heyden. Differential-Algebraic Multiview Constraints. *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)-Volume 01*, pages 159–162, 2006.
- [8] M. Lhuillier and L. Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):418–433, 2005.
- [9] D. G. Lowe. distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.
- [10] B. Lucas and T. Kanade. An iterative technique of image registration and its application to stereo. *Proc. 7th Intl Joint Conf. on Artificial Intelligence*, pages 674–679, 1981.
- [11] D. Nistér. An efficient solution to the five-point relative pose problem. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(6):756–770, 2004.
- [12] D. Nistér. Preemptive RANSAC for live structure and motion estimation. *Machine Vision and Applications*, 16(5):321–329, 2005.
- [13] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual Modeling with a Hand-Held Camera. *International Journal of Computer Vision*, 59(3):207–232, 2004.
- [14] P. Torr and D. Murray. The Development and Comparison of Robust Methods for Estimating the Fundamental Matrix. *International Journal of Computer Vision*, 24(3):271–300, 1997.
- [15] P. Torr and A. Zisserman. MLESAC: a new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78(1):138–156, 2000.
- [16] B. Triggs. Differential Matching Constraints. *matrix*, 15:0.
- [17] T. Zhang and C. Tomasi. Fast, robust, and consistent camera motion estimation. *cvpr*, 01:1164, 1999.
- [18] Z. Zhang. Determining the epipolar geometry and its uncertainty: A review. *Int. J. Comput. Vision*, 27(2):161–195, 1998.