Recovering 3D Shape and Motion from Image Streams using Non-Linear Least Squares

Richard Szeliski
Digital Equipment Corporation
Cambridge Research Lab
One Kendall Square, Bldg. 700
Cambridge, MA 02139

Sing Bing Kang
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213-3890

This paper presents a shape and motion estimation algorithm based on non-linear least squares applied to the tracks of features through time. While our approach requires iteration, it quickly converges to the desired solution, even in the absence of *a priori* knowledge about the shape or motion. Important features of the algorithm include its ability to handle partial point tracks and true perspective, to use line segment matches and point matches simultaneously, and its use of an object-centered representation for faster and more accurate structure and motion recovery.

This paper addresses the problem of extracting both 3D structure (shape) and object or camera motion simultaneously from a given image sequence. Recovering shape and motion is a difficult and important task, and has wide applicability in many areas such as robot navigation and manipulation. Approaches to this problem range from the classical methods which use only two frames and a few points [3] to methods which use many frames and points [7, 6]. Tomasi and Kanade [7] have obtained highly accurate results using a factorization method to extract object-centered shape and motion under orthography. More recently, Taylor, Kriegman, and Anandan [6] developed a non-linear least squares fitting algorithm for 2D shape and motion recovery under perspective using odometry to obtain initial guesses for the camera motion.

Our approach applies a similar non-linear least squares technique to recover 3D shape and motion from *image streams* (the temporal tracks of image features) without *a priori* information about the shape or motion. Least squares guarantees a statistically optimal estimate in the vicinity of the true solution and avoids the potentially unlimited noise amplification which may occur with arbitrary algebraic manipulation. The least squares formulation also enables us to deal easily with perspective or arbitrary camera models, partial and/or uncertain tracks, and even to simultaneously use point and line correspondences. Our results on both synthetic and real data indicate that the algorithm normally converges even when no *a priori* information about shape or motion is given. For a detailed review of previous work and a complete description of our algorithm and results, please see [5].

To formulate the problem, we first write the forward image formation equations using the usual rigid body transformation $\mathbf{x}'_{ij} = \mathbf{R}_j \mathbf{x}_i + \mathbf{t}_j$ (we represent rotation matrices with unit quaternions $\mathbf{q}_j$) followed by a perspective projection onto the image plane. Rather than using the usual projection equation $\mathbf{u}_{ij} = f\mathbf{x}'_{ij}/z'_{ij}$, we use an alternative formula $\mathbf{u}_{ij} = s\mathbf{x}'_{ij}/(1 + \eta z'_{ij})$. This new formulation puts the coordinate system before projection at a distance $t_z = 1/\eta$ along the optic axis; $s = f/t_z$ corresponds to the scaling between world and screen coordinates, and η is a perspective distortion factor. The formulation favors the recovery of structure and motion in *object centered* coordinates, which improves the accuracy of the estimates [7]. We write the complete image formation equations as $\mathbf{u}_{ij} = \mathbf{f}(\mathbf{x}_i, \mathbf{t}_j, \mathbf{q}_j)$.

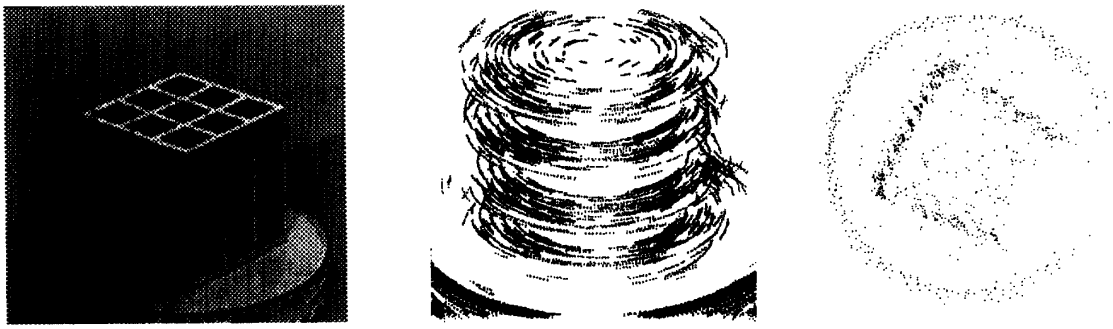
To track point features from frame to frame, we use a relatively simple algorithm based on the *monotonicity operator* [2], which computes the number of neighboring pixels whose intensity is less than that of the central pixel [5]. Figure 1 shows a sample input image and the set of tracks detected in the sequence.

To recover the structure and motion parameter, we use the Levenberg-Marquardt algorithm [4], which iteratively adjusts the unknown shape and motion parameters $\{\mathbf{x}_i\}$ and $\{\mathbf{t}_j, \mathbf{q}_j\}$ to minimize the weighted squared distance between the predicted and observed feature coordinates $C = \sum_{ij} w_{ij} |\mathbf{u}_{ij} - \mathbf{f}(\mathbf{x}_i, \mathbf{t}_j, \mathbf{q}_j)|^2$. The Levenberg-Marquardt algorithm converges more quickly than gradient descent because it approximates an inverse Hessian method. Its implementation requires only the computation of the forward equations mapping current estimates to predicted image positions, and the computation of derivatives with respect to the unknown parameters. To minimize the overall amount of computation, we use sparse matrix techniques based on *skyline storage* and LU decomposition [5]. We have found that solving for shape and motion simultaneously converges more quickly than alternating between these two sets of parameters.

To begin our algorithm, we initialize 3D point locations by projecting the 2D image point locations in the middle frame to a constant depth plane just slightly behind the object coordinate origin, and set the rotation quaternions to unit scalars and the translations to zero. In practice, the Levenberg-Marquardt algorithm then converges quickly to the correct solution (typically 5–8 iterations on small synthetic data sets [5]). This suggests that the region of convergence for the iterative algorithm is quite broad, and that complicated initialization techniques are not required.

During convergence, we do occasionally observe occurrences of depth reversals, especially under weak orthography (narrow fields of view). These are simple to correct, by re-

1063-6919/93 $03.00 © 1993 IEEE



(a) (b) (c)

Figure 1: Results from real image sequence (cube scene, 96 frames): (a) image from sequence (b) traces found (c) top view of reconstructed 3D points.

flecting the shape about a constant depth plane and checking if the image plane error is reduced. Once the algorithm has converged, we can remove or downweight (reduce the w_{ij} of) measurements with large residuals, thus making the recovery of parameters more robust. We can also compute variance and covariance statistics on the parameter estimates.

Our experimental results on both synthetic and real data indicate that the algorithm converges quickly and degrades gracefully with higher levels of noise in the data. For example, for 96 points randomly distributed over a sphere of diameter 100, an incremental rotation of 1° and 8 frames, the algorithm converges after 5 iterations to a 2-D image RMS error of 1.25 and a scaled rigid 3-D RMS error of 0.30 [5]. The algorithm handles pure rotation and pure translation, as well as mixed motions, equally well. The accuracy of our results improve as more frames or larger rotation steps are used.

Figure 1c shows a top-down view of the 3D set of points recovered from the real image sequence shown in Figure 1a. The shape of the box is recovered quite well, although a slight "pinching" implies that the projective structure has been recovered more accurately than the true Euclidean structure. This is a common occurrence in shape-from-motion when the range of viewpoints or the length of the tracks is limited. We expect improvements in our tracking algorithm to reduce this problem, as well as to reduce the noise in individual position measurements.

In [5], we also discuss how our non-linear least squares algorithm can be used to accurately perform camera calibration, using either single or multiple images (we simply fix the $\mathbf{x}_i$ at their known values). We also show how line correspondences can be used in place of (or in addition to) point correspondences, simply by considering only image plane errors perpendicular to the line segment orientations.

We have also begun experiments in recovering projective structure and motion [1]. Our preliminary results indicate that this approach converges more quickly than Euclidean structure recovery [5]. In future work, we plan to investigate a recursive formulation which models the correlation between the structure and motion parameters. From the experimental side, we would like to validate our approach on real data using known 3-D ground truth, and apply our techniques to more complicated scenes.

To summarize, the shape and motion recovery algorithm developed in this paper has several advantages over existing techniques. It can handle arbitrary projection equations, partial and uncertain tracks, and line segment matches in a unified framework. Additional information, such as known calibration points or specific structural constraints can easily be added. It makes optimal and robust use of the data, since measurements can be individually weighted and outliers can be rejected. Solving for the unknowns in a batch fashion leads to optimal estimates, while the computational costs are kept reasonable by using sparse matrix techniques. Recovering object-centered shape is more reliable than camera-centered shape, especially for narrow fields of view. Finally, the iterative recovery of shape and motion without a special bootstrapping stage makes this a particularly simple and general technique for shape recovery.

References

[1] O. D. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig? In *ECCV'92*, pages 563–578, Santa Margherita Liguere, Italy, May 1992. Springer-Verlag.

[2] R. Kories and G. Zimmermann. A versatile method for the estimation of displacement vector fields from image sequences. In *IEEE Workshop on Motion*, pages 101–106, 1986.

[3] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.

[4] W. H. Press *et al. Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, England, second edition, 1992.

[5] R. Szeliski and S. B. Kang. Recovering 3D shape and motion from image streams using non-linear least squares. Technical Report 93/3, Digital Equipment Corporation, Cambridge Research Lab, March 1993.

[6] C. J. Taylor, D. J. Kriegman, and P. Anandan. Structure and motion in two dimensions from multiple images: A least squares approach. In *IEEE Workshop on Visual Motion*, pages 242–248, Princeton, New Jersey, October 1991.

[7] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *Int. J. Comput. Vision*, 9(2):137–154, November 1992.