Shape and Motion without Depth

Carlo Tomasi and Takeo Kanade
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pa 15213

Abstract

Inferring the depth and shape of remote objects and the camera motion from a sequence of images is possible in principle, but is an ill-conditioned problem when the objects are distant with respect to their size. We overcome this problem by inferring shape and motion without computing depth as an intermediate step.

On a single epipolar plane, an image sequence can be represented by the $F \times P$ matrix of the image coordinates of P points tracked through F frames. We show that under orthographic projection this matrix is of rank 3.

Using this result, we develop a shape-and-motion algorithm based on singular value decomposition. The algorithm gives accurate results, without relying on any smoothness assumption for either shape or motion.

Introduction

In principle, the shape of an object can be computed from a sequence of images by first estimating camera motion and depth, and then inferring shape from the depth values.

In practice, however, when objects are distant from the camera relative to their size, this computation is ill-conditioned. First, the translation component along the optical axis is difficult to determine, because the image changes that it produces are small. Second, shape values are very sensitive to noise if they are computed as the small differences between large depth values.

These difficulties can be circumvented by inferring shape directly from variations in the relative position of image features, without computing depth as an intermediate step.

In this paper, we show that shape and camera motion can be inferred precisely from many features and frames, without assuming any model for the motion, and reduce the computation to decomposing a matrix of image measurements.

The resulting algorithm, tested in simple situations, gives remarkably precise motion and shape estimates, without introducing smoothing effects into the result.

For simplicity, we will limit our consideration to one epipolar plane at a time, and assume that motion occurs in that plane. In other words, our images are single scanlines.

Our theory is based on the observation that the incidence relations among projection rays can be expressed as the degeneracy of a matrix that gathers all the image measurements. More specifically, on a single epipolar plane, an image sequence can be represented by the $F \times P$ matrix of the image coordinates of P points tracked through F frames. We show that under orthographic projection this matrix is of rank 3.

In the following, we introduce our scenario, summarize the results, and sketch the relations of our work with previous literature on the subject. The next section introduces the degeneracy principle mentioned above. We then show how to use it to decompose the measurement matrix into shape and camera motion. The experimental results in the following section show the ability of the algorithm to deal with jerky motions without introducing smoothing artifacts in its output.

The Scenario

The world is still, and the camera moves in a plane, within which it can freely rotate and/or translate. P feature points, far away from the camera, are visible in a given scanline, parallel to the plane of motion. Since the frames are taken frequently, it is easy to track the features from frame to frame. As the camera moves, it is panned so as to keep the features in the field of view.

After F frames, an $F \times P$ matrix U of image measurements is available. This matrix is the input to the algorithm.

This scenario approximates what happens with a camera on an airplane, with suitable control mechanisms to align the camera scanlines with the direction of flight, and to keep the same object within the field of view. Because objects are distant from the camera, we can assume orthographic projection.

The Results

This paper first shows that if the measurements are noise-free, the image coordinate matrix U is highly degenerate: its rank is 3. As a result, U can be decomposed into the product of two smaller matrices: an $F \times 3$ matrix that encodes the F camera positions, and a $P \times 3$ matrix that encodes the positions of the P world points.

When noise corrupts the measurements, the rank of U can be defined in an approximate sense, and is still 3.

CH2934-8/90/0000/0091$01.00 © 1990 IEEE

The noisy matrix U is factored by Singular Value Decomposition [3], which is known to be efficient and numerically well behaved. If more points and frames are used than prescribed by equation-counting arguments (which require a minimum of three points and three frames), the effects of noise can be reduced.

The resulting shape and motion algorithm is simple and efficient, and has been implemented and tested on objects as distant as one hundred times their size. The rotation errors are always smaller than one tenth of a degree. The relative precision in the computed shape is of the order of the *relative depth range*, defined as the ratio between the size of the object and its distance from the camera.

Relations with Previous Work

Our algorithm does what photogrammetrists for more than thirty years have done by hand and with two frames at a time [7]. Ullman proposed an automated solution to this problem eleven years ago [10], and called it *structure-from-motion*. He also considered only two frames at once, and as few points as theoretically possible.

Most of the initial efforts in this area have been devoted to finding closed-form solutions with a minimal or nearly-minimal number of points and/or frames.

In general, structure-from-motion is hard to solve, because of the inherent sensitivity of shape and motion to noise in the image, especially when objects are distant. If depth is explicitly represented as an intermediate stage in the computation, performance degrades with reductions in the relative depth range. For instance, the algorithm presented in [9] works very well for close objects (which is the intended goal of that algorithm), but the performance is likely to degrade when objects become more remote, and the relative depth range becomes smaller.

The remedy is to by-pass the computation of depth, as we do in this paper, to remove the main cause of ill-conditioning.

Even with a well-conditioned algorithm, however, noise degrades performance. Few points and/or few frames give bad results, regardless of how good the math is. Our algorithm allows using many frames and many points, thus exploiting redundancy to counteract noise.

Many, tightly spaced frames have been used in [1] and [5], but only for the inference of depth when the motion of the camera is known.

The Decomposition Principle

This section introduces the fundamental principle on which our shape-and-motion algorithm is based: the $F \times P$ matrix of the image coordinates of P points tracked through F frames is higly rank-deficient.

As we stated in the introduction, we consider only one scanline per frame, and assume that the camera moves in a plane parallel to the scanline. In this plane, we define an arbitrary orthogonal system of coordinates (X, Z).

The images are orthographic projections of P points, tracked through F frames.

Let c_f and s_f be the cosine and sine of the angle α_f that frame f forms with the X axis. The projection u_{fp} of point (X_p, Z_p) onto frame f is then given by the equation

$$u_{fp} = c_f X_p + s_f Z_p + t_f \ . \tag{1}$$

The scalar t_f is the projection onto the f-th image of the translation vector between frames 1 and f.

The measurements u_{fp} can be collected in an $F \times P$ matrix U. Then, the $F \times P$ equations (1) can be expressed in matrix form:

$$U = MS \tag{2}$$

where

$$M = \begin{bmatrix} c_1 & s_1 & t_1 \\ \vdots & & \vdots \\ c_F & s_F & t_F \end{bmatrix} \tag{3}$$

is the motion matrix, and

$$S = \begin{bmatrix} X_1 & \cdots & X_P \\ Z_1 & \cdots & Z_P \\ 1 & \cdots & 1 \end{bmatrix} \tag{4}$$

is the shape matrix.

Since M is $F \times 3$ and S is $3 \times P$, we have thus proven the following fact.

The Rank Principle

> *Without noise, the rank of the measurement matrix U is at most three.*

Intuitively, the rank principle expresses the simple fact that the $F \times P$ image measurements are redundant. Indeed, they could all be described more concisely by giving F frame angles and P points, if only these were known.

Geometrically, the rank principle expresses an incidence property. In fact, equation (1) says that the projection lines of point (X_p, Z_p) form a pencil.

In the next section, we show how to use the rank principle to determine the motion and shape matrices M and S.

The Algorithm

When noise corrupts the images, the measurement matrix U will not be exactly of rank 3. However, the rank principle can be extended to the case of noisy measurements in a well-defined manner. The next subsection introduces this extension, using the concept of Singular Value Decomposition (SVD) [3] to introduce the notion of approximate rank.

The rank principle actually determines the matrices M and S only up to an arbitrary affine warping of the plane. We thus introduce the additional constraints needed to complete the solution. Finally, we outline the complete shape-and-motion algorithm.

Approximate Rank

Assuming * that $F \geq P$, the matrix U can be decomposed [3] into an $F \times P$ matrix L, a diagonal $P \times P$ matrix Σ, and a $P \times P$ matrix R,

$$U = L\Sigma R \ , \tag{5}$$

such that $L^T L = R^T R = RR^T = I$, and $\sigma_1 \geq \ldots \geq \sigma_P$. Here, I is the $P \times P$ identity matrix, and the *singular values* $\sigma_1, \ldots, \sigma_P$ are the diagonal entries of Σ. This is the *Singular Value Decomposition* (SVD) of the matrix U.

We can now restate our key point.

The Rank Principle for Noisy Measurements

The first three singular values of the noisy measurement matrix U are much greater than the others:

$$\sigma_1 \ , \ \sigma_2 \ , \ \sigma_3 \gg \sigma_4, \ldots, \sigma_P \ . \tag{6}$$

It can be shown [2] that the rank-3 matrix $U^\star$ that is closest to U in the L_2-norm sense is obtained by setting to zero all the singular values after the third in the decomposition:

$$U^\star = L^\star \Sigma^\star R^\star \ , \tag{7}$$

where $L^\star$ collects the first three columns of L, $\Sigma^\star$ is the first third-order principal minor of Σ, and $R^\star$ gathers the first three rows of R.

The Metric Constraints

Since the rank principle expresses an incidence relation, it only determines the two matrices M and S up to an affine transformation of the plane. In fact, if A is *any* invertible 3×3 matrix, the matrices MA and $A^{-1}S$ are also a valid decomposition of U, since

$$(MA)(A^{-1}S) = M(AA^{-1})S = MS = U \ .$$

The ambiguity can be resolved by noticing that the first two columns of M gather cosines and sines of the frame angles (see equation (3)), and must therefore be normalized. Furthermore, the third row of S contains all ones (equation (4)). These are *metric* constraints, as opposed to the incidence constraints expressed by the rank principle.

Given any initial decomposition $\hat{M}$, $\hat{S}$ of U, it can be shown that the metric constraints uniquely determine a matrix A that transforms $\hat{M}$ and $\hat{S}$ into the actual motion and shape matrices M and S according to

$$\begin{aligned} M &= \hat{M}A^{-1} \\ S &= A\hat{S} \ . \end{aligned} \tag{8}$$

This leads to a simple data fitting problem which, though non-linear, can be solved efficiently and reliably.

*This assumption is not crucial: if $F < P$, everything can be repeated for the transpose of U.

Outline of the Algorithm

Given an image measurement matrix U, the algorithm for computing the motion matrix M and the shape matrix S defined in equations (3) and (4) can be summarized as follows.

1. Compute the singular value decomposition of U:
$$U = L\Sigma R \ .$$
2. Define the initial factors of U as follows:
$$\begin{aligned} \hat{M} &= L^\star(\Sigma^\star)^{1/2} \\ \hat{S} &= (\Sigma^\star)^{1/2} R^\star \ , \end{aligned}$$
where $L^\star$ collects the first three columns of L, $\Sigma^\star$ is the first third-order principal minor of Σ, and $R^\star$ gathers the first three rows of R.
3. Compute the matrix A in equations (8) by imposing the metric constraints (see [8] for details).
4. Compute the motion matrix M and the shape matrix S as
$$\begin{aligned} M &= \hat{M}A^{-1} \\ S &= A\hat{S} \ . \end{aligned}$$

An Experiment

The experiment described in this section illustrates the rank principle and demonstrates the good quality of the results.

The conclusion drawn from this and other experiments is that the relative errors in the computed shape are of the same order as the relative depth range, which we defined as the ratio of the object size along the optical axis and the distance between camera and object.

We put a one-dollar coin (about 4 cm in diameter) approximately 3.5 meters away from a Sony CCD camera with a 300 mm Tokina lens. Thus, the relative depth range was $4/350 \approx 0.011$.

The camera was moved in the plane of the coin, so that only the edge of the coin was visible in every frame. The motion was roughly circular around a point in the vicinity of the coin. The rotation component was controlled with an accurate positioning mechanism, so that precise ground truth was available for performance evaluation. Translation was such as to keep the coin in the field of view, but was otherwise uncontrolled.

The edge of the coin was approximately aligned with the image scanlines, thus yielding easy-to-track image features (the thin vertical notches on the coin's edge). The first 101 frames were taken in steps of 0.1 degrees of camera rotation. After that, 100 more frames were taken at 0.2 degrees per frame. Thus, the overall rotation was 30 degrees. The resulting 201 scanlines are stacked together in figure 1, top to bottom.

The image was filtered with a thirteen-tap finite-impulse-response approximation to the Laplacian of a Gaussian, and

the 104 zero crossings of the result, shown in figure 2, were used as features in the experiment.

The measurement matrix was thus 201 × 104 in size. Its singular values are plotted in figure 5. Notice that the first three singular values are dominant. If it were not for noise, and if the projection were exactly orthographic, the remaining values would be zero.

Figure 3 shows the computed and the true rotation. The difference between the two graphs, hardly visible in figure 3, is enlarged in figure 4. The error is always smaller than one tenth of one degree. The algorithm assumes no motion model, and does no smoothing. As a result, the sharp change in rotational velocity after frame 100 is faithfully preserved in the motion output.

Figure 6 shows the shape results, and the best circular fit to them. The difference between computed and true shape is enlarged in figure 7. The accuracy of shape is of the order of the relative depth range (1 percent), even if variations in depth during the motion of the camera were of the order of the coin size.

Conclusions

The algorithm presented in this paper infers the shape of remote objects and the motion of the camera. It is a *shape* algorithm. It does not compute the depth of the scene.

Algorithms such as the ones described in [4], [6], [9], on the other hand, represent depth explicitly, and compute it from the image sequence. They are *depth* algorithms.

Depth algorithms do not work if objects are very distant from the camera with respect to their size. When the relative depth range is very small, as for instance in aerial cartography and reconnaissance, the completeness of depth algorithms is not only useless, but harmful.

A shape algorithm gives a more stable and accurate answer, because it computes shape and camera motion directly from image deformations, without using depth as an intermediate step.

References

[1] R. C. Bolles, H. H. Baker, D. H. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *IJCV*, 1(1):7–55, 1987.

[2] G. E. Forsythe, M. Malcolm, C. B. Moler. *Computer Methods for Mathematical Computations*. Prentice-Hall, 1977.

[3] G. H. Golub, C. Reinsch. Singular Value Decomposition and Least Squares Solutions. In *Handbook for Automatic Computation* (2):134–151. Springer Verlag, 1971.

[4] J. Heel. Dynamic motion vision. *DARPA IU Workshop*, 702–713, Palo Alto, CA, 1989.

[5] L. Matthies, T. Kanade, R. Szeliski. Kalman filter-based algorithms for estimating depth from image sequences. *IJCV*, 3(3):209–236, 1989.

[6] M. E. Spetsakis, J. Aloimonos. Optimal motion estimation. *IEEE Workshop on Visual Motion*, 229–237, Irvine, CA, 1989.

[7] E. H. Thompson. A rational algebraic formulation of the problem of relative orientation. *Photogrammetric Record*, 3(14):152–159, 1959.

[8] C. Tomasi, T. Kanade. Shape and motion from image sequences: a factorization method - I - The two-dimensional case. Technical report CMU-CS-90-166, Carnegie Mellon University, Pittsburgh, Pa, 1990.

[9] R. Y. Tsai, T. S. Huang. Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces. *PAMI*, 6(1):13–27, 1984.

[10] S. Ullman. *The Interpretation of Visual Motion*. MIT Press, 1979.

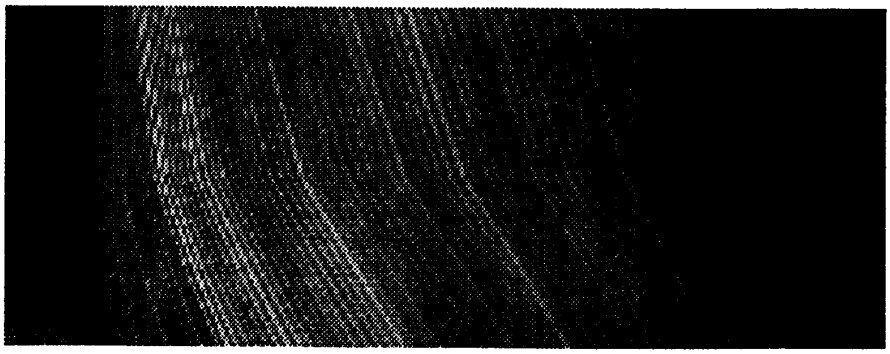

Figure 1: The input to the algorithm; each scanline is a new frame. In [1], this is called an epipolar plane. We use it to recover shape and rotation, rather than depth given known motion.

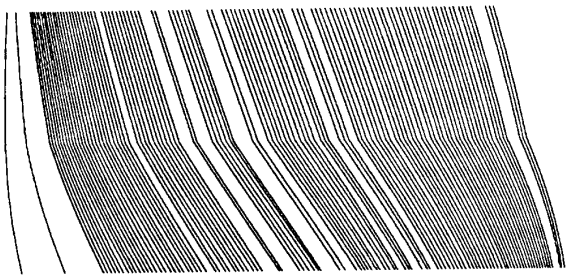

Figure 2: The zero crossings of figure 1 filtered with a Laplacian of a Gaussian.

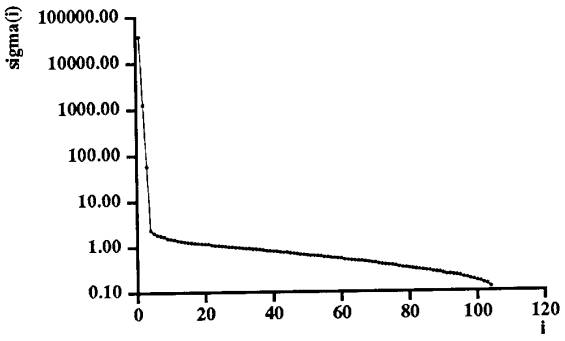

Figure 5: Singular values of the measurement matrix. Notice the logarithmic scale along the ordinate axis.

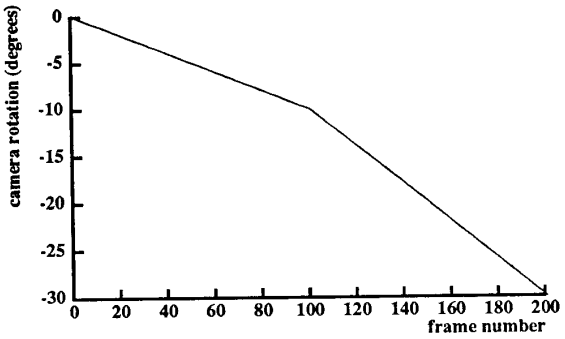

Figure 3: Camera rotation. Computed (solid) and true (dashed) rotation are so close that they can hardly be distinguished (see also figure 4 below).

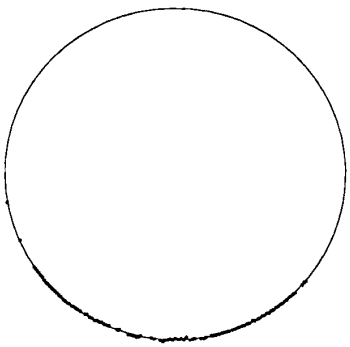

Figure 6: Shape. One hundred points along the edge of a one-dollar coin, as computed by our algorithm (dots), compared to the best fit circle (see also figure 7 below).

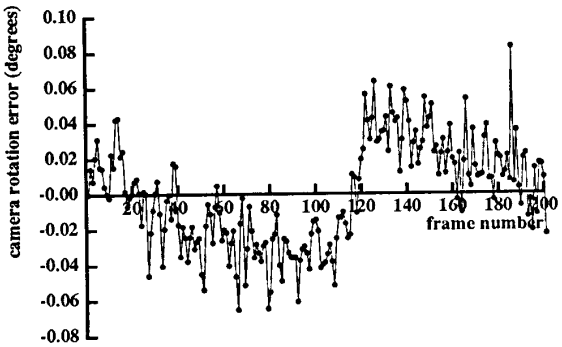

Figure 4: Camera rotation – detail. Blow-up of the difference between the two graphs of figure 3.

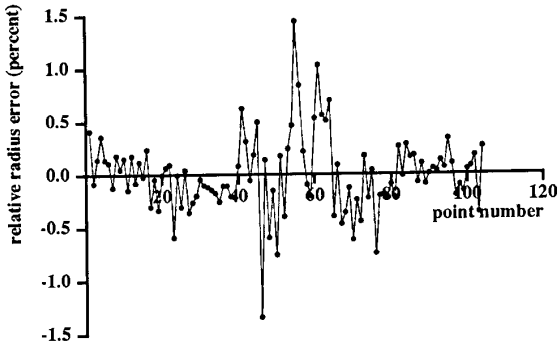

Figure 7: Shape – detail. Blow-up of the radial distance between the dots and the circle of figure 6.