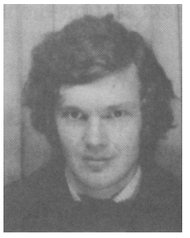
[5] B. K. P. Horn, R. J. Woodham, and W. M. Silver, "Determining shape and reflectance using multiple images," Massachusetts Inst. Technol., Cambridge, AI-Memo. 490, 1978.
[6] R. Jain and H.-H. Nagel, "On the analysis of accumulative difference pictures from image sequences of real world scenes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-1, pp. 206-214, Apr. 1979.
[7] J. Potter, "Scene segmentation by velocity measurements obtained with a cross-shaped template," in *Proc. Int. Joint Conf. Artificial Intell.*, 1975, pp. 803-810.
[8] B. Radig, "Description of moving objects based on parameterized region extraction," in *Proc. Int. Joint Conf. Pattern Recognition*, 1978, pp. 723-725.
[9] D. Waltz, "Understanding line drawings of scenes with shadows," in *The Psychology of Computer Vision*, Winston, Ed. New York: McGraw-Hill, 1975, pp. 19-92.
[10] S. Ullman, *The Interpretation of Visual Motion.* Cambridge, MA: M.I.T. Press, 1979.



Bernd Neumann was born in 1943 in Lüneburg, Germany. He studied electrical engineering in Berlin and Darmstadt and received the diploma in 1967, and was awarded an ESRO/NASA Fellowship for studies in information theory and communications at the Massachusetts Institute of Technology, Cambridge, where he received the M.S. degree in 1968 and the Ph.D. degree in 1971.

During his time at M.I.T. he became increasingly interested in information processing and artificial intelligence. Returning to Germany, he joined the faculty of the Universitaet Hamburg for teaching and research in computer science. As a member of the Vision Group headed by Prof. H.-H. Nagel, he took part in developing the group's experimental system for scene sequence analysis. His research field is image processing and scene analysis, with special interest in early visual processing, motion analysis, and industrial applications.

Determining the Movement of Objects from a Sequence of Images

JOHN W. ROACH, MEMBER, IEEE, AND J. K. AGGARWAL, FELLOW, IEEE

Abstract—**This paper discusses the problem of determining the three-dimensional model and movement of an object from a sequence of two-dimensional images. A solution to this problem depends on solving a system of nonlinear equations using a modified least-squared error method. Two views of six points or three views of four points are needed to provide an overdetermined set of equations when the images are noisy. It is shown, however, that this numerical method is not very accurate unless the images of considerably more points are used.**

Index Terms—**Motion analysis, moving objects, three-dimensional motion analysis.**

INTRODUCTION

COMPUTER image analysis beginning with Roberts' classic paper [20] has concentrated on segmentation, object recognition, and the mathematical analysis required to determine an object's three-dimensional position. The analysis of image sequences of moving objects has received some attention, almost entirely directed to the analysis of the two-dimensional movement of objects. The original motivation for studying two-dimensional motion came from a desire to analyze with a computer the vast quantity of satellite images of clouds (see Endlich *et al.* [8] and Leese *et al.* [13]). An abstract model of cloud movement was examined in Aggarwal and Duda [1]; this work was extended to the analysis of the two-dimensional movement of curvilinear shapes in Chow and Aggarwal [5] and Martin and Aggarwal [16]. In other work on motion analysis, Jain and Nagel [12] used difference pictures to analyze street scene images. Although the images show pedestrians and cars moving in three dimensions, there is no attempt to recover three-dimensional information from the images.

It is clear that past research has mainly been concerned with two-dimensional motion. In part, this is because the interpretation of images of objects moving in three-dimensions is much more complicated than two-dimensional motion since rotation and movement in depth are difficult to analyze. For example, rotation in space is defined to be about a line in three-dimensional space whereas rotation in a plane is defined to be about a single point in the plane. In addition, parts of an object can disappear from view as a result of rotation in space; rotation in a plane does not by itself cause an object to occlude itself. In this paper we shall examine and solve problems involved in determining the three-dimensional motion of objects from a sequence of two-dimensional images.

Analyzing the three-dimensional motion of an object from two-dimensional images requires a mathematical formalization. Psychologists have classically studied movement in terms of texture gradients and other cues that aid human depth perception (see Gibson [11] and Braunstein [3]). These psycholo-

Manuscript received October 1, 1979; revised March 23, 1980. This work was supported in part by the Air Force Office of Scientific Research under Grant AFOSR 77-3190.

J. W. Roach is with the Department of Computer Sciences, University of Texas, Austin, TX 78712.

J. K. Aggarwal is with the Department of Electrical Engineering, University of Texas, Austin, TX 78712.

0162-8828/80/1100-0554$00.75

gists use images of points on the surfaces of objects to study the movement in depth of the objects; Roberts [20] also uses surface points to help determine object depth. By definition, the three-dimensional relationship of points on a rigid object does not change over time. Consequently, changes in the two-dimensional spatial relationship of object surface points between images must be caused by a relative movement between the camera and the object being imaged. In studies involving binocular vision (two cameras or eyes spaced a known distance apart), this change of position of a point between the two images is known as the "disparity" in the images of the point. A simple triangulation argument gives the depth of the point in this case. Thus, the change of position between images of points on a rigid object's surface can be used to formalize the problem of determining the three-dimensional movement of objects in space.

Consider the sequence of images of a moving object given in Fig. 1. This sequence shows a truncated wedge rotating and translating. These simplified line drawings gloss over the difficult low-level processing problems, such as separating out the various objects in an image (segmentation) and extracting the same feature points (or tokens) on an object in each image despite possibly changing illumination conditions. This paper does not address these low-level processing problems; instead, we assume that the points are given to us *a priori*. In images of the real world, however, we know of no general solution to finding the same points on an object's surface in each image.

Once feature points on an object's surface have been extracted in each image, we must determine the correspondence of points between consecutive images. By "correspondence" here we mean the mapping that takes an image of an object point to the image of the same object point in the next image of the object. This is a difficult problem since an image may have more than one moving object and thus many points to choose from. The correspondence problem is further complicated by the disappearance of points on an object due to occlusion from other objects, self-occlusion as points rotate out of view, shadows, etc. The correspondence problem has been examined by Quam *et al.* [18], Ganapathy [10], and Ullman [23]. We shall assume in this paper that the correspondence of points between images is known.

Once the correspondence of points has been established, we can attempt to analyze the motion. Here we are confronted with a basic question: how is the motion to be represented? One method might involve qualitative descriptions such as "moving left and away, rotating to the right," etc. as in Badler [2]. There still remains the question, however, of how the qualitative description is to be calculated. If a more exact mathematical calculation were used to derive the qualitative description, then the more precise mathematical result would also be a valuable characterization. In this paper, we shall take the more quantitative approach and use matrices with "homogeneous coordinates," as Roberts [20] did, to describe the movements of objects. Homogeneous coordinates are an elegant means of representing movement since a 4×4 matrix can represent any rotation or translation. We have to be sure, however, that given these two elementary motion matrices, it is possible to represent any motion an object can have in space. For example, consider a planet traveling about the sun. It is revolving about an axis passing through the sun; at the same time it is rotating about its own polar axis. How is this two-axis rotating movement to be analyzed? What if there are n axes of rotation? The translation and rotation 4×4 matrices are an adequate representation since a theorem from classical mechanics (see Coffin [6]) establishes that any motion, including the rotation within rotation problem just mentioned, can be decomposed into one rotation and one translation. The rotation and translation matrices can be multiplied together to form one matrix useful for predicting the next three-dimensional position of the object. In addition to representing motion, a 4×4 matrix in homogeneous coordinates can be used to model the projection of object points onto the focal plane of the camera.

Now that we have formalized the problem and decided on a means for representing movement we can ask some fundamental questions about analyzing sequences of images of moving objects:

1) whether the multiple images, and thus object motion, help the three-dimensional analysis; and

2) exactly how much of the original three-dimensional information can be recovered from a sequence of two-dimensional images.

The first question is essentially concerned with determining what the value of motion is in analyzing images. The second question concerns the three-dimensional relationship of points on the surface of an object and the entries in a 4×4 matrix that represents the movement of those points. We shall answer these questions in the course of this paper.

Our analysis will depend on certain key assumptions. For example, we assume throughout that all images are from one camera. We assume that there is no *a priori* knowledge of specific objects or their specific motions, that objects in general are rigid, that motion is smooth and continuous, and that central projection is the best geometrical description of the image formation process. By changing this last assumption to parallel projection, an exact model (to within a reflection) for any four noncoplanar points can be derived given three different views of them, as Ullman [23] has shown. Badler [2] used a spherical projection model and was able to predict the point positions in succeeding images of translating objects. We have assumed that central projection is the best model for how cameras work for several reasons: it is the model used in photogrammetry [22] to create maps from real world aerial photographs; and furthermore, it has been the model normally adopted by researchers in image analysis, as in camera calibration; for example, see Sobel [21] or Yakimovsky and Cunningham [25].

The problems in determining the movement of an object from its images are similar in many ways to the problems encountered in optic flow analysis and stereopsis. Stereopsis, or binocular vision, is the problem of determining the depth of objects from two different images. The distance between the two imaging devices is assumed to be known. Optic flow analysis, originated by Gibson [11], depends on a vector field formed by points on object surfaces as a camera or eye moves through its environment. Some recent work by Williams [24]

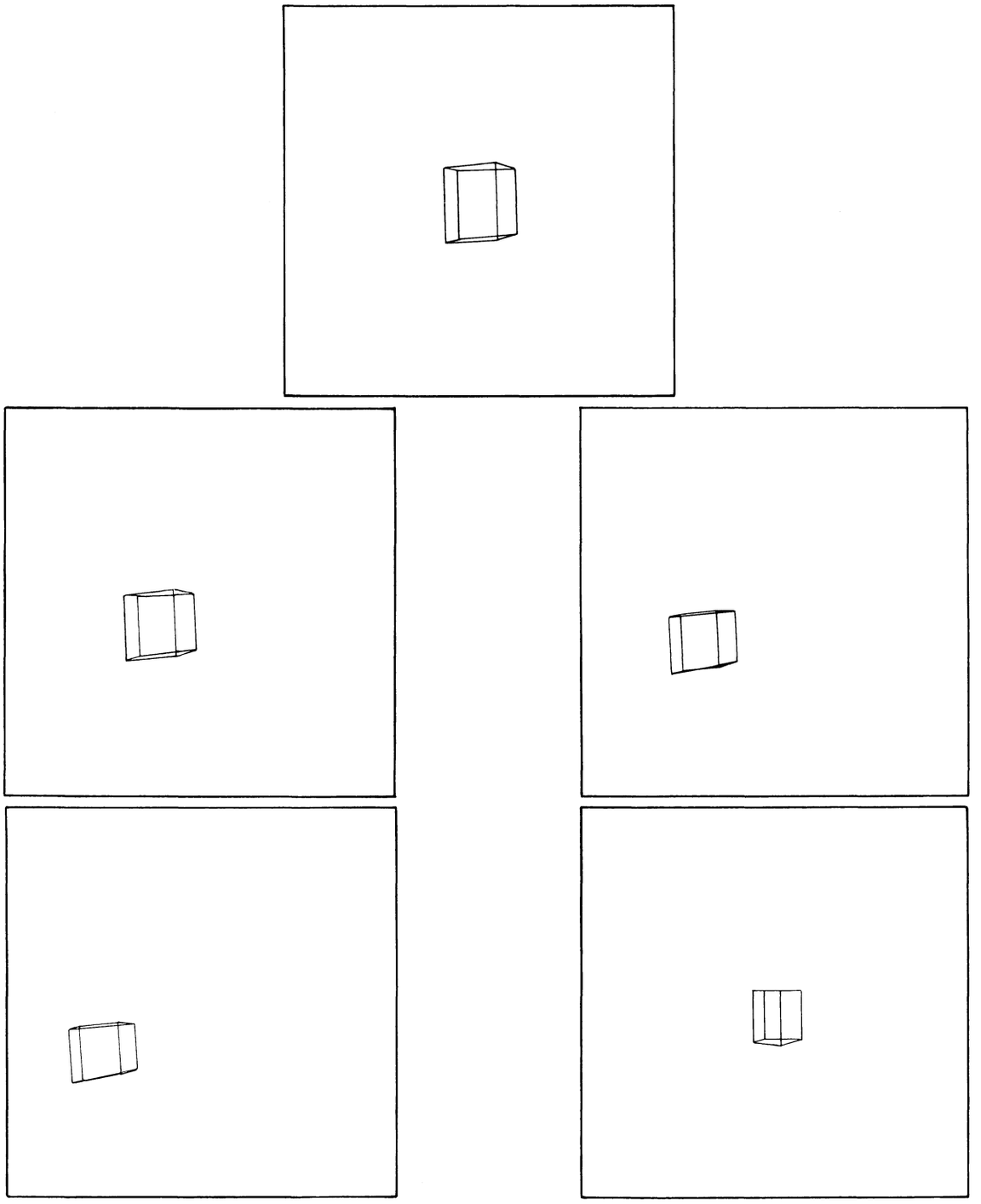


Fig. 1. Images of a truncated wedge moving in space.

concentrates on deriving the "focus of expansion" which lies along the line of sight of the translating camera. In all of these problems separate views of an object are given in which the correspondence of points on the surface of objects must be determined and the displacement of these points used to determine the distance from the camera to the object. There are some differences from our work; for example, binocular studies normally assume that the distance between cameras is known, and optic flow studies assume a moving camera. The problems encountered, nonetheless, are very similar, and the results of this study should be applicable to these other areas of research.

Projection Equations and Their Inverses

Roberts [20] defined scene analysis as the necessary mathematical analysis to locate objects in three-dimensional space from their two-dimensional images. We want to know whether it is possible to determine the position in space and movement (translation and rotation) of an object relative to a fixed three-dimensional coordinate system given the image coordinates of projected object points in a sequence of images. We assume that the images have been segmented, that the same points have been extracted from each image, and that the correspon-

dence of points between images has been determined. In this section, the mathematical equations are given that determine what three-dimensional information can be derived from the projective coordinates of a point. The equations of central projection are also given.

The image of a point under central projection is a function of the point's three-dimensional position, the focal length of the camera, and the location and orientation of the camera's lens relative to the global coordinate system. Information about the camera's position is needed to relate the position of points given in two-dimensional camera coordinates to the global three-dimensional coordinate system. The necessary camera information is the camera focal length F, the orientation angles θ, ϕ, and κ of the camera to the global coordinate system, and the three-dimensional coordinates of the lens center (X_0, Y_0, Z_0). The three angles orient the camera to the global coordinate system as follows (assume for simplicity that the camera lens has been translated to (0, 0, 0) of the global coordinate system): θ is a rotation about the X-axis that brings the optical axis into the X-Z plane, ϕ is a rotation about the Y-axis so that the optical axis is aligned with the Z-axis, and κ is a rotation about the Z-axis so that the x', y' axes of the focal plane are aligned with the global X, Y axes. (Note: the use of primes in this paper in general denotes the focal plane coordinate system.) It is, of course, impossible to determine the original (x, y, z) position of a point from its picture alone. The best we can do is determine a line in space on which the point falls. The following equations are a function of the camera parameters, two-dimensional coordinates (x', y') of the image point, and a free variable z':

$$\begin{aligned} x &= X_0 + F/(F - z')\,(a_{11}x' + a_{21}y' + a_{31}F) \\ y &= Y_0 + F/(F - z')\,(a_{12}x' + a_{22}y' + a_{32}F) \\ z &= Z_0 + F/(F - z')\,(a_{13}x' + a_{23}y' + a_{33}F) \end{aligned} \tag{1}$$

where

$$\begin{aligned} a_{11} &= \cos\phi\cos\kappa \\ a_{12} &= \sin\theta\sin\phi\cos\kappa + \cos\theta\sin\kappa \\ a_{13} &= -\cos\theta\sin\phi\cos\kappa + \sin\theta\sin\kappa \\ a_{21} &= -\cos\phi\sin\kappa \\ a_{22} &= -\sin\theta\sin\phi\sin\kappa + \cos\theta\cos\kappa \\ a_{23} &= \cos\theta\sin\phi\sin\kappa + \sin\theta\cos\kappa \\ a_{31} &= \sin\phi \\ a_{32} &= -\sin\theta\cos\phi \\ a_{33} &= \cos\theta\cos\phi. \end{aligned}$$

These equations (1) give a locus of points that form a straight line in space through (X_0, Y_0, Z_0) and the point (x', y') in the focal plane; each point on the line is determined by a specific value of the free parameter z'. As z' approaches $-\infty$, $F/(F - z')$ approaches 0 and $(x, y, z) = (X_0, Y_0, Z_0)$; if z' is zero, then (x, y, z) is (x', y') given in global coordinates. The consequence of these calculations is that given only the image coordinates of a point on an object we cannot recover the full three-dimensional information about the object point. The best we can do is find a parameterized equation for the ray on which the point falls.

In addition to the equations (1), we also need equations that tell what the image coordinates (x', y') of a point (x, y, z) will be for given camera parameters. The equations are

$$\begin{aligned} x' &= F\frac{a_{11}(x - X_0) + a_{12}(y - Y_0) + a_{13}(z - Z_0)}{a_{31}(x - X_0) + a_{32}(y - Y_0) + a_{33}(z - Z_0)} \\ y' &= F\frac{a_{21}(x - X_0) + a_{22}(y - Y_0) + a_{23}(z - Z_0)}{a_{31}(x - X_0) + a_{32}(y - Y_0) + a_{33}(z - Z_0)}. \end{aligned} \tag{2}$$

Further explanations of the equations in this section may be found in [7], [17], and [22].

Finding the Movement of an Object from a Sequence of Noise-Free Images

We want to know how much of the original three-dimensional information can be recovered given only the images of a moving object. It is possible to show that any sequence of images is inherently ambiguous. That is, there are an infinite number of objects that produce the same sequence of images. The objects are all similar in structure and movement. An example should make this clear. Assume that an object is translating at velocity (p, q, r) and rotating with velocity ω about an axis oriented by angles α and β to the global coordinate system, and point (a, b, c) is on the rotational axis. Assume that a camera has taken a sequence of pictures as the object moved through its field of view. Now the same sequence of images can be produced by another object if it is constructed in the following manner. The new object is twice as far away as the original object and twice as large (from the camera's point of view) so that it gives the same initial image. Let its translational velocity be $(2p, 2q, 2r)$, its rotational velocity be ω, its axis of rotation have orientation angles α and β, and let its point on the axis of rotation be twice as far away (from the camera's point of view) as the point (a, b, c). In general, the scaling factor need not be two, it can be any real number greater than 0 (if it equals one, the new object coincides with the original object). Note that the angular information concerning rotation is not ambiguous. In this section we show how to find the movement and three-dimensional model of points on an object's surface from a sequence of noise-free images up to a scaling factor; that is, by setting the scaling factor to an arbitrary value we can find a particular movement and model for points on the object.

We have assumed that the camera is stationary and the object moving. It is convenient to reformulate the problem such that the object is stationary and the camera moves.

To solve the problem, two views are needed of five points not all in the same plane. The global coordinates of each point are variable, so five points produce 15 variables. The global coordinates and θ, ϕ, κ orientation angles for each camera position are also variable producing 12 more variables. Thus, there are a total of 27 variables in the problem. Each point produces two projection equations [given by (2)] per camera position for a total of 20 nonlinear equations. To make the number of equations and unknowns come out even, seven variables must be known including one variable that will determine the scaling factor. We shall now examine several different ways of setting the seven variables correctly and use the one that is most easily solved. In one problem setup, the three-dimensional coordinates of two points and one of the three coordinates of another point are known: (0, 0, 0), (1, 0, 0), and (?, ?, 0). The

points (0, 0, 0) and (1, 0, 0) fix the X-axis and also the scaling factor; the coordinate system can be rotated about this X-axis until the third point lies in the X-Y plane, thus setting its third coordinate to zero. Although this problem setup is correct, it is difficult to solve numerically since good original estimates for unknown variables (especially camera orientation angles) cannot be determined. Another correct setup for this problem sets all coordinates and orientation angles of the first camera to a value of zero. This sets six variables. The seventh variable and the scaling factor are set by letting the x-coordinate of the second camera be an arbitrary constant (equal to 1.0, say). *Reasonable estimates can be made for the unknown variables,* but the difficulty with this problem setup is that in some special cases the x-coordinate of the second camera should be zero. In these cases, setting the x-coordinate to a nonzero constant is incorrect. In fact, it is possible that the x, y, z coordinates of the second camera are all zero (no camera translation) in which case a different means of setting the scaling factor must be sought. Thus, we need a different formulation for the problem which will now be explained.

We will set the X_0, Y_0, Z_0 position and θ, ϕ, κ orientation angles of the first camera by making all six variables equal to zero. In addition, the z-component of any one of the five points is set to an arbitrary positive constant. We showed earlier that the best result possible in locating the three-dimensional position of a point on an object is to find (sx, sy, sz) where s is an arbitrary scaling factor. By setting the z-component of the position of a point to an arbitrary constant, we are fixing the scaling factor. Once the z-component of a point is known, the x and y components can also be found using the inverse of the projection equations (2) (by determining z' from z and the focal length). The situation is shown in Fig. 2. There are now 18 projection equations in 18 unknowns (actually, there are 20 equations, but two of them have no unknowns); the equations of projection, however, are nonlinear.

Unlike linear equations, there is no developed systematic theory for solving systems of simultaneous nonlinear equations. Finding a closed-form solution for any system of nonlinear equations is rare. Consequently, most nonlinear systems are solved using numerical methods developed to achieve approximate answers. Numerical methods sometimes fail because either they do not converge or converge to the wrong answer. All numerical methods require an initial guess for the unknown parameters. How good the initial guess is normally determines whether the numerical method converges to the correct answer. We have found that the system of nonlinear projection equations explained above can be solved by using a modified finite difference Levenberg-Marquardt algorithm due to Brown [4], [14], [15] without strict descent that minimizes the least-squared error of the 18 equations. This routine is available under IMSL as ZXSSQ [26]. It is a modification of the classical least-squared error technique originated by Gauss at the end of the eighteenth century. Brown's method performs a smoothing operation not available with other techniques which normally permits convergence to the correct answer.

The method employed is iterative and requires an initial guess for each unknown parameter. If we assume that the camera is taking snapshots rapidly, then its position will change

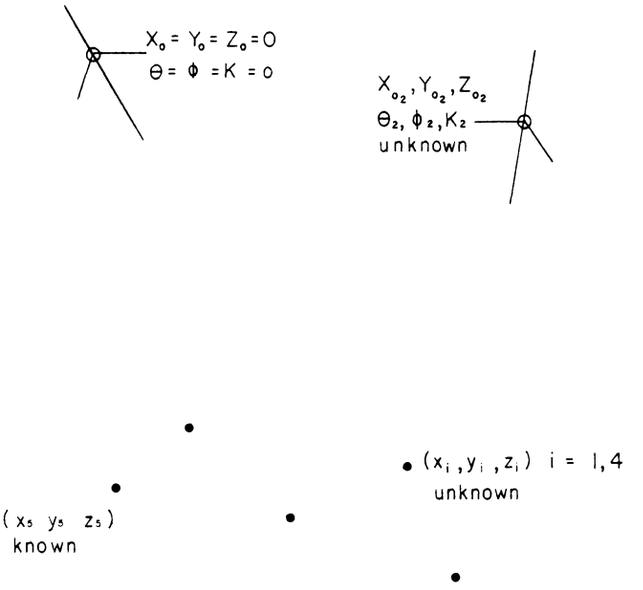


Fig. 2. Problem setup with moving camera.

only slightly between photographs. In particular, we can make the simplifying assumption that $\theta = \phi = \kappa = 0$ for the second camera position and use simple parallax differences to get reasonable initial estimates. The parallax equations are

$$X\text{PARALLAX} = x_1' - x_2'$$

$$X = X_{0_2} \cdot (x_1'/X\text{PARALLAX})$$

$$Y = X_{0_2} \cdot (y_1'/X\text{PARALLAX})$$

$$Z = (X_{0_2} \cdot F)/X\text{PARALLAX}$$

where X, Y, Z are the three-dimensional coordinates of the point to be estimated; x_1', x_2', and y_1' are the x and y photo coordinates from camera images one and two; F is the focal length; and X_{0_2} is the x-coordinate of the second camera position. X_{0_2} is unknown and must itself be estimated before the parallax equations can be applied. This is easily achieved since the three-dimensional coordinates of one point are known: the point for which the z-coordinate is set to an arbitrary constant. Using the known coordinates of this reference point and the parallax equations gives several somewhat different estimates of X_{0_2}.

The possibility of several different values for X_{0_2}, and hence different original estimates for all the points, raises the question of determining which initial estimate is better. Indeed, what is the best means of selecting which of the five object points is to be the reference point whose three-dimensional coordinates are known? We allow each point in turn to become the reference point and then for each guess of X_{0_2} estimate the three-dimensional position of each of the other four points as well as Y_{0_2} and Z_{0_2} (using the inverse of the central projection equations). The least-squared error of this set of estimates is calculated by substituting the estimates for unknown variables in the equations for central projection (assuming for camera two that $\theta_2 = \phi_2 = \kappa_2 = 0.0^\circ$) and the answers differenced with the observed projective coordinates of the points. The errors are all squared and added together; the set of esti-

mates with the least-squared error is taken as the best initial guess. There are only 15 sets of initial guesses so this preprocessing time is quite small, and without it, the initial estimate is not always good enough to allow convergence to the right answer.

Brown's method without strict descent was very successful in converging to the correct answer in a large number of trials with analytically correct data. (The data were generated by a computer program that takes as input the three-dimensional position of points on an object, the movement of the object, and the position of the camera; the program moves the object and mathematically projects an image at prescribed time intervals. The photocoordinates are used to ten places of accuracy.) The method failed to find the exact answer only in cases where a moving object was rotating (no translation) and the axis of rotation passed through the lens of the camera. The answer that this numerical method computes gives both the object's movement since the movement of the camera is calculated and the three-dimensional model for the points on the object. By negating all values of the solution, a second solution is attained. This solution amounts to setting the z coordinate of the reference point to the negation of the original arbitrary scalar. The method typically converges to the correct answer in 15 s on a Cyber 170/75 and hence is reasonably efficient.

It should be noted that a large number of other methods were tried on this system of equations: Brown's algorithm with strict descent, Newton's method, normal least-squared error, Fletcher-Powell [9], and several other methods designed to deal with simultaneous nonlinear equations. Newton's method and least squares were also used with many different step lengths. Some of these methods converge to answers that are close to correct, but others do not converge at all. Only Brown's method without strict descent regularly converges to the correct answer.

Our work is somewhat like the camera calibration systems of Sobel [21] and Yakimovsky and Cunningham [25]. In their work multiple images of points together with a central projection model and numerical methods are used to determine camera parameters such as focal length, position, and orientation. These studies, however, have considerably more information about the three-dimensional positions of points than we are assuming. Thus, the problems being solved and the information given for the calibration systems are different from our work.

Finding Answers from Noisy Images

We have been implicitly making two very important assumptions: that the objects being observed are rigid and that the images of the object are noise free and thus completely accurate. The first assumption is an important restriction since the problem solution presented does not work with images of moving, highly nonrigid objects. The second assumption is not reasonable. In this section we shall explore the problems introduced when the images are noisy.

In effect, when noise is added to the images of the object it is equivalent to taking perfect photos of an object that is not quite rigid—jello-like is an apt metaphor. To test the effect of sensor noise and digitization error on the numerical method described in the previous section, from one to four pixels were randomly added or subtracted from the exact photocoordinate data for a moving object.

One of the main reasons for using a least-squared error technique to solve a problem is to make adjustments to observations that contain error (noise). Adjustment is only possible, however, when there are more equations than unknowns. Two views of five points are therefore inadequate for noisy data since there are the same number of equations as unknowns. Two views of six points or three views of four points produce 22 equations in 21 unknowns using the same problem model discussed in the previous section. Examination of experimental runs using overdetermined systems of equations shows that minimal overdetermination is not very accurate. It is only with considerable overdetermination (two views of 12 or even 15 points; three views of seven or eight points) that the results become accurate. Clearly, attaining good accuracy depends on considerable overdetermination. It should be noted that in some examples we tried, Brown's method without strict descent would not converge using a reference point that produced the minimum squared error from its original estimate. By trying a different reference point, and thus a different initial estimate, that produced a small squared error, convergence to a satisfactory result was achieved. The Appendix has a graphical comparison of some of the overdetermined experiments. For the two views case, the model of the object improves considerably, and the camera position improves somewhat as the number of points increases. For three views, the opposite effect seems to hold: the camera positions' accuracy improves considerably and the model points' positions (originally fairly good) improves somewhat.

In addition to synthetic data, an experiment was run using laboratory images, 108 X 108, from a rather noisy image dissector camera. The images used appear in Roach and Aggarwal [19]. In general, these images contain too few points to assure accurate results. The images of three different objects were used with three views of four points and three views of five points. The results verify our findings that too few points result in an inaccurate answer.

Conclusions

This work is directly related to the problems encountered by researchers in optic flow studies, as mentioned in the introduction, although the findings here have no bearing on finding the "focus of expansion" for a translating camera. Gibson's texture gradients [11], however, seem to be very much related to the work presented here.

Roberts [20] originally used least-squares analysis together with simple models and a support assumption to locate objects in space given one image of six points on the object. This study shows that it is possible with more than one view to dispense with the model. It is possible, in fact, to determine the three-dimensional relationship of the points on the object as well as its movement (up to a scale factor) from the multiple views. This is also the conclusion of Ullman's recently available dissertation [23] which uses a novel closed-form solution for restricted motions of objects from noise-free central projection images. For unrestricted motion, his method requires an im-

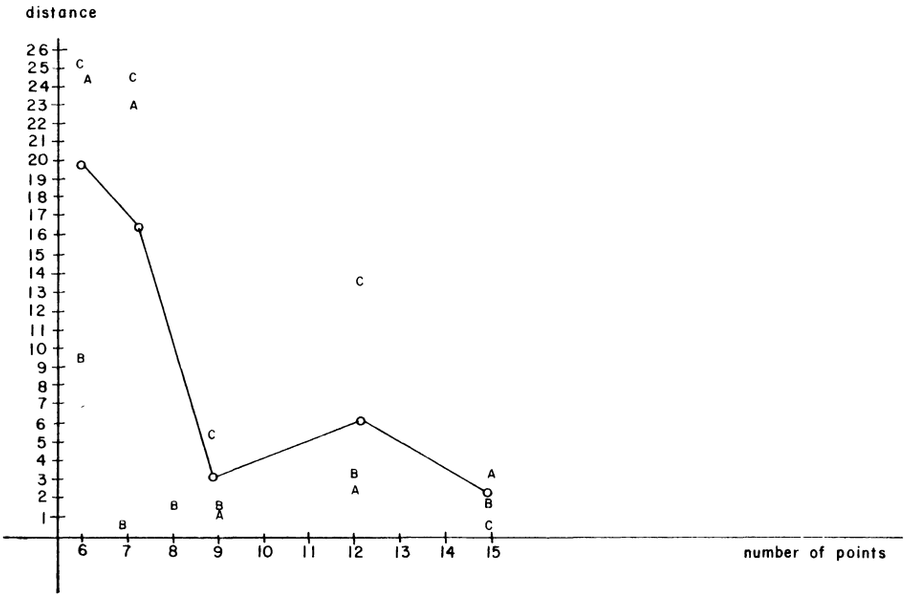


Fig. 3. Average distance from computed model point to correct model point versus number of points.

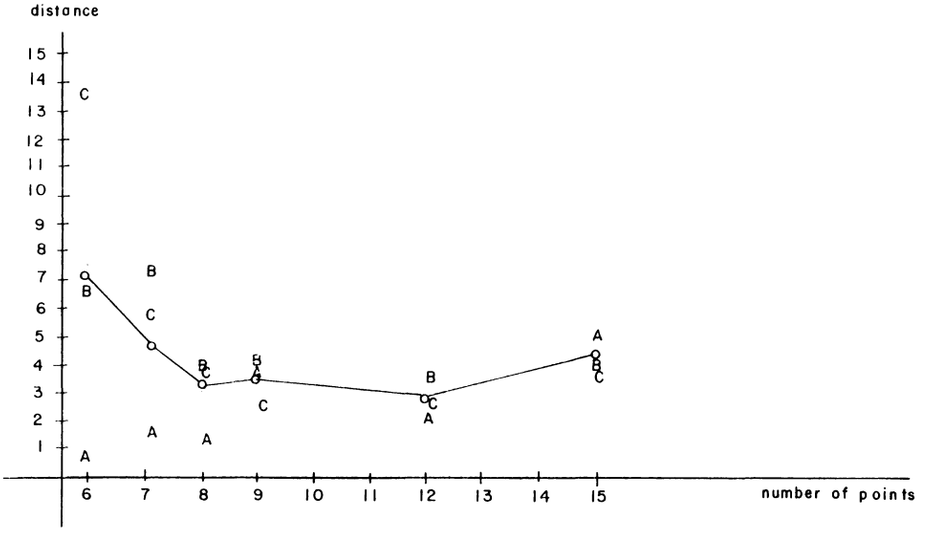


Fig. 4. Average distance from computed camera position to correct camera position versus number of points.

practical computation in a very large (infinite) search space. Another problem with central projection according to Ullman is that perspective effects are often small (especially for small objects), and thus noise makes central projection an unsuitable model for determining three-dimensional structure and movement. Obviously, no method can succeed in the case where noise effects overwhelm image changes due to motion. We have shown in this paper that given noisy central projection images, the movement and three-dimensional relationship of points can be attained only when there are considerably more equations than unknowns. We have assumed that noise does not overwhelm image changes caused by the movement of objects. We chose to use synthetic data originally to ensure that the numerical model adopted converges to the correct answer and later as a control over the accuracy of the answer when noise was added to the data. The need for considerable overdetermination and thus the images of many points reemphasizes a problem that no one has solved since Roberts' paper: how to find the correct tokens on the surface of the object in images (in our case, the same tokens in every image). Note that we are *not* referring to the well known correspondence problem. Without the ability to determine reliably the same feature points in each image, the whole analysis scheme fails. Finding points on blocks as in figure one is easy; finding identical points in each image of a sequence from the real world is considerably more difficult. Future research will have to devise a reliable low-level processing solution for this problem.

Appendix

Here we present graphical evidence that considerable overdetermination improves the computation of both the model of the object and the camera position.

Figs. 3, 4 and 5 show graphs derived from the three experiments, labeled *A*, *B*, and *C*, run with two views of a varying

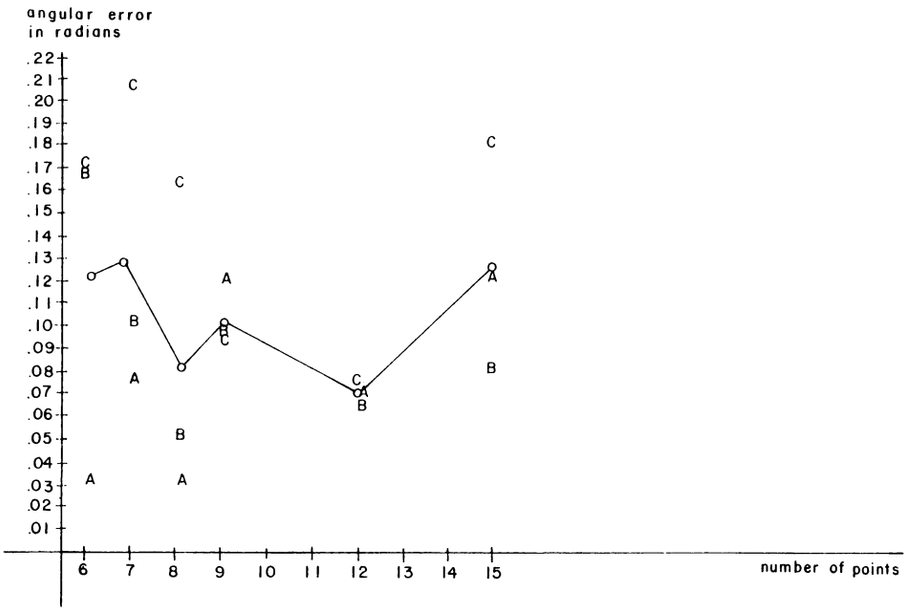


Fig. 5. Average angular error given in radians versus number of points.

number of points. Each experiment was run using a different object that had its own movement. The z-distance to the reference point in each example was set to 10. The typical distance to the reference point, therefore, was on the order of 15-20 units. Three sets of statistics were kept for each experiment: average distance from computed model points to correct model points, distance from the computed camera position to the correct camera position, and average angular error of θ, ϕ, and κ.

Each graph shows an average value for each of the three experiments. The "A" token denotes the value for the A experiment, etc. In addition, to indicate the trend of the data, the average of the three experiments is connected by a solid line in the graphs.

The first graph shows the average distance between the model points for the computed answer given noisy data and the correct answer assuming no noise. For the 6, 7, or 8 point cases, there is one point whose computed coordinates are extremely poor. Note that there is no average distance for the A or C experiments with eight points since their averages were too large to fit on the graph. In general, two views of fewer than nine points result in a numerically unstable model of an object.

The second graph shows the improvement in the second camera position as the number of points increases. The improvement does not appear to be very great mainly due to the unusually good camera position for the A experiment when six or seven points were used.

The third graph shows the average error for the angles θ, ϕ, and κ. There does not seem to be any improvement here as the number of points increases. The average error is about 0.1 rad.

In conclusion, the model of the object showed a marked improvement as the number of points increased, the camera position showed a modest improvement, and the angular orientation of the camera showed little or no improvement.

Experiments with three views showed the model of the object to be fairly good with minimal overdetermination, but the camera positions were poor until there were three views of at least seven points.

References

[1] J. K. Aggarwal and R. O. Duda, "Computer analysis of moving polygonal images," *IEEE Trans. Comput.*, vol. C-24, pp. 966-976, Oct. 1975.

[2] N. Badler, "Temporal scene analysis: Conceptual descriptions of object movements," Ph.D. dissertation, Univ. Toronto, Toronto, Ont., Canada, TR80, 1975.

[3] M. L. Braunstein, *Depth Perception Through Motion*. New York: Academic, 1976.

[4] K. M. Brown and J. E. Dennis, "Derivative free analogues of the Levenberg-Marquardt and Gauss algorithms for nonlinear least squares," *Numer. Math.*, vol. 18, pp. 289-297, 1972.

[5] W. K. Chow and J. K. Aggarwal, "Computer analysis of planar curvilinear moving images," *IEEE Trans. Comput.*, vol. C-26, pp. 179-185, Feb. 1977.

[6] J. Coffin, *Vector Analysis*, 2nd ed. New York: Wiley, 1911.

[7] R. O. Duda and P. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley-Interscience.

[8] R. M. Endlich, D. E. Wolf, D. J. Hall, and A. E. Brain, "Use of a pattern recognition technique for determining cloud motions from sequences of satellite photographs," *J. Appl. Meteorol.*, vol. 10, pp. 105-117, Feb. 1971.

[9] R. Fletcher and M. J. D. Powell, "A rapidly convergent descent method for minimization," *Comput. J.*, vol. 6, pp. 163-168, 1963.

[10] S. Ganapathy, "Reconstruction of scenes containing polyhedra from stereo pair of views," Ph.D. dissertation, Stanford Univ., Stanford, CA, AIM-272, Dec. 1975.

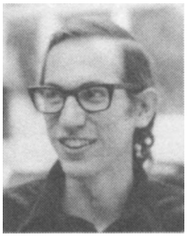
[11] J. J. Gibson, *The Perception of the Visual World*. Boston, MA: Houghton and Mifflin, 1950.

[12] R. Jain and H.-H. Nagel, "On the analysis of accumulative difference pictures from image sequences of real world scenes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-1, pp. 206-213, Apr. 1979.

[13] J. A. Leese, C. S. Novak, and V. R. Taylor, "An automated technique for obtaining cloud motion from geosynchronous satellite data using cross-correlation," *J. Appl. Meteorol.*, vol. 10, pp. 118-132, Feb. 1971.

[14] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *Quart. Appl. Math.*, vol. 2, pp. 164-168, 1944.

[15] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *J. SIAM*, vol. 11, no. 2, 1963.
[16] W. Martin and J. K. Aggarwal, "Computer analysis of dynamic scenes containing curvilinear figures," *Pattern Recognition*, vol. 11, pp. 169–178, 1979.
[17] W. M. Newman and R. F. Sproull, *Principles of Interactive Computer Graphics*. New York: McGraw-Hill, 1973.
[18] L. H. Quam and M. J. Hannah, "Stanford automatic photogrammetry research," Stanford Univ., Stanford, CA, AIM-254, Dec. 1974.
[19] J. Roach and J. K. Aggarwal, "Computer tracking of objects moving in space," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-1, pp. 127–135, Apr. 1979.
[20] L. G. Roberts, "Machine perception of three-dimensional solids," in *Computer Methods in Image Analysis*. J. K. Aggarwal, R. O. Duda, and A. Rosenfeld, Eds. New York: IEEE Press, 1977, pp. 285–323.
[21] I. Sobel, "On calibrating computer controlled cameras for perceiving 3-D scenes," *AIJ*, vol. 5, no. 2, pp. 185–198.
[22] M. M. Thompson, *Manual of Photogrammetry*, 3rd ed. Falls Church, VA: Amer. Soc. Photogrammetry, 1966.
[23] S. Ullman, "The interpretation of structure from motion," M.I.T. Artificial Intell. Lab., A.I. Memo 476, Oct. 1976.
——, *The Interpretation of Visual Motion*. Cambridge, MA: M.I.T. Press, 1979.
[24] T. D. Williams, "Region analysis for a moving scene," presented at the Workshop on Comput. Anal. of Time-Varying Imagery, Philadelphia, PA, Apr. 1979.
[25] Y. Yakimovsky and R. Cunningham, "A system for extracting three-dimensional measurements from a stereo pair of TV cameras," *Comput. Graphics Image Processing*, vol. 7, no. 2, pp. 195–210.
[26] ZXSSQ, Int. Math. Statist. Libraries, Houston, TX, version 7, 1979.



John W. Roach (S'75–M'79) was born in Boulder, CO, in 1948. He received the B.A. degree in Plan II in 1970, the M.A. degree in computer sciences in 1974, and the Ph.D. degree in computer sciences in 1980, all from the University of Texas, Austin.

As a graduate student, he taught undergraduates for four years and did research on robot planning and image processing.

J. K. Aggarwal (S'62–M'65–SM'74–F'76), for a photograph and biography, see this issue, p. 494.