International Journal of Computer Vision, 3, 33–49 (1989)

Generalizing Epipolar-Plane Image Analysis on the Spatiotemporal Surface

H. HARLYN BAKER
ROBERT C. BOLLES
Artificial Intelligence Center, SRI International, 333 Ravenswood Avenue Menlo Park, CA 94025

Abstract

The previous implementations of our *Epipolar-Plane Image Analysis* mapping technique demonstrated the feasibility and benefits of the approach, but were carried out for restricted camera geometries. The question of more general geometries made the technique's utility for autonomous navigation uncertain. We have developed a generalization of our analysis that (a) enables varying view direction, including variation over time (b) provides three-dimensional connectivity information for building coherent spatial descriptions of observed objects; and (c) operates sequentially, allowing initiation and refinement of scene feature estimates while the sensor is in motion. To implement this generalization it was necessary to develop an explicit description of the evolution of images over time. We have achieved this by building a process that creates a set of two-dimensional manifolds defined at the zeros of a three-dimensional spatiotemporal Laplacian. These manifolds represent explicitly both the spatial and temporal structure of the temporally evolving imagery, and we term them *spatiotemporal surfaces.* The surfaces are constructed incrementally, as the images are acquired. We describe a tracking mechanism that operates locally on these evolving surfaces in carrying out three-dimensional scene reconstruction.

Introduction

1.1 Epipolar-Plan Image Analysis

In an earlier publication in this journal [1], we described a sequence analysis technique developed for use in obtaining depth estimates for points in a static scene. The approach bridged the usual dichotomy of passive depth sensing in that its large number of images led to a large baseline and thus high accuracy, while rapid image sampling gave minimal change from frame to frame, eliminating the correspondence problem. Rather than choosing quite disparate views and putting features into correspondence by stereo matching, with this technique we chose to process massive amounts of similar data, but with much simpler and more robust techniques. The technique capitalized on several constraints we could impose on the image acquisition process, namely:

1. The camera moved along a linear path.
2. It acquired images at equal spacing as it moved.
3. The camera's view was orthogonal to its direction of travel.

With these constraints, we could guarantee that

1. Individual scene features would be observed in single epipolar planes over the period of scanning.
2. Images of these planes could be constructed by collecting corresponding image scan lines in successive frames.
3. The motion of scene features in these images would appear as linear tracks.

We termed these image planes *epipolar-plane images,* or EPIs, and the process *Epipolar-Plane Image Analysis.*

1.2 Problems with the Previous Approach

In that earlier paper we commented on our previous dissatisfactions with the approach, and the limitations that would restrict its usefulness. Summarizing, the limitations were

L_1 Orthogonal viewing would preclude many of the camera attitudes one would expect to be necessary for an autonomous vehicle—notably that attitude in which the vehicle is looking along its direction of motion, or when it is to track some particular feature and follow it while moving across the scene.

L_2 A constant rate of image acquisition would be difficult to guarantee, and probably not be desirable in a general context. Sampling rates will be affected heavily by computational demands on the system, and vehicle velocities may be dictated by higher-level concerns.

L_3 A linear path would be an unacceptable or highly improbable trajectory in most every situation except extended flight.

L_4 Static scenes are the *least* likely—winds blow, clouds move; often a moving object in a scene is the one of most interest.

The dissatisfactions were

D_1 The analysis should proceed sequentially as the imagery is acquired. To insist that all data be available before scene measurement can begin would eliminate one of the principal goals of the process—to provide timely information for a vehicle in motion.

D_2 The EPI partitioning, through its selection of the temporal over the spatial analysis of images, could not provide spatially coherent results. It produced point sets. We attempted clustering operations on these, but were never satisfied with such a post hoc approach. The proper approach to obtaining spatial coherence in our results would begin with not losing it in the first place.

2 New Approach to EPI Analysis

2.1 Generalizations

We have developed generalizations to our earlier approach that enable us to resolve L_1, L_2, D_1, and D_2. Arbitrary and varying camera attitudes and velocities are permissible in our new formulation, and we process the data sequentially as acquired, forming estimates, of increasing precision, descriptive of spatial contours rather than points. The generalizations also suggest a mechanism for dealing with the nonlinear path issue of L_3. Although we have not pursued this as yet, in section 3.4 we outline an approach consistent with our EPI analysis.

L_4 rises as an incompatibility between our performance desires for a vision system and our definition of the task we choose to address. We wish to build three-dimensional descriptions of scenes, and it is inappropriate to expect this to be possible if our view of the scene is undergoing change unrelated to our active pursuit of observations. In our previous publication we discussed this motion issue, and suggested means to recognize its presence in a scene. Once distinguished from the static elements, it would be possible to invoke higher-order models and filters to estimate these objects' dynamics (as done by Broida and Chellappa [2] and Gennery [3], but our current interest is in modeling static structure.

It is worth repeating this to clarify our goals in the current work. We are not working with changing scenes, nor is our aim to build descriptions of moving or deforming objects. Our camera is all that moves, and any changes in the imagery arise strictly from this movement. Our goal is to model the geometry of a real static scene through which the camera is moving. This distinguishes us from most of the current efforts in spatiotemporal analysis that use image-plane velocities for measuring arbitrary flows (for example Heeger [7]), or that combine the measured flow with assumptions of constant 3D motion and rigidity for estimating known-order analytic surfaces (i.e., Waxman and Wohn [4], Waxman et al. [5], and Subbarao [6]).

In common with our earlier work, our new approach involves the processing of a very large number of images acquired by a moving camera. The analysis is based on three constraints:

1. The camera's movement is restricted to lie along a linear path.
2. The camera's position and attitude at each imaging site are known.

3. Image capture is rapid enough with respect to camera movement and scene scale to ensure that the data is, in general, temporally continuous.

Within this framework, we generalize from the traditional notion of epipolar *lines* to that of epipolar *planes*—a set of epipolar lines sharing a property of transitivity (which we discuss in section 2.2). We formulate a tracking process that exploits this property for determining the position of features in the scene. This tracking occurs on what we term the *spatiotemporal surface*—a surface defining the evolution of a set of scene features over time. Critical to visualizing this space-time approach is obtaining an understanding of the geometry of the sensing situation, and this is described in the next section.

2.2 Geometric Considerations of Camera Path and Attitude

Figure 1 shows the geometry pertaining to our analysis, indicating several imaging positions

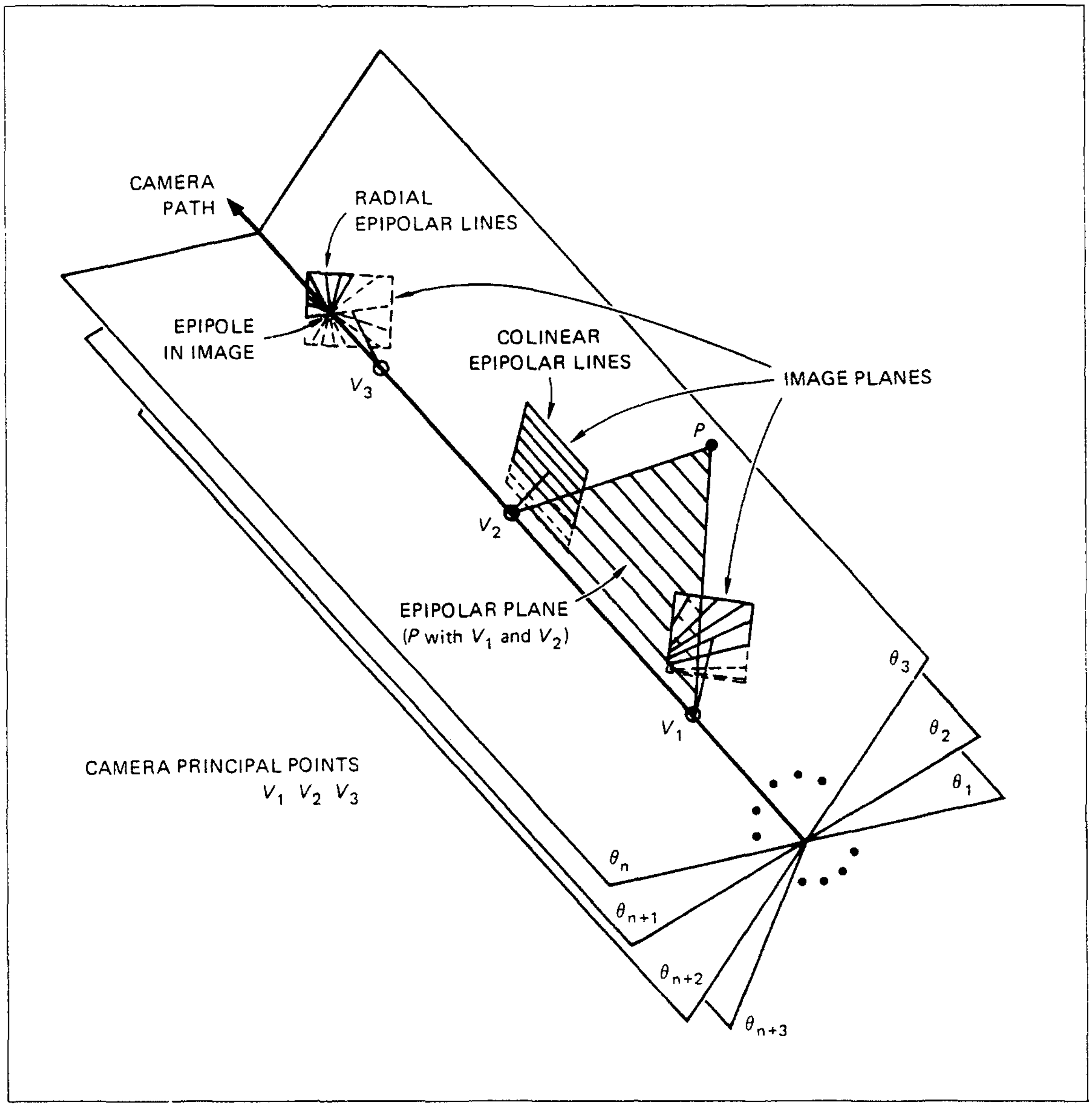

Fig. 1. General epipolar configuration.

and attitudes along a straight path. The camera is modeled as a pin-hole with image plane in front of the lens. For each feature P in the scene and two viewing positions such as V_1 and V_2, there is an *epipolar plane* that passes through P and the line joining the two lens centers. This plane intersects the two image planes along corresponding *epipolar lines* (note that, here, intersection and projection are, in a sense, equivalent). An *epipole* is the intersection of an image plane with the line joining the lens centers. In motion analysis, an epipole is often referred to as the *focus of expansion* (FOE) because the epipolar lines radiate from it. The camera moves in a straight line, and the lens centers at the various viewing positions lie along this line. Notice that the FOE is the image of the camera path. This structuring divides the scene into a pencil of planes passing through the camera path, several of which are sketched (θ_1, θ_2, θ_3, $\theta_n, \ldots, \theta_{n+3}$). This pencil is crucial to our analysis. We view the space as a cylindrical coordinate system with axis the camera path, angle defined by the epipolar plane, and radius the distance from the axis. Note that a scene feature is restricted to a single epipolar plane, and any scene features at the same angle (within the discretization) share that plane. This means that, as in our earlier work, the analysis of a scene can be partitioned into a set of analyses, one for each plane, and these planes can be processed independently. In section 3 we describe how we organize the data to exploit this constraint.

With viewing direction orthogonal to the direction of travel, as depicted at V_2 in figure 1, the epipolar lines for a feature such as P are horizontal scan lines, and these occur at the same vertical position (scan line) in all the images. This is the camera geometry normally chosen for computer stereo vision work. Each scan line is a projected observation of the features in an epipolar plane. The projection of P onto these epipolar lines moves to the right as the camera moves to the left. If one were to take a single epipolar line (scan line) from each of a series of images obtained with this camera geometry and compose a spatiotemporal image, with horizontal being spatial and vertical being temporal, one would see a pattern as in the EPI of figure 2. For this type of motion, feature trajectories are straight lines, as can be seen. This is the case handled by our previous analysis. If, on the other hand, the camera were moving with an attitude as shown at V_3 in figure 1, the set of epipolar lines would form a pattern as shown in figure 3. For this type of motion, feature trajectories are hyperbolas. Notice that the epipolar lines are no longer scan lines—they are oriented radially and pass through the FOE. Allowing the camera to vary its attitude along the path gives rise to spatiotemporal images as shown in figure 4. Here, the epipolar line pattern is not fixed from frame to frame, and the paths of

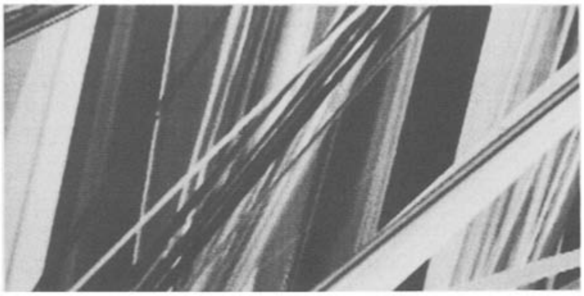
Fig. 2. Orthogonal viewing.

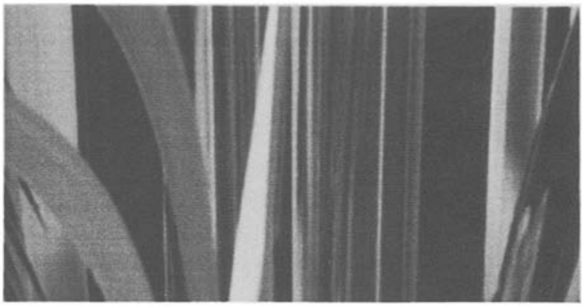
Fig. 3. Fixed, nonorthogonal viewing.

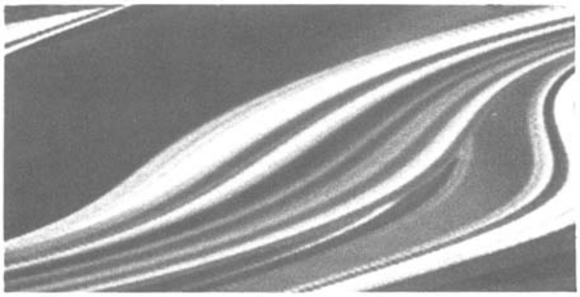
Fig. 4. View direction varying.

features in the EPI are neither linear nor hyperbolic—in fact they are arbitrary curves.

The transitivity property mentioned in section 2.1 arises from the fact that any pair of lines selected from the set form a corresponding pair. That is, for the set of epipolar lines $E^\theta = (e^\theta_0, e^\theta_1, \ldots, e^\theta_n)$ from epipolar plane θ over images I_0 through I_n, any two members comprise a pair of corresponding epipolar lines—e^θ_0 with e^θ_1, e^θ_3 with e^θ_7, et cetera. This occurs because the camera's linear path guarantees that a single pencil of planes defines the epipolar mapping over the entire sequence. Thus, any mapping done on the basis of e^θ_0 with e^θ_1 and then e^θ_1 with e^θ_2 implies the mapping of e^θ_0 with e^θ_2. A similar argument holds for all pairs of mappings in E^θ, and the transitivity follows. If the camera path were nonlinear, no single pencil of planes could be defined, and no such set E^θ could be formed. The only complicating detail with the varying-attitude case (as indicated in figure 4) is that the pattern of epipolar lines changes from image to image: For a fixed camera attitude the pattern is the same for all images in the sequence.

2.3 Keeping the Problem Linear

Recall that our goal is to determine the position of stationary features in the scene: We do this by tracking their appearance over time as they project onto these epipolar planes. Obviously in the case of orthogonal viewing (e.g., as in figure 2 and at V_2 in figure 1), the tracking is linear. For general camera attitudes, including varying, it is nonlinear. Computational considerations make it extremely advantageous for the tracking to be posed as a linear problem. To maintain the linearity regardless of viewing direction, we find not linear feature paths in the EPIs (figures 2 through 4), but linear paths in a *dual space.* The insight here (introduced by Marimont [8]) is that no matter where a camera roams about a scene, for any particular feature, the *lines of sight* from the camera's principle point through that feature in space all intersect at the feature (modulo the measurement error). A line of sight is determined by the line from the principal point through the point in the image plane where the projected feature is observed. From mathematical duality, the duals of these lines of sight lie along a line whose dual is the scene point (see figure 5); fitting a point to the lines of sight is a linear problem. This, then, gives us a metric for linear tracking of features: We map feature image coordinates to lines of sight, and use an optimal estimator to determine the point that minimizes the variance from those lines of sight.

Our estimation is done in the scene Cartesian space, not the dual space, because the error metric, nonlinear in the dual space, has more intuitive meaning and better behavior in scene space. The estimated error in each observation is a function of the size of the Gaussian filter employed and the distance of the feature from the camera. We currently model only these uncertainties in image-plane observations, and not others related to the strength of the feature signal or uncertainty in the position of the camera. These others will have to be modeled in a complete solution.

2.4 Transformations Required

Having decided on a representation that restores the linearity of our estimator, we must now demonstrate a mechanism for extracting the feature observations from the individual images in which they occur and grouping them by epipolar plane. Only in the case of viewing angle orthogonal to the motion is this grouping simple (figure 2), and this was the case our earlier work addressed. To obtain this structuring in the general cases, we could take one of two approaches. The first is to *transform the images* from the Cartesian space in which they are sampled to an epipolar representation (as has been done by Baker et al. [9] and by Jain et al. [10]). Because of aliasing effects (particularly on the observation variances) and nonlinearities in the mapping (it is singular when the FOE is in the image, and could require an infinite imaging surface for the reprojection), we prefer to avoid this transformation. Probably the best solution would be to use a sensor that delivers the data directly in the epipolar form—a spherical sensor having meridian scanning would accomplish this (the flow geometry that is the basis for such a sensor was discussed by Gib-

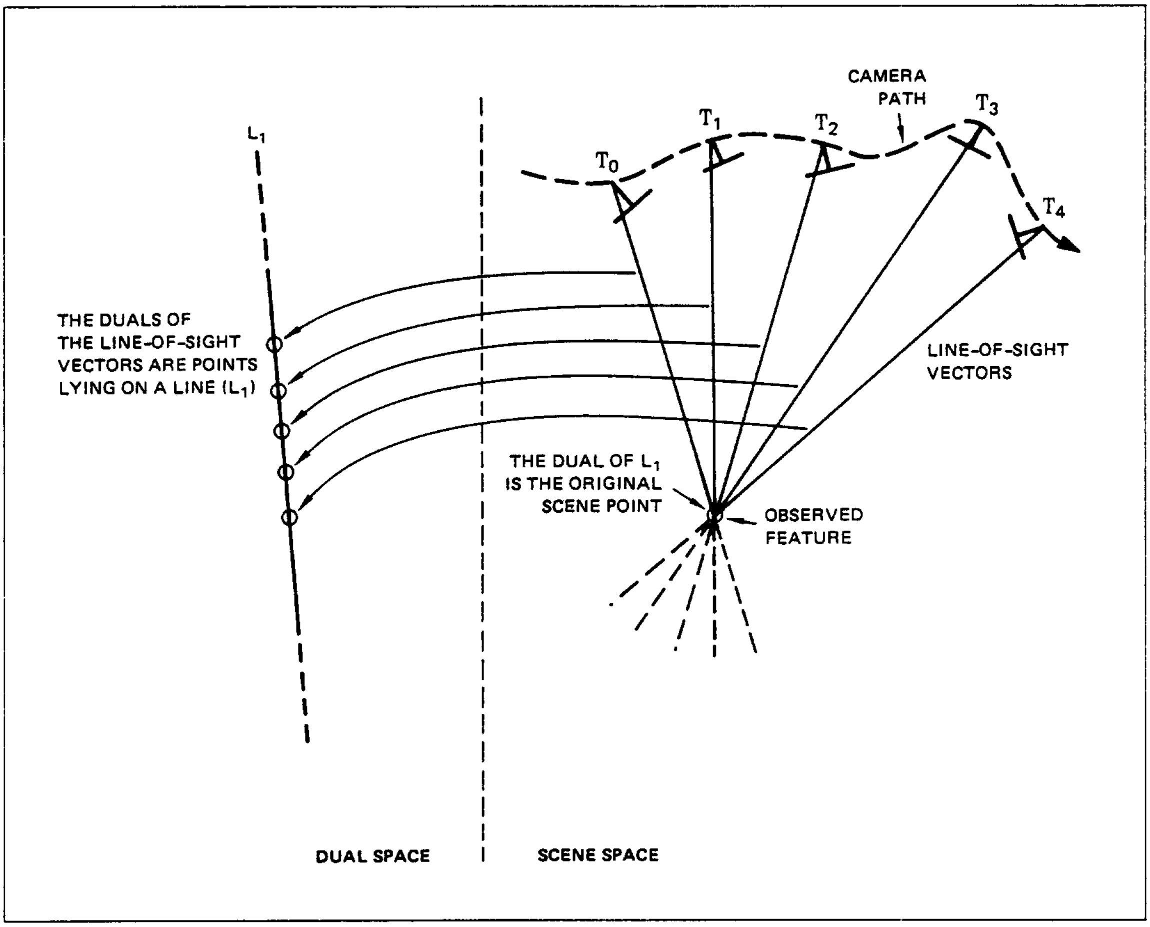
CAMERA PATH
T_0
T_1
T_2
T_3
T_4
L_1
THE DUALS OF THE LINE-OF-SIGHT VECTORS ARE POINTS LYING ON A LINE (L_1)
LINE-OF-SIGHT VECTORS
THE DUAL OF L_1 IS THE ORIGINAL SCENE POINT
OBSERVED FEATURE
DUAL SPACE
SCENE SPACE

Fig. 5. Line-of-sight duality.

son [11] nearly forty years ago). Because such a sensor is not yet available, we choose an alternate approach: to *transform the features* we detect in image space to the desired epipolar space, the cylindrical coordinate system of figure 1. Here the singularity at the FOE presents no problem, and the observation variances are uniform. The structure we have developed for implementing this transformation brings us several other advantages, as the next section describes.

3 The Spatiotemporal Surface

3.1 Structuring the Data—Spatiotemporal Connectivity

We collect the data as a sequence of images, in fact stacking them up as they are acquired into a spatiotemporal volume, as shown in figure 6. As each new image is obtained, we construct its *spatial* and *temporal* edge contours. These contours are three-dimensional zeros of the Laplacian of a chosen three-dimensional Gaussian (Buxton and Buxton [12] and Heeger [7] also use spatiotemporal convolution over an image sequence), and the construction produces a spatiotemporal *surface* enveloping the signed *volumes* (note that, in two dimensions, edge contours envelop signed *regions*). The *spatial* connectivity in this structure lets us explicitly maintain object coherence between features observed on separate epipolar planes; the *temporal* connectivity gives us, as before, the tracking of features over time. See the companion paper in this issue [13] for a description of how these surfaces are constructed.

The need for maintaining this spatial connec-

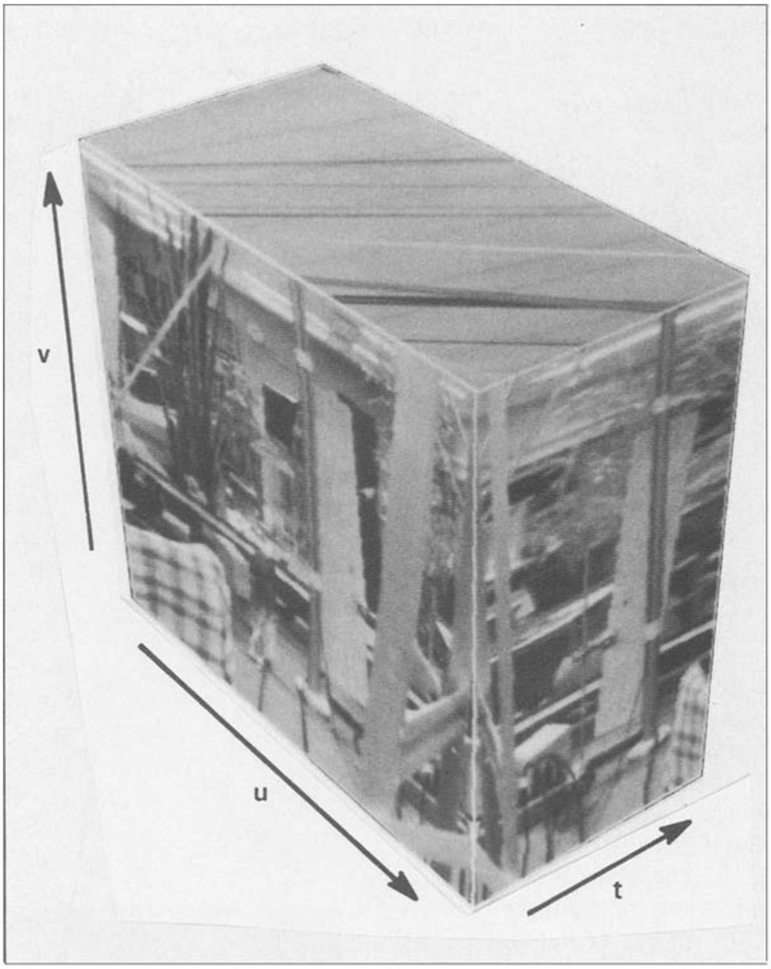

Fig. 6. Spatiotemporal volume.

tivity can be observed by viewing our earlier results [1], one set of which is shown in figure 7. There, in processing the EPIs independently, we obtained separate planes of isolated scene feature estimates. Wishing to exploit the fact that there should be some spatial coherence between these sets of points, we used proximity of the resulting estimates on adjacent planes to filter outliers. Features not within the error (covariance) ellipses of those above or below them (i.e., those which could not be joined into a 3-space contour) were discarded. The remaining point field (figure 7) was sparse and fragmented, and not really representative of the continuous solid surfaces visible in the scene. The problem, however, did not lie with this post hoc filtering but with the loss of spatial connectivity in the first place. Our separation of the data into EPIs, and then subsequent independent processing of these, lost the spatial connectivity apparent in the original images. We maintained instead the temporal connectivity that was critical to the feature tracking. For spatial connectivity in the scene reconstruction, spatial connectivity in the imagery must be preserved. The next two figures present a simplified example of this spatial and temporal connectivity. Figure 8 shows a sequence of simulated images depicting a camera zooming in on a set of rectangles; figure 9 shows a rendered view of the spatiotemporal surfaces arising from this motion. The spatial and temporal interpretation of these surfaces should be quite apparent.

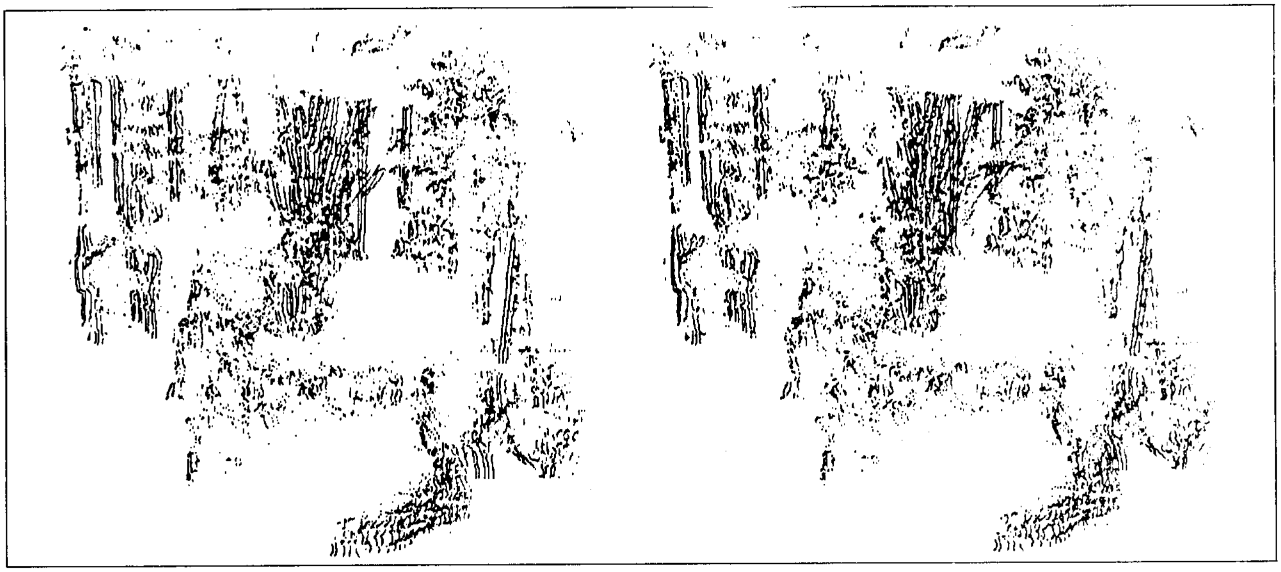

Fig. 7. Orthogonally viewed scene: Results (displayed for crossed-eye viewing).

In our spatiotemporal-surface representation, feature observations bear (u, v, t) coordinates, and are spatiotemporal *voxel facets.* Figure 12 shows a mesh description of the facets for the spatiotemporal surfaces associated with the forward-viewing sequence whose first and last images are depicted in figure 10. These images are much more complex than those of figure 8. Let us reemphasize that the surface is defined at the zeros of a Laplacian of a 3D Gaussian applied over the sequence: There is no thresholding, and the features are simply zero crossings. In the interest of clarity, the surface representations we

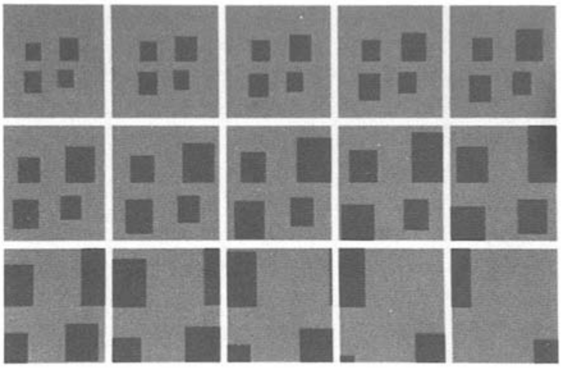

Fig. 8. Simulation: Linear path, motion toward center.

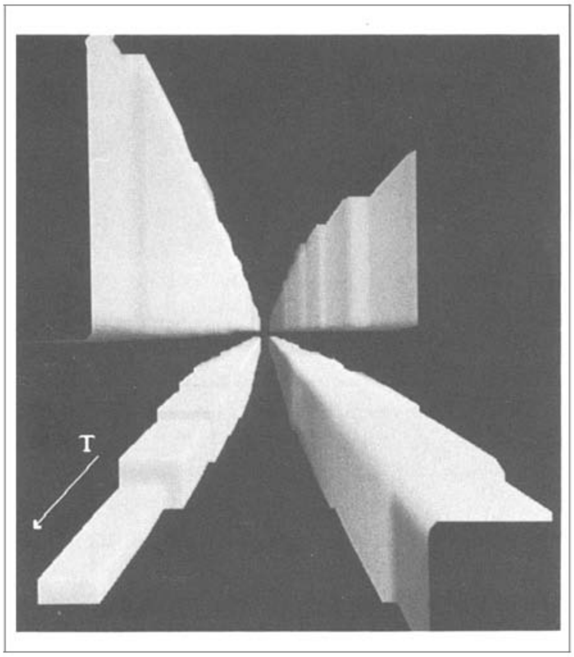

Fig. 9. Surfaces of figure 8 rendered for display.

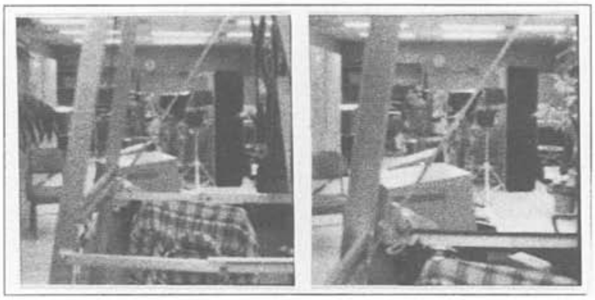

Fig. 10. Sequence 1st and 128th images.

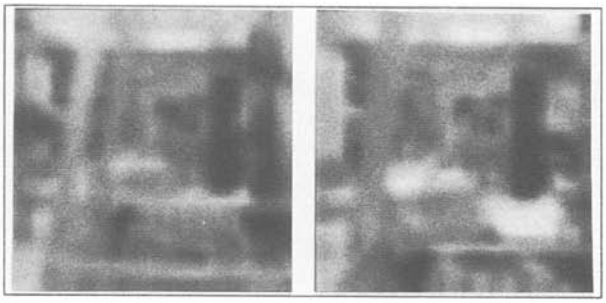

Fig. 11. 1st and 128th images at 1/8 resolution.

will show in the remaining figures are based on a simplified version of this imagery—one-eighth the linear resolution of the originals. Figure 11 shows these two frames at the reduced resolution.

Others have addressed this problem of combining spatial and temporal information, although no one has either built surfaces such as these or attempted to maintain explicit track of the temporal change. Perhaps the closest is Waxman [14], who discusses the use of *evolving contours*—isolated 2D contours whose projections over time can be used in deriving the shape of a restricted class of analytic surfaces. He provides no method for tracking the contours through time, however, or for extracting them from real images—nor does he develop a methodology for utilizing the temporal evolution of individual components of the contours over multiple frames. Later work by Waxman and colleagues [15], presenting *convected activation profiles*, involves spatiotemporal convolution of Gaussian gradients applied at features detected in the individual spatial images by a DOG operator. In this, estimates of image-plane velocities are formed from quotients of the spatiotemporal gradients. There is, however, no estimate of scene motion, and no notion of motion associated with specific objects in the field of view—motion

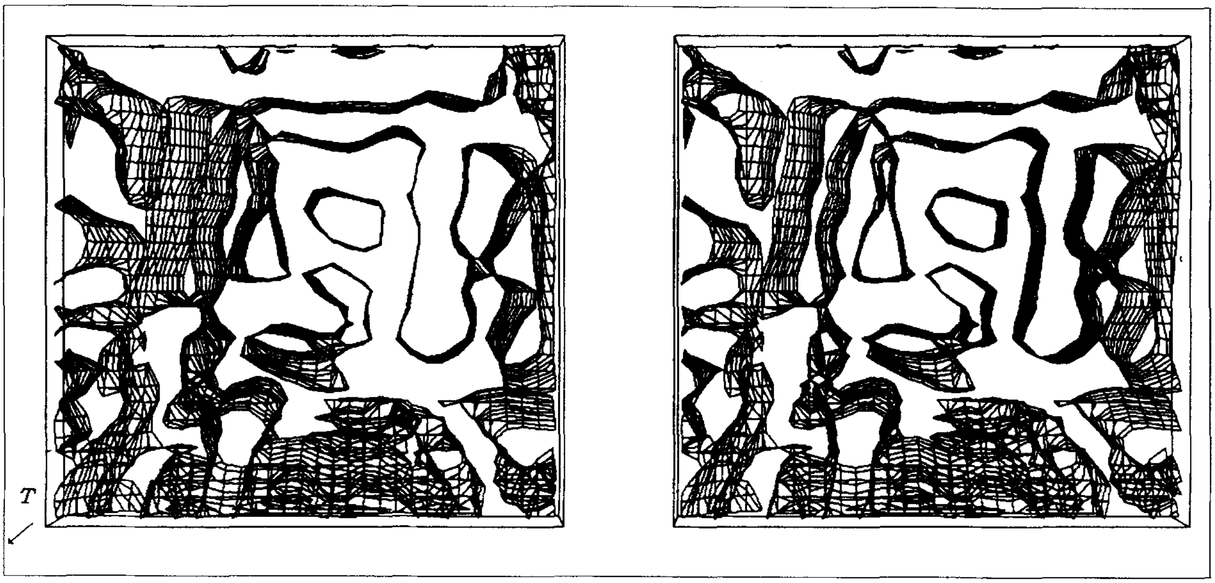

Fig. 12. Spatiotemporal-surface representation, first 10 frames.

is ascribed to pixels in the plane. Others, for example Hildreth and Grzywacz [16], who work with velocity point sets, and Negahdaripour and Horn [17], who determine relative motion of a plane from image gradients, also do not address these issues of local shape, establishing correspondence over time, associating movement with objects, or extracting the measures from real images. Although we have directed our efforts only at ego motion, our space-time surfaces provide a complete representation of these other projective velocity measures, and maintain a continuous track relating them to their underlying scene features. Our work in the future will include looking into using the surface representation for this more general form of motion analysis.

3.2 Structuring the Data—Epipolar-Plane Representation

As mentioned in the previous section, for nonorthogonal viewing directions, epipolar lines are not distinguished by the spatial v scan-line coordinate. To obtain this necessary structuring we develop within the spatiotemporal-surface representation an *embedded* representation that makes the epipolar organization explicit. Over each of the sequential images, we transform the (u, v, t) coordinates of our spatiotemporal zeros to (r, h, θ) *cylindrical coordinates* (θ indicates the epipolar-plane angle ($\theta \in [0,2\pi]$); the quantized resolution in θ is a supplied parameter; and the transform for each image is determined by the particular camera parameters). In this new coordinate system, we build a structure similar to our earlier EPI edge contours, but dynamically organized by epipolar plane. This is done by *intersecting* the spatiotemporal surfaces with the pencil of appropriate epipolar planes (as figure 1). We weave the epipolar connectivity through the spatiotemporal volume, following the known camera viewing direction changes. Figure 13 shows a sampling of the spatiotemporal surfaces as they intersect the pencil of epipolar planes (every fifth plane is depicted). You will notice the obvious radial flow pattern away from the epipole (FOE). Figure 14 shows seven of these surface/plane intersections, along with the associated bounding planes (refer to figure 1). The edge that all share is the camera path (the epipole). These seven planes show exactly the contours one would detect in spatiotemporal intensity images such as depicted in figure 3.

In figure 15 we isolate a single surface from the top left of figure 12, and shows its spatiotemporal structure. Figure 16 shows the same surface structured by its epipolar-plane components. The

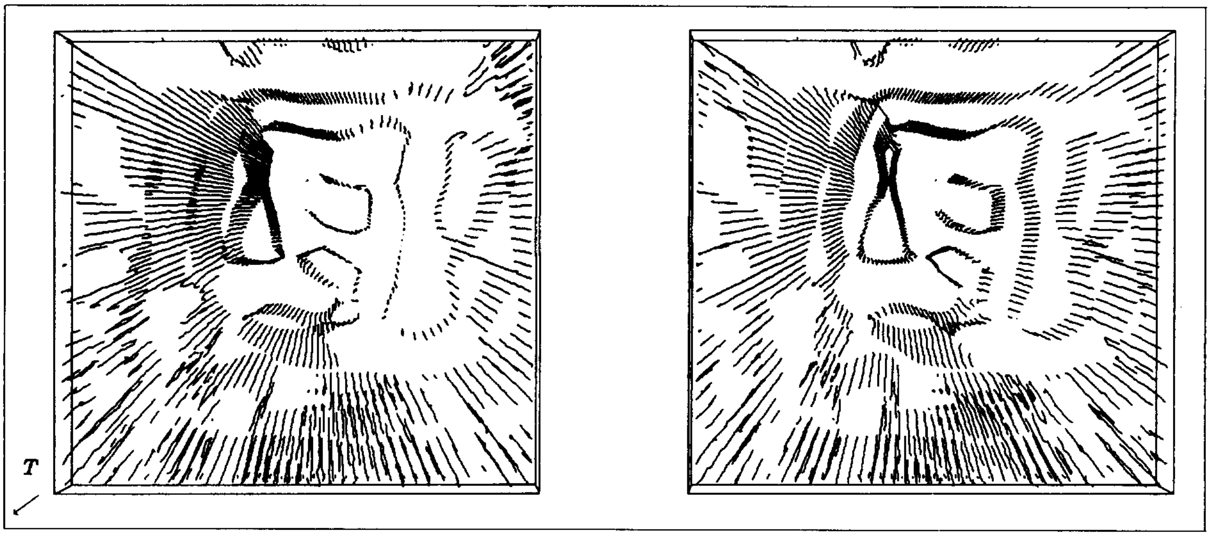

Fig. 13. Epipolar-plane surface representation.

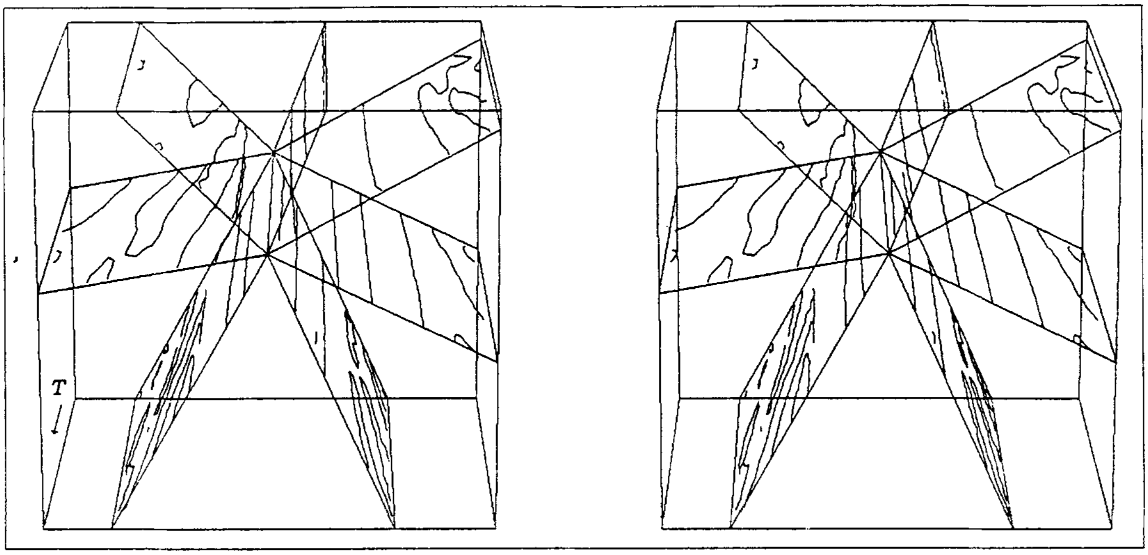

Fig. 14. Intersection: 7 epipolar planes, spatiotemporal surfaces.

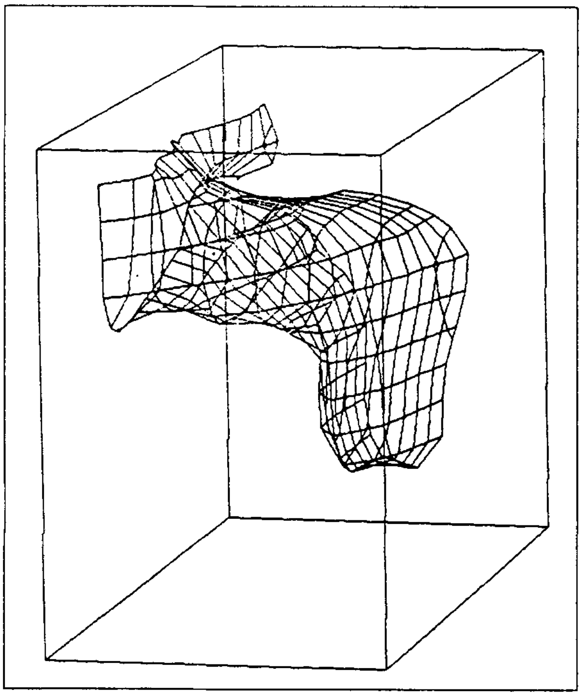

Fig. 15. Spatiotemporal surface.

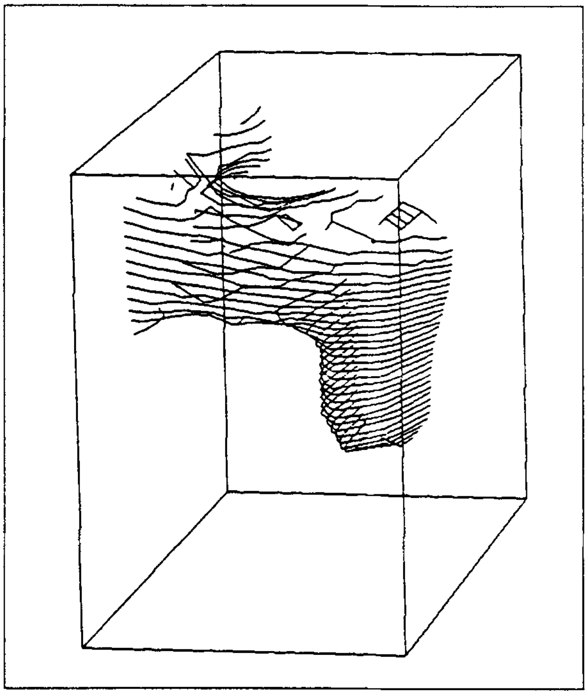

Fig. 16. Epipolar planes.

companion paper [13] gives details of the intersection operation on the spatiotemporal surface. Recall that the displays are in space-time image coordinates. If the camera had been allowed to vary its attitude, the planes depicted in figure 14 would appear skewed, perhaps helical, mirroring the migration of epipolar lines as they are projected on the imaging surface. They might vary in a manner similar to that in which figure 4 varies from figure 2, and for similar reasons. To facilitate presentation, we have not demonstrated this more general camera movement; it is, however, covered by our analysis and implementation.

3.3 Feature Tracking and Estimation

Our approach to scene reconstruction involves tracking scene features as they move in space-time, and to use techniques from estimation theory in approximating and maintaining estimates of their position. This is in distinction with, for example, the work of Hildreth and Grzywacz [16], Waxman and Wohn [4], and others who do not utilize this particular mathematics. Researchers who have built tracking systems using estimation theory include Broida and Chellappa [2] and Gennery [3], as mentioned, Matthies et al. [18], Dickmanns [19], and Hallam [20]. The latter two describe vehicle navigation controllers that work sequentially (as does ours), utilizing Kalman and other filters for estimating motion parameters. Our tracker is a sequential linear estimator, and is implemented as a Kalman filter without the extrapolation phase. Extrapolation is unnecessary since the camera constraints and the space-time surface tell us where each feature will move from frame to frame—there is no need to extrapolate and verify this. Notice that this also makes it clear that there is no *aperture problem* in our approach. The work of Matthies et al. [18] has similarities to ours in its pursuit of scene depth from the analysis of image sequences, but lacks several important elements. These include the generality with respect to view angle that comes with our use of the line-of-sight formulation, and the explicit use of spatial connectivity—they obtain only scene point estimates (as we had with our earlier approach), rather than higher-level descriptors such as scene contours. Furthermore, they must establish feature correspondence via correlation between frames, and this is not necessary with the spatiotemporal surface. On the other hand, we do not aim currently at producing the dense depth maps that they do. Their depth maps are obtained through a combination of tracking and regularization: When we attempt full-surface reconstruction, we will do so with analysis over scale (as discussed in the companion paper [13]), and through the use of inference on the computed free space (the determination of scene free space was shown in our earlier paper [1]).

Figure 17 shows the tracking of scene features on the spatiotemporal surfaces in the vicinity of the surface of figure 15. The tracking occurs along paths such as those shown in figure 16. The final pair shows, in crossed-eye stereo form, the result of the tracking after 10 frames. The coding is as follows: initiation of a feature tracking is marked by a circle; the leading observation of a feature (active front) is shown as an ×; lines join feature observations; 5 observations (an arbitrary number, 2 may be sufficient) must be acquired before an estimate is made of the feature's position—at that point an initial batch estimate is made, and a Kalman filter (discussed by Gelb [21] and Mikhail [22]) is turned on and associated with the feature—this initiation of a Kalman filter is coded by a square; where two observations merge, the tracking is stopped and the features are entered into the data base—this is coded by a diamond.

As mentioned earlier, observations are expressed as line-of-sight vectors, and these are represented in the epipolar plane by the homogeneous line equation $ax + by - c = 0$ (its dual is the point (a, b, c)—see the description of duality in section 2.3). For the initial batch estimation, the coordinates (X) of the feature are the solution of the normal equations for the weighted least-squares system: $\mathbf{X} = (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}\mathbf{C}$. $\mathbf{H}$ is the $m\times2$ matrix of (a_i, b_i) observations; $\mathbf{C}$ is the vector of c_i; and $\mathbf{W}$ is the diagonal matrix of observation weights, determined by σ of the Gaussian, the distance from the camera to the observed feature at observation position i, and the focal distance. We estimate $\mathbf{X}$ first without weights, then compute the weighted solution and

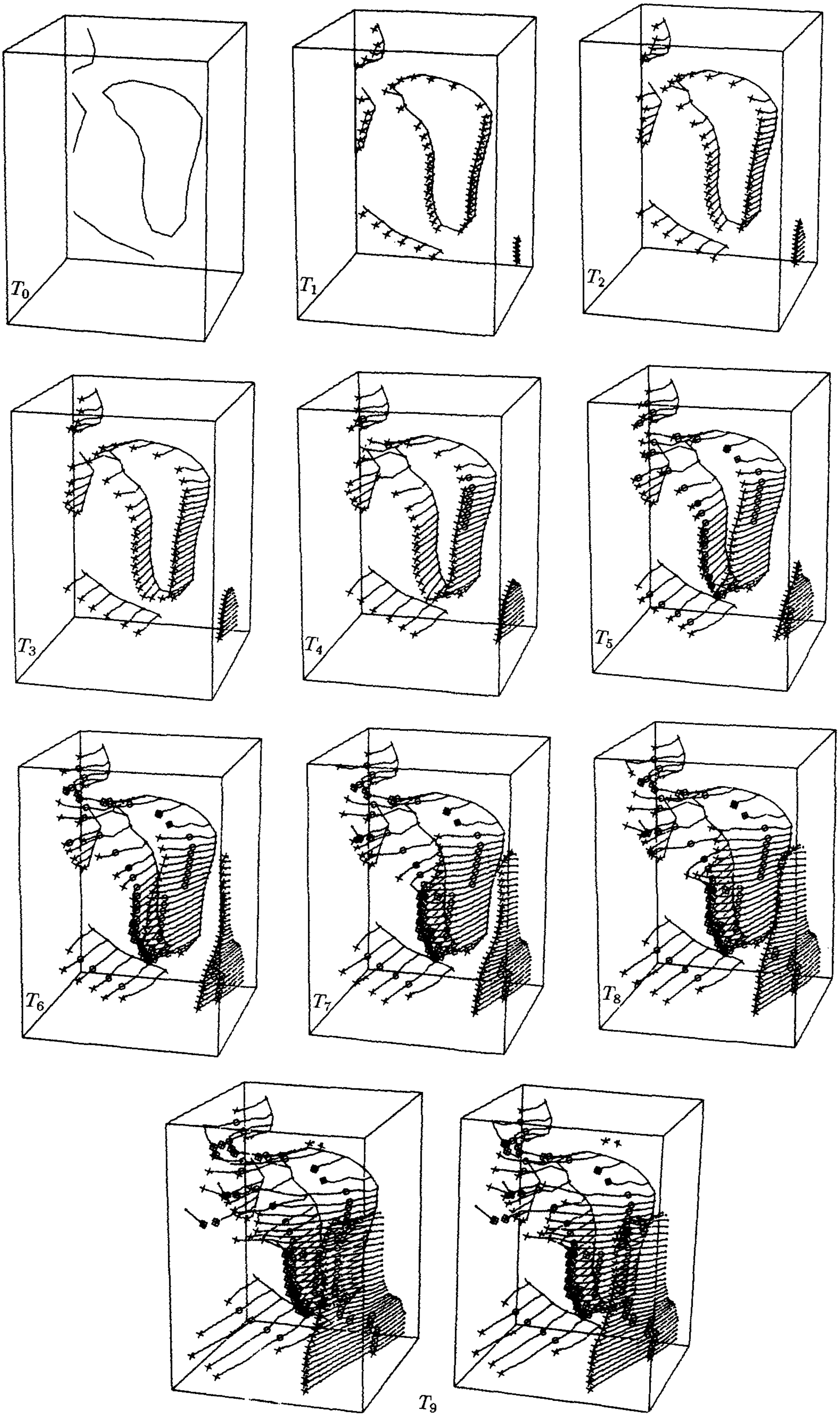

Fig. 17. Sequential feature tracking on the spatiotemporal surface.

the desired covariance matrix, V. Given a current estimate X_{i-1} and covariance V_{i-1}, the Kalman filter at observation *i* updates these as

$$K_i = V_{i-1}H_i^T/[H_iV_{i-1}H_i^T + w_i]$$
$$V_i = [I - K_iH_i]V_{i-1}$$
$$X_i = X_{i-1} + K_i[c_i - H_iX_{i-1}]$$

K_i is the 2×1 Kalman gain matrix, and w_i is the observation weight, a scalar, dependent on the distance from the camera at observation position *i* to the estimate X_{i-1}.

The tracking of an individual feature is depicted in figure 18. The camera path runs across the figures from the lower left. Lines of sight are shown from the camera path through the observations of the feature at the upper right. As the Kalman filter is begun (T_4), an estimate (marked by an ×) and confidence interval (the ellipse) are produced. As further observations are acquired, the estimate and confidence interval are refined. Tracking continues until either the feature is lost, or the error term begins to increase—suggesting that observations not related to the tracked feature are beginning to be included. This could arise because, among other reasons, the zero crossing is erroneous, the feature is not stationary, or the feature is on a contour rather than being a single point in space. Note that although a single feature is presented in this tracking depiction, it is part of a spatiotemporal surface. This means that we have explicit knowledge of those other features to which this is spatially adjacent. Figure 19 shows a contour—a connected set of features on such a surface—observed over time as its shape evolves. Such contours are being construct-

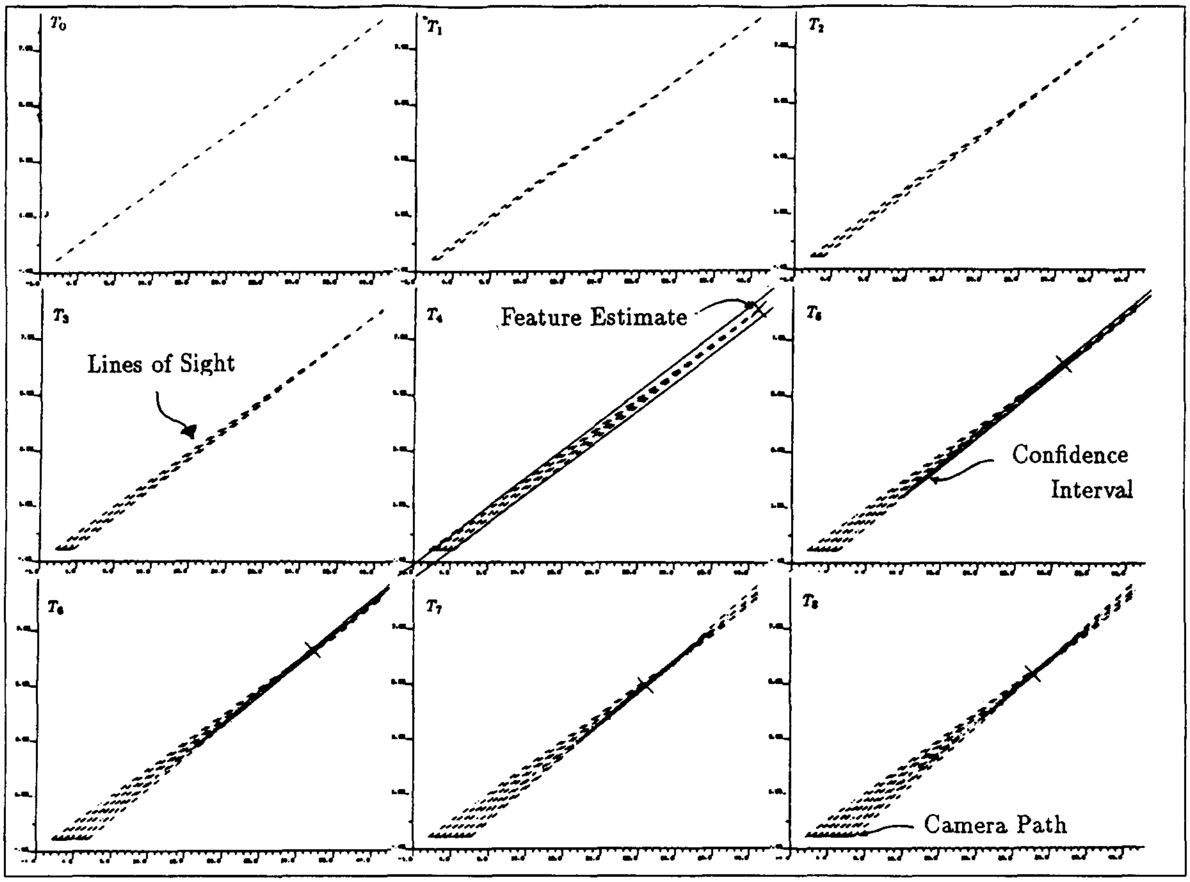

Fig. 18. Sequential estimation.

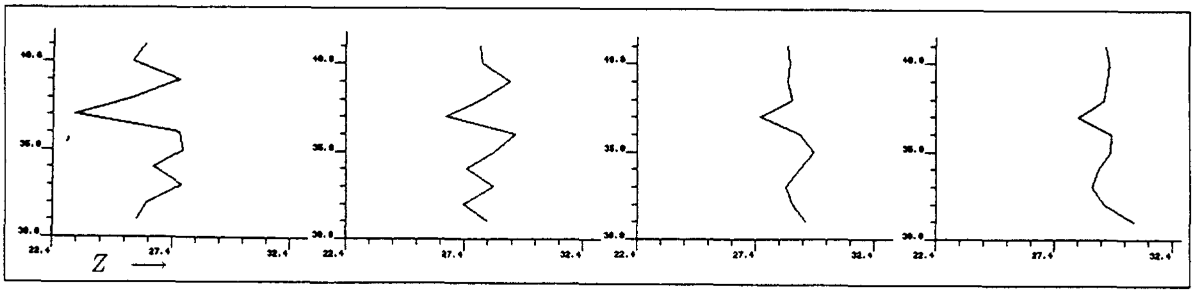

Fig. 19. Contour evolving over time.

ed and refined over the entire image as the analysis progresses. Our current representation of scene structure is based on these evolving contours.

3.4 Generality from the Spatiotemporal Surface

A crucial constraint of the current epipolar-plane image analysis is that having a camera moving along a linear path enables us to divide the analysis into planes, in fact, the pencil of planes of figure 1 passing through the camera path. With this, we are assured that a feature will be viewed in just a single one of these planes, and its motion over time will be confined to that plane. Another crucial constraint is the one we generalized from the orthogonal viewing case—we know that the set of line-of-sight vectors from camera to feature over time will all intersect at that feature, and determining that feature's position is a linear problem. The linearity of the estimator does not depend upon the linearity of the camera path. In fact, the problem would remain linear even if the camera meandered in three dimensions all over the scene.

This knowledge gives us a possibility of removing the restriction that limits us to a linear camera path. All that the linear path guarantees is that the problem is divisible into epipolar planes. If we lose this constraint, then we cannot restrict our feature tracking to separate planes. The observations will, however, still form linear paths in the space of line-of-sight vectors (not to be confused with the (u, v, t) observation space): This is because the lines of sight will all pass through the single feature point. The motion of these observations will give us *ruled* surfaces in this space—visualize pick-up-sticks jammed in a box, with the sticks being the rules. The rules can be used in the same way they have been with the linear path constraint, to determine the positions of features in the scene. The difference is that the linearities must be located—and the spatiotemporal surface is just the place for doing this. It would also be possible to track using the epipolar constraints that apply pairwise between images—that the constraints are limited to pairwise use arises because, for a nonlinear path, the images will not have the transitivity property we cited earlier.

It is equally worth noting that, when the camera attitude and position parameters are not provided, the spatiotemporal surface contains everything that is necessary for determining them. This is, of course, another problem, but one that must be addressed for a realistic vision system. Our initial work in this involves locating distinctive projective features on the spatiotemporal surface—dihedrals selected using Förstner's measure [23]—and tracking them. Depending upon knowledge of the features chosen, these can enable estimation of both relative and absolute camera parameters [24].

This generality suggests there is even broader application for the technique than we had initially thought—it seems quite adaptable to nonlinear camera paths; and should be usable equally in refining the camera model or solving for its unknown parameters.

4 Conclusions

We showed, in our earlier work, the feasibility of extracting scene depth information through

Epipolar-Plane Image Analysis. Our theory applies for any motion where the lens center moves in a straight line, with the earlier implementation covering the special case of camera sites equidistant and viewing direction orthogonal to the camera path.
The generalizations obtained through spatio-temporal-surface analysis bring us the advantages of

- Incremental analysis
- Unrestricted viewing direction (including direction varying along the path)
- Spatial coherence in our results, providing connected surface information for scene objects rather than point estimates structured by epipolar plane
- The possibility of removing the restrictions that fix us to a known linear path

The current implementation, running on a Symbolics 3600, processes the spatiotemporal surfaces at a 1-KHz voxel rate. The associated intersecting, tracking, and estimation procedures bring this rate down to about 150 Hz, 75 percent of which is consumed in the surface intersection (the surface intersection would not be required if we had a sensor of the appropriate geometry). Both the feature tracking and the surface-construction computations are well suited to MIMD (perhaps SIMD) parallel implementations. With these considerations, and the process's inherent precision and robustness, spatiotemporal-surface-based epipolar-plane image analysis shows great promise for tasks in real-time autonomous navigation and mapping.

Acknowledgements

This research has been supported by DARPA Contracts MDA 903-86-C-0084 and DACA 76-85-C-0004. David Marimont, currently with Xerox PARC, was crucial in the development of this work, providing insights for both the geometry and mathematics of the tracking process and the design of the surface builder. Lynn Quam has provided excellent image manipulation and graphics tools, and, whenever required, thoughtful assistance.

References

1. R.C. Bolles, H.H. Baker, and D.H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *Intern. J. Computer Vision* 1:7-55, June 1987.
2. T.J. Broida and R. Chellappa, "Kinematics and structure of a rigid object from a sequence of noisy images," *Proc. Workshop on Motion: Representation and Analysis,* IEEE Computer Society, Kiawah Island, SC, pp. 95-100, May 1986.
3. D.B. Gennery, "Tracking known three-dimensional objects," *Proc. Nat. Conf. Artif. Intell.,* Pittsburgh, pp. 13-17, August 1982.
4. A.M. Waxman and K. Wohn, "Contour evolution, neighborhood deformation, and global image flow: Planar surfaces in motion," *Intern. J. Robotics Research* 4:95-108, Fall 1985.
5. A.M. Waxman, B. Kamgar-Parsi, and M. Subbarao, "Closed-form solutions to image flow equations for 3D structure and motion," *Intern. J. Computer Vision,* 1:239-258, October 1987.
6. M. Subbarao, "Interpretation of image motion fields: A spatio-temporal approach," *Proc. Workshop on Motion: Representation and Analysis,* IEEE Computer Society, Kiawah Island, SC, pp. 157-165, May 1986.
7. D.J. Heeger, "Depth and flow from motion energy," *Proc. 5th Nat. Conf. Artif. Intell.,* Philadelphia, pp. 657-663, August 1986.
8. D.H. Marimont, "Projective duality and the analysis of image sequences," *Proc. Workshop on Motion: Representation and Analysis,* IEEE Computer Society, Kiawah Island, SC, pp. 7-14, May 1986.
9. H.H. Baker, T.O. Binford, J. Malik, and J.F. Meller, "Progress in stereo mapping," *Proc. DARPA Image Understanding Workshop,* Arlington, VA, pp. 327-335, June 1983.
10. R. Jain, S.L. Bartlett, and N. O'Brien, "Motion stereo using ego-motion complex logarthmic maping," *IEEE.* PAMI 9:356-369, May 1987.
11. J.J. Gibson, *The Perception of the Visual World.* Houghton Mifflin: Boston, 1950.
12. B.F. Buxton and H. Buxton, "Monocular depth perception from optical flow by space time signal processing," *Proc. Roy. Soc. London,* Ser. B, 218:27-47, 1983.
13. H.H. Baker, "Building surfaces of evolution: The weaving wall," *Intern. J. Computer Vision* (this issue).
14. A.M. Waxman, "An image flow paradigm," *Proc. Workshop on Computer Vision: Representation and Control,* IEEE Computer Society, Annapolis, MD, pp. 49-57, April 1984.
15. A.M. Waxman, J. Wu, and F. Bergholm, "Convected activation profiles and the measurement of visual motion, *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* Ann Arbor, MI, pp. 717-723, June 1988.
16. E.C. Hildreth and N.M. Grzywacz, "The incremental recovery of structure from motion: Position vs. velocity based formulations," *Proc. Workshop on Motion: Representation and Analysis,* IEEE Computer Society, Kiawah Island, SC, pp. 137-143, May 1986.

17. S. Negahdaripour and B.K.P. Horn, "Direct passive navigation," *IEEE Trans.* PAMI 9:168-176, January 1987.
18. L. Matthies, R. Szeliski, and T. Kanade, "Incremental estimation of dense depth maps from image sequences," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* Ann Arbor, MI, pp. 366-374, June 1988.
19. E.D. Dickmanns, "An integrated approach to feature based dynamic vision," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* Ann Arbor, MI, pp. 820-825, June 1988.
20. J. Hallam, "Resolving observer motion by object tracking," *Proc. 8th Intern. Joint Conf. Artif. Intell.,* Karlsruhe, West Germany, pp. 792-798, August 1983.
21. A. Gelb (ed.), *Applied Optimal Estimation.* Written by the Technical Staff, The Analytic Sciences Corporation, MIT Press, Cambridge, MA, 1974.
22. E.M. Mikhail, with F. Ackerman, *Observations and Least Squares.* University Press of America, Lanham, MD, 1976.
23. W. Förstner, "A feature based correspondence algorithm for image matching," *Proc. Symp. "From Analytical to Digital,"* Intern. Archives of Photogrammetry and Remote Sensing, vol. 26-III, Rovaniemi, Finland, August 1986.
24. W. Förstner, "Reliability analysis of parameter estimation in linear models with applications to mensuration problems in computer vision," *Computer Vision, Graphics, and Image Processing* 40:273-310, December 1987.