Motion Tracking on the Spatiotemporal Surface

H. H. Baker T. D. Garvey
Artificial Intelligence Center
SRI International
Menlo Park, CA 94025.

Abstract

The spatiotemporal (ST) surface has been shown to be a useful representation of projected scene dynamics. Our previous use of this representation has focused on geometric recovery of scene static structure from the analysis of relative motions on the moving image plane. That earlier work exploited the implicit partitioning of motions along epipolar lines to enable search-free feature tracking and position estimation. The ST manifolds provide explicit information about feature 3D contiguity, and their use leads to the recovery of feature 3D position, object 3D contours, and scene 3D surfaces. We have recently turned our attention to the task of interpreting non-static scenes, and track and estimate motions of independently moving objects and background by their appearance and behavior on the ST surface. Selecting the most reliable and discriminating information in the scene, the system demonstrates robust feature tracking over a large range of feature sizes and velocities. When coupled with the more mature Epipolar-Plane Image Analysis system, this motion analysis capability will enable camera solving, dynamics tracking, and scene reconstruction within a unified framework.

1 Tracking over an Image Sequence

The problem of tracking particular objects through a series of images has proved to be a challenging one. The most common tracking techniques include edge (or more generally, feature) tracking, centroid tracking, correlation tracking, and gradient-based optic flow analysis. Each suffers from significant disadvantages. Edge tracking is problematic because it is difficult to make a robust association between a particular group of edges[1] and the object being tracked. Centroid tracking is difficult for related reasons: there is no clear association between scene objects and computable centroids. Correlation tracking is problematic due to changing aspect of the target with respect to the tracker; the tracked object can rotate while translating, changing its image appearance from one frame to the next. Gradient-based optic flow relates differential changes in reflectance with orientation or motion of surfaces in the scene. This relationship is approximate for short spatial or temporal baselines and quite inappropriate for long baselines.[2] Feature analysis has advantage in that it focuses processing at the most discriminable parts of the imagery with the greatest localization and provides robustness through lower sensitivity to projective difficulties such as occlusion and illumination effects.

This paper describes our efforts at utilizing feature tracking on the space-time surface [2] for motion analysis. The principal distinction of this space-time-manifold approach to motion analysis is that it unifies the representation of scene features over space and time. In feature tracking this alleviates the major difficulty of feature-based analysis – the correspondence problem. For EPI analysis it will resolve independent motions within the same framework as solving for scene geometric structure. In discussing our approach to motion analysis we will begin by summarizing our earlier research in recovering scene structure from motion (Epipolar-Plane Image (EPI) Analysis, described in [3] and [4]), connect the techniques used there with the more general problem of unknown scene dynamics, and then discuss our use of the ST surface for motion tracking.

1.1 ST Manifolds for Scene Structure

In the scene reconstruction task, our use of the spatiotemporal surface involved tracking features as they moved under known constraints in space-time, and approximating and maintaining estimates of their positions through the sequence. The approach bridged the usual dichotomy of 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 and, with camera knowledge, eliminated the correspondence problem. Within this framework, we generalized from the traditional notion of epipolar *lines* to that of epipolar *planes*. We then formulated a tracking process that exploited the above constraints in determining the position of features in the scene.

Our tracker was a sequential linear estimator, implemented as a Square Root filter without the extrapolation phase. Extrapolation was unnecessary since the camera constraints and the space-time surface told us

[1] some of which may be only artifacts of position or illumination

[2] Notice the mapping and resampling necessary in Heel's work [9] to make optic flow coherent across time.

TH0390-5/91/0000/0340/$01.00 © 1991 IEEE

where each feature moved from frame to frame (there was no 'aperture problem'). The work of Matthies[10] had similarities to ours in its pursuit of scene depth from the analysis of image sequences, but lacked several important elements, including the generality with respect to view angle that came with our use of a line-of-sight formulation, the explicit use of spatial connectivity that provides higher-level scene contour descriptors, and our match-free tracking.

While simplifying the problem through the use of three assumptions – the camera movement was linear, its position and attitude were known, data capture was sufficiently rapid that the imagery was temporally coherent – we developed a system that could a) work for any camera attitude, b) acquire images at varying rates, c) operate sequentially in time, and d) provide spatially coherent results — 3D contours.

Critical to this was the development of a unique process that constructed, in parallel as the frames were obtained, a 3D space-time description of the evolving imagery. This specifies fully the temporal and projective relationships between scene objects and the sensor. When the scene was stationary and observed by a moving camera, the representation provided simple, direct and robust estimates of scene structure. In extending the analysis of the ST surface representation for more general dynamic analysis, the major difference is that we cannot rely on known camera motion for our tracking, but must actually do the matching – addressing the correspondence problem. One of the benefits of the ST manifold is that it greatly simplifies this problem.

1.2 ST Manifolds for Scene Dynamics

Several issues arise in the move from static to dynamic scenes. Since we have to decouple sensor-induced motion from scene motion, we must be able to solve for the camera. For distinguishing moving from stationary objects, we must be able to discriminate real from sensor-induced motion (moving objects versus the background), – we must be able to model the scene static structure. Motion analysis and scene reconstruction should operate together, with the estimated scene geometry aiding in the camera solving (using known stationary features) and being used to discriminate object motion (by providing a 'background').

The approach we have taken to motion tracking is built on our scene structure estimation process within this unified framework. It's processing is based upon a multi-stage scheme involving feature detection, selection, grouping, and motion classification. First, we represent the spatiotemporal structure of scene dynamics – this is handled by the ST manifold. Using a localization measure on the space-time surface, we then isolate features of interest. Propagating from maxima of the localization measure, we determine the paths of features through time. Finally, a simple linear estimator characterizes feature velocity.

Feature 'edges' detected in 2D images become surface 'facets' in 3D. The connectivity of these facets gives us our tracking mechanism. We described earlier [2] how we locate and parameterize those individual 3D elements — the facets — and structure them together through time. In brief, we define the manifolds (sheets in space-time) that separate image features. These manifolds are 2D surfaces embedded in the 3D space-time dimensions of our data, and are positioned at the extrema of smoothed brightness gradient in the imagery – zero crossings of the Laplacian of a Gaussian (LOG). By following localizable 'features' on these surfaces we track them in time. The following sections describe our use of these features and provide details of our tracking method.

Although we have processed a variety of image sequences with this tracking process, display of detailed analyses of large data sets is difficult in panchromatic reproduction. Our displays here will be limited to simple local indications of the processing – more detail will be presented at the workshop, including various displays of the moving data; space-time surface building; representation of space-time surface localization; the trackers; the extracted 'interesting' features; superposition of reticles grouping observations on individual images in the sequences; and extraction of movement from the 'background.'

2 Selecting Features for Tracking

Figure 1 left shows frames from a synthetic motion sequence of a rotating square. The motion is described by the zero-crossings of a 3D LOG over these data,[3] as shown in Figure 2, with time progressing out of the figure.[4] Figure 2 right shows a side view of these surfaces, oriented so that the temporal structure is more visible. What should be noticed is that the connectivity captured by the surface-building process is an explicit representation and grouping of the motion in the scene. A spatial cut through the 3D zero crossings would produce a 2D spatial, single-image feature description. The temporal facets provide connectivity information for the space-time dimension of the data.

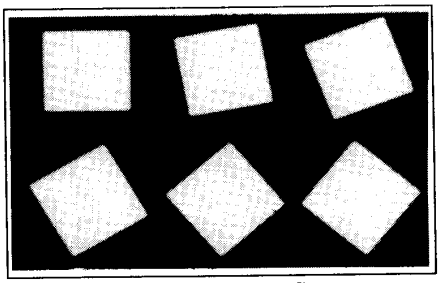

Fig. 1: Frames of Rotating Square

[3] 3D convolution is a standard means of incorporating temporal information (for example, see [1], [6] and [8]). In general, others have not attempted to utilitize or represent the temporal zero crossings.

[4] It is true; many of the figures presented here are so small and detailed as to seem unintelligible. Being of 3D data, larger single figures provide considerably less to appreciate and, in fact, viewing these in stereo gives very good assessment.

Tracking requires determining the correspondence between features in successive views. If we know the direction in which an object is moving (or conversely, the direction in which the sensor is moving through a static scene), then we can use this knowledge in determining their positions in space (as demonstrated by our EPI work). On the other hand, if we have no knowledge of the motion of the sensor, or if the scene can contain objects exhibiting independent motion, then these constraints do not apply. To track a feature we must be able to recognize it from frame to frame and distinguish it from the other features around it (this raises the aperture problem). Only a small percentage of the features we have selected with our 3D detection process can be adequately distinguished for this. For example, if the object to be tracked happens to have a square shape, then the only discriminable parts of it will be the corners. We must determine a measure to use on the images to locate features that *are* discriminable — features that can be reliably tracked from frame to frame.

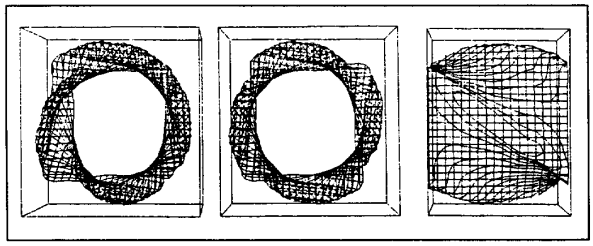

Fig. 2: Spatiotemporal Surfaces

2.1 The Autocorrelation Function

The autocorrelation function – convolving a small window of the image over some larger subset – provides such a measure. Where the window and the subset are identically aligned, the convolution will indicate a high correlation; elsewhere, the correlation will be poorer. A unimodal and highly peaked autocorrelation distribution indicates good localization, whereas a flat profile indicates ambiguity. Autocorrelation is quite expensive to compute, involving evaluation of order mn at every location in the image subset. Interpreting the autocorrelation structure is problematic.

2.2 Förstner's Measure

A variety of corner-detecting analogues to autocorrelation have been suggested, and we work with one developed by Förstner [7]. Here, a simple measure based on a quotient of the determinant and the trace of the covariance matrix related to a planar fit to the window specifies the localizability of the feature at the center of the window. In its full development, the measure determines a confidence ellipse in which the feature can be expected to be localized. Three parameters of the measure define the major and minor axes of the ellipse and its orientation. These three parameters are mapped to a single value (FM).

Figure 3 shows a square and, beside it, the image of its FM. Figure 4 shows an amplified sampling of the localization measures for this image, and similar measures for the image rotated. These are oriented ellipses whose minor axis indicates the most reliable localization direction and whose axis magnitudes show the quality of the localization.

We have used two modifications to Förstner's measure in this work. We do not perform the center-of-gravity refinement, and we normalize the measure by the local gradient. The center-of-gravity computation improves the reliability of the estimation, especially in the vicinity of sharp corners where the simpler measure can produce dual peaks to the sides rather than a stronger peak at the vertex. Since we wish to develop a local computation mechanism for tracking, we are forgoing this correction in our initial studies.

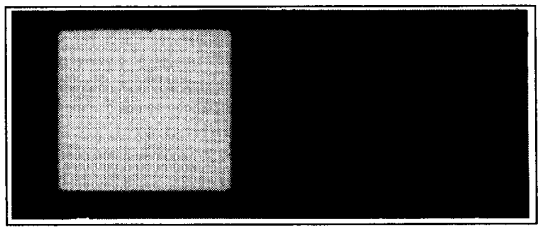

Fig. 3: Image Förstner Measure

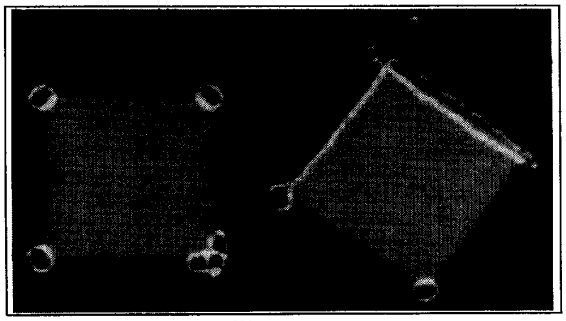

Fig. 4: Localization Ellipses

Figure 5 shows the structure of the space-time surface at the top right corner of the data, with dotted lines indicating the temporal connectivity. Figure 6 shows the relative strengths of the FM along this rotating corner, coded with dots and solid lines in increasing strength. It is clear that the corners of the square are the discriminable features, and the sides are poorly localized.

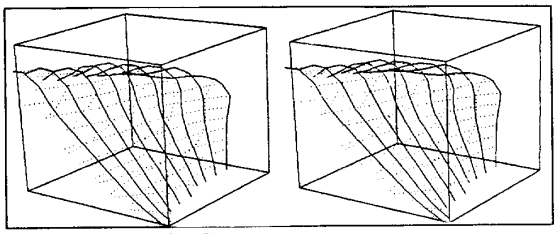

Fig. 5: Top Right Corner

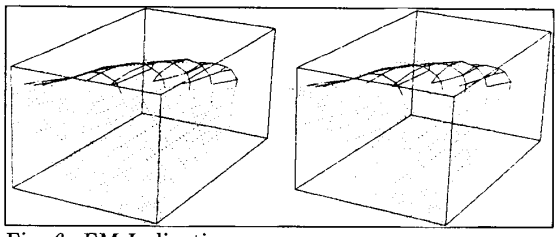

Fig. 6: FM Indications

2.3 Tracking and Velocity Estimation

The display in Figure 6 shows the features we will track — it does not show an actual tracking of features. Relating the observations indicated together over time must still be demonstrated. In forming these observations into unified trackers, we connect *local maxima* of these measures. The velocity estimation will then occur at this level of the analysis — as FM-maxima observations are associated through time, a sequential filter will be updated and refined with the new information.

2.3.1 Feature Tracking

The maxima can move in any direction between frames of a sequence, and can move an arbitrary distance depending on the velocity (translational and rotational) of the objects of which they are part. This means that while the spatiotemporal surface can be defined by feature connectivity, and a particular maximum will be seen to move along a single spatiotemporal surface, the matching of maxima cannot be defined strictly on the basis of proximity. For one thing, they need not be adjacent from frame to frame; for another, if maxima are fairly dense, then there may be significant difficulty in unambiguous assignment when they come close to one another. With large motions or repeated fine patterns, accurate tracking could be difficult.

As well as being accurate, our tracking mechanism must be designed to work within the framework of the surface-building process – it must be able to operate at a local level and be amenable to parallel implementation. The tracking mechanism we have designed satisfies these criteria. It uses a propagation mechanism, with each maximum at time T spreading itself forward to neighbors on the spatiotemporal surface and each maximum at time $T+1$ reaching back to neighbors on the spatiotemporal surface to see if there is a maximum from which it may have descended. When only one predecessor can be found, the tracking assignment can simply use this pairing, and can deduce that whatever the history of the feature at time T, this new observation at time $T+1$ shares that history and affects the estimation of that motion. When there are multiple choices for the assignment then an adjudication must be made to select the most likely. The tracking process and the adjudication are described in the next section.

2.3.2 Tracking Maxima on ST Surfaces

The principal intermediary in the tracking propagation is the set of *temporal* facets. When a contour is stationary, it's spatial facets will be adjacent over time – there will be no spatial motion requiring temporal representation. However, with spatial motion of a contour, the temporal facets serve to join these observations. At time $T+1$, the spatiotemporal surface from T to $T+1$ must be built, the local FM maxima be determined, and then be associated with previous FM maxima at T via the intervening temporal facets.

Several considerations are involved in establishing these temporal associations. If there is no ambiguity in the assignment (only one at each time is being considered for matching with the other), then the association is made and the tracking propagated. If more than one is in contention, then the values of the FM, the local spatial normal to the ST surface,[5] the established velocity, and the distance of motion are all considered. If one pairing is clearly better in position, orientation, and localization measure, then it is chosen, otherwise multiple pairings are maintained as being possible. In the latter case, all contending tracks will be retained, with the final judgement being made later when more knowledge of the behavior of the features is available and a choice minimizing the ambiguity is possible. Another measure being evaluated is a correlation score (SSD) evaluated at contending ST facets. Figure 7 shows the tracking results for this rotating square.

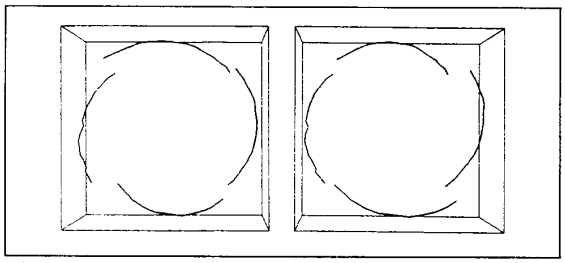

Figure 7: Trackers on Rotating Square

2.3.3 Motion Interpretation

Given a single feature moving in an image sequence, the best we can do, without other information (such as stereo or a DTM), is to determine its velocity in an image-based coordinate system. In our demonstrations here we will estimate only this image-plane velocity. As an initial approximation we will model feature motion as piecewise linear in time — that is, piecewise constant velocity.[6]

Image-plane motion is determined by solving a system of two linear equations defining the velocity vector in space-time. A sequential least-squares filter is set

[5] Notice that the 3D surface normal also suggests the direction of feature motion, as would the eigenvector of the largest eigenvalue of a 3D version of Förstner's measure.

[6] See [5] for inferring a rotating and translating 3D model of object motion.

up for this estimation, and velocities are determined for each feature tracked. Velocity is a relative measure, and its interpretation depends on the activity around it. We must determine from the image-plane velocities which features are in motion with respect to others and which should be considered of importance for tracking – we must establish a background frame for velocity reference.

Since observed velocity depends on range, the geometric structure of the scene, if known, can be used to distinguish moving features from the background. Our earlier EPI work will provide the depth information for static components of the scene when it is integrated with this tracking system. In the meantime, our determination of the 'background' for these studies has been quite simple — we presume it to be planar and select as 'interesting' features those lying outside of one standard deviation from that estimated plane.

In the interests of demonstrating some preliminary object-like groupings, in the video demonstrations we isolate features moving with respect to the 'background' and group together those which are spatially connected. The reticles displayed over the imagery, indicating tracked features, enclose these grouped features. This demonstrates a primitive form of object tracking — features are tracked together when they are seen to be spatially related. Similar results were obtained by grouping features using their projective velocities. Our objective is to use behavior (dynamics) and shape (statics) to couple object tracking with object identification.

3 EXPERIMENTAL RESULTS

Figure 8 shows several frames of part of a low resolution IR sequence of a moving car[7] – notice the vehicle moving down to the left while the background slides right because of a camera panning action. Figure 9 shows the major-contrast ST manifolds from these data. The next figure presents a side view, where the dots show temporal surface elements joining the solid lines of the spatial zero crossings. Figure 11 shows the final estimations, depicting velocities as dots through dashes to solid lines, in units of velocity standard deviation.

Figure 8: Moving Car

[7]These data are part of the workshop database.

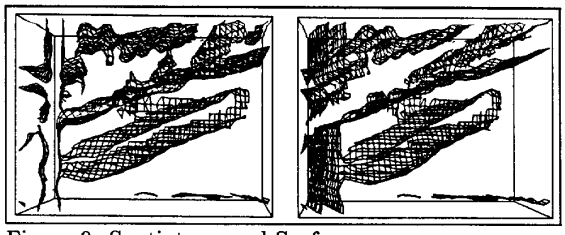

Figure 9: Spatiotemporal Surfaces

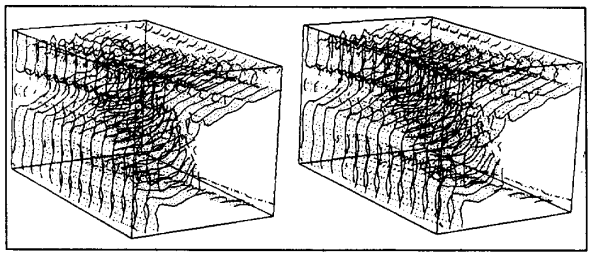

Figure 10: Side View of ST Surfaces

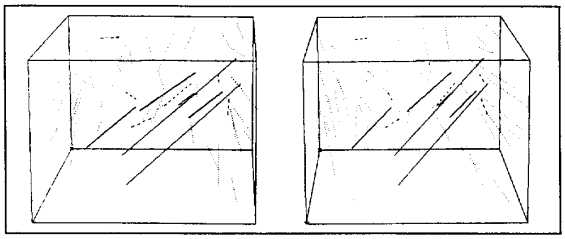

Figure 11: Velocity Estimates

3.1 Tracking Parameters

Recall that our first filtering of the imagery locates zero crossings of the Laplacian of a Gaussian (LOG) of the imagery. This operation selects the locally-largest intensity differences — the 'edges' in the images. The size of this filter determines the range of contrasts and the spatial extent of these edges. Although in this study we have selected large values for the filter size to facilitate display, these should be determined automatically by the tracking process – and determined differentially across the imagery depending on the character of the features observed. [8]

From the set of edges determined by the LOG operation, we select those whose gradients are greater than one standard deviation from the mean of gradients over all contours. This ensures that the features (edges) have 'significance,' i.e., are not likely to be artifacts of the detection process or noise. From among these gradient-selected features we choose those which are most localizable. In selecting from among the most-localizable features that are tracked for those we wish to consider of 'interest,' we have again used some

[8]We are addressing the notion of scale filtering on the scale-space manifold, and are finding that analysis over a range of resolutions can lead to selecting the 'best' filter size for each feature.

a priori assumptions. Notable among these is the assumption that such features are moving with respect to the background (our EPI analysis will handle the converse). We also require, for reliable tracking, that we have enough observations of a feature to enable an accurate and consistent estimate of its velocity to be made. This means that we discard (for our displays) tracked features that are not viewed for a sufficient duration.

3.2 Performance Issues

The current surface-building process constructs the spatiotemporal representations of the imagery at a rate of about 1000 pixels per second. Evaluation of Förstner's measure and the tracking of FM maxima reduce this to roughly 500 pixels per second. Gaussian and Laplacian convolution are not included in these figures since we compute the filtered images in an off-line fashion before studying a data set. These convolution computations are quite simple, however, as they are decomposable into a total of eleven 1D convolutions, and could be computed in a realtime pipeline. The tracking has been designed with parallel-processing in mind, and most computations require only a small local support. Such parallelism could provide sufficient performance for realtime analysis.

3.3 Concluding Remarks

Our intented use of this tracking process begins with estimation of the dynamics of objects in motion and their subsequent recognition based on behavioral and shape characteristics. We will also be integrating the tracker with the original EPI analysis. Features determined to be stationary will be used for camera solving, and this will enable processing of non-linear camera motions. In an intriguing combination of the two, we are investigating use of derived groupings and rigid motion interpretations to run the EPI analysis in reverse over the space-time surfaces (using inverses of the observed motion parameters), and compute the 3D shape of tracked objects even while they are undergoing independent and arbitrary motion.

The important element to note in this work is not that we can track features through a sequence – there are a variety of techniques that can do this, more or less successfully – but that we can utilize the ST surfaces to track and estimate feature motions, distinquish moving from stationary objects, and inform EPI analysis of features to use for camera solving in its geometric recovery. This is a critical component of enabling EPI analysis to be used on non-linear motion trajectories through dynamic scenes.

Acknowledgement

This study has been supported by a research contract from Fujitsu System Integration Laboratories, Kawasaki, Japan.

References

[1] Adelson, E.H., and J.R. Bergen, "Spatiotemporal energy models for the perception of motion," *Journal of the Optical Society of America A*, **2**:2 (1985), 284–299.

[2] Baker, H.H., "Building Surfaces of Evolution: The Weaving Wall," *International Journal of Computer Vision*, Vol.3, No.1 (1989), 51–72.

[3] Baker, H.H., and R.C. Bolles, "Generalizing Epipolar-Plane Image Analysis on the Spatiotemporal Surface," *International Journal of Computer Vision*, Vol.3, No.1 (1989), 33–50.

[4] Bolles, R.C., H.H. Baker, and D.H. Marimont, "Epipolar-Plane Image Analysis: An Approach to Determining Structure from Motion," *International Journal of Computer Vision*, Vol.1, No.1 (1987), 7–55.

[5] Broida, T.J., and R. Chellappa, "Estimating the Kinematics and Structure of a Rigid Object from a Sequence of Monocular Images," *IEEE PAMI*, Vol.13, No.6, (1991), 497–513.

[6] Buxton, B.F., and H. Buxton, "Computation of Optic Flow from the Motion of Edge Features in Image Sequences," *Image and Vision Computing*, **2**:2 (1984), 59–75.

[7] Förstner, W., "Reliability Analysis of Parameter Estimation in Linear Models with Applications to Mensuration Problems in Computer Vision," *Computer Vision, Graphics, and Image Processing*, **40** (1987), 273–310.

[8] Heeger, D.J., "Depth and Flow from Motion Energy," *Proceedings of the Fifth National Conference on Artificial Intelligence*, Philadelphia (1986), 657–663.

[9] Heel, J., "Temporally Integrated Surface Reconstruction," *Proceedings of the Third International Conference on Computer Visionn*, IEEE Computer Society, Osaka, Japan (1990), 292–295.

[10] Matthies, L., R. Szeliski, and T. Kanade, "Incremental Estimation of Dense Depth Maps from Image Sequences," *Proceedings of the Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Ann Arbor, Michigan (1988), 366–374.