Detecting Moving Objects Using the Rigidity Constraint

William B. Thompson, Pamela Lechleider, and Elizabeth R. Stuck

Abstract—**We describe a method for visually detecting moving objects from a moving camera using point correspondences in two orthographic views. The method applies a simple structure-from-motion analysis and then identifies those points inconsistent with the interpretation of the scene as a single rigid object. It is effective even when the actual motion parameters cannot be recovered. Demonstrations are presented using point correspondences automatically determined from real image sequences.**

Index Terms—**Motion, moving object detection, outlier detection, robust estimation, segmentation.**

I. Introduction

The ability to visually detect moving objects is important in a wide variety of circumstances. Simple temporal differencing suffices if the camera is known to be stationary and lighting well controlled [1]. The problem becomes significantly more difficult if the camera is also moving since now the task is to detect objects moving with respect to the environment and not the camera. Differencing is of little value as all visible surfaces are likely to be moving with respect to the camera in a manner that will generate noticeable changes throughout the image. If camera motion is known to consist only of translation, a relatively straightforward analysis of the optical flow field can be used to find moving objects. Translational motion produces flow radiating out from a focus of expansion at the image plane location corresponding to the direction of gaze equal to the direction of motion. If either the direction of motion is known or the focus of expansion can be accurately estimated, any flow vectors with an inconsistent direction are due to moving objects (e.g., see [2]).

When general camera motion is possible, more sophisticated analysis is required [3], [4]. In a previous paper, we reported a family of methods for detecting moving objects [3]. Each technique used partial information about the camera motion and/or scene structure to develop constraints on optical flow that should be satisfied by all visible background points. Moving object detection was based on searching for patterns of flow that violated these constraints. Results were demonstrated for four situations: known rotation (but unknown direction of translation), active tracking of objects of interest, object motion constrained to a smooth surface, and combined use of stereo and motion.

In this correspondence, we present an alternate approach based on the rigidity constraint: A scene containing moving objects can be thought of as undergoing a particular sort of nonrigid motion with respect to the camera. Structure-from-motion techniques that are sensitive to the presence of such nonrigid motions can thus be used to detect moving objects.

Ullman was one of the first to recognize that the recovery of the 3-D structure and motion giving rise to a particular time-varying image sequence is greatly aided by the assumption that the structure is rigid [5]. In fact, the method he developed gives an indication of whether or not a set of point correspondences actually has an interpretation as the projection of a moving but otherwise rigid object. Ullman points out that a combinatorial search could therefore be used to find separately moving objects in a scene. Heeger and Hager [6] and Zhang *et al.* [7] suggest that moving objects be detected by recovering the relative motion between camera and environment and then detecting image regions incompatible with this motion. Both methods require *a priori* estimates of camera motion. In addition, Zhang *et al.* use calibrated stereo input.

The method we describe below is particularly simple. We use the linear structure-from-motion (SFM) algorithm given in [8] to solve for partial parameters governing the relative motion of camera and background.[1] As with most other SFM algorithms, ill-conditioning leads to poor estimates due to the noise inherent in dealing with real imagery. In our case, however, we are only interested in those image points not consistent with a rigid interpretation. These can often be found even when the motion and structure of the rigidly moving background cannot be recovered. The method is demonstrated on real image sequences for which point correspondences have been determined in a fully automated manner.

II. Method

Huang and Lee revisit Ullman's original problem of determining ⁻otion and structure of four noncoplanar object points from image oint correspondences over three distinct orthographic views [8]. ɹhey show that under orthographic projection, two views are insufficient to determine a unique solution to the problem no matter how many point correspondences are available. They go on to develop a solution by using the two-view case to determine constraints on object

Manuscript received December 20, 1990; revised February 28, 1992. This work was supported by the National Science foundation under Grant IRI-8722576. Recommended for acceptance by Associate Editor E. Hildreth.

W. B. Thompson is with the Department of Computer Science, University of Utah, Salt Lake City, UT 84112.

P. Lechleider and E. R. Stuck are with the Computer Science Department, University of Minnesota, Minneapolis, MN 55455.

IEEE Log Number 9205821.

[1] We adopt the convention that the coordinate system is fixed to the camera; therefore, camera motion is equivalent to the environment moving by a stationary camera.

rotation between each pair of images in the three-view situation and then find the unique motion making these constraints consistent.

Adopting the notation of [8], the following holds:

$$(x_i, y_i, z_i) = \text{object-space coordinates of point } P_i \text{ at } t_1$$
$$(x'_i, y'_i, z'_i) = \text{object-space coordinates of } P_i \text{ at } t_2$$
$$(X_i, Y_i) = \text{image-space coordinates of } P_i \text{ at } t_1$$
$$(X'_i, Y'_i) = \text{image-space coordinates of } P_i \text{ at } t_2$$

For rotation R and translation T

$$\begin{bmatrix} x'_i \\ y'_i \\ z'_i \end{bmatrix} = R \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} + T,$$

$$\text{where } R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \text{and } T = \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix}.$$

Because the projection is orthographic, $(X_i, Y_i) = (x_i, y_i)$, and $(X'_i, Y'_i) = (x'_i, y'_i)$. Set the origin in each of two views to coincide with a particular corresponding pair of points (e.g., $(X_1, Y_1) = (X'_1, Y'_1) = (0, 0)$) and adjust all of the other point coordinates appropriately. Then, the following constraint holds:

$$r_{23}X'_i - r_{13}Y'_i + r_{32}X_i - r_{31}Y_i = 0, \; i = 2, \cdots, n. \quad (1)$$

If $n \geq 4$ and the points are not coplanar, this defines a system of equations that can be solved for r_{23}, r_{13}, r_{32} and r_{31} up to a scale factor:

$$Pr = 0, P = \begin{bmatrix} X'_2 & -Y'_2 & X_2 & -Y_2 \\ & \vdots & & \\ X'_n & -Y'_n & X_n & -Y_n \end{bmatrix}, r = \begin{bmatrix} r_{23} \\ r_{13} \\ r_{32} \\ r_{31} \end{bmatrix}. \quad (2)$$

Presuming that the correspondences between (X_i, Y_i) and (X'_i, Y'_i) are correct, it is also the case that

$$r_{13}^2 + r_{23}^2 = r_{31}^2 + r_{32}^2. \quad (3)$$

The system defined by (2) will have an exact solution if there is no error in point displacements and if no independently moving objects are present in the field of view. In the presence of noise, such systems can be solved by using a standard least squares approach that minimizes $||Pr||$ subject to the constraint that $||r|| = 1$. (Clearly, any scalar multiple of this solution also satisfies (2).) Moving objects can introduce inconsistencies such that (1) has no meaningful solution since the parameters of motion relating the moving object and the imaging system need not have a relationship to r_{23}, r_{13}, r_{32} and r_{31}. One way to deal with this problem is to consider points associated with moving objects as *outliers*—removing such points from (1) again allows a solution. Researchers in computer vision have recently become very interested in applying methods from robust statistics to problems such as structure-from-motion that are inherently unstable and thus sensitive to both outliers and even small amounts of noise (e.g., see [9]–[11]). We also use such an approach. Our interest, however, is not in actually solving (2) but only in identifying the outliers that make the solution difficult.

Outliers in (2) are found using a modification of the least median squares (LMedS) algorithm presented in [11] and [12]. Standard least-squares techniques find an approximate solution to the system of linear equations $Pr = b$ by defining a residual vector $e = [e_i] = [Pr - b]$, and then solving for the r then minimizes the sum of squares of the e_i's. LMedS is similar, except that the r that minimizes the *median* of the squares of the e_i's is found. This is more appropriate than least-squares optimization when outliers are likely. Unfortunately, least median squares methods are quite computationallly intensive. Some efficiency is possible, however, by adopting a Monte Carlo approximation.

In our case, $b = 0$, and the solution procedures described in [12] cannot be applied directly. Instead, we use a residual vector defined as

$$Pr = e = \begin{bmatrix} e_2 \\ \vdots \\ en \end{bmatrix} \quad (4)$$

and search for the r minimizing the median over i of e_i^2 (or equivalently of $|e_i|$). To approximate the optimal r, we successively choose random sets of three image points. For each three-point sample, we find a solution to (2), subject to the constraint that $||r|| = 1$. These solutions are exact and, except in degenerate cases, unique. For each such set, we compute the median residual over all matched feature points in the image, using the values of r just found. The process is continued, keeping track of the particular r resulting in the smallest median residual so far. A relatively small number of trials gives reasonable assurance that the best r is a good approximation of the LMedS solution. For the problem described here, we do not care about the actual value of r; we only care about the outliers in (2). These are found by using the best r found through the random trials, computing the individual residuals, and then flagging as potential outliers any points resulting in a value of $|e_i|$ greater than

$$t_{\text{outlier}} = C \cdot med|e_i|. \quad (5)$$

Although developed for a different formulation of the problem, the formula for C given in [12] proved to be sufficient for this purpose.

III. Experimental Results

We present the results of this method applied to two image sequences. Both sequences consist of 13 frames taken by a conventional television camera and digitized at a resolution of 240 by 320 pixels. A variety of objects are scattered over a table. In both cases, the camera is looking down on the scenes from an oblique angle and rotating around an axis near the center of the visible objects. The camera rotation between consecutive frames is approximately constant within each sequence. The per frame rotation varies between the two sequences but is less than 1°/frame in each case. The axis of rotation is perpendicular to the table surface (i.e., the camera is moving over a plane parallel with the table). The camera is approximately 3.4 m away from the objects, and a lens with focal length 102 mm is used to approximate orthographic projection. The field of view is approximately 7 by 5°. No effort was made to correct for geometric distortions.

Figs. 1 and 2 show the first and last frames of a sequence in which, in addition to the camera motion, an object is rolling along the table. For this sequence, the total camera rotation is 7.5°s. Figs. 3 and 4 are the first and last frames of a similar sequence in which one of the objects is translating over the table. Total camera rotation for the second sequence is 7.0°.

The feature point matching method described in [13] was applied to each consecutive frame pair in both sequences.[2] No parameter tuning was performed other than to adjust the number of feature points to a reasonable value. A linking algorithm was used to track points from frame to frame. In both examples, over 200 points could be tracked through the complete sequence. Intermediate frames were then discarded and the corresponding locations in the first and last frames used to define discrete point correspondences. Twenty-five sets of three feature points were randomly selected from each sequence

[2] The feature points used were local extrema in difference of Gaussian images.

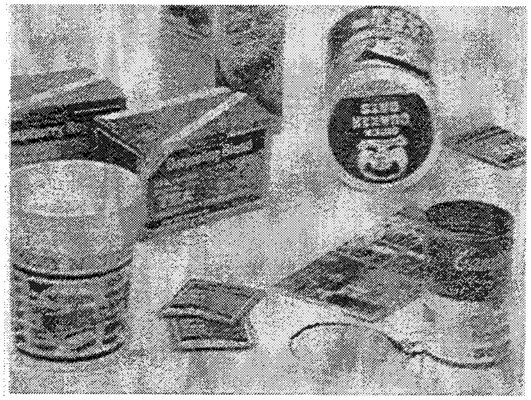

Fig. 1. First frame of sequence with rotating object.

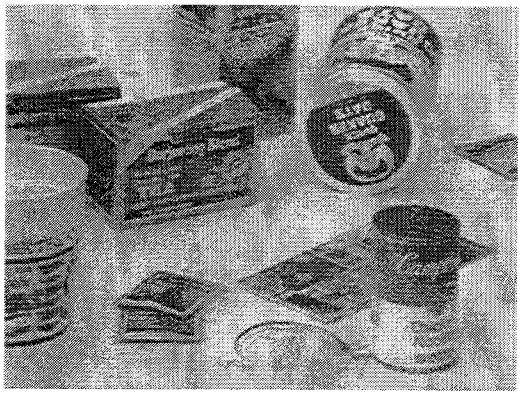

Fig. 2. Last frame of sequence with rotating object.

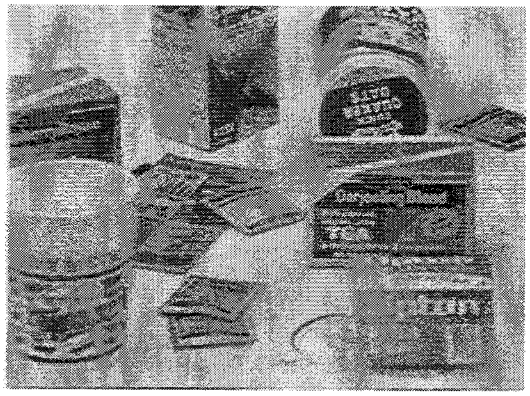

Fig. 3. First frame of sequence with translating object.

and used in the outlier detection process. Empirical testing later showed that solutions were reasonably stable for any sample larger than about 12 sets of points. The particular set of three features leading to the minimum median residual over the entire collection of feature point correspondences was found. Those points in the complete set with a residual value in excess of the threshold given by (5) were marked as outliers inconsistent with a single, rigid motion of the background. The values of C for the first and second sequence were 3.794332 and 3.788867, respectively. The value of C in (5) was defined prior to any experimental results based on a method similar to that described in [12]. No subsequent tuning of the parameter was performed. Subsequent experimentation has shown that results are relatively stable for thresholds within ± 20% of this value.

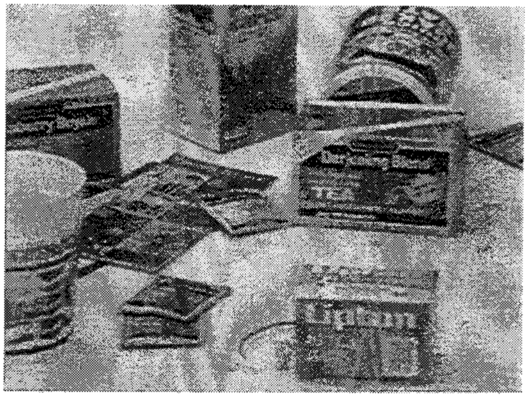

Fig. 4. Last frame of sequence with translating object.

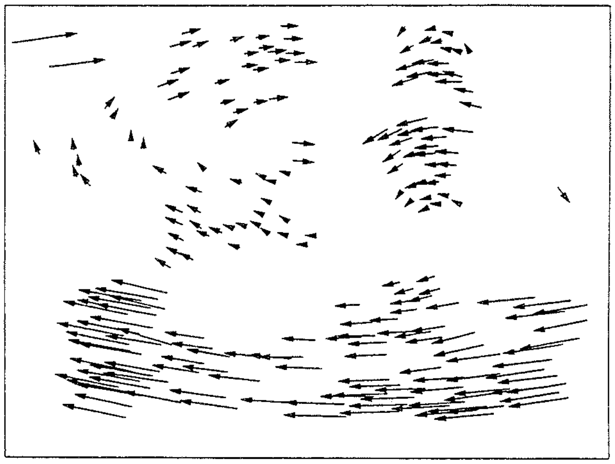

Fig. 5. Feature point disparities for Figs. 1 and 2.

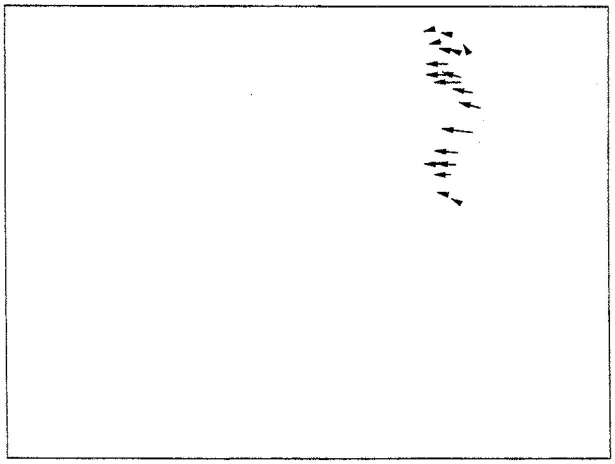

Fig. 6. Detected outliers in Fig. 5.

Fig. 5 shows the disparities found for the sequence starting with Fig. 1. Fig. 6 includes only those disparity values in Fig. 5 that were flagged as outliers. Two hundred and fourteen feature points were matched across the complete sequence. Of these, 46 were actually on the moving object. The outlier detection method flagged 19 of these moving object points correctly. None of the background points was erroneously flagged. Figs. 7 and 8 give the disparity vectors and detected outliers for the sequence starting with Fig. 3. Two hundred and twenty eight feature points were tracked over the sequence, and

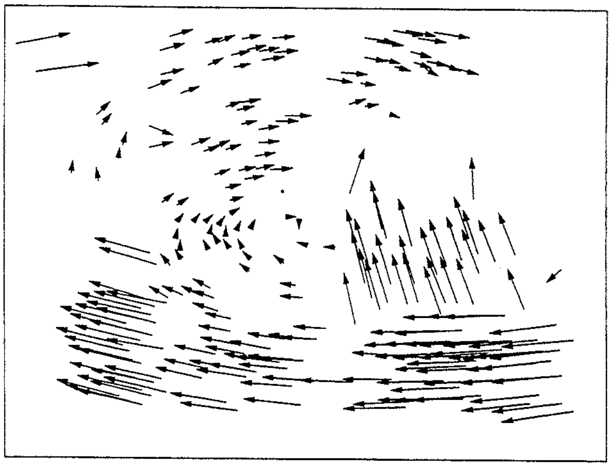

Fig. 7. Feature point disparities for Figs. 3 and 4.

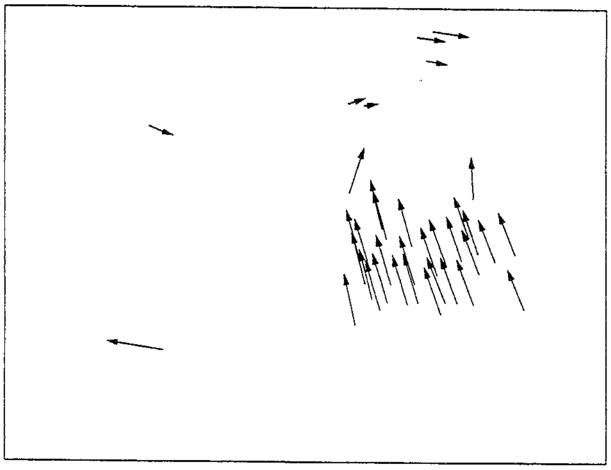

Fig. 8. Detected outliers in Fig. 7.

28 of them were on the moving object. All 28 of the moving object points were detected. An additional eight background points were also flagged (4% of the total background points).

IV. Discussion

Few of the structure-from-motion algorithms proposed in the literature have been demonstrated to work on real imagery. The method we present is able to perform well using automatically determined point correspondences that are rather noisy and contain a number of outright mismatches. The assumption of orthographic projection is only approximated in our imaging setup. The linear algorithm is simple to describe and implement but is likely to be less stable than many nonlinear formulations. The algorithm depends only on the point correspondences between two frames and involves no assumptions about the camera motion. The method works as well as it does because it embodies principles articulated in [14], including the principle of least effort. We do qualitative analysis that looks for large effects—the presence of outliers—rather than trying to accurately determine quantitative parameters that depend on subtle differences in the input data. In fact, the method does estimate quantitative motion parameters as an intermediate step. With simulated data, these parameters can be recovered accurately. With real data (or simulations involving substantial noise), the recovered rotation values may bear little if any relationship to the actual camera motion, but moving objects are still recoverable. This result is observed for the experimental data presented here.

It is important to note the limitations of the specific technique we have presented. Perhaps most importantly, many situations involving independently moving objects are compatible with a single rigid motion observed in only two orthographic views. The large number of moving object points not flagged in Fig. 6 is an example. These are points with disparities that are plausibly consistent with the motion of the background, given the limited power of the contraints used in the specific method we have described. Such situations are not detectable by this method unless information about depth is also available. In addition, this technique relies on the selection of a "distinguished point," which is assumed to lie on the rigid object. For effective use of this method, consistency checks would be necessary to ensure that the chosen point is not from the moving object. One such test would be (3) since a poor choice for the distinguished point yields an estimate vector that does not correspond to a feasible rotation in some cases. Further tests would also be desirable. In addition, detection can almost certainly be improved by considering perspective projection (more than two frames) and/or continuity of motion over longer intervals, although the nonlinear nature of such a problem will significantly complicate the construction of reliable outlier detection procedures. Our goal, however, is not to argue for a specific algorithm but instead to point towards a general approach we feel is likely to be successful.

The question of computational complexity is more difficult to address. Our approach, together with most of the other "robust" methods for computer vision, involves substantial computation. Clearly, parallel implementations are appropriate and can produce significant speedups. We suspect that a more fundamental answer may be to move away from median-based optimization methods that require some form of sorting and towards mode-based estimations. Under appropriate circumstances, distribution modes are an efficiently computed robust estimator. The success of many "relaxation labeling" and "connectionist" algorithms for specific vision problems can be better understood by recognizing that the essential component is the extraction of the mode of some simple function of the input data.

Acknowledgment

B. Schunck first pointed out to us the value of least median squares solutions to problems of this sort.

References

[1] R. Jain, D. Militzer, and H. -H. Nagel, "Separating non-stationary from stationary scene components in a sequence of real world TV images," *Proc. Fifth Int. Joint Conf. Artif. Intell.*, 1977, pp. 425–428.

[2] R. Jain, "Segmentation of frame sequences obtained by a moving observer," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. PAMI-6, pp. 624–629, Sept. 1984.

[3] W. Thompson and T. Pong, "Detecting moving objects," *Int. J. Comput. Vision*, vol. 4, pp. 39–57, Jan. 1990.

[4] R. Nelson, "Qualitative detection of motion by a moving observer," in *Proc. IEEE Conf. on Comput. Vision Patt. Recogn.* (Lahaina, HI), 1991, pp. 173–178.

[5] S. Ullman, *The Interpretation of Visual Motion.* Cambridge, MA: MIT Press, 1979.

[6] D. Heeger and G. Hager, "Egomotion and the stabilized world," in *Proc. Second Int. Conf. Comput. Vision*, 1988, pp. 435–440.

[7] Z. Zhang, O. Faugeras, and N. Ayache, "Analysis of a sequence of stereo scenes containing multiple moving objects using rigidity constraints," in *Proc. Second Int. Conf. Comput. Vision*, 1988, pp. 177–186.

[8] T. Huang and C. Lee, "Motion and structure from orthographic projections," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 11, pp. 536–540, May 1989.

[9] R. Haralick *et al.*, "Pose estimation from corresponding point data," *IEEE Trans. Syst. Man Cybern.*, vol. 19, pp. 1426–1446, Nov./Dec. 1989.

[10] *Proc. Int. Workshop Robust Comput. Vision* (Seattle, WA), 1990.

[11] P. Meer, D. Mintz, A. Rosenfeld, and D. Kim, "Robust regression methods for computer vision: A review," *Int. J. Comput. Vision*, vol. 6, pp. 59–70, Apr. 1991.

[12] P. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection.* New York: Wiley, 1987.

[13] S. Barnard and W. Thompson, "Disparity analysis of images," *IEEE Trans. Patt. Anal. Machine Intell*, vol. PAMI-2, pp. 333–340, July 1980.

[14] W. Thompson and J. Kearney, "Inexact vision," in *Proc. Workshop Motion: Representation Analysis*, 1986, pp. 15–21.