3D Structure and Motion Estimation from 2D Image Sequences†

T. N. Tan, K. D. Baker and G. D. Sullivan
Intelligent Systems Group
Department of Computer Science
University of Reading, ENGLAND

Abstract

Two novel algorithms are presented in this paper for depth estimation using point correspondences and the ground plane constraint. One is a direct non-iterative method, and the other a simple well-behaved iterative technique where the choice of initial value is straightforward. The algorithms are capable of handling any number of points and frames as well as points which become occluded. Once the point depths are determined, motion parameters can be obtained by a linear least squares technique. Extensive test results are included which show that the proposed algorithms are robust to noise, and perform satisfactorily using real outdoor image sequences.

1 Introduction

In previous work [16-17], we have shown that the ground plane constraint (the fact that objects, such as road vehicles, are often confined to move on the ground surface) can be used to develop simple and robust structure from motion (SFM) algorithms using point correspondences from pairs of image frames. In this paper, we discuss the use of multiple (more than two) image frames. We call SFM algorithms that use multiple frames the Multiple Frame SFM or simply MFSFM algorithms. A MFSFM algorithm can be either recursive or batch in nature depending on whether it processes one frame at a time or all frames simultaneously. In general, batch approaches have been shown to be both more accurate and stable [7]. The algorithms presented in this paper belong to the batch group.

Many MFSFM algorithms have been reported [1, 2-12]. These algorithms, however, have a number of limitations: unrealistic assumptions about object and/or camera motion [3-5, 7, 9-11]; requirement of good initial guesses to initialise the iteration process; high computational complexity [2, 7, 9, 11-12]; failure to handle feature occlusion; and unknown performance in real image data [3-4, 6, 8, 11]. These difficulties are mostly due to the scale of the task of solving the six degrees of freedom non-linear problem allowed by the existing MFSFM algorithms.

Many practical tasks in vision need be concerned with fewer degrees of freedom, and object motion is often subjected to physical constraints, such as the commonly occurring ground plane constraint. We show in this paper that the ground plane constraint can be used to develop simple and robust MFSFM algorithms which avoid the

†. This work was carried out as part of the ESPRIT project P2152 (VIEWS).

BMVC 1992 doi:10.5244/C.6.8

above problems. The work presented in this paper has mainly been motivated by the desire to apply machine vision in automatic monitoring and surveillance in airport and road traffic, but is also applicable to a wide range of potential industrial applications. With autonomous vehicles, for example, the ground plane constraint is equivalent to assuming that the camera is at a known fixed height, tilt and roll. This is frequently the case, at least for brief periods. The algorithms therefore provide robust and efficient methods for the recovery of unknown obstacles for robots moving on a flat surface.

The ground plane constraint is defined in the next section. Section 3 describes two different techniques for recovering point depths using multiple frames and the ground plane constraint. Section 4 outlines an algorithm for optimal 3D motion parameter estimation. Experimental results are presented in Section 5.

2 The ground plane constraint

The scenes considered in this paper concern airport or road traffic, where objects (e.g., aeroplanes, vehicles, etc.) are confined to move on the ground surface, which is, at least in the local region of our interest, approximately planar. We represent the ground surface by the X-Y plane of a WCS whose Z-axis points upwards. The movement of an object only has three degrees of freedom: translations (T_x and T_y) along the X and Y axes on the ground plane, and rotation (θ) about the vertical Z axis. The other three motion parameters, i.e., the rotations (α and β) about the X and Y axes, and the vertical translation (T_z), are all zero:

$$\alpha, \beta, T_z = 0 \tag{1}$$

We call this the *ground plane constraint* (GPC). We observe that when object motion is expressed in the camera-centred frame (as is usually the case in the existing SFM algorithms), then the number of unknown motion parameters under the GPC cannot, in general, be reduced to less than four (although the unknowns have to satisfy one or more equation computable from the GPC). This simply means that the GPC can be used most effectively only by SFM algorithms (such as those presented in this paper) that are defined in the WCS.

The GPC ensures that points on the object are constrained to move in planes parallel to the ground plane. With known camera parameters, there is a one to one correspondence between any such plane and the image. Hence if we know the depth from the camera of a point in one frame, then the plane on which the point is confined to move is uniquely determined, and the depth of the same point in any other frame can easily be computed. In fact it can be shown that under the GPC, the depth λ_j of a point in the jth frame is related to its depth λ_i in the ith frame by [18]

$$\lambda_j = (W_i / W_j)\, \lambda_i \tag{2}$$

where W_i and W_j are terms computable from known camera parameters and image coordinates.

In the subsequent discussions, we assume that the motion of an image sequence of an object is described by the motions of the object w.r.t. its pose in an arbitrarily chosen frame (we call it the reference frame). Further discussions on the GPC and its use in

model-based object pose recovery are described in a companion paper [20].

3 Depth estimation

We now discuss the estimation of point depths (structure parameters) from given point correspondences. We define the following symbols:

$S_F = \{F_0, F_1, F_2, \ldots, F_{M-1}, F_M\}$: the set of $M+1$ frames in which points have been detected and matched, and F_0 is used as the reference frame;

$S_P = \{P_1, P_2, \ldots, P_{N-1}, P_N\}$: the set of points appearing in S_F;

$S_{Pm} = \{P_{m1}, P_{m2}, \ldots, P_{mN_m}\}$: the set of points present in frame F_m, i.e., $S_{Pm} \in S_P$.

S_{Fi}: the set of frames in which point P_i is present;

$S_{Fij} = S_{Fi} \cap S_{Fj}$: the set of frames in which both P_i and P_j appear.

We do not require $S_{Pm} = S_{Pn}$, $m \neq n$, thus point occlusions are allowed. For convenience, we assume all points are present in the reference frame, i.e., $S_{P0} = S_P$. Let the 3D structure of an object be defined by the depths $\lambda_1, \lambda_2, \ldots, \lambda_N$ of N points in the reference frame. The problem to be solved is: Given S_F and $S_{Pm}, m \in \{0, 1, 2, \ldots, M\}$, determine $\lambda_1, \lambda_2, \ldots, \lambda_N$. Two solutions to this problem are given.

3.1 The direct non-iterative solution

We first consider two points P_1 and P_2 in two frames F_0 and F_m ($F_m \in S_{F12}$). According to the distance invariance property [14] of the rigidity constraint [15], the distance between P_1 and P_2 in F_0 is the same as the distance between the two points in F_m. From (2), this gives the following second-order polynomial equation on the depths λ_1 and λ_2 (both associated with F_0) of P_1 and P_2 [16-17]:

$$A_{m1}\lambda_1^2 + B_{m12}\lambda_1\lambda_2 + A_{m2}\lambda_2^2 = 0 \tag{3}$$

where subscript m signifies F_m, and A_{m1}, B_{m12} and A_{m2} are terms computable from known parameters such as image coordinates and extrinsic camera parameters. Their expressions can be found in [16-17]. By considering the two points in F_0 and each of the other frames in S_{F12}, one at a time, a set of second-order polynomial equations on λ_1 and λ_2 can be obtained:

$$A_{m1}\lambda_1^2 + B_{m12}\lambda_1\lambda_2 + A_{m2}\lambda_2^2 = 0, \quad \forall m, F_m \in S_{F12} \tag{4}$$

The number of equations in (4) equals to the number of frames in S_{F12} or $\#S_{F12}$. Since all constraint equations in (4) are homogeneous in λ_1 and λ_2, depth can only be solved from (4) up to a global scale. We therefore arbitrarily choose $\lambda_1 = 1$, and (4) becomes a set of quadratic equations in λ_2:

$$A_{m2}\lambda_2^2 + B_{m12}\lambda_2 + A_{m1} = 0, \quad \forall m, F_m \in S_{F12} \tag{5}$$

which can easily be solved for each equation separately using the standard formula. Let L_2 denote the set of all positive roots obtained from (5) (note: according to definition, λ_2 must be positive). Then the task is to derive a suitable solution for λ_2 from L_2. Each equation in (5) produces up to two positive depth solutions. If an equation in (5) does have two distinct positive roots, then one is valid, and the other is due to the reflection caused by the use of the distance invariance property in deriving the depth constraint equations. Therefore L_2 can be divided into two subsets L_{2T} and L_{2F}, with L_{2T} representing the set of physically valid solutions, and L_{2F} the false solutions. We thus first detect L_{2T} from L_2 (for a simple technique, see [18]), and then define the median of L_{2T} as the final solution for λ_2:

$$\lambda_2^{(1)} = \textbf{median}\,(L_{2T}) \tag{6}$$

where superscript (1) indicates that the depth solution was obtained by using P_1 as the reference point (i.e., the point whose depth was initially set to 1). By maintaining $\lambda_1 = 1$, we can compute depths of all other points in S_P in a similar way. We write all these solutions collectively as $(\lambda_1^{(1)}, \lambda_2^{(1)}, \ldots, \lambda_i^{(1)}, \ldots, \lambda_{N-1}^{(1)}, \lambda_N^{(1)})$ with $\lambda_1^{(1)} = \lambda_1 = 1$.

These solutions have been obtained by treating the point P_1 as a reference point. If this point is disturbed by noise, then the resulting depths of points $P_2, P_3, \ldots, P_N$ will be in error. To avoid this bias towards P_1, we repeat the above process using each P_i as the reference point independently. This generates N sets of depths for the given N points in S_P:

$$\{\,(\lambda_1^{(n)}, \lambda_2^{(n)}, \ldots, \lambda_i^{(n)}, \ldots, \lambda_{N-1}^{(n)}, \lambda_N^{(n)}) :\ n = 1, 2, \ldots, N\,\} \tag{7}$$

where $\lambda_n^{(n)} = 1, n = 1, 2, \ldots, N$, and the superscript n indicates the depths computed under reference point P_n. The depths of each set in (7) may be normalised with respect to (say) the first depth of the set to give

$$\{\,(\bar{\lambda}_1^{(n)}, \bar{\lambda}_2^{(n)}, \ldots, \bar{\lambda}_i^{(n)}, \ldots, \bar{\lambda}_{N-1}^{n}, \bar{\lambda}_N^{(n)}) :\ n = 1, 2, \ldots, N\,\} \tag{8}$$

where $\bar{\lambda}_1^{(n)} = 1, n = 1, 2, \ldots, N$. Then the final solution for the depths of the N points in the reference frame F_0 is defined as

$$\lambda_i = \textbf{median}\,\{\bar{\lambda}_i^{(n)}, n = 1,2,\ldots,N\}, i = 1, 2, \ldots, N \tag{9}$$

(9) is justified by the fact that all sets of normalized depths in (8) describe the same relative structure of the given N points.

3.2 The non-linear minimization solution

Given an initial value for λ_2, the depth constraint equations in (5) may also be solved simultaneously using the standard non-linear least squares technique. Then the steps

described in (7)-(9) can be followed to get the final depth solutions. For detailed descriptions, the reader is referred to [18].

Several remarks can be made at this point. Under normal viewing conditions, the depth range within an object (i.e., the maximum depth difference of points on the object) is much smaller than the nominal depth of the object. Therefore, the depth value assigned to the reference point provides a good initial guess for the depths of all other points. This makes the choice of initial guesses in this non-linear minimization approach a trivial matter indeed. Since the iteration process involves only one unknown and is provided with a good initial guess, its convergence to the correct solution is extremely fast. The total number of iterations required is typically three and rarely exceeds five.

Once the point depths in the reference frame are determined, those in other frames can easily be obtained using (2). If required, the 3D world coordinates may be computed from known image coordinates and the determined depths [18].

4 Estimation of 3D motion parameters

The motion parameters to be determined consist of the translational and rotational parameters of all frames w.r.t. the reference frame. Under the GPC, the motion between the reference frame F_0 and frame F_m is characterized by three independent motion parameters (expressed in the WCS): the translations T_{xm} and T_{ym} on the ground plane, and the rotation angle θ_m about the Z-axis. It can be shown that using the 3D world coordinates of the points in the reference frame computed in the preceding section and the given 2D image coordinates in frame F_m of the N_m points in F_m, a set of $2N_m$ constraint equations on T_{xm}, T_{ym} and θ_m can be derived [18]:

$$\begin{cases} D_{mi1}\cos\theta_m + E_{mi1}\sin\theta_m + F_{mi1}T_{xm} + G_{mi1}T_{ym} = H_{mi1} \\ D_{mi2}\cos\theta_m + E_{mi2}\sin\theta_m + F_{mi2}T_{xm} + G_{mi2}T_{ym} = H_{mi2} \end{cases}, \forall i, P_i \in S_{Pm} \qquad (10)$$

where D, E, F, G and H are terms computable from known image and world coordinates. By regarding $\cos\theta_m$ and $\sin\theta_m$ as two independent unknowns, (10) can be solved using the standard linear least squares technique to get $\cos\theta_m$, $\sin\theta_m$, T_{xm} and T_{ym}. θ_m is then computed as $\theta_m = \tan^{-1}(\sin\theta_m / \cos\theta_m)$. The correct quadrant of θ_m is determined from the senses of $\cos\theta_m$ and $\sin\theta_m$. Motion parameters of other frames in S_F can be obtained similarly.

5 Experimental results

The two proposed algorithms have been tested using both synthetic and real outdoor image sequences.With the synthetic data, Monte Carlo simulations were conducted as follows. An object was specified by N points randomly chosen from within a cuboid. A sequence of frames was then generated by moving the object on the ground plane. The ideal image coordinates of the points in each frame were perturbed by noise. Relative estimation errors were recorded during simulation. The relative error in a motion parameter was obtained by computing the average absolute relative error in the parameter over all frames in a trial, and then calculating the mean of this over all trials. The accuracy of the recovered 3D structure was measured by the *standard scene error*

(SSE) defined as the average Euclidean distance in the reference frame between the original and the reconstructed points. The SSE was computed at each trial, and its mean over all trials was divided by the diameter of the synthetic cuboid model to yield the relative SSE measure.

5.1 Robustness against image data noise

Noise was simulated by adding zero-mean, uniformly distributed random values to the ideal image coordinates of all points in all frames, the level of noise given by ΔE (in pixels) defining a uniform distribution interval $[-\Delta E, +\Delta E]$. Monte Carlo simulations were performed to investigate the noise robustness of the proposed algorithms using a fixed number of points (=10) in a fixed number of frames (=10). The results are summarized in Fig.1. It can be seen [18] that the overall performances of the two

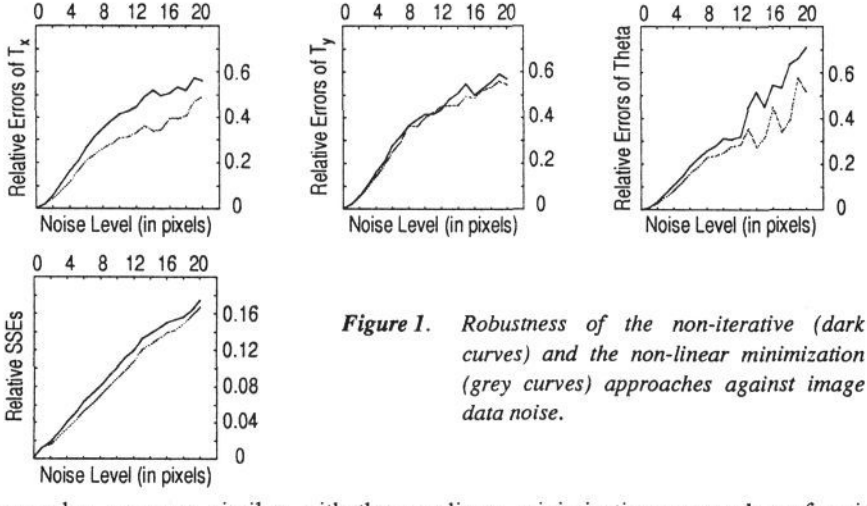


Figure 1. *Robustness of the non-iterative (dark curves) and the non-linear minimization (grey curves) approaches against image data noise.*

approaches are very similar, with the non-linear minimization approach performing slightly better than the direct technique. Both algorithms are very robust against image data noise. The relative errors in the motion parameters rarely exceed 60%, and the relative SSE is always less than 18% even using unrealistically high noise levels of ±20 pixels.

5.2 Effectiveness of using more frames in noise reduction

Monte Carlo simulations were also carried out to study the benefits of using longer image sequences (i.e., more frames) in noise reduction. The number of points used was fixed at 10, and the noise level was maintained at $\Delta E = 5$ pixels. The results are given in Fig.2. The robustness of the two algorithms is consistently improved by using longer image sequences, with most improvement when the number of frames increases from 3 to 6. Further increase in the number of frames beyond 15 results in barely noticeable improvement.

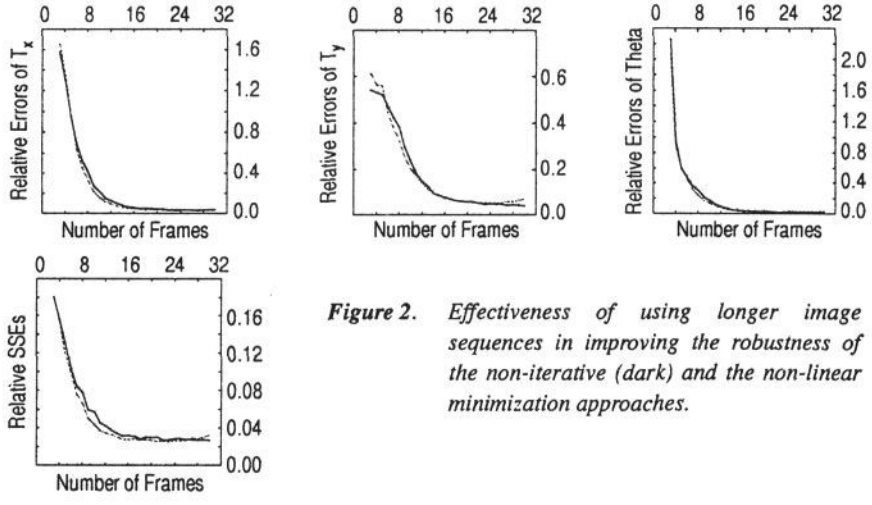


Figure 2. *Effectiveness of using longer image sequences in improving the robustness of the non-iterative (dark) and the non-linear minimization approaches.*

5.3 Effectiveness of using more points in noise reduction

In a further experiment, we kept unchanged the number of frames used (=10) and the level of noise involved (ΔE = 5.0 pixels), and varied the number of points used to examine the effect of the number of points in combating noise. Monte Carlo simulation results are summarized in the plots in Fig.3. The results show that the robustness of the

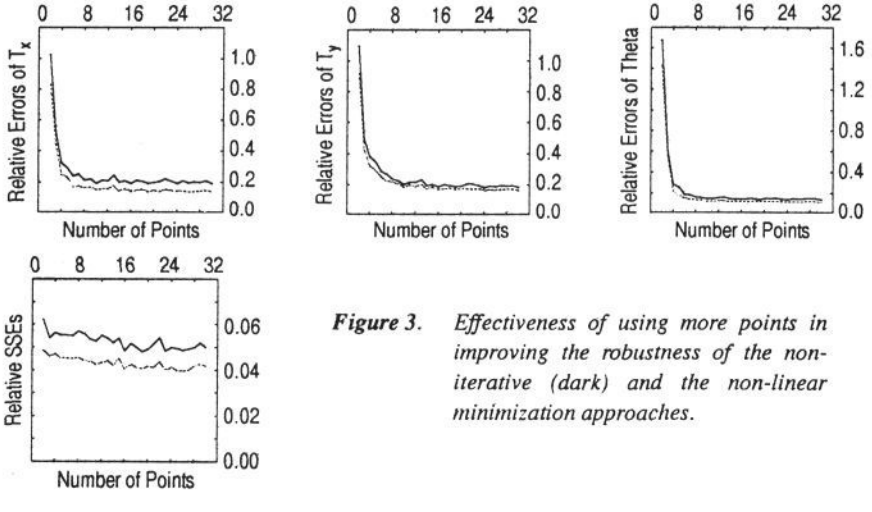


Figure 3. *Effectiveness of using more points in improving the robustness of the non-iterative (dark) and the non-linear minimization approaches.*

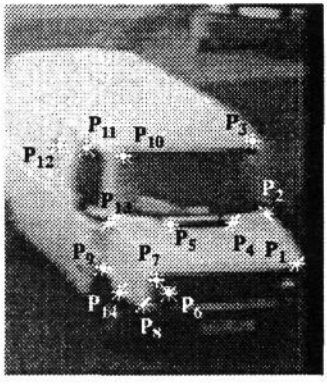
two algorithms is consistently improved by using more points. The improvement is most dramatic when the number of points is increased from 2 to 4, and is only marginal for additional points.

5.4 Performance using real outdoor image sequences

The proposed algorithms have also been tested using outdoor image sequences. The Plessey corner finder [19] was applied to detect corner points on a slowly moving (reversing) van (see Fig.4). The image plane trajectories of 14 detected corners over 10 frames were then used as the input to the MFSFM algorithms. Because of the smoothly curved body of the van, the physical corners are not well defined, as well as the limited accuracy of the corner finder, the corner trajectories are subject to significant measurement errors. Table 1 lists the heights (in meters) of the 14 points recovered by

Table 1. Recovered heights of 14 van points

Point	P_1	P_2	P_3	P_4	P_5
Height	0.74	1.08	1.85	1.10	1.10
Point	P_6	P_7	P_8	P_9	P_{10}
Height	0.45	0.57	0.40	0.59	1.73
Point	P_{11}	P_{12}	P_{13}	P_{14}	
Height	1.67	1.72	1.03	0.42	



the non-linear minimization algorithm, where the global scale has been set by assuming P_8 to be 0.4m high. Since the ground truth is not available, the quantitative measures used for synthetic data cannot be calculated. Qualitatively, however, the figures in Table 1 as a whole are consistent with our perception (e.g., relative heights) of the object. The results from the non-iterative approach are very close to those given in Table 1.

The performance of the algorithms using the outdoor image sequence can be further appreciated from Fig.4. The figure shows the originally detected (marked by x) and the reconstructed (marked by +) corner points overlaid on four consecutive frames of the van sequence. If no x is shown near a +, then the corner point marked by the + was not detected in the corresponding frame by the corner finder, but was "predicted" by the algorithm based on the recovered structure and motion. It can be seen that both the detected and the "missing" corners were reconstructed fairly accurately.

6 Conclusions

Novel algorithms have been presented in this paper for 3D structure and motion estimation from 2D image sequences using the ground plane constraint. It has been shown that the depth parameters can be computed using either a non-iterative direct approach or a simple well-behaved non-linear minimization approach, and that the motion parameters can be estimated using the standard linear least squares technique.

The algorithms possess a number of desirable characteristics. They are very robust and perform satisfactorily with real outdoor image sequences. They do not require excessively large numbers of points and/or frames for satisfactory performance and are

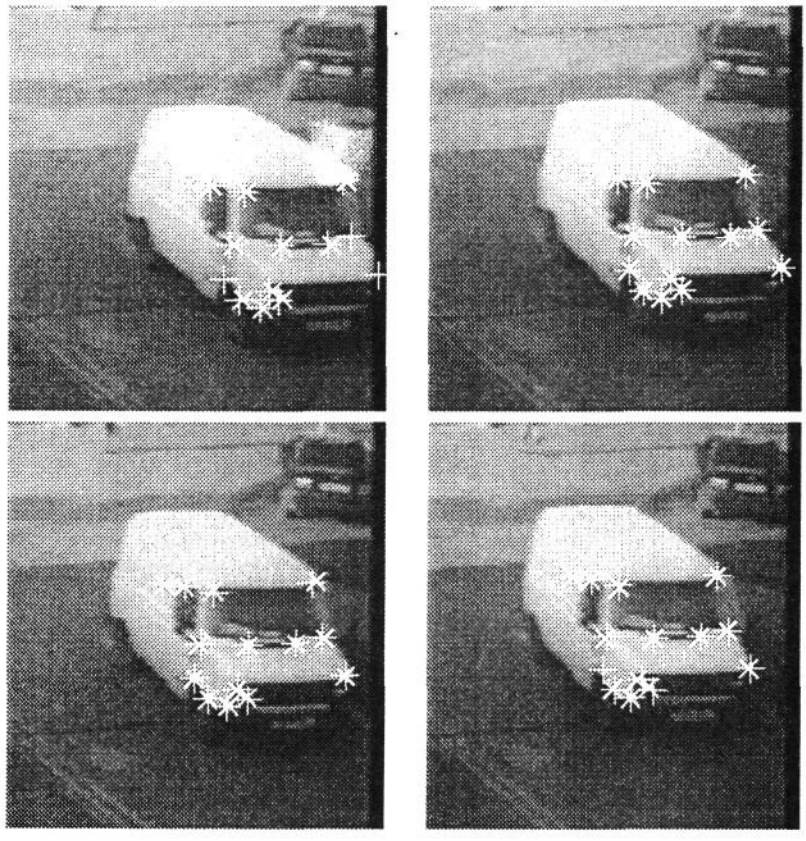


Figure 4. *Originally detected (marked by x) and reconstructed (marked by +) corners overlaid on four frames of a van sequence.*

capable of handling any number of points and/or frames as well as point occlusions. The algorithms are computationally very simple and highly parallel in nature. All these make the algorithms very desirable where applicable. They may be used for a wide range of potential industrial applications.

References

[1] J. K. Aggarwal and N. Nandhakumar, On the Computation of Motion from Sequences of Images - A Review, Proc. of IEEE, vol.76, 1988, pp.917-935.

[2] L. Dreschler and H. -H. Nagel, Volumetric Model and 3D Trajectory of a Moving Car Derived from Monocular TV Frame Sequences of a Street Scene, CGIP, vol.20, 1982, pp.199-228.

[3] Y. Yasumoto and G. Medioni, Robust Estimation of Three-Dimensional Motion Parameters from a Sequence of Image Frames Using Regularization, IEEE Trans. PAMI, vol.8, 1986, pp.464-471.

[4] J. Y. Weng, T. S. Huang and N. Ahuja, 3-D Motion Estimation, Understanding, and Prediction from Noisy Image Sequences, IEEE Trans. PAMI, vol.9, 1987, pp.370-389.

[5] H. Shariat and K. E. Price, Motion Estimation with More Than Two Frames, IEEE Trans. PAMI, vol.12, 1990, pp.417-434.

[6] C. Jerian and R. Jain, Polynomial Methods for Structure from Motion, IEEE Trans. PAMI, vol.12, 1990, pp.1150-1166.

[7] T. J. Broida and R. Chellappa, Estimating the Kinematics and Structure of a Rigid Object from a Sequence of Monocular Images, IEEE Trans. PAMI, vol.13, 1991, pp.497-513.

[8] M. Spetsakis and J. Aloimonos, A Multi-frame Approach to Visual Motion Perception, Inter. J. Comput. Vision, vol.6, 1991, pp.245-255.

[9] R. Kumar, A. Tirumalai and R. C. Jain, A Non-linear Optimization Algorithm for the Estimation of Structure and Motion Parameters, Proc. of CVPR'89, June 4-8 1989, San Diego, USA, pp.136-143.

[10] H. S. Sawhney, J. Oliensis and A. R. Hanson, Description and Reconstruction from Image Trajectories of Rotational Motion, Proc. of ICCV'90, December 1990, Osaka, Japan, pp.494-498.

[11] G. S. Young, R. Chellappa and T. H. Wu, Monocular Motion Estimation Using a Long Sequence of Noisy Images, Proc. of IEEE Inter. Conf. on ASSP, 1991, pp.2437-2440.

[12] N. Cui, J. Y. Weng and P. Cohen, Extended Structure and Motion Analysis from Monocular Image Sequences, Proc. of ICCV90, December 1990, Osaka, Japan, pp.222-229.

[13] J. K. Aggarwal and A. Mitiche, Structure and Motion from Images: Fact and Fiction, The 3rd IEEE Workshop on Vision: Representation and Control, October 1985, pp.127-128.

[14] A. Mitiche and J. K. Aggarwal, A Computational Analysis of Time-Varying Images, in Handbook of Pattern Recognition and Image Processing, T. Y. Young and K. S. Fu, Eds., New York: Academic Press, 1986.

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

[16] T. N. Tan, G. D. Sullivan, and K. D. Baker, Structure from Constrained Motion Using Point Correspondences, British Machine Vision Conference 1991, P. Mowforth Ed., Springer-Verlag, 1991, pp.301-309.

[17] T. N. Tan, G. D. Sullivan, and K. D. Baker, Structure from Motion Using the Ground Plane Constraint, Proc. of ECCV-92, G. Sandini Ed., LNCS-Series Vol.588, Springer-Verlag, 1992.

[18] T. N. Tan, 3D Structure and Motion Estimation from 2D Image Sequences Using the Ground Plane Constraint, ESPRIT II project (P.2152) research report, RU-03-WP•T2122-TNT-03, University of Reading, April 1992.

[19] J. A. Noble, Finding Corners, Proc. of 3rd Alvey Vision Conf., University of Cambridge, England, 15-17 September 1987, pp.267-274.

[20] T. N. Tan, G. D. Sullivan and K. D. Baker, Linear Algorithms for Object Pose Estimation, Proc. of British Machine Vision Conference 1992, Springer-Verlag, 1992.