Volumetric Model and 3D-Trajectory of a Moving Car Derived from Monocular TV-Frame Sequences of a Street Scene

L. Dreschler and H.-H. Nagel
Fachbereich Informatik
Universitaet Hamburg
Schlueterstrasse 70
2000 Hamburg 13
Germany

Abstract

A polyhedral approximation for the volumetric description of a moving rigid object from a real-world scene is derived, based on measurements in monocular TV-frame sequences. The trajectory and attitude of the object motion relative to the camera is simultaneously determined up to the same factor which scales the object description. Results from one street scene sequence are presented. The approach is compared to related ones reported in the recent literature.

1. Introduction

Technological progress throughout the past decade has made it feasible to record, digitize, store, and evaluate image sequences of real-world scenes with moving objects. A comprehensive survey of application-oriented research which attempts to exploit such possibilities has been given by Nagel [30]. It provides clear evidence from numerous application areas that progress in the evaluation of image sequences is closely related to improvements in modeling the depicted scene and its temporal variations.

Nagel [29] outlined an approach to derive descriptions of rigid objects in motion from monocular TV-frame sequences. The current contribution reports results from efforts to implement this approach.

Despite a recent burst of publications about image sequence analysis, no directly comparable results are known to us, yet. Numerous investigators have been concerned with problems which appear as subproblems in our context.

Object tracking in TV-frame sequences from real-world scenes has been reported recently by Gilbert et al. [16, 17], Radig [38], Yachida et al. [46], Gerlach [15], Landzettel and Hirzinger [23], Hirzinger et al. [18], Korn and Kories [21], Bers et al. [5]. Although some of these systems have provisions to cope with partial or complete occlusion of the object to be tracked and, therefore, are more robust than our approach as far as tracking is concerned, none of them actually attempts to derive a 3D description of the moving object.

Various descriptors, ranging from the mere 2D position of a dot in a binary image up to complex relational structures have been used to characterize the image of the same space point in each frame of an image sequence.

Depending on the type of descriptors, different methods for matching descriptors from one image frame to the next have been developed in order to solve the correspondence problem (Duda and Hart [14]). We modified a relaxation approach reported by Barnard and Thompson [3]. Other approaches are discussed by, e.g., Ullman [42], Radig et al. [39], Kraasch et al. [22], Jacobus et al. [19] and Cheng and Huang [8].

Using such techniques, a series of corresponding image coordinates can be extracted from an image sequence for a set of points which are hypothesized to be located on the same rigid moving object. Provided certain conditions are satisfied, the 3D configuration of these points as well as their space trajectory can be determined from such measurements. In the case of orthographic projection, this is guaranteed by the "structure-from-motion" theorem of Ullman [42]. The determination of translation and rotation for the perspective case has been studied by Nagel [32], Meiri [25], and Nagel and Neuinmann [33]. Similar investigations, but based on optica] flow, have been retorted earlier by Prazdny [36, 37]. Special situations in optical flow analysis for the derivation of 3D descriptions have been studied by Clocksin [9, 10, 11], by Lawton [24] and by Williams [45].

Neumann [34] used simulated data in a search approach which attempts to solve simultaneously the problems of grouping image points together as representing the same object, of finding the correspondences between images of the same point from different frames and of determining the interframe translation and rotation parameters. A numerical minimization approach towards 3D reconstruction investigated with simulated data has been reported by Roach and Aggarwal [41]. Our results have been obtained with a minimization approach described by Bonde and Nagel [7, 31].

Since we are restricting ourselves to rigid 3D point configurations, recent work by Rashid [40], Asada et al. [1], as well as Webb and Aggarwal [44] on jointed objects is outside our

scope. We exclude, too, model-based approaches such as those of O'Rourke and Badler [35] or Wallace and Mitchell [43]. These authors assume that object-specific knowledge in the form of 3D prototypes is available whereas we strive for deriving 3D descriptions based solely on the observed image sequence and on general assumptions like the rigidity hypothesis. Moreover, all approaches quoted in this paragraph employ simulated input data rather than image sequences from real-world 3D scenes as in our case.

This highly condensed survey of the recently published literature should illustrate the conditions which characterize our approach: a model-free derivation of 3D descriptions based on perspective projections of moving rigid objects from real-world scenes as recorded in real-time by a monocular TV-frame sequence.

2. Decomposition of our Approach

The discussion of the relevant literature in the preceding section roughly reflected the main steps in our system:

(i) isolation of nonstationary image areas around the image of a moving object;
(ii) extraction of descriptors from the subimage corresponding to one object;
(iii) solution of the correspondence problem by matching descriptors between consecutive frames;
(iv) derivation of a rigid 3D point configuration and its frame-to-frame motion relative to the camera;
(v) determination of the convex hull for this 3D point configuration as an approximation to a volumetric model of the moving object.

In order to assess the result, the convex hull can be projected back into each image frame using computer graphic methods including hidden line removal and surface shading. This backprojection can be compared visually with the original object image in the corresponding frame.

Step (i) is based on an approach described by Jain and Nagel [20]. Details about the descriptor extraction and the interframe matching of descriptors will be discussed in subsequent sections. The fifth section will outline the derivation of 3D descriptions and present various results. The final section will discuss areas where further research is required in order to obtain a more robust approach, to refine the description or to extend its range of applicability.

3. Extraction of Descriptors for Object Point Candidates

Despite extensive experimentation, no segmentation algorithm is known to us which could reliably decompose digitized TV-images from real-world scenes into semantically meaningful regions without recourse to object-specific knowledge. Radig and coworkers segmented images from blocksworld scenes, fitted straight line segments to region boundaries, and selected an intersection of such straight line segments as candidate for a vertex image of a block [22, 39]. An extension of such an approach to images of - for example cars did not appear to be immediately feasible.

Inspired by the work of Barnard and Thompson [2, 3], we set out to select an image point with a high greyvalue variance in four directions - a "corner point" - as a candidate for the image of an identifiable point on the object surface. Barnard and Thompson [2, 3] employed an operator proposed and used by Moravec [26, 27, 28]. Detailed investigations showed however, that this operator could not cope adequately with data such as ours.

We began therefore, to investigate the location of extreme values of the local greyvalue curvature as a function of the raster plane coordinates. Beaudet [4] described operators to compute these curvatures from digitized images. Since the Gaussian curvature is the product of the two main curvatures, it appeared attractive initially to select local extrema of the Gaussian curvature as the characteristic of candidates for images of important object points. Experience has shown, however, that such extrema of Gaussian curvature appear often along lines of pronounced greyvalue transitions: one of the two main curvatures is very large due to the steep greyvalue transition, but the other one (along the direction of a linear transition front) is rather small and may change its value more or less due to noise. Nevertheless, the product of these two can give rise to sizable extrema.

The characteristic greyvalue distribution around an image point of interest to us may be likened to a promontory with steep slopes, extending into the sea with an acute angle of the shore line - see, e.g., the left upper corner of the rear car window in figs, Ib or Ib: the dark greyvalues of the window correspond to the promontory and the surrounding bright car body to the sea. A magnification of this image area is given in fig. 2 where bright pixels correspond to small greyvalues and dark pixels to large ones (greyvalues extend between 0 and 255).

At the pinnacle of the promontory, we have a local minimum of both main curvatures because there the intensity drops towards three directions (north, west, and south). Around the shoreline, we expect a local minimum (see remark above) of one main curvature - the one corresponding to the bend in the shoreline. Moreover, the second main curvature must exhibit an extremum of opposite sign in this image area because the steep slope of the promontory cliff has to flatten out into the

surrounding sea. Thus, we expect a local minimum of the Gaussian curvature somewhere in the sea close to the tip of the promontory - the product of both main curvatures which have opposite sign in this situation.

Fig. 2 presents the numerical values of the relevant parameters for all pixels in the enlarged section. The greyvalue is given in the upper left corner. The top number at the right side of each pixel area represents the more positive main curvature, the second number from above the smaller one (i.e. more towards the negative) of the two main curvatures. The third number represents the Gaussian curvature, i.e. the product of the two numbers above it. The line within each pixel area indicates the orientation of the more positive main curvature. Gaussian curvatures printed white on dark background represent the local extrema selected by the algorithm. We now select as a candidate for the image of an important object point the pixel with the steepest greyvalue slope along the line which connects the location of the maximum with the location of the minimum of the Gaussian curvature. Only one of the two main curvatures can change its sign if we trace along this line from the minimum to the maximum of the Gaussian curvature. The steepest greyvalue slope occurs where this main curvature crosses zero. We stipulate two additional checks that we have indeed combined a proper pair of extrema for the Gaussian curvature:

a) the orientation of the main curvature which changes its sign between the two extrema must indeed point into the direction of the associated extremum with opposite sign of the Gaussian curvature-
fa) the greyvalue at the location of the maximum of the Gaussian curvature must be larger (i.e. darker in this example - pinnacle 1) than the greyvalue at the associated location of the Gaussian curvature minimum (extremum with negative sign - at the base of the cliff corner 1).

A little white square is entered in fig. 2 at the center of the pixel area which corresponds to these conditions. As will be seen, there are other extrema of the Gaussian curvature. A cutoff radius of D=4 pixels is used to exclude extrema from being tentatively combined to form a candidate point. The curvatures have been computed with the 5x5 pixel operators of Beaudet [4]. Since this resolution appeared to be slightly too coarse, we refined the candidate locations by repeating the search with 3x3 operators. Since such a smaller operator mask is more sensitive to noise, it is applied only in environments where the 5x5 operator responded strongly. The result of this refinement step is presented in fig. 3.

In order to avoid misunderstanding, we do not claim to have a method by which we can reliably detect the images of prominent points on a 3D object surface. Our experience only indicates that we can select a reasonable set of candidates for such points. Supporting evidence from other frames is required to filter out those candidates which appear to be indeed images from 3D object points. Such evidence can be obtained by finding corresponding candidates in neighboring frames - and by a successful 3D description which appears to be compatible with observations throughout an image sequence.

4. Interframe Matching of Candidate Descriptors

We used the relaxation approach described by Barnard and Thompson [3]. Since our descriptors catch more of the relevant greyvalue structure than the Moravec operator employed by these authors, we could exclude candidate matches where the sign of the main curvatures at the corresponding extrema of the Gaussian curvature did not match. The details of the relaxation procedure used here differ from that described by Barnard and Thompson [3], see [12, 13].

Every other frame between those shown in figs, la and lb, i.e. a total of 12 frames has been treated in this manner. The descriptors matched through the relaxation procedure between consecutive pairs of frames are chained together. Fig. 4 depicts the resulting chains of descriptor locations, anchored at the car image from fig. lb. Operator interaction could be used to delete chains which appear unreliable, for example very short ones or those connected to an cluttered image area.

5. 3D Description and Results

A minimization approach has been developed which is based on the hypothesis that the image locations of all descriptors must be explained by a rigid 3D configuration of points which translate and rotate from frame to frame relative to the camera (for details see [31]). As explained in section 2, a convex hull is computed for the 3D point configuration in order to better visualize their 3D arrangement. The edges of the convex hull corresponding to the measurements discussed earlier have been projected back into the image of fig. la. This is possible because the relative location and attitude of the 3D point configuration with respect to the camera coordinate system have been determined in the course of the minimization for each frame used. This knowledge enables the suppression of edge lines which are hidden by the object itself in the depicted position - see fig. 5. It should be realized that the backprojection may contain vertices which need not have been extracted from the current frame. The vertices of the convex 3D hull represent the global description of all observations throughout the evaluated fraction of the image sequence.

Since the location and attitude of the convex hull relative to the camera is known for each

frame time, it is possible to depict the convex hull with a hypothetical illumination provided one assumes a reflectivity law for the faces of the convex hull. A simple choice is to assume Lambert's law. Fig. 6 depicts the convex hull corresponding to fig. 5, but illuminated by a hypothetical light source behind the camera. Results from additional frames of this sequence as well as from other sequences cannot be presented due to space limitations.

6. Conclusion

We have presented an approach for the extraction of 3D descriptors for moving rigid objects from monocular TV-frame sequences of real-world scenes. Currently, we approximate a volumetric representation of the moving object by the convex hull of the 3D point configuration which can be obtained by the methods described above. It is obvious that the 3D object description should be improved by extracting more information about it from the frames than merely the greyvalue "corners".

There are several areas where we want to improve our approach. One subproblem is a more reliable and more accurate estimate of the mask covering the image of the moving object in each frame. Furthermore, we have to study our approach for extraction of descriptors more extensively. There is the question, by which methods it might be possible to select those descriptors which are indeed images of identifiable points on the object surface. It is tempting to exploit approximate knowledge about the object and its motion as obtained from an initial evaluation of an image sequence in order to reevaluate the same image sequence. Such knowledge could be exploited to segment subimages containing the moving object and to use inference processes to exclude descriptors incompatible with the segmentation results. An alternative would be to employ such knowledge to evaluate an extension of the image sequence covering additional observation periods of the same scene with the same moving object.

Another problem area is the technique used for interframe matching of descriptors. Alternative approaches have to be investigated here. We would like, too, to develop algorithms which screen the descriptor chains already before we use their 2D coordinate measurements for the derivation of a 3D point configuration.

Despite all these open subproblems it is encouraging to know that one may learn at least an approximate 3D object description merely by observing the moving object. It thus appears possible to design a system where methods of model-based image analysis and understanding can be employed without providing object-specific models a priori.

7. Acknowledgement

We gratefully acknowledge many fruitful discussions with our collegues G. Hilie, B. Neumann, B. Radig, and H. Westphal. The opportunity to test our ideas on image sequences from real-world scenes has been provided by a long cooperation of all present and former members of our research group. In addition to those mentioned above, we gladly remember the continuing help of I. Heer, H. Kemen, K. Kleemann, W. Benn, H. Faasch, B. Fischer, St. Shafer (on visit from CMU) and many others who contributed to the hardware and software facilities of our laboratory. We thank Mrs. R. Jancke for her help in editing this text. We gratefully acknowledge a grant of the Deutsche Forschungsgemeinschaft which partially supported these investigations.

8. References

[1] Three Dimensional Motion Interpretation for the Sequence of Line Drawings
M. Asada, M. Yachida, and S. Tsuji
ICPR-80, pp. 1266-1273

[2] Disparity Estimation Using Feature Point Matching, S.T. Barnard and W.B. Thompson
WCATVI-79, p. 2

[3] Disparity Analysis of Images
S.T. Barnard and W.B. Thompson
IEEE Trans. Pattern Analysis and Machine Intelligence, PAMI-2 (1980) 333-340

[4] Rotationally Invariant Image Operators
P.R. Beaudet, IJCPR-78, pp. 579-583

[5] Object Detection in Image Sequences
K.H. Bers, M. Bohner, and H. Gerlach,
ICPR-80, pp. 1317-1319

[6] Untersuchungen zur dreidimensionalen Modellierung bewegter Objekte durch Analyse von Formveraenderungen der Objektbilder in TV-Aufnahmefolgen
T. Bonde, Diplomarbeit (Januar 1979)
FB Informatik, Universitaet Hamburg

[7] Deriving a 3-D Description of a Moving Rigid Object from Monocular TV-Frame Sequences, T. Bonde and H.-H. Nagel
WCATVI-79, pp. 44-45

C 8] Algorithms for Matching Relational Structures and their Application to Image Processing, J.-K. Cheng and T.S. Huang,
TR-EE 80-53 (December 1980)
School of Electrical Engineering
Purdue University, West Lafayette/IN

[9] Determining the Orientation of Surfaces from Optical Flow, W.F. Clocksin, Proc. AISB/GI-78 on Artificial Intelligence
Hamburg, July 18-20, 1978, pp. 93-102

[10] The Effect of Motion Contrast on Surface Slant and Edge Detection, W.F. Clocksin,
Proc. AISB-80 Conference on Artificial Intelligence, St. Hardy (ed.), Amsterdam, July 1-4, 1980

[11] Perception of Surface Slant and Edge Labels from Optical Flow: A Computational Approach, W.F. Clocksin,
Perception 9 (1980) 253-269

[12] Ermittlung markanter Punkte auf den Bildern bewegter Objekte und Berechnung einer 3D-Beschreibung auf dieser Grundlage
L. Dreschler, FB Informatik, Universitaet

Hamburg, Hamburg/Germany (in preparation)
[13] On the Frame-to-Frame Correspondence between Greyvalue Characteristics in the Images of Moving Objects L. Dreschler and H.-H. Nagel Proc. GI-workshop on AI (in press) Bad Honnef/Germany, January 26-30,1981
[14] Pattern Classification and Scene Analysis R.O. Duda and P.E. Hart, John Wiley & Sons, New York, 1973
[15] Digitale Bildfolgenauswertung zum Wiederfinden von Objekten in natuerlicher Umgebung, H. Gerlach, in: [47], pp. 199-207
[16] A Real-Time Video Tracking System Using Image Processing Techniques, A.L. Gilbert M.K. Giles, G.M. Flachs, R.B. Rogers, and Y. Hsun U, IJCPR-78 , pp. 1111-1115
[17] A Real-Time Video Tracking System A.L. Gilbert, M.K. Giles, G.M. Flachs, R.B. Rogers, and Y. Hsun U IEEE Trans. Pattern Analysis and Machine Intelligence PAMI-2 (1980) 47-56
[18] Automated TV Tracking of Moving Objects - The DFVLR-Tracker and Related Approaches G. Hirzinger, K. Landzettel, and W. Snyder, ICPR-80, pp. 1255-1261
[19] Motion Detection and Analysis of Matching Graphs of Intermediate-Levels Primitives C.J. Jacobus, R.T. Chien, J.M. Selander IEEE Trans. Pattern Analysis and Machine Intelligence, PAMI-2 (1980) 495-510
[20] On the Analysis of Accumulative Difference Pictures from Image Sequences of Real World Scenes, R. Jain and H.-H. Nagel, IEEE Trans. Pattern Analysis and Machine Intelligence, PAMI-1 (1979) 206-214
[21] Motion Analysis in Natural Scenes Picked up by a Moving Optical Sensor, A. Korn and R. Kories, ICPR-80, pp. 1251-1254
[22] Automatische Dreidimensionale Beschreibung bewegter Gegenstaende, R. Kraasch, B. Radig, W. Zach, in: [47], pp. 208-215
[23] Konzept und Realisierung eines mit Kontra8tauswertung arbeitenden TV-Trackers K. Landzettel and G. Hirzinger, in: [47], p. 222
[24] Constraint-Based Inference from Image Motion, D.T. Lawton, Proc. AAAI-80, pp. 31-34
[25] On Monocular Perception of 3-D Moving Objects, A. Zvi Meiri, IEEE Trans. Pattern Analysis and Machine Intelligence PAMI-2 (1980) 582-583
[26] Towards Automatic Visual Obstacle Avoidance,. H.P. Moravec, IJCAI-77, p. 584
[27] Visual Mapping by a Robot Rover H.P. Moravec, IJCAI-79, pp. 598-600
[28] Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover H.P. Moravec, Ph.D. Thesis, Department of Computer Science, Stanford University available as CMU-RI-TR-3 (September 1980) Robotics Institute, Carnegie-Mellon University, Pittsburgh/PA
[29] Analysing Sequences of TV-Frames: System Design Considerations, H.-H. Nagel IJCAI-77, p. 626, IfI-HH-B-33/77 (March 1977), FB Informatik, Univ. Hamburg
[30] Image-Sequence Analysis: What Can we Learn from Applications ?, H.-H. Nagel FBI-HH-M-79/80 (September 1980) FB Informatik, Univ. Hamburg/Germany to appear in: T.S. Huang (ed.) Image Sequence Analysis, Springer Verlag Berlin-Heidelberg-New York 1981
[31] From Digital Picture Processing to Image Analysis, H.-H. Nagel, Proc. International Conference on Image Analysis and Processing Pavia/Italy, October 22-24, 1980 pp. 27-40
[32] On the Derivation of 3D Rigid Point Configurations from Image Sequences, H.-H. Nagel IEEE PRIP-81 (to appear)
[33] On 3D Reconstruction from Two Perspective Views, H.-H. Nagel and B. Neumann, IJCA1-81
[34] Motion Analysis of Image Sequences for Object Grouping and Reconstruction B. Neumann, ICPR-80, pp. 1261-1265
[35] Model-Based Image Analysis of Human Motion Using Constraint Propagation J. O'Rourke and N.I. Badler, IEEE Trans. Pattern Analysis and Machine Intelligence PAMI-2 (1980) 522-536
[36] Motion and Structure from Optical Flow K. Prazdny, IJCAI-79, pp. 702-704
[37] Egomotion and Relative Depth Map from Optical Flow, K. Prazdny, Biological Cybernetics 36 (1980) 87-102
[38] Description of Moving Objects Based on Parameterized Region Extracting B. Radig, IJCPR-78, pp. 723-725
[39] Matching Symbolic Descriptions for 3-D Reconstruction of Simple Moving Objects, B. Radig, R. Kraasch, and W. Zach, ICPR-80, pp. 1081-1084
[40] Towards a System for the Interpretation of Moving Light Displays, R.F. Rashid IEEE Trans. Pattern Analysis and Machine Intelligence, PAMI-2 (1980) 574-581
[41] Determining the Movement of Objects from a Sequence of Images, J.W. Roach and J.K. Aggarwal, IEEE Trans. Pattern Analysis and Machine Intelligence PAMI-2 (1980) 554-562
[42] The Interpretation of Visual Motion S. Ullman, The MIT Press, Cambridge/Mass., 1979
[43] Analysis of Three-Dimensional Movement Using Fourier Descriptors, T.P. Wallace and O.R. Mitchell, IEEE Trans. Pattern Analysis and Machine Intelligence PAMI-2 (1980) 583-588
[44] Observing Jointed Objects, J.A. Webb and J. K. Aggarwal, ICPR-80, pp. 1246-1250
[45] Depth from Camera Motion in a Real World Scene, T.D. Williams, IEEE Trans. Pattern Analysis and Machine Intelligence PAMI-2 (1980) 511-516
[46] Automatic Motion Analysis System of Moving Objects from the Records of Natural Processes, M. Yachida, M. Asada, and S. Tsuji, IJCPR-78, pp. 726-730
[47] Angewandte Szenenanalyse, J.P. Foith (ed.) Informatik Fachberichte 20, Springer Verlag, Berlin-Heidelberg-New York 1979

a)

b)

Fig. 1: First (a) and last (b) of a series of 22 TV-frames showing a moving car.

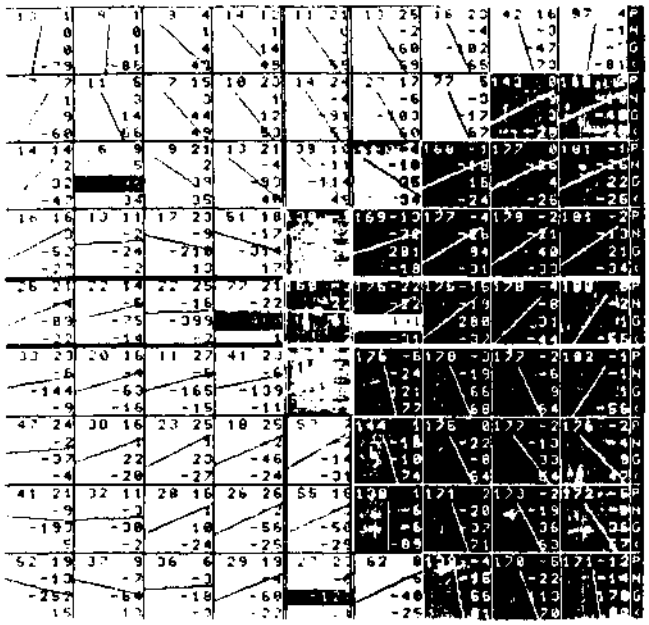

Fig. 2: Enlarged section (9*9 pixels) of the upper left corner of the rear car window (explanation of data see section 2).

Fig. 3: Candidates for prominent points (see section 3). Some of the prominent points at the rear part of the car are caused by the branches of a tree without leaves (recorded in early spring).

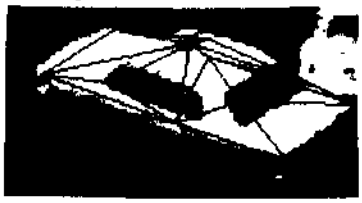

Fig. 5: Visible Edges of the convex hull projected back into the car image from fig. 1a.

Fig. 4: Chains of matched descriptors connecting the descriptor locations of every second frame of the series from fig. 1.

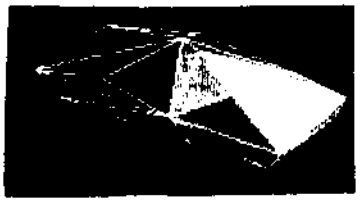

Fig. 6: The visible surfaces of the convex hull (corresponding to fig. 5) shaded according to Lambert's reflectivity law.