SURFACE RECONSTRUCTION FROM OUTDOOR IMAGE SEQUENCES

Debra Charnley and Rod Blissett

Plessey Research Roke Manor, United Kingdom

This paper describes the results of a study aimed at inferring the surface structure of outdoor scenes from video data acquired by a vehicle-mounted TV camera. The method is based on a decomposition of each frame of the video sequence into a dense set of feature points. Through the application of a Structure-from-Motion algorithm, the 3D locations of the time-consistent features are estimated and sequentially updated as each new frame is processed. The resultant point-based 3D representation has been found to be reliable, accurate and rich enough to enable the surface structure of the viewed scene to be reconstructed. We present here examples of the surfaces obtained by this 'bottom-up' approach, extracted from a 12 second video sequence. The implications of these results for autonomous vehicle navigation are discussed.

INTRODUCTION

The ability of a robot to sense surface structure is fundamental if it is to interact with its environment. In addressing this problem, vision has formed a major theme of robotics research. The motivation behind our work stems from the requirement of an autonomous vehicle to navigate within an unstructured outdoor environment. To date, success has been achieved in vehicle autonomy via road following[1], provided that the road is free from traffic or other obstacles and has clearly defined boundaries. This capability has been impressively demonstrated by the application of two-dimensional (2D) vision techniques.

With regard to tasks of more complexity, such as obstacle avoidance and navigation along poorly defined roads and tracks, the need to deduce the 3D structure of the environment becomes apparent. For these tasks, it is important to identify regions ahead of the vehicle that are safe to drive through in that they do not contain upstanding objects. Hence, recovery of height information is a basic requirement. It is difficult, if not impossible, to discriminate between an obstacle on the road and shadows or surface markings by relying on 2D processing alone.

A vision system has been developed at Roke Manor based on passive monocular sensing and the Structure-from-Motion principle[2]. The system was originally conceived to enable a robot manipulator to pick up a known object of arbitrary pose located on a workbench. We have now evaluated the vision system using outdoor video data taken with a TV camera which was mounted on the roof of a moving vehicle. It is clear that we do not drive using Structure-from-Motion alone and we are not proposing that an autonomous vehicle should do so either. Nevertheless it is informative to ask "What is the quality of the 3D information that can be obtained?" and "How can it best be used for vehicle guidance?".

Our main evaluation criteria were:

(i) the ability of the system to handle outdoor scenes with a significant natural vegetation content,

(ii) whether sufficient numbers of features could be detected and reliably matched to provide a dense enough 3D description (ie to enable surfaces to be reconstructed),

(iii) how accurately this description could be obtained,

(iv) whether the system could reliably deduce the camera ego-motion without additional sensor information,

(v) whether this information could be exploited for vehicle guidance.

This paper continues with a description of our approach to reconstructing surfaces from a set of 3D feature locations and highlights some of the specific problems faced when dealing with outdoor imagery.

STRUCTURE-FROM-MOTION USING OUTDOOR IMAGERY

The vision system, that has been developed at Roke Manor (and known colloquially as DROID), was designed to be a versatile algorithm for processing general extended visual image sequences and as such, has needed no fundamental adaptation from that used to derive the 3D structure of objects for a robot manipulator at a workbench[2]. The essential difference in extracting structure from natural outdoor imagery lies in the density of information to be tracked throughout the sequence, affecting primarily the control parameters relating to corner detection and matching. Other significant modifications have arisen solely in the ultimate use of the 3D scene description provided by DROID; the surface reconstruction techniques presented in this paper are more appropriate to the navigational requirements of an autonomous vehicle than the object recognition and pickup tasks considered in the original robotics application[3].

In addition to the input of grey-level images and estimates of the camera or vehicle motion during image capture, the operation of the DROID vision system is determined by a series of user-defined parameters relating to the corner detection, matching and camera ego-motion algorithms. Crucial to the formation of a detailed 3D scene description is the number of two-dimensional image tokens, or corners, introduced into the system. Image sequences of both semi-structured and unstructured scenes have been processed using a corner detector recently developed at Roke Manor[4]. The trade off between extracting sufficient detail in the 3D scene, but not confusing the feature matching process with too many potential candidates, has led to the selection of a corner threshold giving, typically,

AVC 1988 doi:10.5244/C.2.24

250 points in semi-structured images and perhaps twice that number in those of an unstructured scene containing a significant proportion of natural vegetation. This threshold has generally produced an even distribution of 2D corners over the image (important for stability of the ego-motion calculation) except in blander regions of the image such as road surfaces. As the Structure-from-Motion algorithm depends critically upon tracking these corners from image to image, the temporal consistency of the corner operator is particularly important. Processing of the two extended image sequences described in this paper has confirmed that the Plessey operator is consistent in detecting both the well-defined corners on buildings and manmade road-markings as well as the less obvious textural structure of bushes and hedges. It should be stressed that dentical sets of control parameters, defining the corner detection, matching and ego-motion algorithms, were used to process these two outdoor image sequences.

Feature Matching

The large number of corners detected in unconstrained outdoor imagery necessitates careful specification of the matching parameters. The search for potential candidate matches is focused, in the BOOT and RUN phases of the vision system[2], by employing search regions typically less than 10 pixels wide. Use of these comparatively small regions is encouraged by close frame spacing, as the observational error contribution to their size is correspondingly reduced. Under these circumstances, approximately 75% of the corners detected in any given frame are matched either to existing 3D features or 'limbo matched'[2] to any remaining unmatched corners detected in earlier frames.

Attributes of each detected corner, based on local grey-level image properties, are used to give an indication of match quality. Small frame-to-frame spacing of the image sequence, and the consequent similarity of adjacent images, allows the attribute mismatch threshold, above which potential matches between corners are rejected, to be kept low. A high proportion of the accepted matches are therefore good, most notably in areas of the image with a high density of corners, such as bushes and hedges. The ability of the vision system to consistently detect and track features even in natural vegetation, clearly demonstrates the robustness of both the corner detection and tracking algorithms.

Feature Retirement

The detection and matching of large numbers of corners from the initial image pair of each sequence generates several hundred 3D features which must be maintained, i.e. considered as match candidates and updated as necessary, throughout the remainder of the sequence. The subsequent incorporation of new features, via the limbo matching procedure, as scene structure approaches or new objects enter the field of view, soon leads to a rapid increase in the total number of features to be maintained. In order to control this growth, a technique known as feature retirement has been introduced into the DROID system. As the sequence progresses, many of the known features may come to reside behind the camera, or at least have passed out of the field of view of the camera some time ago. Such features are periodically extracted from the feature list and held in a separate 'retirement' list; they are not considered further by the vision system but remain available for end-use applications such as the 3D surface reconstruction described below. This retirement technique successfully ensures the tracked 3D feature list is kept to a maximum of around 500 features, considerably reducing the processing effort involved in matching and ego-motion calculation.

EVALUATION

In order to evaluate the potential of Structure-from-Motion techniques for vehicle navigation, we have been provided with video data captured using a Landrover and the experimental autonomous vehicle, shown in Figure 1, at the Royal Signals and Radar Establishment, Malvern. Two sequences were taken with the vehicles under manual control and travelling at approximately eight miles per hour. Alternate video frames from a twelve second period were processed, giving sequences of 150 256 x 256 8-bit grey-level images, generated during approximately 50 metres of vehicle travel.

Figure 1. RSRE experimental autonomous vehicle

The two sequences differed mainly in the proportion of natural vegetation present in the viewed scene. The first sequence, of a semi-structured outdoor environment containing regular man-made objects, a well-defined road and limited natural vegetation, generally produced fewer corners, and hence 3D features, than did the visually more complex second sequence taken whilst travelling down a narrow country lane bounded by hedges and banks. Typical frames from each sequence are shown in Figure 2, together with detected corners. The road in the semi-structured sequence is particularly bland and gave rise to few corners except from its distinctive markings; the building and foreground bushes on the other hand generate many corners, the majority of which are tracked consistently throughout the sequence. Natural structures also produced many trackable features in the second sequence, where the road is well-defined by corners from extensive small-scale detail.

Figure 2(a). Original image with detected corners: semi-structured environment.

Figure 2(b). Original image with detected corners: unstructured environment.

Figure 3 shows the number and distribution of features available for matching, together with typical search regions. Those that are matched are used to deduce a new estimate of the camera ego-motion, more consistent with the visual data[2]. All six rotational and translational degrees of freedom are permitted, enabling the small variations in vehicle motion due to road irregularities and camber to be reproduced. A selection of the ego-motion parameters for the first sequence, shown in Figure 4, illustrate the camera motion as the vehicle travels around the building seen in Figure 2(a). Figure 4(a) shows the variation of the frame-to-frame displacement of the camera (an indication of vehicle speed) during data capture. Significant changes in the direction of vehicle travel are manifest primarily as yaw (Figure 4(b)), while the small scale pitching motions of the camera (Figure 4(c)) are the result of an uneven road surface, amplified by a lightly damped vehicle suspension system.

Figure 3. Tracked 3D features with associated search regions.

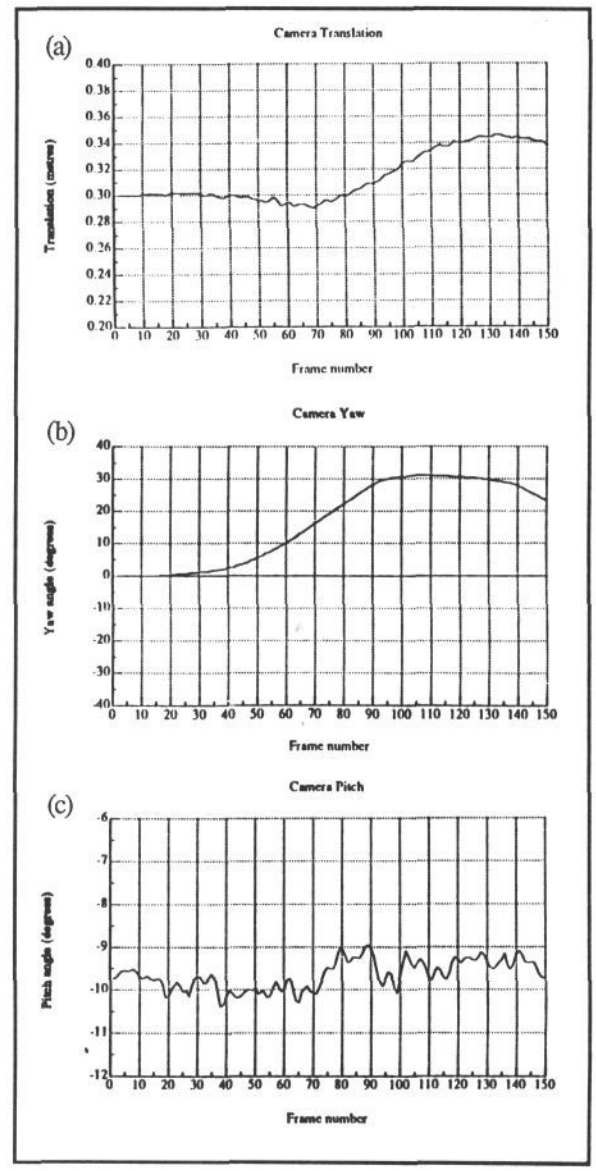

Figure 4. Derived motion parameters: semi-structured sequence.

Several thousand 3D features have been instantiated and tracked throughout each of the two sequences. Many of these features have been tracked over considerable proportions of the entire sequence, increasing the accuracy and confidence in their positional information with each observation[2]. Figure 5 shows the extent and quality of the features from the first sequence; the regular road markings are clearly evident to the right of the vehicle path for example.

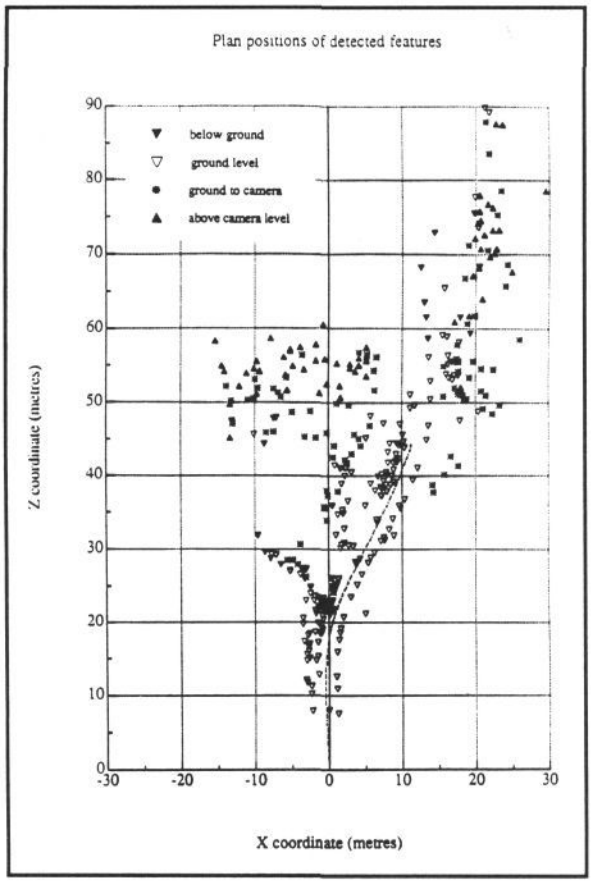

Figure 5. Plan view of all tracked 3D features, segmented by height, and vehicle path: semi-structured sequence

SURFACE RECONSTRUCTION

To further demonstrate the accuracy and detail of the scene description available from the DROID vision system, the feature information has been used to reconstruct a 3D surface of the perceived environment. This is performed on a frame by frame basis, interpolating between features with triangular planar facets. The triangulation is achieved, using an algorithm of Delaunay[5], between the projections of feature points onto the current image-plane. This algorithm seeks to generate a set of compact triangles on the image plane, subject to the criterion that a circumscribing circle constructed through the three vertices of a triangle must not encompass any other vertex. The resulting triangular regions are introduced into 3D, using the known positions of the associated vertices, to form the required 3D surface description of the scene. It must be noted, however, that not all of the available 3D features are included in the triangulation; those features only recently instantiated, or those which have not been matched and updated regularly in preceding frames, are considered unreliable and should not therefore be used to determine surface structure.

Figure 6 illustrates the derived surface for a typical frame of the semi-structured sequence. The contours represent the intersection of the modelled surface with vertical planes, 3 metres apart, parallel and perpendicular to the camera at the beginning of the sequence. The displayed contours are confined to a depth of 48 metres ahead of the vehicle simply to clarify the image; 3D information is, of course, available beyond this distance.

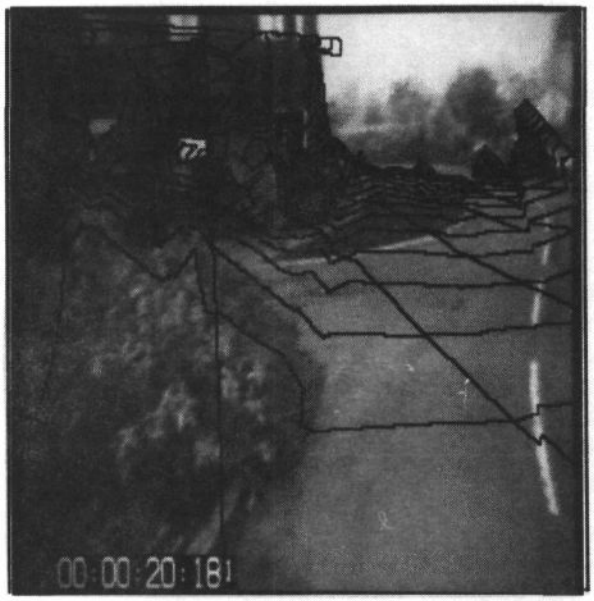

Figure 6. Surface contouring: semi-structured sequence.

The bush in the foreground is clearly distinguished by the contours rising around it, as is the building. Closely packed contours, such as those from the top of the shrubs to the building behind, or those surrounding the white sign near the centre of the image, generally indicate strong depth discontinuities and demonstrate unambiguously the ability of the DROID vision system to extract 3D structure.

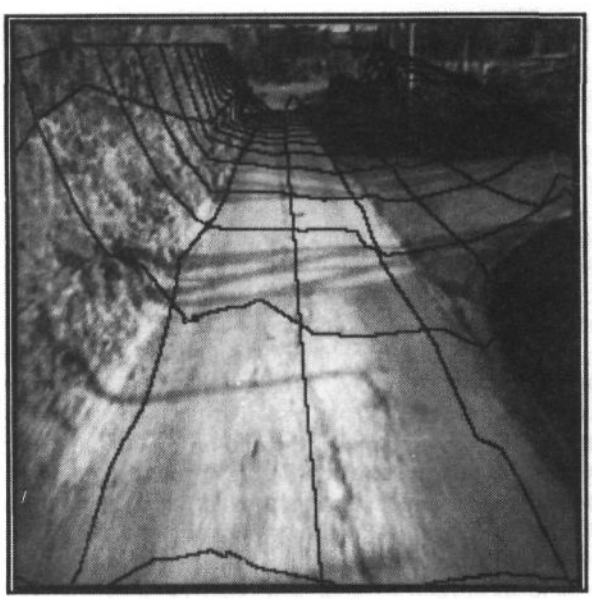

Figure 7. Surface contouring: unstructured sequence.

Figure 7 shows a typical contour image of the unstructured environment, but with the transverse contours at only 1 metre spacing.

The slopes of the hedge and bank have been well reproduced while the indentation of the bank by the side road to the right is seen to become more pronounced as the vehicle approaches the junction. Displaying extended sequences of the modelled surface contours alone, with the contour 'grid' fixed relative to the ground, illustrates the temporal consistency of the 3D structure provided by the vision system. Objects initially perceived in the far distance can be visually tracked, and gradually adopt a more stable structure as they are approached. Particularly noticeable throughout these sequences is the presence of a well-defined flat region in the surface contouring, such as that present in Figure 8.

Figure 8. Contour image with flat, potentially navigable region: semi-structured environment.

Algorithms to reliably distinguish these potentially navigable regions (an example of which is shown in Figure 9) are currently being developed as an aid to vehicle guidance and obstacle avoidance.

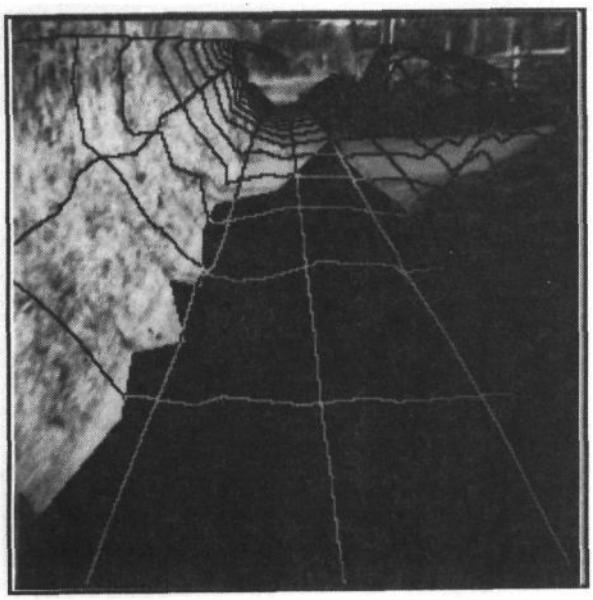

Figure 9. Contour image with navigable region: unstructured environment.

CONCLUSIONS AND DISCUSSION

The application of the DROID vision system to natural outdoor image sequences has provided us with some interesting results. We have demonstrated in this paper that plausible surfaces can be generated from a points-only Structure-from-Motion vision system. This has been made possible because the feature-point detector employed at the front end of our vision system has proved highly successful in generating dense and time consistent 3D information. We have been particularly encouraged with the success rate of feature matching from frame to frame, especially in highly textured regions. The large number of features that have been consistently matched from frame to frame has aided greatly in stabilising the camera ego-motion computations. This has provided us with a reasonably richly sampled 3D description of the scene under surveillance.

We have shown that it is possible to take the sampled description and reconstruct a 'bottom-up' surface representation of the scene. The generated surface shows good qualitative agreement with the assumed scene structure. We have reported elsewhere[6], that it is possible to identify upstanding objects and navigable regions ahead of the vehicle from this information. This is clearly necessary to combat the limitations of the current techniques for autonomous vehicular guidance1.

We are encouraged that Structure-from-Motion can provide reliable and useful data for vehicle guidance. Our efforts are now directed towards two objectives; a real-time implementation of our basic vision system and the extension of the current approach to include topological as well as metrical information.

The development of a real-time demonstrator will allow us much more effectively to assess performance against a wider set of environments and enable us to optimise such system factors as camera focal length, camera look-down angle and image pixel size. Work is already underway on the development of dedicated hardware for the front end image processing.

The inclusion of topological information and in particular, the instantiation of edges into 3D, will further enrich the information extracted via Structure-from-Motion. Related work is published elsewhere in these proceedings on a combined point and edge detector[4] and a Region-Edge-Vertex graph 3D scene representation[7]. These developments promise to significantly enhance the capabilities of our current vision system.

ACKNOWLEDGEMENTS

This work was carried out with the support of the Procurement Executive, Ministry of Defence. The authors gratefully acknowledge the assistance of J Sherlock and G Edwards of RSRE, Malvern for making available the experimental autonomous land vehicle and for setting up the data gathering trials. The grey-level images presented in this paper, kindly supplied by RSRE, are subject to the following copyright: Copyright © Controller HMSO London 1988.

REFERENCES

1. **Dickmanns, E.D. and Zapp, A.** "Guiding land vehicles along roadways by computer vision". Congres Automatique 1985 - 'Des outils pour demain', Toulouse, pp. 233-243 (October 1985).

2. **Harris, C.G. and Pike, J.M.** "3D Positional Integration from Image Sequences". AVC '87, Cambridge, pp. 233-236 (September 1987).

3. **Harris, C.G., Pike, J.M. and Stephens, M.J.** "The Plessey Computer Vision System: Demonstration of Object Pick-up". Plessey Technical Note 72/88/043N (February 1988).

4. **Harris, C.G. and Stephens, M.J.** "A Combined Corner and Edge Detector". AVC '88, Manchester (August 1988).

5. **Boissonnat, J-D.** "Representing 2D and 3D shapes with the Delaunay triangulation". Proceedings of the 7th Int. Conference on Pattern Recognition, Montreal, Canada, pp. 745-748 (1984).

6. **Blissett, R.J., Charnley, D. and Harris, C.G.** "Towards Robot Mobility through Passive Monocular Vision". International Symposium on Teleoperation and Control, University of Bristol (July 1988).

7. **Stephens, M.J. and Harris, C.G.** "3D Wire-Frame Integration From Image Sequences", AVC '88, Manchester (August 1988).