COMPUTER GRAPHICS AND IMAGE PROCESSING (1973) 2, 281–297

Scene Analysis in Support of a Mars Rover

DOUGLAS A. O'HANDLEY
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91103

Communicated by A. Rosenfeld

Received July 23, 1973

The National Aeronautics and Space Administration/Jet Propulsion Laboratory breadboard mobile robot is a test system for developing adaptive variably autonomous capabilities in support of future missions to the planets. A vision subsystem is a part of this breadboard development.

Scene analysis software has been developed and experiments conducted in order to segment a scene with irregular objects in it and to develop an algorithmic definition of the scene and relationships in that scene.

The results at present have used simulated real-time TV inputs. The scenes which have been used are both contrived and natural. The results presented in this paper will form the basis for implementation of the vision software on the mobile robot.

INTRODUCTION

The projected mission plan of the National Aeronautics and Space Administration (NASA) includes a mission to Mars in the 1980's in which a mobile robot vehicle is to be used [1]. The development of concepts and research directed toward an early demonstration of this robot rover in a breadboard configuration are underway at the Jet Propulsion Laboratory (JPL) of the California Institute of Technology under the Research program "Artificial Intelligence for Integrated Robot Systems" sponsored by the Office of Aeronautics and Space Technology of NASA. This robot is to have variable autonomy in the goal-directed coordination of the subsystems.

Three scenarios of progressive complexity have been written. Initially, the robot will be tethered and will operate in a laboratory test area which simulates an outdoor environment characterized by high visual contrast and rocks located in random positions. The second scenario is to take place in a prepared outdoor environment. In the third scenario, all constraints on the operating environment are removed (Fig. 1).

The robot will operate in environments characterized by two important categories of natural objects: (a) scientific samples and (b) obstacles. The former will consist of small objects having maximum cross-sectional dimensions of approximately 10 cm. The obstacles may be such terrain features as craters, deep canyons, large rocks, steep hills, or unknown or unapproachable objects.

The rover will be equipped with tactile bumpers to protect against unexpected contacts. It will have one arm (or possibly two) for manipulating objects. The operation of the arm will incorporate tactile sensing for purposes of closed-loop operation. It will also have a laser range finder and dual TV for

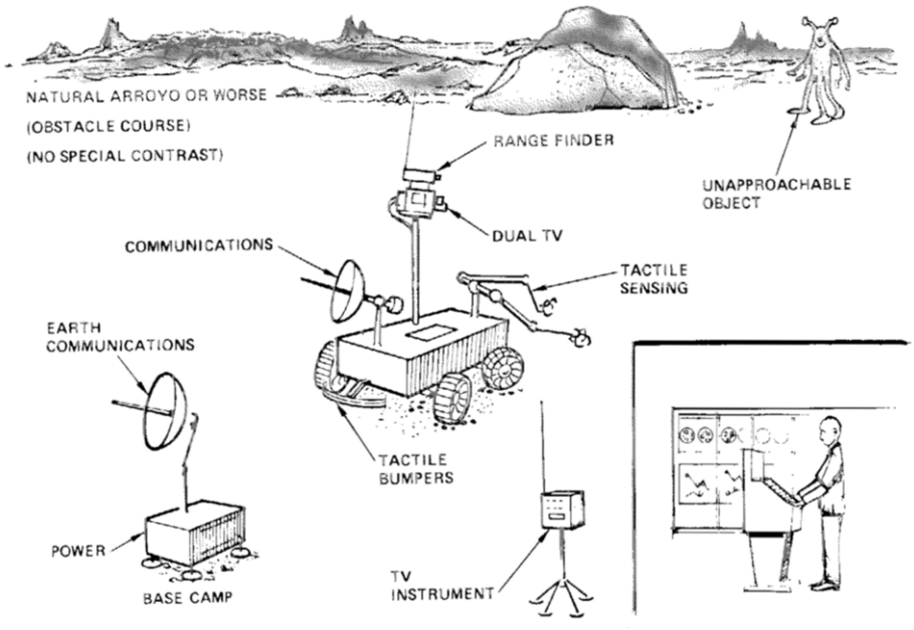

FIG. 1. Ultimate scenario. The Mars rover has the capabilities of perception, manipulation, and mobility. The environment is characterized by a natural scene which includes craters, mountains, hills, valleys, and small and large rocks or protrusions.

depth perception and will be provided with a radio link for communication with the base camp transmitter. Also indicated in Fig. 1 is the man/machine interaction which sets goals, assays communications from the rover, and provides simulation on the ground for the expected behavior of the rover.

Figure 2 shows the components of the robot software system. Central to "variable autonomy" is an adaptive computer program (Robot EXecutive) (REX) capable of planning and executing goals based upon stored knowledge of the robot's world. The robot's world includes not only the environment but

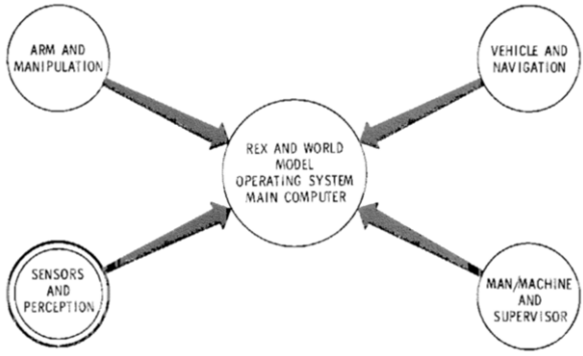

FIG. 2. Software and hardware relationships. The integration of software and hardware is the paramount objective. The Robot Executive and world model are central to this.

a model of the physical robot, with the dynamic components continuously sensed. Peripheral to REX are the navigation, manipulation, man/machine interaction, sensor, and perception subsystems. A more comprehensive documentation of the JPL arm subsystem is found in [2]. Many of the other subsystems are now in the breadboard or conceptual phase and are therefore not yet to be found in the literature. The emphasis of this paper is the research in "Sensors and Perception".

PERCEPTION

Research is now under way in computer perception (or scene analysis). The objective of this research is to perceive a three-dimensional scene, using both stereoscopic television and a laser range finder, and then describe that scene symbolically.

Initially, the test environment is to be an indoor laboratory area which is characterized by high visual contrast; the walls and floor will be of a uniform light color, and dark-colored rocks will be scattered over the floor in random positions. For navigation, the descriptors provided to REX must allow avoidance of large rocks; for manipulation, the descriptors must provide information that will allow the objects of interest (i.e., small rocks) to be picked up.

The term "scene analysis" refers here to the description by algorithmic techniques of a three-dimensional scene by using two-dimensional images and data from a laser range finder. Typically, a relational graph like our generic model (Fig. 3) is the basis for describing the scene. The nodes represent objects in the scene and arcs represent relationships between the nodes [3]. This graph is used by REX as the major component of the world model of its environment.

Our research on scene analysis has been divided into three operations: (i) segmentation of objects in the scene; (ii) depth perception, which establishes the three-dimensional parameters of the object and also contributes in segmentation; and (iii) feature analysis, which assembles the information from segmentation and depth perception into descriptors which are then passed to the world model.

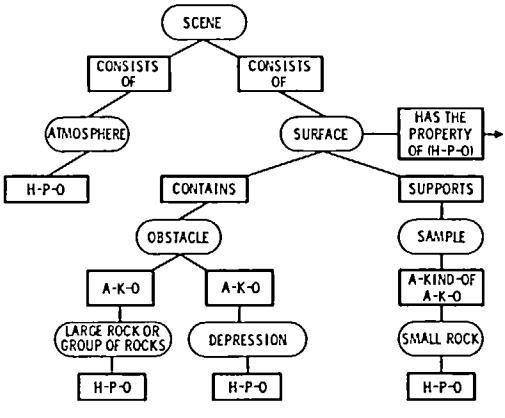

FIG. 3. Generic model. A preliminary data structure for describing a scene.

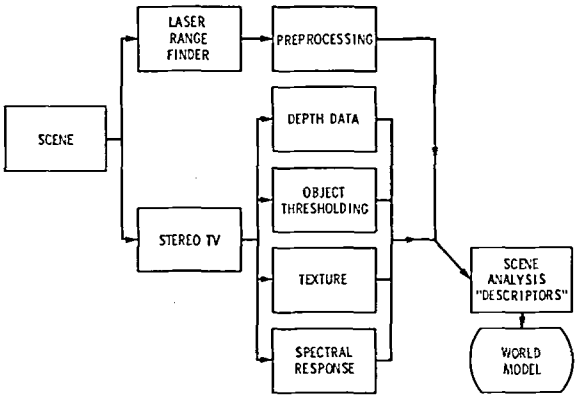

FIG. 4. Perception. A scene is imaged by both a laser range finder and stereo TV. The process of segmentation, identification of relationships, and ultimate formulation of a descriptor for transmittal to the world model result from the data collection shown here.

In Fig. 4 is shown a functional diagram of perception as applied to the laser range finder and stereo TV. The laser range finder requires a computer for control of its operation and to preprocess the data. The result is positional information for use in the world model descriptors. The stereo TV also yields depth data and, in addition, object thresholding, texture analysis, and spectral response. The stereo TV is also under computer control. These components are integrated to provide a complete descriptor.

In the example shown in Fig. 5, the object (a rock) is distinguished and a

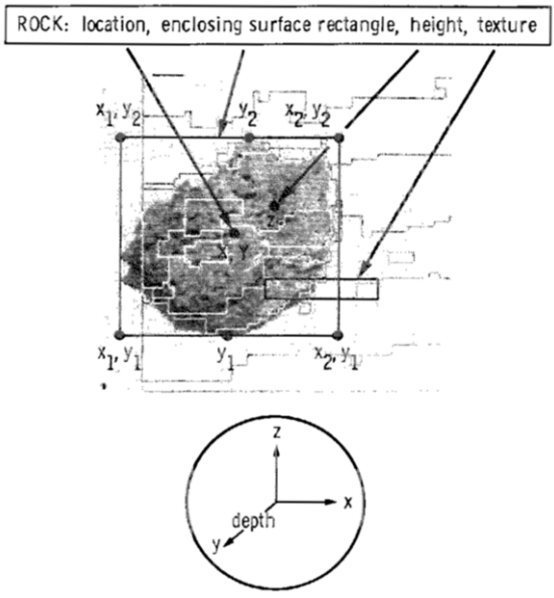

FIG. 5. Scene analysis—world model descriptors. The descriptor identifies the object, the location of that object, the size (by means of an enclosing rectangle), and texture.

location noted at the center of gravity; an enclosing rectangle is identified in which the X coordinates of the object designate the visible extremities of the object in width. The Y coordinates are the tangency locations of the closest (and lowest) edge and farthest (and, by definition, highest) perceivable point. The height of the object is, then, the height of this farthest perceivable point. The texture represents a measure of the variations in gray level on the surface.

SEGMENTATION

Our initial step in processing images is to separate the objects from the background (i.e., rocks from the ground) (see Fig. 6). The segmentation process uses only the right image of the stereo pair. No *a priori* assumptions are made with respect to size, shape, location, or numbers. In the first scenario, all objects will be of high contrast with reference to the background. The scene of Fig. 6 is an archetype of that environment.

Figure 6 shows our earliest attempts at segmentation [4]. This digitized playback is a grid of M by N (where $M = N = 500$) values of individual gray levels or pixels. Subsequently, this set of 250,000 values was subdivided into an 8×8 matrix or subset called a sector. A histogram of gray levels was determined for each of these 64 sectors. These histograms of each sector could consist of a single mode or be bimodel, depending on the gray level of the objects within the sector. The basic assumption was that the gray level of the overall background was less than the gray level of the object to be segmented. As a consequence, the algorithm segmented objects from the background on the basis of thresholding or position with relationship to the dark end of the histogram.

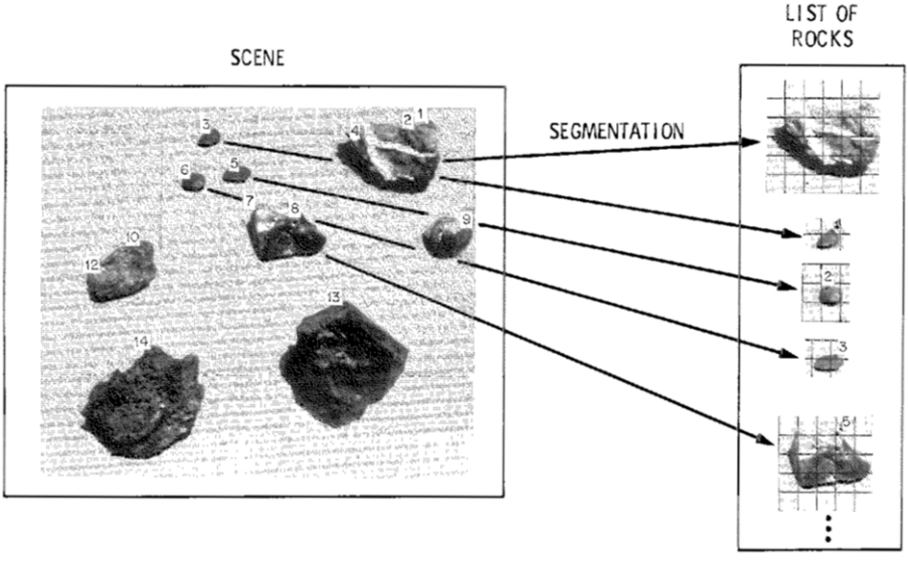

FIG. 6. Object segmentation. This scene has been segmented using thresholding techniques, and a list of rocks has been made. Object decomposition is apparent.

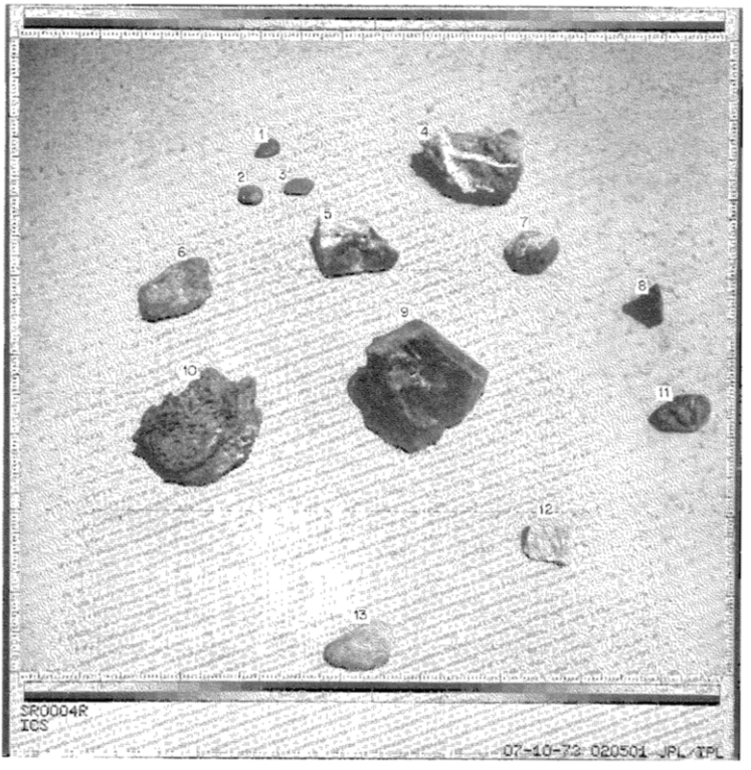

FIG. 7. Adaptive object segmentation. This scene has been segmented using multinoded histograms with logical searches and decisions.

In this figure, the rock in the upper right-hand corner of the scene was segmented into components 1, 2, and 4. This happened because (a) the light striations were lighter than the assumed background defined by the histogram and (b) the resolution of the matrix was too coarse.

The problems associated with segmentation of this scene were overcome through (a) increasing the resolution to 256 sectors or 16 × 16 matrix and (b) providing bimodal analysis with additional constraints. The results of correct segmentation of the scene in Fig. 6 using bimodal analysis is seen in Fig. 7. The bimodal analysis consisted of smoothing the histogram of the matrix element with a nine-point filtering and 16 × 16 sectors were the result of experimentation. This segmentation technique was found to be sentitive to the smoothing applied to the histogram. In addition, the 16 × 16 sectors were allowed to have a single mode or multimodes. Additional logic allows the peak associated with the background to be identified or a phantom peak assumed. The background peak is found through constraining the one-dimensional search to a range predicted by the location of the background peak in previously determined adjoining sectors.

An indoor laboratory scene, again typical of the environment for the initial scenario, is shown in Fig. 8a. In this scene, there is nonuniform lighting, rocks

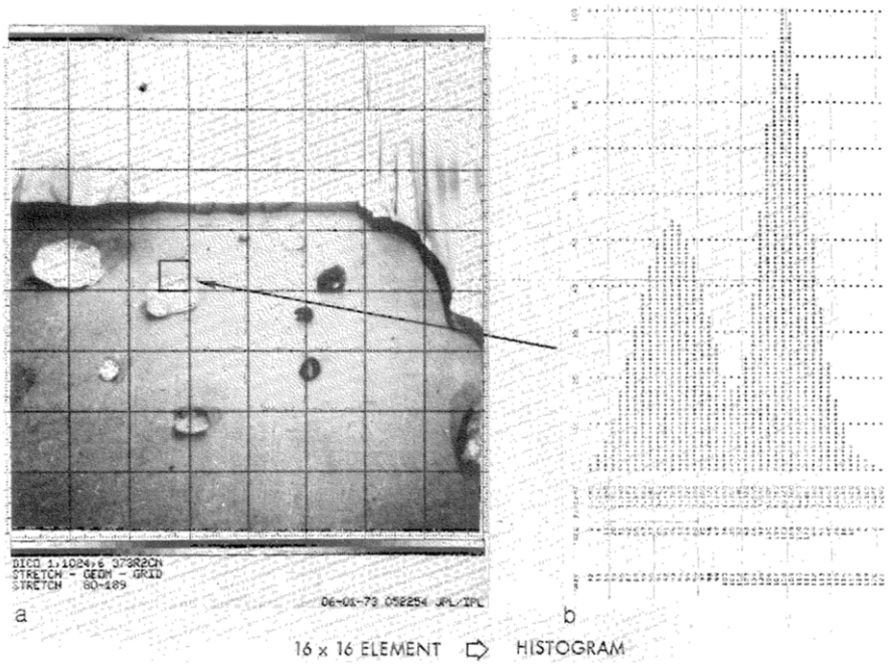

FIG. 8. Indoor laboratory. (a) Element of 16 by 16 matrix showing a rock and background. (b) The bimodal histogram associated with the rock and background.

of varying contrast with the background, and drapes instead of a uniformly painted light wall. The nonuniform lighting made multiple shadows on the floor. This characteristic frustrated much of our analysis which had been successful in segmenting the previous scene where the sun provided the source of light. The rocks, in addition to their multiple shadows, were both lighter and darker than the surrounding background. As a consequence, additional parameters were added to our segmentation algorithms which separate objects from the background.

In Fig. 8b is seen the nine-point smoothed histogram for the accentuated portion of Fig. 8a. This bimodal histogram contains a background which is darker than the rock. Here the background would be chosen as the right node because of continuity constraints applied from adjoining elements. In Fig. 9a is shown the same indoor scene with a different area accentuated. In Fig. 9b, we have a nine-point smoothed trimodal histogram. The floor is now the center node and is identified through continuity considerations, the border of the drapes is the node to the right, and the left node represents the white portion of the drape. The final result of segmentation is seen in Fig. 10 where all objects have been segmented correctly. A more thorough discussion of this segmentation process is found in [5].

DEPTH DATA

Depth data or range is pertinent to the problem of object isolation and also provides relationships between objects. This research has assumed that input

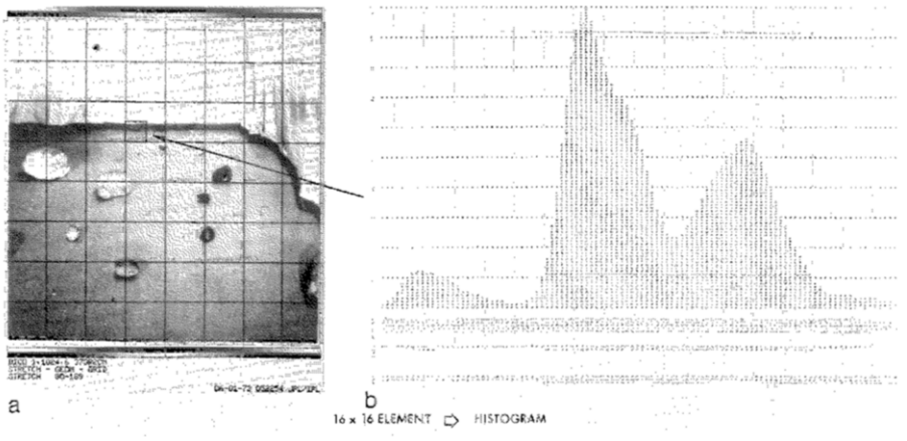

FIG. 9. Indoor laboratory. (a) Element of 16 by 16 matrix showing three objects: (1) floor, (2) curtain, and (3) curtain border. (b) Trimodal histogram resulting from the element. The curtain is to the right, floor is center, and curtain edge is to the left.

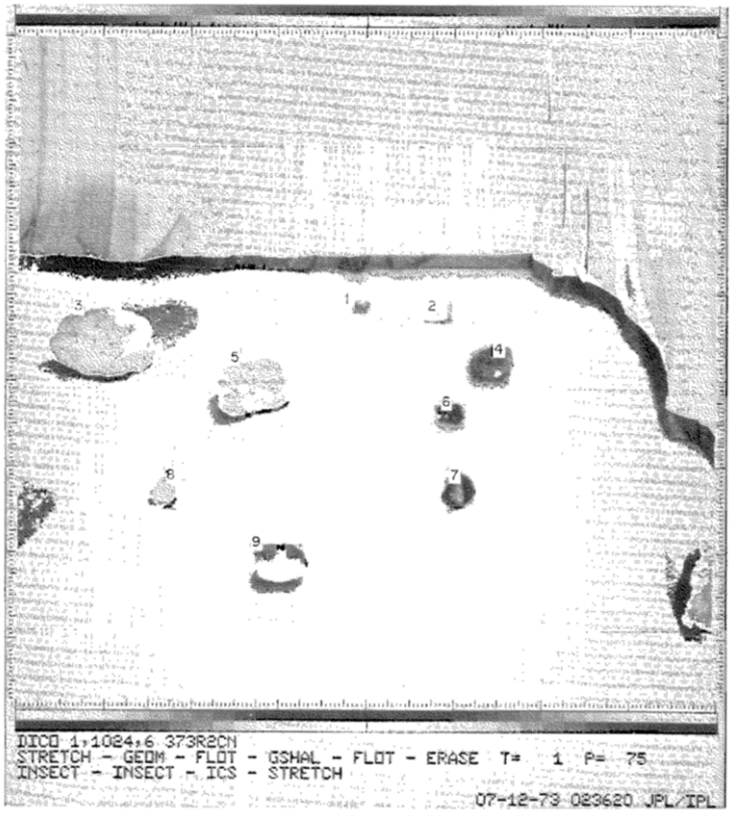

FIG. 10. Indoor laboratory, segmented. Successful segmentation is shown with identification of the 10 rocks on the floor.

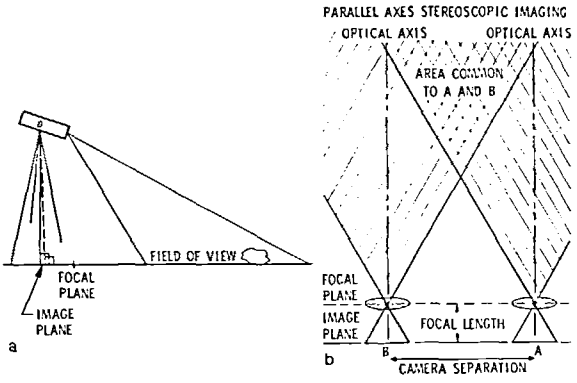

FIG. 11. Depth data, geometrical properties. (a) The camera mounted approximately 1 m above the ground plane, with the perspective indicated. (b) Parallel axes stereoscopic imaging system, with range defined as the horizontal distance from a point in the scene to the image plane.

is to be comprised of two digitized images obtained by two identical and aligned optical systems with parallel optical axes (Fig. 11). Range is defined as the horizontal distance from a given point to the image plane.

The right-hand digitized image is taken as the reference image for the stereo pair; the objective then is to determine depth for each point in the image matrix. This requires the algorithmic matching of identical points in the left and right images, the so-called correspondence problem [6]. The solution to the problem of determining range using stereo TV had to be efficient with respect to computer time and also context independent. To this end, an adaptive correlation window is incorporated as an aid in the solution of the correspondence problem. The correspondence problem is visualized in Fig. 12.

The digitized stereo images of the sidewalk scene, taken with a Stereo Realist camera and played back through the Video Film Converter, are seen in Fig. 13. The separation of the optical axes is 7 cm and provides adequate reso-

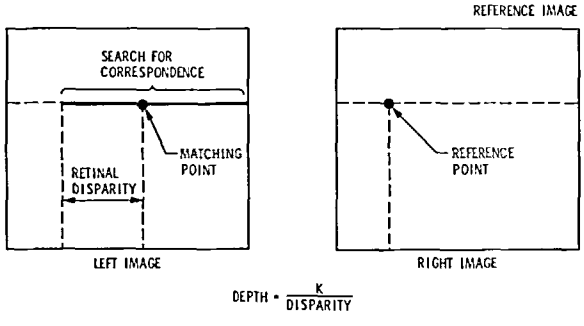

FIG. 12. The correspondence problem. The right and left images are visualized with a particular reference point indicated in the right image and matching point and resultant information shown in the left image. Depth is defined as K (a constant related to number of samples in each image and separation of the optical axes) divided by the retinal disparity (in terms of pixels of digital elements).

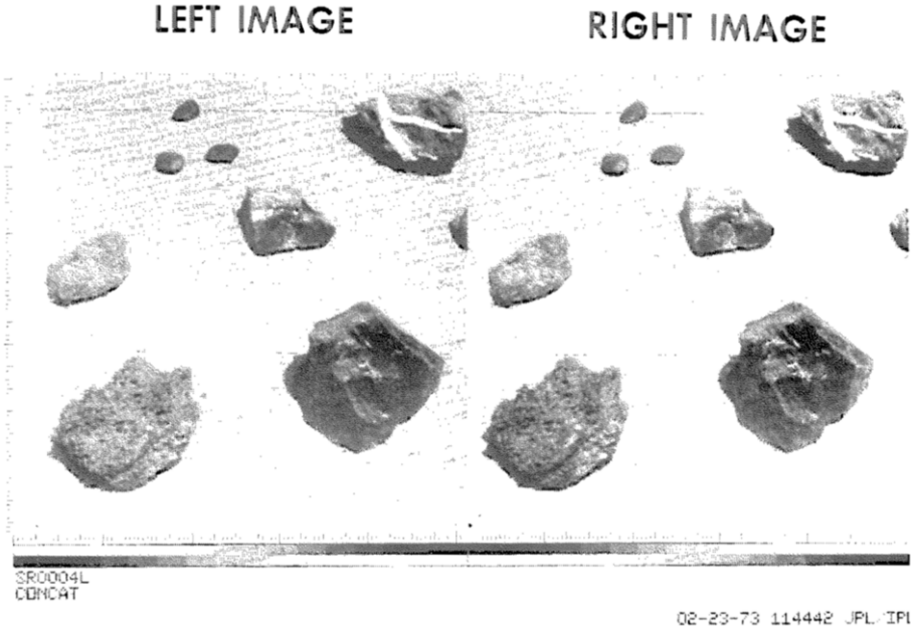

FIG. 13. Sidewalk scene. Stereo pair of images taken with Stereo Realist camera and played back through the Video Film Converter.

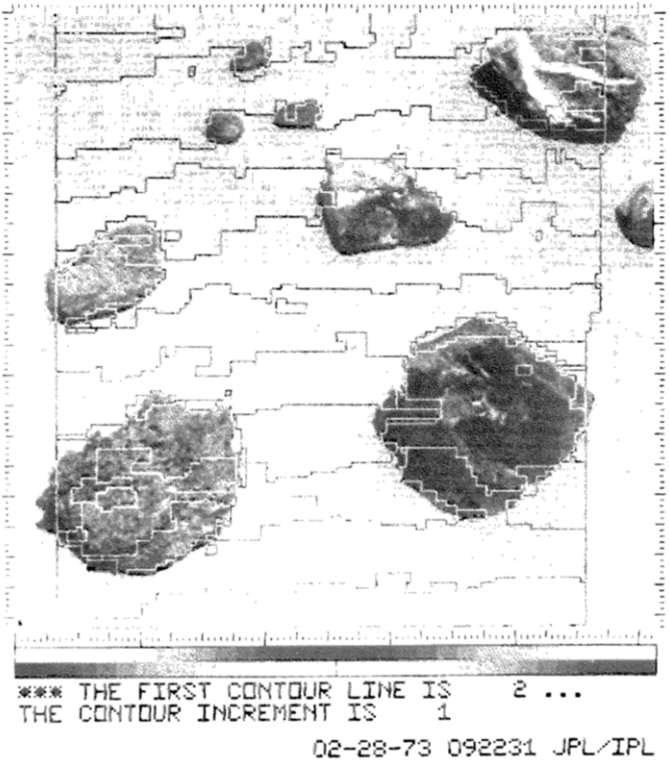

FIG. 14. Depth map of sidewalk scene. Contours of equal distances are superimposed on the reference image.

lution at a 1-m range (the manipulation distance). In Fig. 14, the contours of range are displayed, superimposed on the right reference picture. The relatively flat areas of the sidewalk surface appear as lines across the images. The nonuniformity of the lines across the sidewalk is a result of the mode of execution of the algorithm which provides the adaptive correlation window. The results are considered satisfactory. In this computation, 2000 reference points were used and required 11 min of IBM 360/44 computer time for the determination of this depth map.

To demonstrate the accuracy of this computation, Fig. 15 (the left-hand picture) shows the left minus the right image of the stereo pair. The white outline to the left of each rock is indicative of the retinal disparity from the foreground to the background. The image in the right-hand picture was developed by constructing a pseudo-image of the actual left image, wherein the right image has been moved by the amount of computed retinal disparity so that it would appear to be an actual left image. The left minus pseudo-left configuration shows a lack of detail and indicates the precision with which these algorithms perform.

In Fig. 16 is seen our initial attempt at using a Martian landscape surrogate. It is a stereo pair of photographs taken in the Arroyo Seco, a canyon situated near the Jet Propulsion Laboratory. These stereo images were taken with a single Rolleiflex camera which, between pictures, was moved a prescribed distance to achieve the stereo effect. This test data has been obtained through an approximately aligned system with parallel axes. The pictures have also been output from the video film converter. The overall image matrix of the pictures is a 500 × 500 element array.

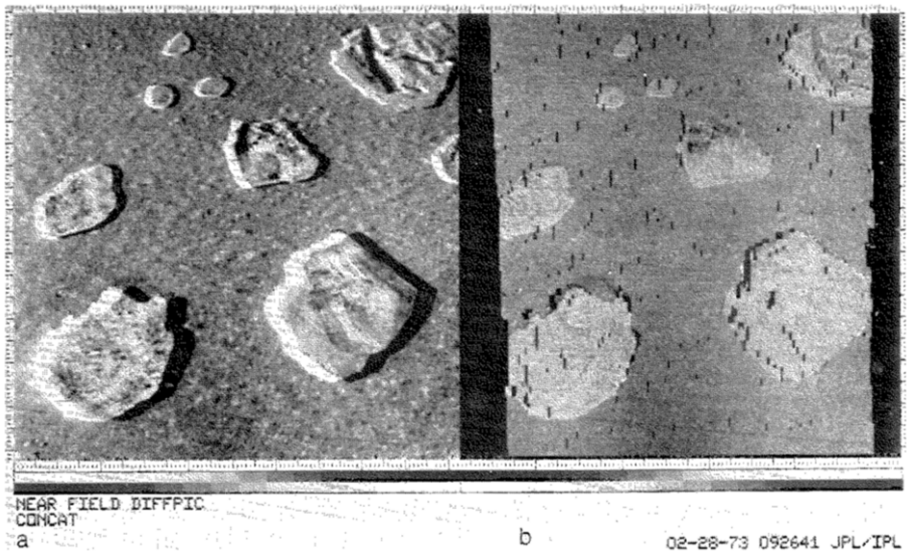

FIG. 15. Left minus right/left minus pseudo-left. (a) The left and right digitized images have been subtracted and retinal disparity appears as the white outline to the left of the rocks. (b) The right image has been rectified, using the depth information, to produce a pseudo-left image. The left minus pseudo-left now reflects the accuracy of the depth map.

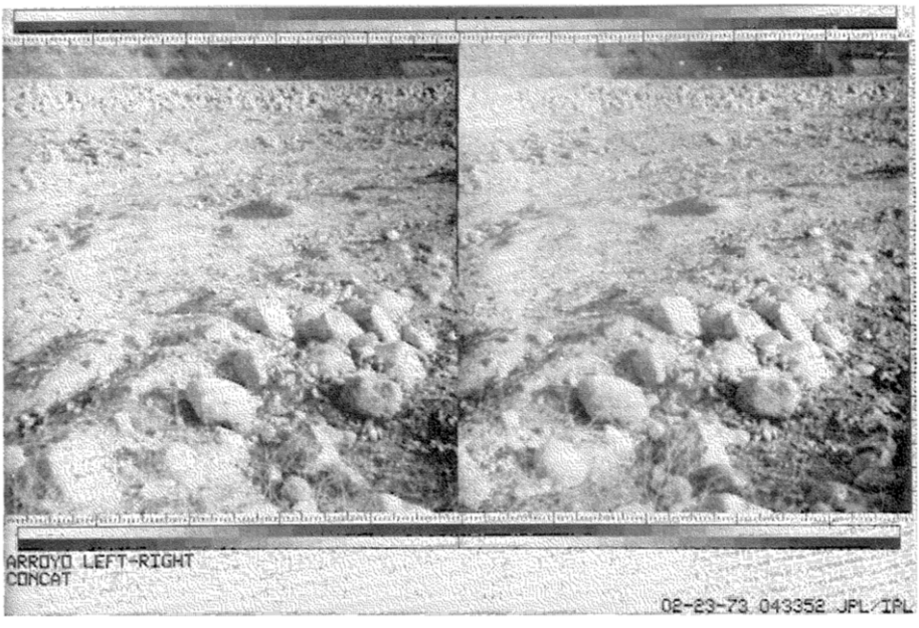

FIG. 16. Arroyo Seco, Martian surrogate. Stereo pair of images taken with a single Rolleiflex camera moved a prescribed distance between shots to achieve stereo.

In Fig. 17 are shown the contours of range superimposed on the reference image. This range computation required about 15 min of IBM 360/44 computer time. The contours cluster together in the middle of the scene where there is a steep dropoff and also in the top third of the picture. The faces of the large rocks near the observation point approximate vertical planes and as such exhibit a constant range value, as shown in the contour of the surface. Because of the manner in which the computation takes place, the contour lines will have significant resolution in the near scene whereas the lines will tend to be hyperbolically distributed with increasing distance toward the back of the scene. To demonstrate the performance of these algorithms, we have also provided a test for this scene. Each point in the right image is displaced to the left by the retinal disparity and a pseudo-left image has been created. Again, we now difference the left minus the pseudo-left image and the departures from accuracy will be seen in terms of detail. As is shown in Fig. 18, the departures are minimal and probably result from improper focus and nonuniform shading. A more thorough discussion of the computation of depth with stereo images can be found in [7].

CONCLUSIONS

As an example of three typical scenes to be analyzed in the near future, we show the following figures. Figure 19 is a rock scene devoid of vegatation taken in the Mojave Desert. This particular scene was taken with an optically aligned system with an 8/10 of a meter separation between the optical axes. There are a number of difficulties. Initially, the separation is such that the ob-

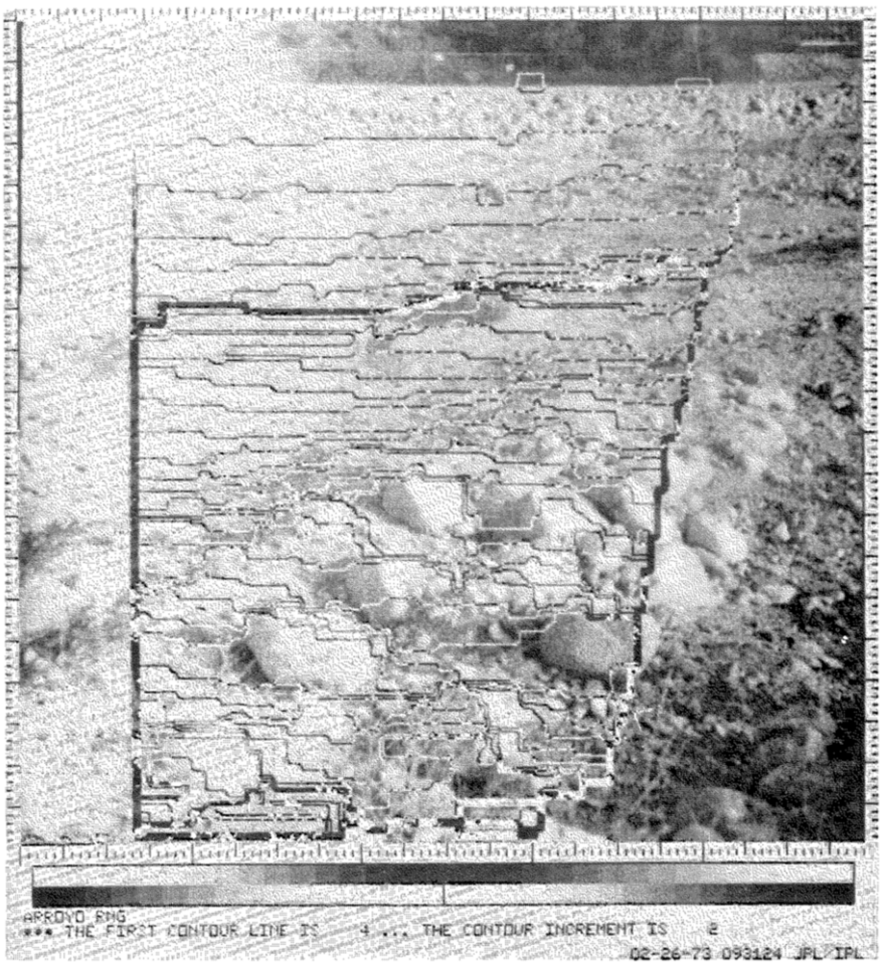

FIG. 17. Depth map of Arroyo Seco. Contours of equal distance are superimposed on the reference image. Note clustering of contours in middle of scene where there is a steep dropoff and also in the top third of the picture.

jects in the left-hand image are significantly rotated in perspective from those in the corresponding reference image. An object appearing in the middle of the rock scene on the left-hand side of the right image disappears in the corresponding collection of rocks in the left-hand image. Such analyses will require iterative techniques of assumptions of possible descriptors followed by verification through subsequent observations made from different perspectives. Examination of this scene will also require spectral response and a greater dependency on texture analyses. It may be true that the 8/10 of a meter separation is optimal for navigation, but it does not appear appropriate for manipulation.

In 1976, the United States will place a Viking lander on the surface of Mars. This lander will have a stereo imaging system with an 8/10 of a meter separa-

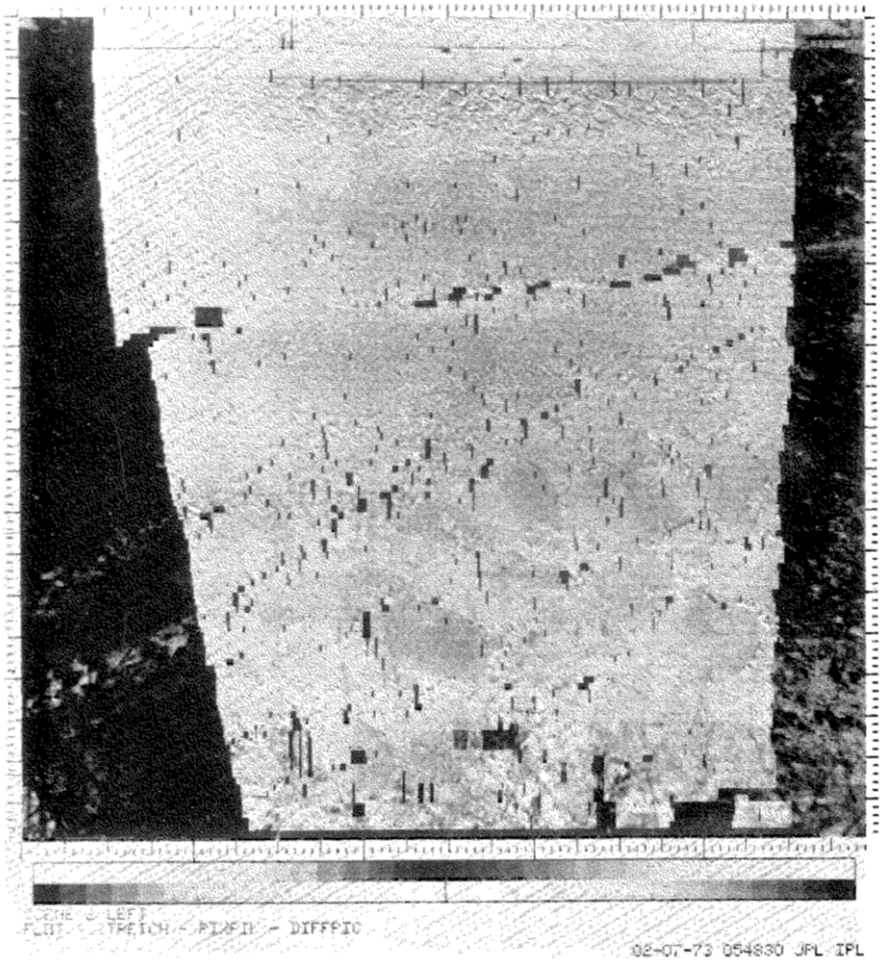

FIG. 18. Arroyo Seco, left minus pseudo-left. Resolution of depth map is demonstrated through subtraction of the left digitized image from the pseudo-left digitized image (e.g., Fig. 15b).

tion. Because the first lander on Mars will have such a stereo imaging system, research will be carried out to determine the accuracy with which we can produce a depth map for purposes of manipulating the scoop on the Viking lander.

Two additional scenes are shown, more to show the difficulties they present than to demonstrate a solution. In our first scene (Fig. 20) are seen sand dunes in which the detail is represented by wind striations on the surface. As of yet, we have not performed any depth mapping of this particular image. It is not certain that one could do any segmentation of this particular image aside from the separation of sky and background range of mountain. In Fig. 21 we see a lava field in which there is a great deal of texture and a great many small rocks. This particular image might overwhelm the computer with data and require

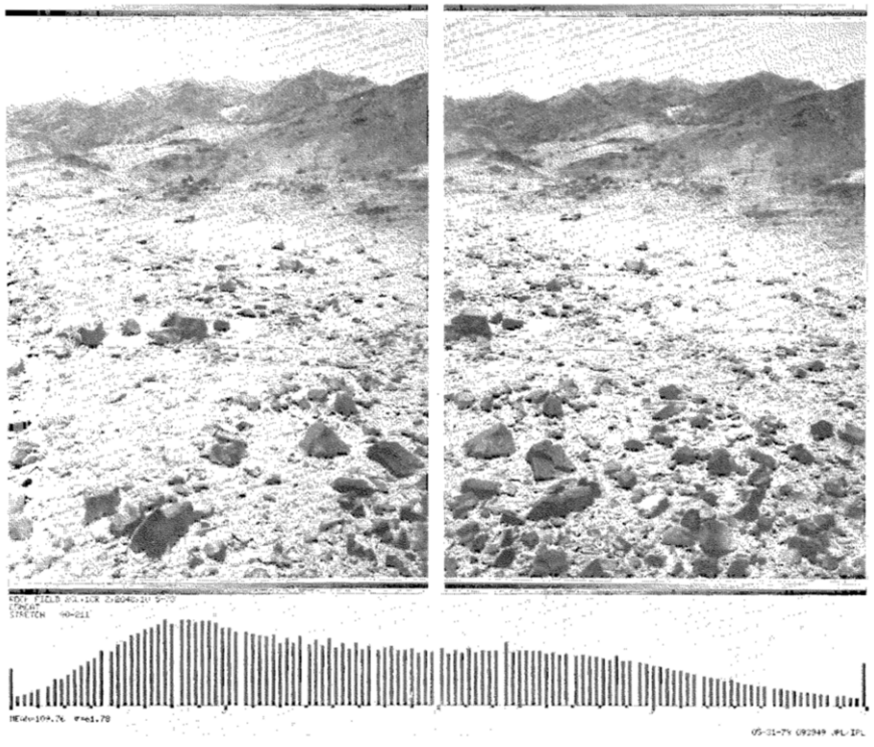

FIG. 19. Digitized stereo images of a rock scene in the Mojave Desert.

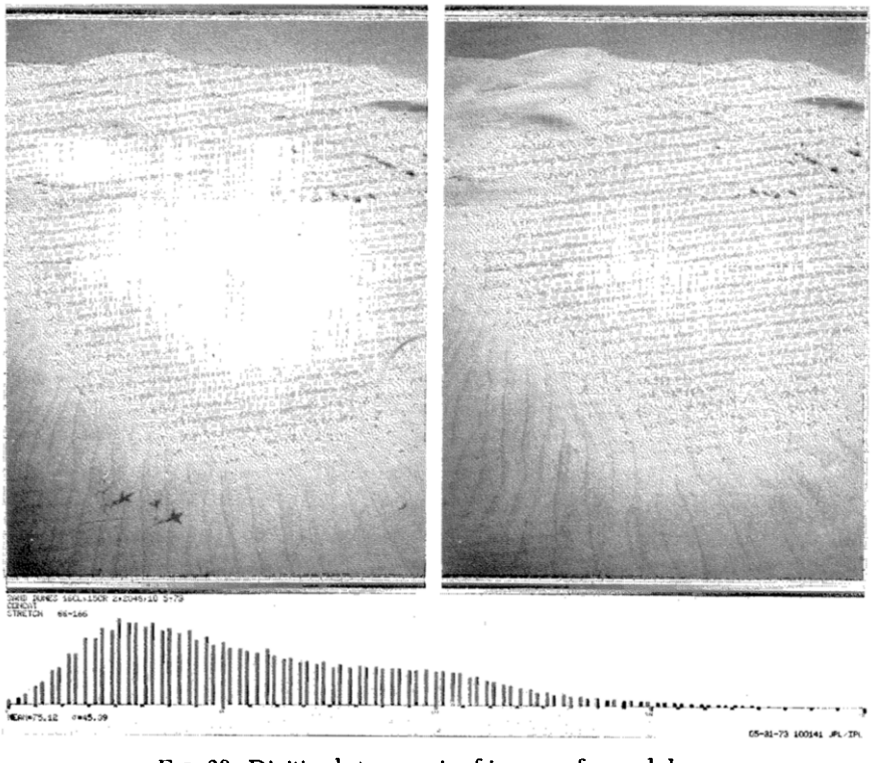

FIG. 20. Digitized stereo pair of images of a sand dune.

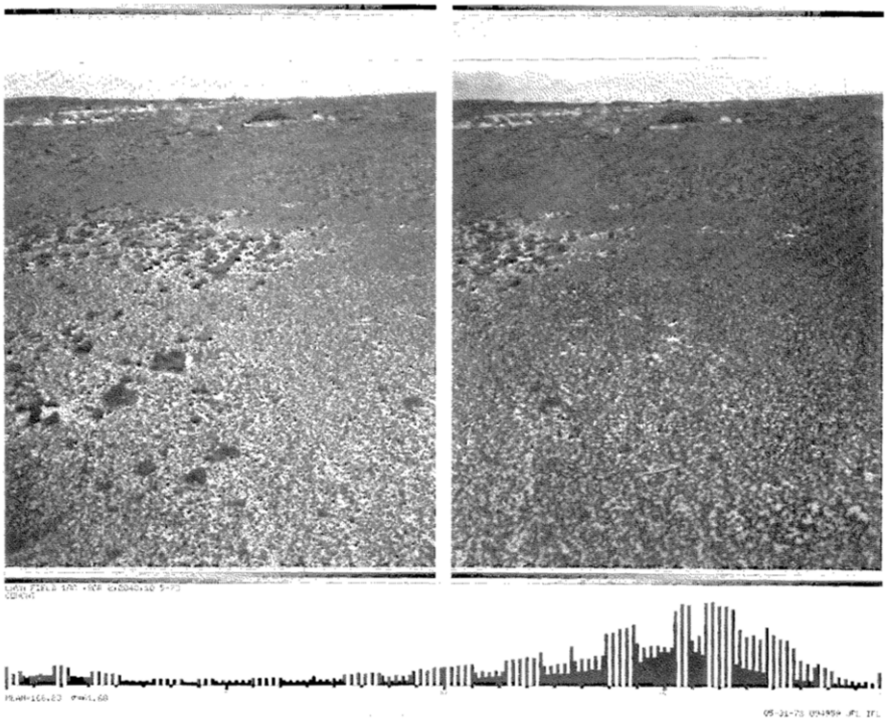

FIG. 21. Digitized stereo pair of images of lava field.

some heuristic instructions that will cause it to realize that the lava ash is simply very fine dust and should be treated in the same manner as sand.

The Martian environment does not represent the only hostile environment to which vehicles or rovers might be sent. The stereo imaging activity might also be used on an underwater vehicle in attempting to locate objects or crevices underwater. Stereo TV for depth perception has been used in our research, but this does not indicate a bias in favor of stereo imaging for depth determination or against the use of a laser range finder. It appears that, for navigation and for distances exceeding 30 m, the power required for a laser range finder might be in excess of that which could be adequately provided. Research will continue in the coming years on both systems. It is hoped that during this coming year a system incorporating a laser range finder, a stereo television system, and a copy of the Stanford arm will be implemented and will provide a basic test of scene analysis on irregular objects in order to demonstrate the integration of perception and manipulation. In the following year, the perception and manipulation will be integrated with locomotion to provide a variably autonomous integrated robot.

ACKNOWLEDGMENTS

The author acknowledges significant work done in this study by Dr. Martin Levine, currently on sabbatical leave from McGill University, and the dedication with which he has guided much of this research. Mr. Gary M. Yagi has

given much of his time and effort to this research. The work described here is the result of efforts of many members of the Image Processing Laboratory staff which, although not directly related to the program in artificial intelligence, have nevertheless provided a base from which scene analysis research could be started and continued in the rapid manner in which it has proceeded.

This paper presents the results of one phase of research carried out at the Jet Propulsion Laboratory, California Institute of Technology, under Contract NAS7-100, sponsored by the National Aeronautics and Space Administration.

REFERENCES

1. J. W. MOORE, Toward remotely controlled planetary rovers, *Astronautics and Aeronautics* 10, No. 6, 42–48, 1972.
2. R. A. LEWIS, AND A. K. BEJCZY, Planning Considerations for a Roving Robot with Arm, Third International Joint Conference on Artificial Intelligence, Stanford, California, August 20–24, 1973, paper 11–4.
3. D. O'HANDLEY, Reality of robots, presented at the Milwaukee Symposium on Automatic Control (Application of Mini-Computers to Control and Robotic Systems), March 10, 1973.
4. D. A. O'HANDLEY, Recent developments in digital image processing at the Image Processing Laboratory of JPL, *Remotely Manned Systems (Exploration and Operation in Space)* (E. Heer, Ed.), California Institute of Technology, 1973.
5. M. D. LEVINE, Scene analysis for a breadboard Mars robot functioning in an indoor environment, JPL – Technical Memorandum 33-645, 1973.
6. R. O. DUDA, AND P. E. HART, *Pattern Classification and Some Analysis*, Wiley–Interscience, New York, 1972.
7. M. D. LEVINE, D. A. O'HANDLEY, AND G. M. YAGI, Computer determination of depth maps, *Computer Graphics and Image Processing*, to be published.