2P

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Technical Memorandum 33-645

Scene Analysis for a Breadboard Mars Robot Functioning in an Indoor Environment

Martin D. Levine

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

September 1, 1973

46

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Technical Memorandum 33-645

Scene Analysis for a Breadboard Mars Robot Functioning in an Indoor Environment

Martin D. Levine

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

September 1, 1973

Prepared Under Contract No. NAS 7-100
National Aeronautics and Space Administration

PREFACE

The work described in this document was performed by the Space Sciences Division of the Jet Propulsion Laboratory.

ACKNOWLEDGMENT

The author would like to acknowledge the contributions made to this research by H. Alsberg, F. C. Billingsley, D. A. O'Handley, and G. M. Yagi, and to thank the American Society of Engineering Education for the ASEE-Ford Foundation Fellowship, which made his stay at JPL possible.

CONTENTS

FIGURES

CONTENTS (contd)

FIGURES (contd)

ABSTRACT

This report deals with the problem of computer perception in an indoor laboratory environment containing rocks of various sizes. The sensory data processing is required for the NASA/JPL breadboard mobile robot that is a test system for an adaptive variably-autonomous vehicle that will conduct scientific explorations on the surface of Mars. Scene analysis is discussed in terms of object segmentation followed by feature extraction, which results in a representation of the scene in the robot's world model.

I. INTRODUCTION

Research in computer perception aimed at satisfying the needs of a simulated NASA/JPL Mars robot is concerned with two main problems. One arises out of the requirement that a breadboard version of a mobile robot be constructed and experimented with in a laboratory, while the second deals with the test vehicle in an outdoor environment. The primary emphasis of the work until now has been on the first problem, although not to the complete exclusion of the second (Ref. 1). Reference 1 describes a procedure for obtaining, by means of stereoscopic vision, the depth maps required by the robot for both indoor and outdoor environments.

The robot, shown in Fig. 1, will exhibit an adaptive variably-autonomous behavior and is aimed at conducting scientific explorations of the surface of Mars. Such a vehicle is required both to cope with the long communication delays between Earth and Mars, and to enhance scientific yield with respect to flight cost. The two primary modes of operation of the robot are navigation and manipulation. The former must address itself to the problem of scene analysis in a dynamic situation, the main objective being to avoid obstacles during a traverse. From a visual point of view, manipulation is concerned with detecting interesting samples and evaluating their shape. A detailed discussion of the robot is beyond the scope of this report, but some discussion related to it is available elsewhere (Refs. 2, 3, 4, and 5).

The navigation and manipulation modes of operation are also the concern of the robot in the laboratory environment. We shall further assume that both obstacles and samples are rocks and are distinguishable only by their size, the large ones being obstacles and the small ones samples. We note that on the surface of Mars other obstacles such as steep cliffs, rock fields, and soft sand dunes will exist. Also interesting samples may be characterized by their texture and color pattern in addition.to size. Figure 2 is an outdoor scene of rocks lying on a sidewalk exhibiting a complexity and image contrast that one would expect for the manipulative task. Ultimately, the robot laboratory is planned to have uniformly painted walls, floor, and ceiling, and to contain high-contrast rocks, thus greatly simplifying the scene analysis. However, in our tests we used the more complex picture shown in Fig. 3, which is an indoor laboratory scene of generally low contrast containing a large variety of rocks, both in size and albedo. This has allowed us to experiment with our segmentation methods, in particular to determine their strengths and weaknesses. The general scenario might be typical of what the robot could expect during the navigation phase. These two scenes should be contrasted with Figs. 4, 5, and 6, which were taken in the California desert and give a more realistic indication of what the Martian scene analysis problem will entail. We categorically state that the approach discussed in this paper is applicable to the indoor environment only and is not extensible to outdoor scenes! Indeed a completely different and more powerful approach is necessary for the latter, and this will be discussed briefly in Section IV.

The visual input for the robot is comprised of dual parallel television cameras, which are of the silicon vidicon type. A sequential column digitizer has been constructed for the pair and acts as the interface to the computer. Controlled pan and tilt will also be made available. Binocular input data is required for depth perception as described in Ref. 1, but the right camera, R, will henceforth be considered as the reference image input.

The objective of the scene analysis system is to interpret the three-dimensional environment as input by the stereo image digitizers and laser range finder, to output a suitable representation to the world model according to the process shown in Fig. 7. The world model of the robot acts as the depository of all sensory information. It also contains the complete status of the robot system and plays the role of a switchboard for information that it routes to the appropriate subsystems. If additional data is called for by one of these subsystems, the world model can task the scene analysis program to achieve the appropriate goal.

The world model may be considered to be comprised of two major components. The first is referred to as a model of assertions or generic model because it represents in terms of a relational graph our postulations concerning the scenes of interest. In nearly all scene analysis programs to date, this model is merely implicit in the design and development process. However, there is an increasing feeling that this model should be explicitly defined and incorporated as a semantic memory and this is the approach taken with the NASA/JPL robot.[1] The second component is the so-called model of assignments, which essentially is a description, in terms of a graph, of the scene presently under consideration by the robot. Obviously all assignments are made in terms of the generic model. The work of Winston (Ref. 6) on the learning of structural descriptions is of interest in this regard. The model of assertions for the indoor laboratory scenes such as Fig. 3 is shown in Fig. 8. We may distinguish two categories of information plus the robot in this model: the rocks that are the data, and the walls and ceiling that constitute the background. Compared with this simple situation, Fig. 9 shows a relational graph for outdoor environments of the kind shown in Figs. 4, 5, and 6. Considerable research is required to complete and supply the missing details in this graph.

The scene analysis or computer perception program may be fragmented into three distinct steps. The first entails the process of image segmentation, which is concerned with the problem of isolating as entities the object subsets in the image. This classification of the points of an image into background and data point sets is the precursor to the ensuing feature analysis of the objects in the scene. The problem of segmentation in the indoor environment is discussed in Section II. The second step relates to the perception of the depth of each point in the object set, and this has been

[1] Udupa, S. M., private communication in the form of the unpublished report, Data Structure and World Model. California Institute of Technology, Pasadena, Calif., January 26, 1973.

discussed in detail by Levine, et al. (Ref. 1). Finally, Section III is concerned with step three: the object description using both the depth and segmented image data.

II. OBJECT SEGMENTATION

The process of image digitization is followed by its segmentation into objects that might be either obstacles or samples. No a priori assumptions can be made regarding the number, size, location, or shape of the rocks in the scene. It is only known that they may be found on the floor of the laboratory in a random position.

Let the actual three-dimensional scene perceived by a human be denoted by S and let R denote the digitized two-dimensional image of S as observed by the right camera of the stereo pair mounted on the robot. We further let R be a picture function defined on an $M \times N$ grid Π that constitutes the raster scanned by the digitizing hardware.[2] The set Π consists of picture elements π located at the coordinates (α, γ) where $1 \leq \alpha \leq M$ and $1 \leq \gamma \leq N$ define the coordinates of any particular image point. The lower left-hand point in any image will be referenced as $(\alpha, \gamma) = (0, 0)$ and processing will proceed row by row from left to right ($\gamma = 1$ to $\gamma = 1000$) and bottom to top ($\alpha = 1$ to $\alpha = 1000$). The optical density or gray level is given by $R(\gamma, \alpha)$ and can take on any integer value in the set $(1, \cdots, 256)$. In all cases we have obtained $R(\alpha, \gamma)$ by digitizing a photograph.

The segmentation problem can now be stated as follows: given $R(\alpha, \gamma)$, classify each point $\pi(\alpha, \gamma) \epsilon \Pi$ as either belonging to the set B of background points or the set D of data points. We then define objects Ψ_i as subsets of the points in the set D that are connected according to the property of four-connectivity (Ref. 7). The set of all segmented objects in the image R is given by $O = (\Psi_1, \Psi_2, \cdots,)$. The feature analysis of the isolated objects Ψ_i in O that represent the rocks in the scene S as seen by the robot will be discussed in the next section.

The procedure by which the set $\Pi = (B, D)$ is segmented will be referred to as adaptive histogram analysis. This is because the gray-level

[2] In our case $M = N = 1000$.

thresholds that are used to classify a given point into either B or D are determined by an analysis of the local histogram. To avoid the problems associated with constructing a histogram with an inadequate number of samples, each histogram was smoothed with a nine-point filter with weights (1, 2, 3, 4, 5, 4, 3, 2, 1) centered at 5. The segmentation technique is sensitive to the degree of smoothing since too little smoothing yields many local peaks in the histogram, while a large amount eliminates many of them. Experimentation with the type of images under study is required to determine the scope of the filter.

Suppose that we make the assumption that the background is uniformly shaded. Lighting conditions will of course result in variations in the scene and it is for this reason that a local analysis is required. Further we shall assume that the background is represented in the picture histogram by a Gaussian distribution function or the major part of such a distribution. In actuality this function will be corrupted by the effect of objects in the picture. Let μ_B and σ_B define the background mean and standard deviation, respectively, as shown in Fig. 10. Let Γ_B be the full-width at half-maximum, the so called half-width. Then it can be shown that

$$\Gamma_B = 2.354\ \sigma_B$$

Given that the peak of the distribution has been obtained by search, Γ_B and thence $2\sigma_B$ can easily be computed. Thresholds $T_L = -2\sigma_B$ and $T_R = 2\sigma_B$ can be arbitrarily set to isolate and erase those points in the image that are part of the background. We shall refer to this as the two-sigma assumption.

Let the grid Π be subdivided into m × n submatrices for which a local histogram may be calculated.[3] It is desirable to detect that peak in the histogram that is attributable to the background. In general, however, there will exist one peak or more. We ensure that the background peak is

[3] In our case m = n = 16 resulting in the picture containing 256 submatrices. It is in this sense that we use the term local.

the one found by constraining the one-dimensional search to a range predicated on the location of the peaks already detected. Linear interpolation for the μ_B's using the matched filter shown in Fig. 11(a) is used to predict the probable location of the peak. The ensuing one-dimensional search of the histogram is restricted to plus or minus δ from this predicted gray-level value. In our case $\delta = 16$ was found to be optimum. Given that μ_B has been found, then the two-sigma assumption is invoked to calculate the thresholds T_L and T_R. If no background peak is found we may deduce that it is masked by the data peak or that the image contains only data in this neighborhood. In this circumstance, the thresholds for the submatrix centered at the point (α, γ) under consideration are estimated by averaging the values of the pertinent thresholds of the four closest neighboring submatrices using the matched filters shown in Fig. 11(b).

The situation may occur in which we cannot calculate Γ_B because of adjoining peaks that eliminate the point of half-maximum. The three possible cases are shown in Fig. 12. For any given side of the background distribution, if this situation occurs, the threshold is placed half-way between μ_B and the adjoining peak. Thus in Fig. 12, (a) and (b) the two-sigma assumption is valid only for one side of the background distribution, while in Fig. 12(c) it is invalid for both sides.

Although T_L and T_R are only computed from the actual histograms for the 256 submatrices, we do estimate their value for each element in the set Π. This estimation is carried out, after the complete image matrix has been processed, by means of linear interpolation among the computed thresholds. We note that for any background probability distribution, the points in the submatrix that contribute to its tails are most likely located on the extremities of the region delineated by the given submatrix. Large variations in background shading would presumably cause a large standard deviation σ_B. Therefore it would be expected that the two-sigma assumption would yield the best results for points near the center of the region. By estimating T_L and T_R for the other points in this region we are essentially tracking the changes in μ_B and thus contributing to a more accurate segmentation.

Figure 13 shows the segmented image of the sidewalk scene in Fig. 2. The two-sigma assumption was used to determine the thresholds for the background and only the set D is displayed. Small particles less than a certain perimeter are then erased and the result is shown in Fig. 14. The objects Ψ_i, where the label i is shown next to the rock, are seen to be well isolated. The result of this processing is therefore the set of objects $O = (\Psi_1, \Psi_2, \cdots, \Psi_{13})$.

In comparison to the high contrast sidewalk scene, the low contrast indoor laboratory scene shown in Fig. 3 is much more problematical. This scene was illuminated by two floodlamps, resulting in certain areas on the floor with shadows cast by either lamp or a combination of the two. Because of the existence of the three types of shadows it was not possible to track their peaks in the histogram with any confidence. However, given better pictures, it would seem to be within the realm of possibility to predict which peak belonged to the shadow, which to the background per se, and thus to isolate the objects. Certain semantic information could be brought to bear such as that the shadows' peak must be darker than the floor's (background) peak. We also note in Fig. 3 that there is a tremendous variation in shading on the rocks with parts of some of them barely visible. No effort was addressed to the problem of detecting the drapes hanging on the two walls, but it will be observed in the results that the bottom of the drapes could be easily detected using a simple edge-detector.

Figures 15 and 16 show the segmented scene before and after, respectively, the erasure of small insignificant objects. The existence of two shadows is quite evident for object 3, while other objects such as 2 and 8 have only one shadow. Rocks 5 and 10 have been broken into more than one part although the parts could be easily connected using a local operator. A large part of rock 9 has been lost; however the object descriptor to be discussed in the next section requires only the extremities of rocks and would therefore be relatively insensitive to this type of loss. It is unclear from the pictures whether the same holds true for rock 2. A significant amount of experimentation has been carried out with the parameters associated with the method, and any attempt to bring out greater portions of the rock results in some of the background being assigned to the set D.

Considering the poor quality of the image, the results are encouraging and await a test on the higher contrast images that can be obtained with the actual robot television cameras.

We have shown in this section that the adaptive histogram analysis, as an approach to picture segmentation, yields good results when a reasonably high contrast image is available. Even with considerable image degradation and poor lighting, the performance of the technique is quite adequate. The resulting set of objects O in the picture is analyzed and descriptors computed as discussed in the next section.

III. OBJECT DESCRIPTION

A. Point Transformation

The set O consists of subsets of points Ψ_i, each of which represents the set of points constituting a particular object, in our case a rock. The problem to be discussed in this section deals with transforming a particular point $\pi \in \Psi_i$ in the picture domain into an equivalent point in a global coordinate system. That is, given a point in the picture on a rock, what are its coordinates in the scene S?

The mapping of a point in a picture to a point in three-dimensional space will be achieved by means of a perspective transformation with the analysis and notation following quite closely the approach of Duda and Hart (Ref. 8). The latter refer to this transformation as "the natural first-order approximation to the process of taking a picture;" a detailed discussion of the model used is presented in Ref. 8 and will not be repeated here.

Let us consider a global rectangular coordinate system (x, y, z) in which the object point p corresponding to the image point π is represented. The origin of this coordinate system or base point is fixed by the robot world model and used as a reference for all distance and location computations. It is assumed that measuring devices affixed to the robot are capable of computing the distance and direction of travel from the base point. In this way, the gimbal center $\underline{v}_o = (x_o, y_o, z_o)^t$, which is the origin of the gimbal frame for the stereo cameras, can be tracked in the global coordinate system and is therefore assumed to be known.

In addition to the global system, we shall define the picture coordinate system referenced as before to the right camera image, R, of the stereo pair. Henceforth all references to picture points in the picture coordinate system will carry this interpretation. Let the symbols

$$\begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}$$

denote the coordinates in the picture coordinate system of the point π.

We wish to show that given π in the picture coordinate system and the corresponding retinal disparity, d, the coordinates of p in the global coordinate system can be computed. It is assumed that the retinal disparity for any point π in the right image can be found using either the algorithm of Levine, et al. (Ref. 1), which was concerned with the computation of a depth map, and therefore the disparity, for the complete image array, or the laser range finder. An example of the result of the stereoscopic processing algorithm is shown in Fig. 17, where the computed depth contours are superimposed on the original reference image shown in Fig. 2. Obviously the disparity need only be computed for points belonging to the set O and this will be discussed in greater detail in the next section.

Figure 18 shows the physical situation for the stereoscopic TV input scanners with the appropriate coordinate systems. It is assumed that both cameras are panned through identical angles θ measured counterclockwise from the y axis, and tilted through identical angles ϕ, measured positive in an upward direction. Both θ and ϕ measurements are assumed to be always available to the computing system. There exists a constant offset vector $\underline{\ell} = (\ell_1, \ell_2 + f, \ell_3)^t$, where f is the focal length between the camera gimbal references and the origin of the image plane, and this offset is assumed to be identical for both cameras. Figure 19 shows both the left and right gimbal reference vectors $\underline{v}_{o_L}$ and $\underline{v}_{o_R}$, respectively, as well as the gimbal center. If we define $\underline{\Delta}$ as the baseline vector, then it is easily seen that the gimbal references are given by:

$$\underline{v}_{o_R} = \underline{v}_o + \underline{\Delta}_R$$

$$\underline{v}_{o_L} = \underline{v}_o + \underline{\Delta}_L$$

where $\underline{\Delta} = \underline{\Delta}_R - \underline{\Delta}_L$. Note that $\underline{v}_{o_L}$, $\underline{v}_{o_R}$, and $\underline{\Delta}$ are constants of the system and are therefore assumed to be known.

We are now ready to state the result as derived in Ref. 8. The coordinates (x, y, z) of the point p in the global coordinate system, given the picture coordinates (α, 0, γ) in the right image and the disparity, d, for this point, are computed by:

$$\underline{p} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \frac{1}{2}\left[a_o\,\underline{u}_L + b_o\,\underline{u}_R + \underline{\Delta}\right] + \underline{k}_L$$

where

$$\underline{u}_L = \frac{\underline{c}_L - \underline{k}_L}{\|\underline{c}_L - \underline{k}_L\|} , \quad \underline{u}_R = \frac{\underline{c}_R - \underline{k}_R}{\|\underline{c}_R - \underline{k}_R\|}$$

$$\underline{c}_L = \underline{v}_{o_L} + \begin{bmatrix} (\alpha + d + \ell_1)\cos\theta - (f + \ell_2)\cos\phi\,\sin\theta + (\gamma + \ell_3)\sin\phi\,\sin\theta \\ (\alpha + d + \ell_1)\sin\theta + (f + \ell_2)\cos\phi\,\cos\theta - (\gamma + \ell_3)\sin\phi\,\cos\theta \\ (f + \ell_2)\sin\phi + (\gamma + \ell_3)\cos\phi \end{bmatrix}$$

$$\underline{c}_R = \underline{v}_{o_R} + \begin{bmatrix} (\alpha + \ell_1)\cos\theta - (f + \ell_2)\cos\phi\,\sin\theta + (\gamma + \ell_3)\sin\phi\,\sin\theta \\ (\alpha + \ell_1)\sin\theta + (f + \ell_2)\cos\phi\,\cos\theta - (\gamma + \ell_3)\sin\phi\,\cos\theta \\ (f + \ell_2)\sin\phi + (\gamma + \ell_3)\cos\phi \end{bmatrix}$$

$$\underline{k}_L = \underline{v}_{o_L} + \begin{bmatrix} \ell_1 \cos\theta - \ell_2 \cos\phi \sin\theta + \ell_3 \sin\phi \sin\theta \\ \ell_1 \sin\theta + \ell_2 \cos\phi \cos\theta - \ell_3 \sin\phi \cos\theta \\ \ell_2 \sin\phi + \ell_3 \cos\phi \end{bmatrix}$$

$$\underline{k}_R = \underline{v}_{o_R} + \begin{bmatrix} \ell_1 \cos\theta - \ell_2 \cos\phi \sin\theta + \ell_3 \sin\phi \sin\theta \\ \ell_1 \sin\theta + \ell_2 \cos\phi \cos\theta - \ell_3 \sin\phi \cos\theta \\ \ell_2 \sin\phi + \ell_3 \cos\phi \end{bmatrix}$$

$$a_o = \frac{(\underline{u}_L \cdot \underline{\Delta}) - (\underline{u}_L \cdot \underline{u}_R)(\underline{u}_L \cdot \underline{\Delta})}{1 - (\underline{u}_L \cdot \underline{u}_R)^2}$$

and

$$b_o = \frac{(\underline{u}_L \cdot \underline{u}_R)(\underline{u}_L \cdot \underline{\Delta}) - (\underline{u}_R \cdot \underline{\Delta})}{1 - (\underline{u}_L \cdot \underline{u}_R)^2}$$

This computation is not as imposing as it might seem at first glance. First, $\underline{k}_L$ and $\underline{k}_R$ need only be computed once during the analysis of a particular image. Second, for a fixed γ, that is for any row in the image, the vectors $\underline{c}_L$ and $\underline{c}_R$ are related according to

$$\underline{c}_L(\alpha_2) = \underline{c}_L(\alpha_1) + \begin{bmatrix} \left\{(\alpha_2 - \alpha_1) + (d_{\alpha_2} - d_{\alpha_1})\right\} \cos\theta \\ \left\{(\alpha_2 - \alpha_1) + (d_{\alpha_2} - d_{\alpha_1})\right\} \sin\theta \\ 0 \end{bmatrix}$$

$$\underline{c}_R(\alpha_2) = \underline{c}_R(\alpha_1) + \begin{bmatrix} \left\{(\alpha_2 - \alpha_1)\right\} & \cos\theta \\ \left\{(\alpha_2 - \alpha_1)\right\} & \sin\theta \\ 0 & \end{bmatrix}$$

where α_1 and α_2 are the coordinates of two picture points within the same image row.

In the next section we shall assume that the coordinates in the global coordinate system of any point π in the picture coordinate system can be computed as discussed above.

B. Feature Extraction

For the navigation and manipulation tasks proposed for the robot in the indoor setting described in Section I, the rock descriptions may be kept quite simple. Indeed we shall confine ourselves to a location and a shape description. Thus each object i in the three-dimensional scene can be represented by the following LEAP (Ref. 9) data structure:[4]

NAME ⊗ ROCK ≡ (i)

DESCRIPTION ⊗ ROCK ≡ (LOCATION DESCRIPTOR,
SHAPE DESCRIPTOR)

The location descriptor is a single point in (x, y, z) space that approximates the center of gravity (g) of the rock. We may describe the shape descriptor as the minimum perceivable enclosing polygonal cylinder (MPEPC). Each of these descriptions will now be discussed in detail.

It is important to realize that because of occlusion not all points on a given rock may be perceivable at any one time, or for that matter at any time. Nevertheless, an attempt is made to enclose each rock of interest

[4] The basic form in LEAP is given by:
ATTRIBUTE ⊗ OBJECT = VALUE

within a box with a polygonal cross section and height equal to the highest observed point on the obstacle. Given all the information available concerning the rock, this description turns out to be a conservative one. However, it is wholly adequate for the purposes of navigation and manipulation in the indoor laboratory environment. In the former we are concerned with planning a path and are interested only in avoiding obstacles with complete certainty. In the other case, the tilt angle will be chosen sufficiently negative so that the complete surface cross section is visible by the camera. Both of these conform to human practice.

Under certain circumstances a minimum perceivable enclosing rectangular parallelepiped (MPERP) might be adequate thereby reducing the computer storage requirements for the world model. Indeed the shape descriptor procedure computes the MPERP for the objects in the scene at each point in time. If the MPERP was previously computed for a particular rock from a different vantage point, then these are combined to form the MPEPC.

Let us first consider the MPERP as the shape descriptor and define it by means of the five points $p_k = (x_k, y_k, z_k)^t$, where $k = 2, 4, 6, 8$ (as shown in Fig. 20), and 9. How these points are determined utilizing the segmented objects Ψ_i will be described below. The shape descriptor is given by the following quintuple:

$$\text{SHAPE DESCRIPTOR} \otimes \text{ROCK} \equiv \Big((x_2, y_2), (x_4, y_4), (x_6, y_6), (x_8, y_8), z_9\Big)$$

We note that the first four elements prescribe the points that define the rectangular cross section and the fifth the height of the parallelepiped. To compute p_2, p_4, p_6, and p_8 it is first necessary to find the points p_1, p_3, p_5, and p_7. These are found by searching in the manner described below for certain points $\pi \epsilon \Psi_i$ where π has the coordinates $(\alpha, 0, \gamma)$ in the picture coordinate system and i refers to the name of the rock.

Let π_3 and π_7 be mapped into the points $p_3 = (x_3, y_3, z_3)^t$ and $p_7 = (x_7, y_7, z_7)^t$ in the global coordinate system. Then π_3 and $\pi_7 \epsilon \Psi_i$ are defined such that

$$\alpha_3 \geq \alpha \quad \forall \quad \pi \in \Psi_i$$

$$\alpha_7 \leq \alpha \quad \forall \quad \pi \in \Psi_i$$

This is shown in Fig. 21.

Whereas π_3 and π_7 are determined solely by the information in the α-γ plane, p_1 and p_5 require depth information. Let π_1 and $\pi_5 \in \Psi_i$ correspond to $p_1 = (x_1, y_1, z_1)^t$ and $p_5 = (x_5, y_5, z_5)^t$, respectively. These points are chosen so that they map into points in the global coordinate system that represent the closest and farthest points on the rock with respect to the baseline vector. We define these points such that

$$d_1 \geq d \quad \forall \quad \pi \in \Psi_i$$

$$d_5 \leq d \quad \forall \quad \pi \in \Psi_i$$

where d_1 and d_5 represent the retinal disparities for π_1 and π_5, respectively.[5]

The highest point z_9 on the rock is determined by searching for π_9, which maps into the point $p_9 = (x_9, y_9, z_9)^t$. Thus π_9 is chosen such that

$$z_9 \geq z \quad \forall \quad \pi \in \Psi_i$$

If the height z_9 is below some threshold, the object may be ignored in that the robot vehicle will be capable of driving over it without endangering its safety. Figure 22 shows a picture with a possible placement of the five points.

[5] Parenthetically we note that in addition to the stereoscopic method described in Ref. 1, the closest and farthest points on the rock may also be obtainable by means of the laser range finder shown in Fig. 1. By viewing the point of laser light in both the left and right images, the set Ψ_i can be searched to determine the disparities d_1 and d_5. The conceptual integration of this system has not yet been accomplished.

The points p_1, p_3, p_5, and p_7 are employed to compute p_2, p_4, p_6, and p_8, which are the elements of the shape descriptor. Figure 20 shows an MPERP and the relationship between the points. A line is drawn through each of p_1 and p_5 parallel to the baseline vector $\underline{\Delta}$, while two others are drawn through p_3 and p_7 perpendicular to the baseline. The resulting rectangle is the MPERP. If the points are labelled as in Fig. 20, then it can be shown using geometrical arguments that (x_k, y_k), k = 2, 4, 6, 8 is given by:

$$x_k = \frac{x_{k-1}\, t^2 + (y_{k+1} - y_{k-1})t - x_{k+1}}{(t^2 - 1)}$$

$$y_k = \frac{y_{k+1}\, t^2 + (x_{k-1} - x_{k+1})t - y_{k-1}}{(t^2 - 1)}$$

where if

$$0 \le \theta \le \frac{\pi}{2} \quad \text{or} \quad \pi \le \theta \le \frac{3\pi}{2}$$

then

$$t = \tan\theta \quad \text{for} \quad p_2 \text{ and } p_6$$

$$t = \cot\theta \quad \text{for} \quad p_4 \text{ and } p_8$$

if

$$\frac{\pi}{2} \le \theta \le \pi \quad \text{or} \quad \frac{3\pi}{2} \le \theta \le 0$$

then

$$t = \cot\theta \quad \text{for} \quad p_2 \text{ and } p_6$$

$$t = \tan\theta \quad \text{for} \quad p_4 \text{ and } p_8$$

The MPERP is an approximation, given the parts of the rock that are visible, to the convex hull of the rock. A better approximation might be obtained by having the robot move away from its original point of observation and reanalyze the scene from a new spot. The MPERP obtained in this way can be combined with any previous MPERPs for the same rock to determine the MPEPC. Figure 23 shows the surface cross section of a rock in the x-y plane with the MPERP evaluated at two different observation points. We define the cross section of the MPEPC for a given rock as the union of the cross sections of all the MPERPs found for that rock and represent it by a list of points that can be joined to form a piecewise linear approximation to the union. The height of the MPEPC is set equal to the maximum height of all the MPERPs.

The above definitions are arbitrary but simple to implement. The MPERP will probably be adequate for navigational purposes even though it may be a poor approximation to the complete shape of the rock. In any event, more information is clearly available that could be used to better approximate the shape of the rock at the obvious expense of memory in the world model. With respect to manipulation, the tilt angle ϕ must be chosen so that the MPERP cross section completely encloses the rock sample as shown in Fig. 24. In this case the α-γ picture plane is close to parallel to the x-y plane and therefore the boundary of Ψ_i in the picture plane may be considered to be a good approximation to the cross-sectional boundary of the rock in the x-y plane. This boundary could be described more accurately if desired by chain encoding techniques (Ref. 8). Suitable features for manipulation, such as optimum gripping points, could then be calculated using this more detailed description.

We now turn to the location descriptor. The result of the above analysis for shape features is a space map, essentially a crude, three-dimensional memory of space allocation in the scene. Let us reference each MPERP in the scene and consequently in the space map by its center of gravity g, which is easily obtained:

LOCATION DESCRIPTOR ⊗ ROCK ≡ (CENTER OF GRAVITY)

$$\text{COORDINATES} \otimes \text{CENTER OF GRAVITY} \equiv \left(\left(x_4 + \frac{1}{2}(x_8 - x_4)\right), \left(y_8 + \frac{1}{2}(y_4 - y_8)\right), \frac{1}{2} z_9\right)$$

Figure 25 depicts such a space map for the scene shown in Fig. 3. The g for an MPEPC is found by averaging the g's of all the MPERPs that constitute it.

Essentially we observe that the space map is referenced by means of the location descriptor. For most cases, a particular object can be located using a nearest neighbor classification procedure (Ref. 8). In this way, as the robot traverses a path, obstacles that have been encountered previously are easily pinpointed. A certain degree of forgetting is necessary as the memory allocated for the space map will necessarily be limited. Possibly some combination of short- and long-term memory is called for, but a detailed discussion of this issue is beyond the scope of this report.

IV. DISCUSSION AND CONCLUSIONS

The problem environment considered above is an indoor laboratory with uniformly painted walls, floor, and ceiling. Rocks varying in size from a few centimeters to about 75 cm are to be placed in a random configuration on the floor to act as both samples and obstacles. If the breadboard robot is to inhabit such an area, it will be required to observe these scenes in a dynamic fashion and build up an appropriate representation in terms of a world model. In this report we have been concerned with this perceptual component of the robot's cognitive machinery and have addressed ourselves to the feasibility of its algorithmic implementation.

The scene analysis that accomplishes this task proceeds in three sequential stages after the image has been digitized. The first stage is concerned with object segmentation, the second with depth analysis, and the third with feature extraction. The last ultimately results in a representation in the world model of all the rocks in the scene. We emphasize that the approach is exclusively bottom-up in nature and depends heavily on the fact that the segmentation procedure outputs objects that are meaningful and can be named by humans. Complete dependence is made on background uniformity as a means of isolating the objects in the image.

The question arises of the degree of extensibility of this work to scenes of the kind shown in Figs. 4, 5, and 6. The experience with scene analysis in outdoor environments is extremely limited with only two references cited in the literature (Refs. 10 and 11). Both of these approaches are knowledge-based in that the semantics of the problem domain are embodied in the world model and are employed extensively during the scene analysis. The attribution of computer understanding of meaning is primarily based on this process. The analysis is goal-directed resulting in a complex combination of both bottom-up and top-down programming. It is therefore not necessary for the initial segmentation to yield objects that are nameable. In fact, one would be hard-pressed to distinguish the segmentation procedure from the rest of the scene analysis. This type of analysis may be likened to hypothesis generation and verification followed by global analysis and reorientation until the ultimate goal of representing the information in the scene is achieved.

We conclude that an approach that embodies these general principles is necessary to cope with the natural environments shown in Figs. 4, 5, and 6. Obviously the entities in these scenes are not as easily delineated as those in Fig. 3, and the segmentation procedure described in Section II would necessarily fail. A robot that will function on Mars will require an understanding of its environment far in excess of that of the indoor breadboard model.

REFERENCES

1. Levine, M. D., O'Handley, D. A., Yagi, G. M., "Computer Determination of Depth Maps," Computer Graphics and Image Processing (accepted for publication).

2. Bejeczy, A. K., Remote Manipulator Systems, Technology Review and Planetary Operation Requirements, Report 760-77. Jet Propulsion Laboratory, Pasadena, Calif., July 1, 1972. (JPL internal report).

3. Choate, R., Jaffe, L. D., Science Aspects of a Remotely Controlled Mars Surface Roving Vehicle, Report 760-76. Jet Propulsion Laboratory, Pasadena, Calif., July 1, 1972. (JPL internal report).

4. Heer, E., Advance Teleoperator/Robot Studies for Planetary Surface Roving Vehicles, Summary Report 760-75. Jet Propulsion Laboratory, Pasadena, Calif., July 1, 1972. (JPL internal report).

5. Lewis, R. A., Bejczy, A. K., "Planning Considerations for a Roving Robot with Arm," Paper 11.4, presented at the Third International Joint Conference on Artificial Intelligence, Stanford, Calif., August 1973.

6. Winston, P. H., Learning Structural Descriptions From Examples, Ph.D. thesis, Report MAC TR-76. Massachusetts Institute of Technology, Cambridge, Mass., September 1970.

7. Rosenfeld, A., Picture Processing by Computer, Academic Press, New York, N.Y., 1969.

8. Duda, R. D., Hart P. E., Pattern Classification and Scene Analysis, Wiley, New York, N.Y., 1973.

9. Feldman, J. A., Rovner, P. D., "An Algol-Based Associative Language," Communication of the ACM, Vol. 12, No. 8, pp. 439-449, August 1969.

10. Yakimovsky, Y., Feldman, J. A., "A Semantics-Based Decision Theoretic Region Analyzer," Paper 21.5, presented at the Third International Joint Conference on Artificial Intelligence, Stanford, Calif., August 1973.

11. Preparata, F. P., Ray, S. R., An Approach to Artificial Nonsymbolic Cognition, Information Sciences, Vol. 4, No. 1, pp. 65-86, January 1972.

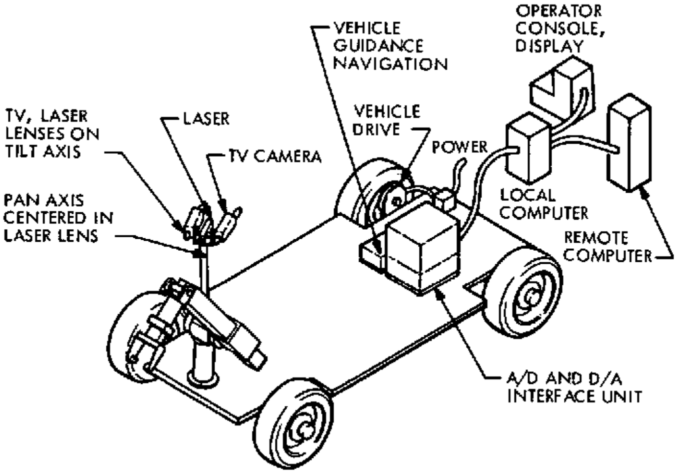

OPERATOR CONSOLE, DISPLAY
VEHICLE GUIDANCE NAVIGATION
TV, LASER LENSES ON TILT AXIS
LASER
VEHICLE DRIVE
POWER
TV CAMERA
PAN AXIS CENTERED IN LASER LENS
LOCAL COMPUTER
REMOTE COMPUTER
A/D AND D/A INTERFACE UNIT

Fig. 1. The robot breadboard hardware configuration showing the two identical and aligned optical systems and scanners

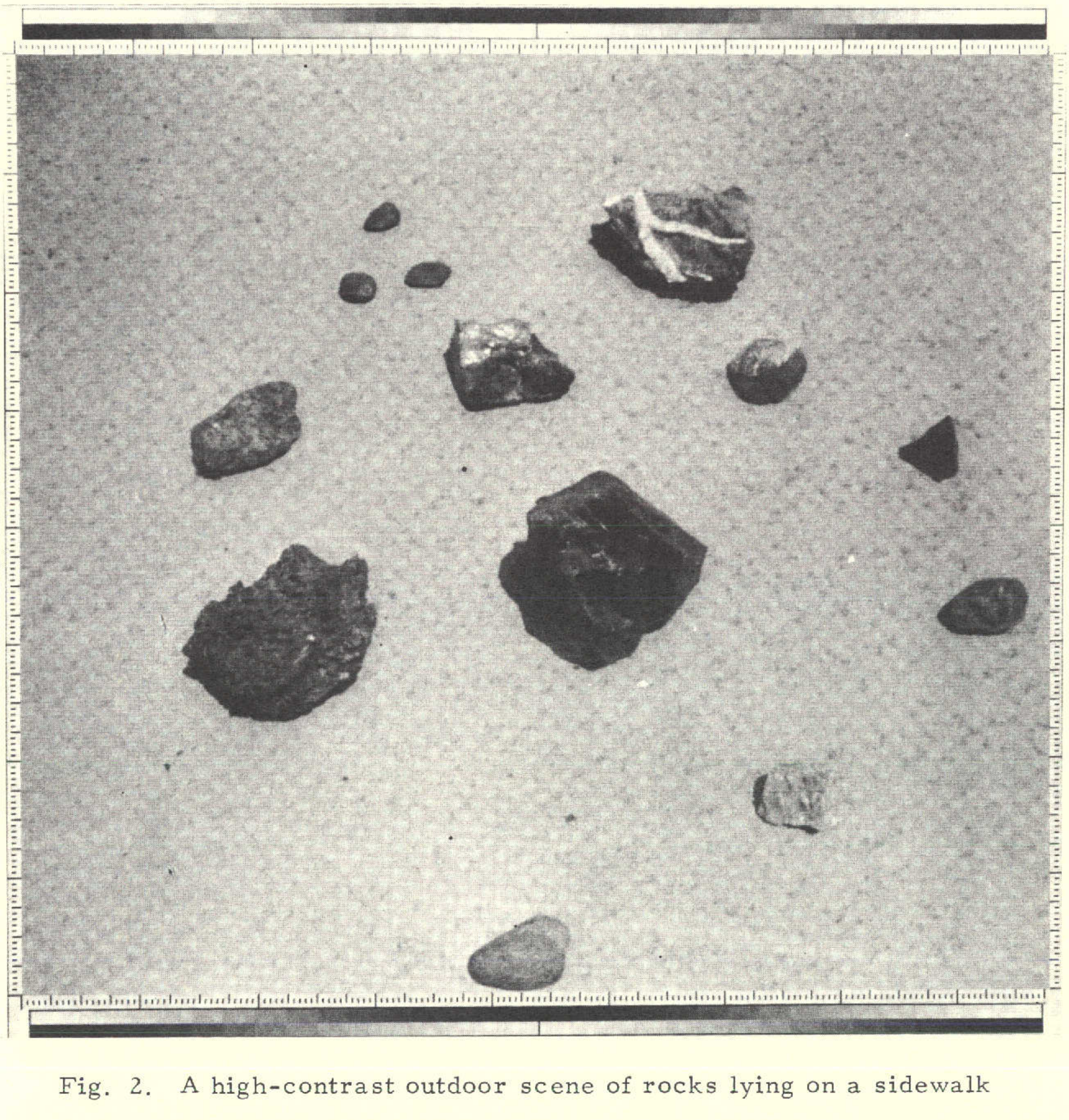


Fig. 2. A high-contrast outdoor scene of rocks lying on a sidewalk

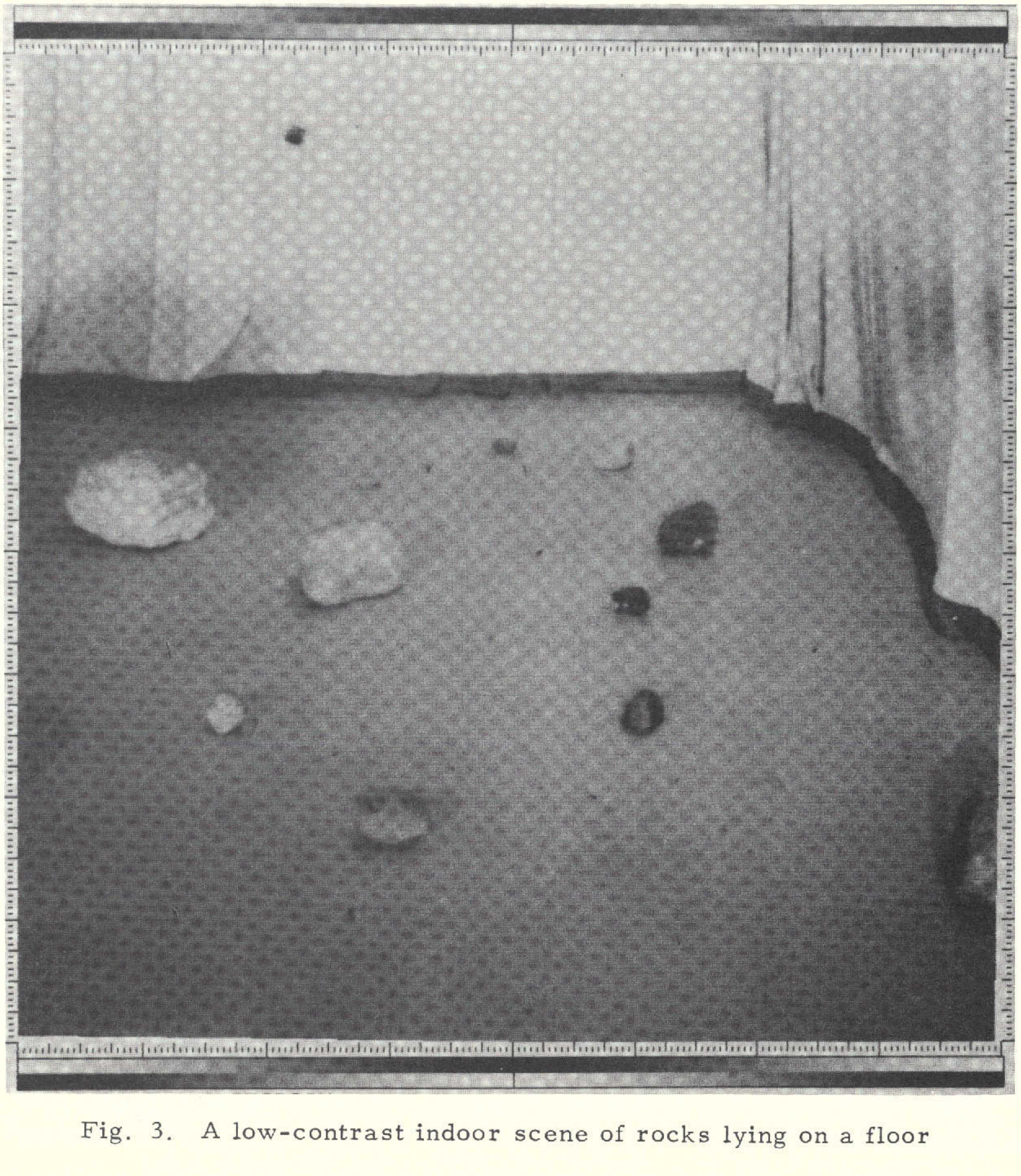


Fig. 3. A low-contrast indoor scene of rocks lying on a floor

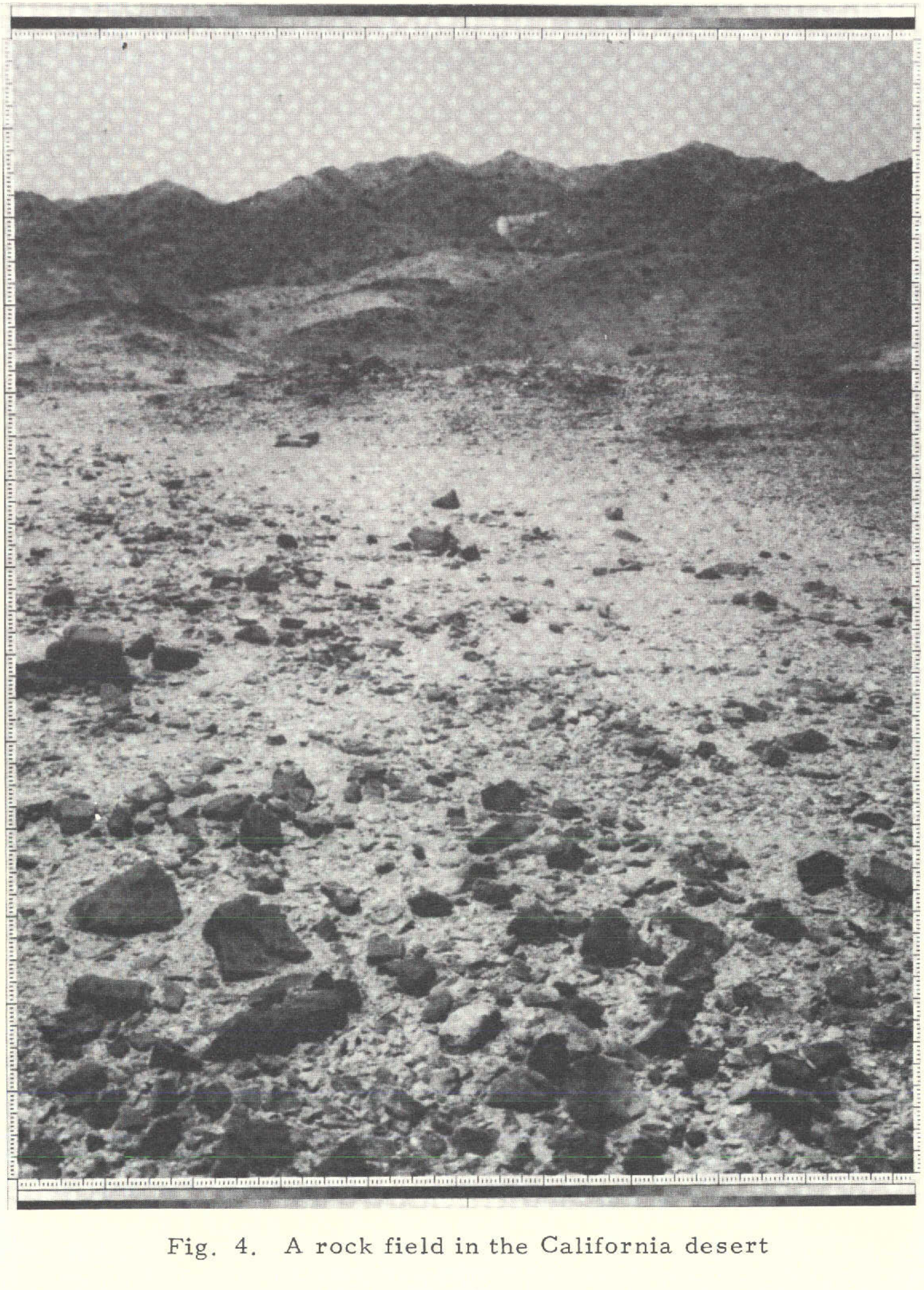


Fig. 4. A rock field in the California desert

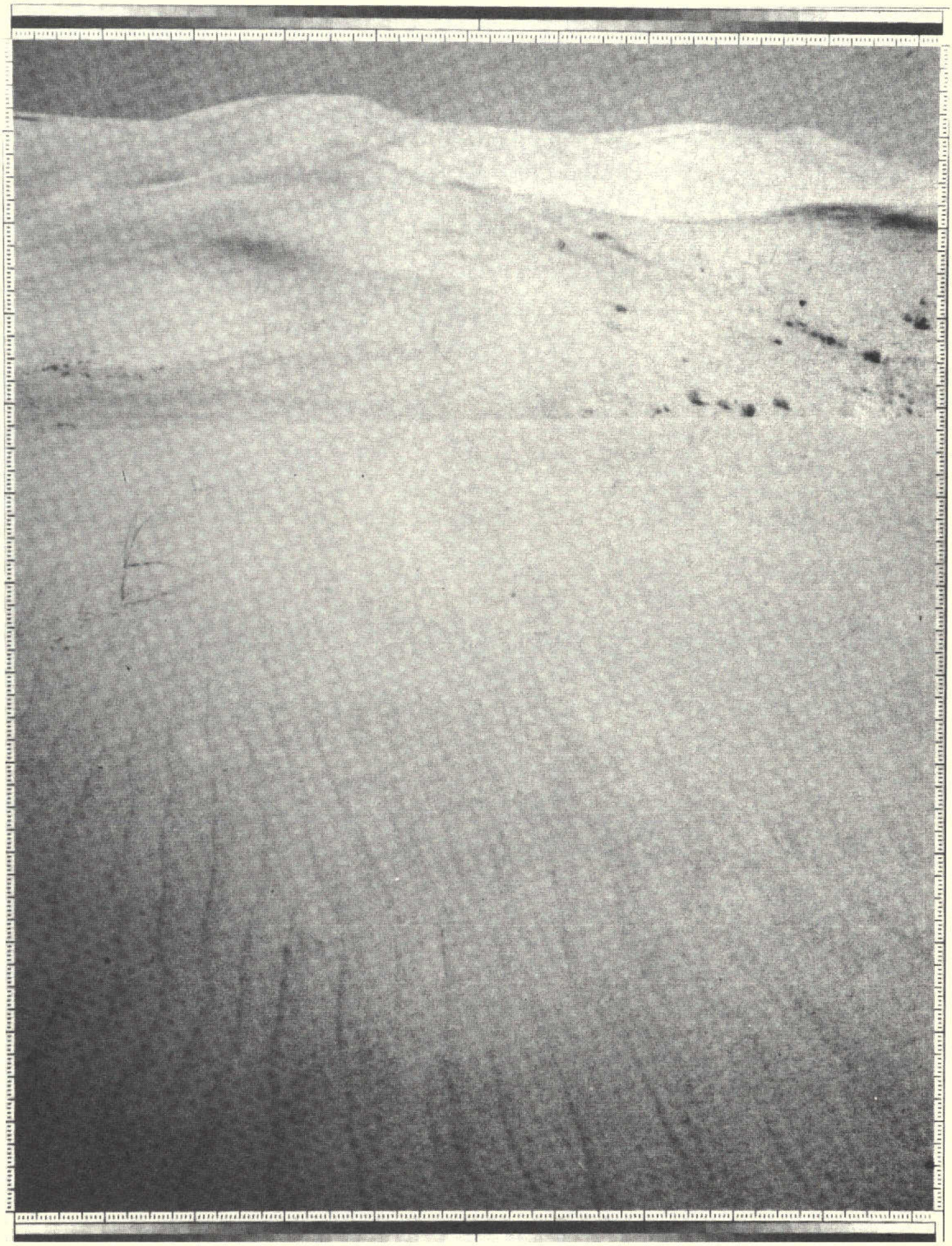


Fig. 5. An outdoor scene showing sand dunes

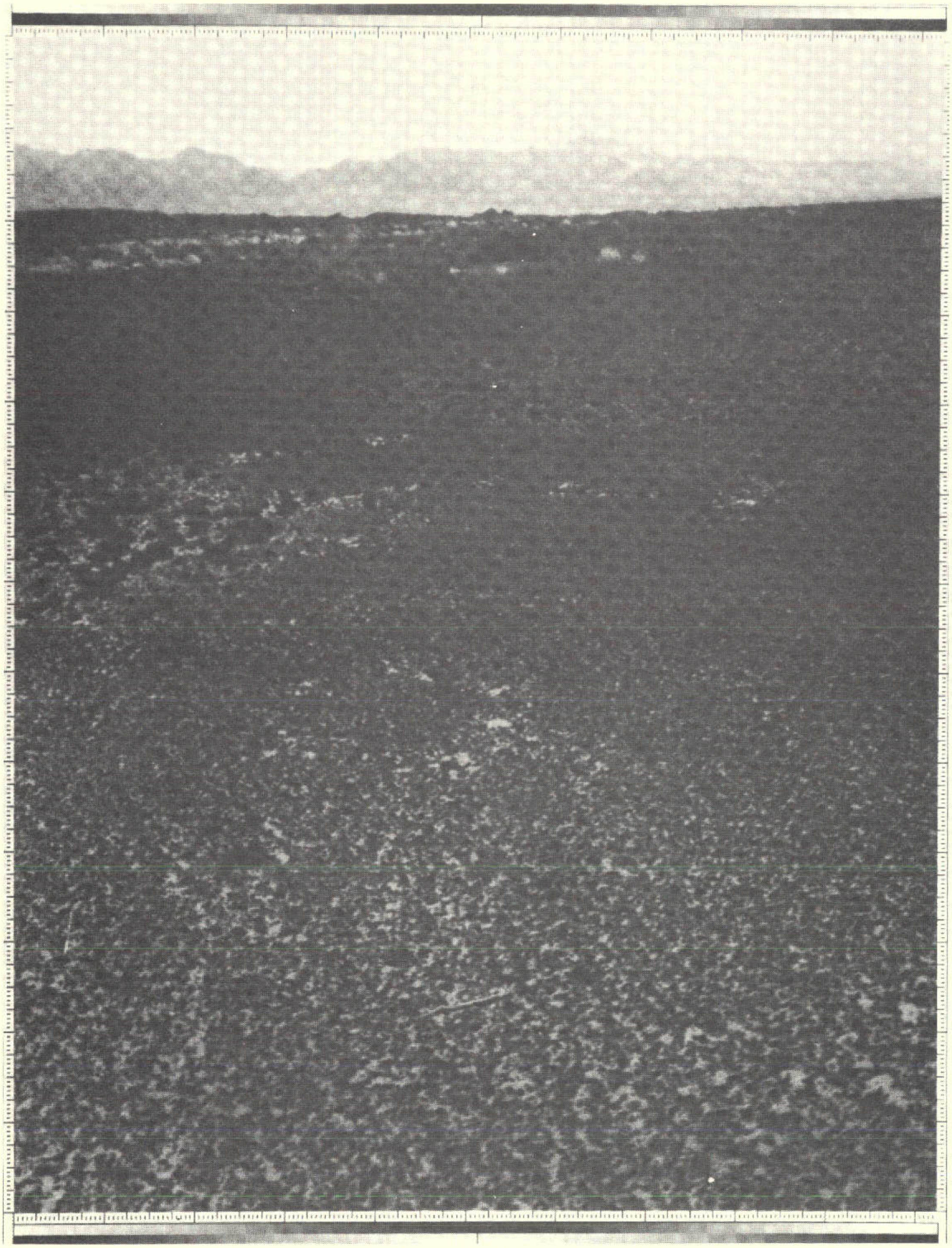


Fig. 6. An outdoor scene depicting the results of lava flow

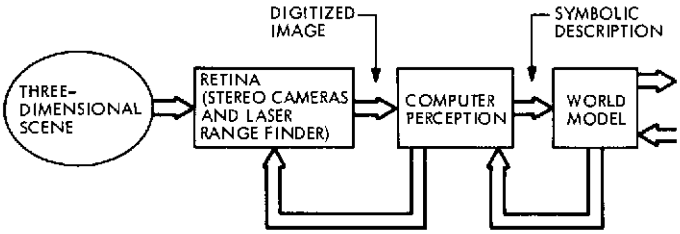


Fig. 7. Information flow for the process of scene analysis

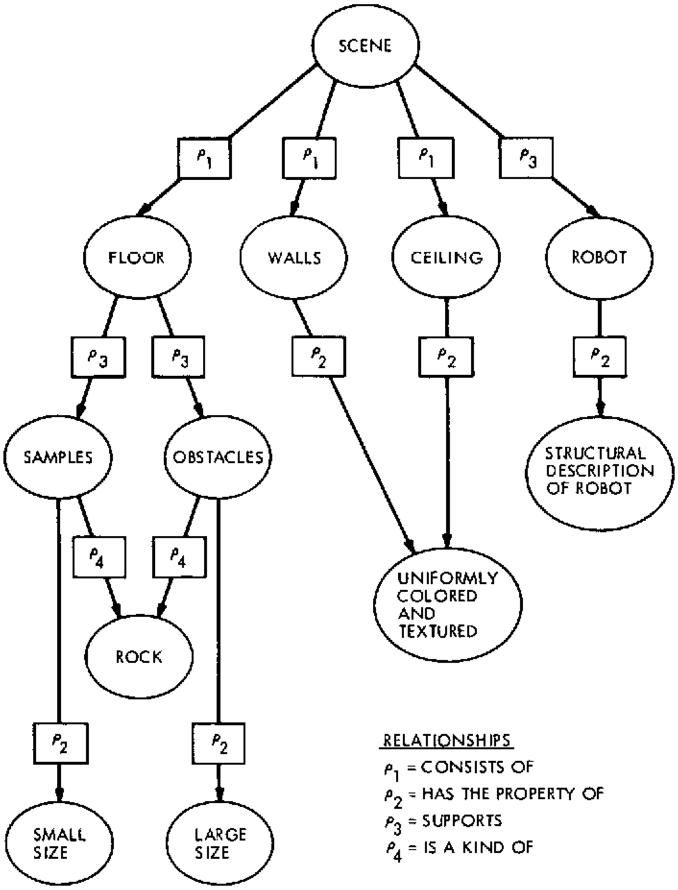


Fig. 8. The model of assertions for a typical indoor scene

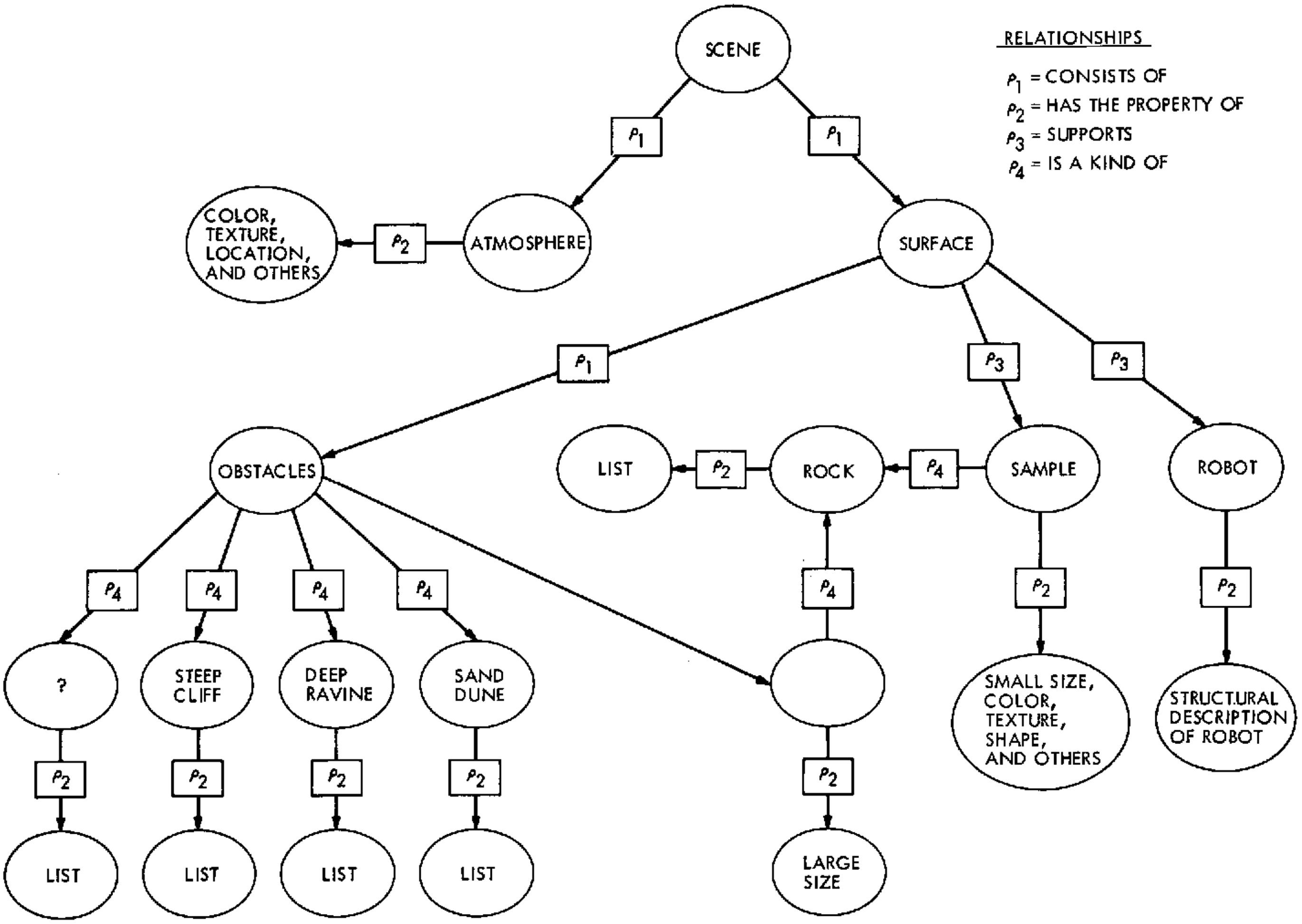


Fig. 9. The model of assertions for an outdoor scene

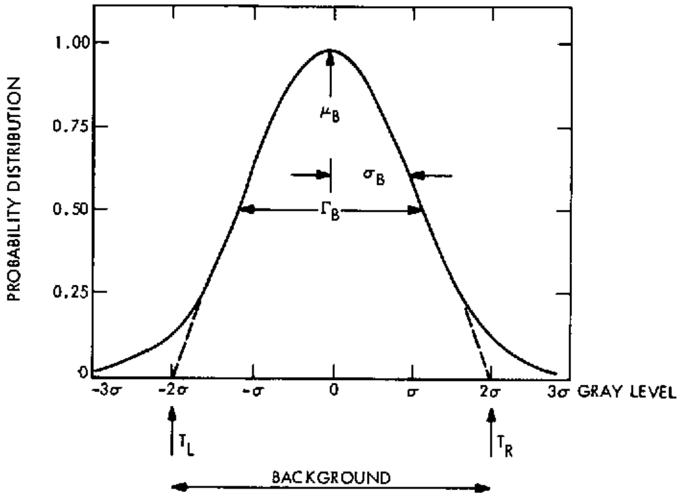


Fig. 10. The local histogram for the background information

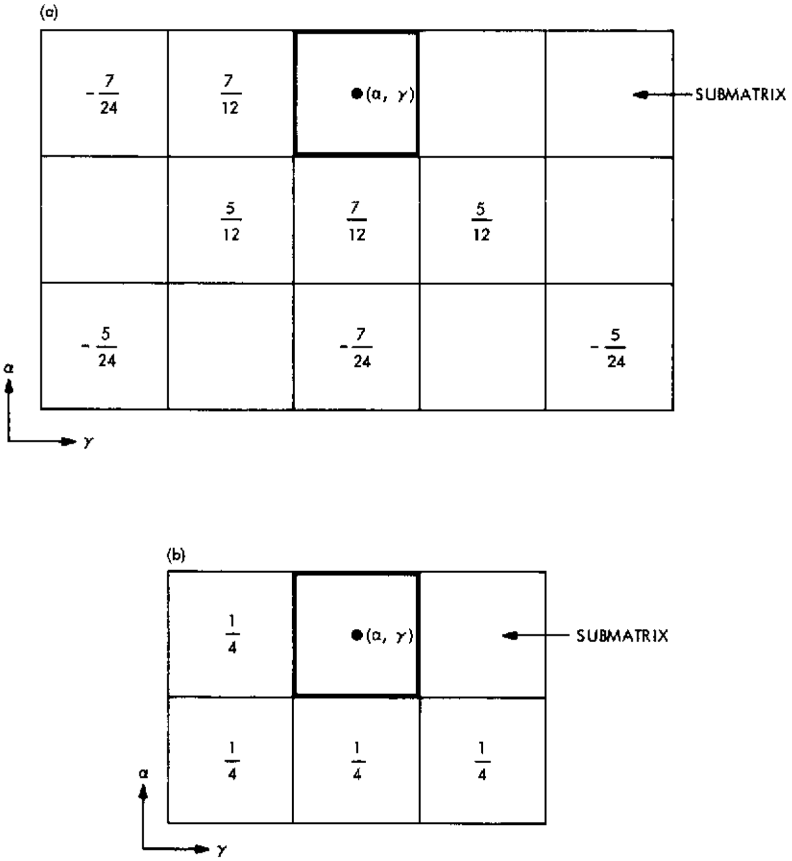


Fig. 11. Matched filters: (a) matched filter for the interpolated estimate of u_B; (b) matched filter for obtaining the thresholds T_L and T_R when a background peak is not found by search

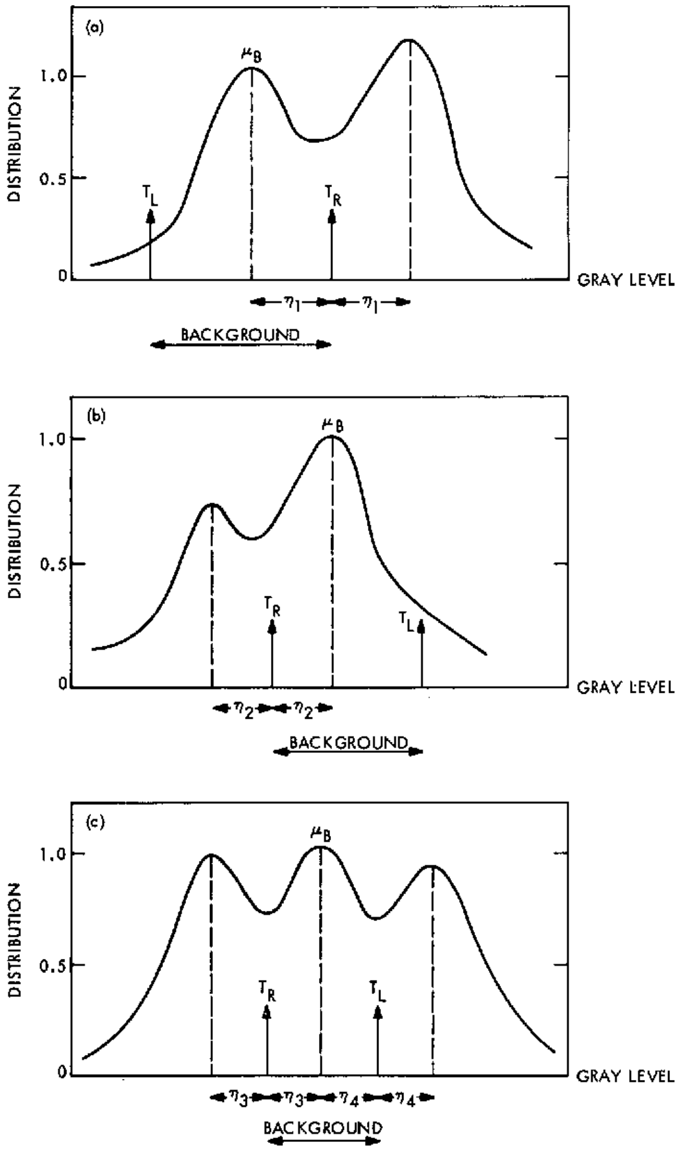


Fig. 12. Special cases where the two-sigma assumption does not hold completely

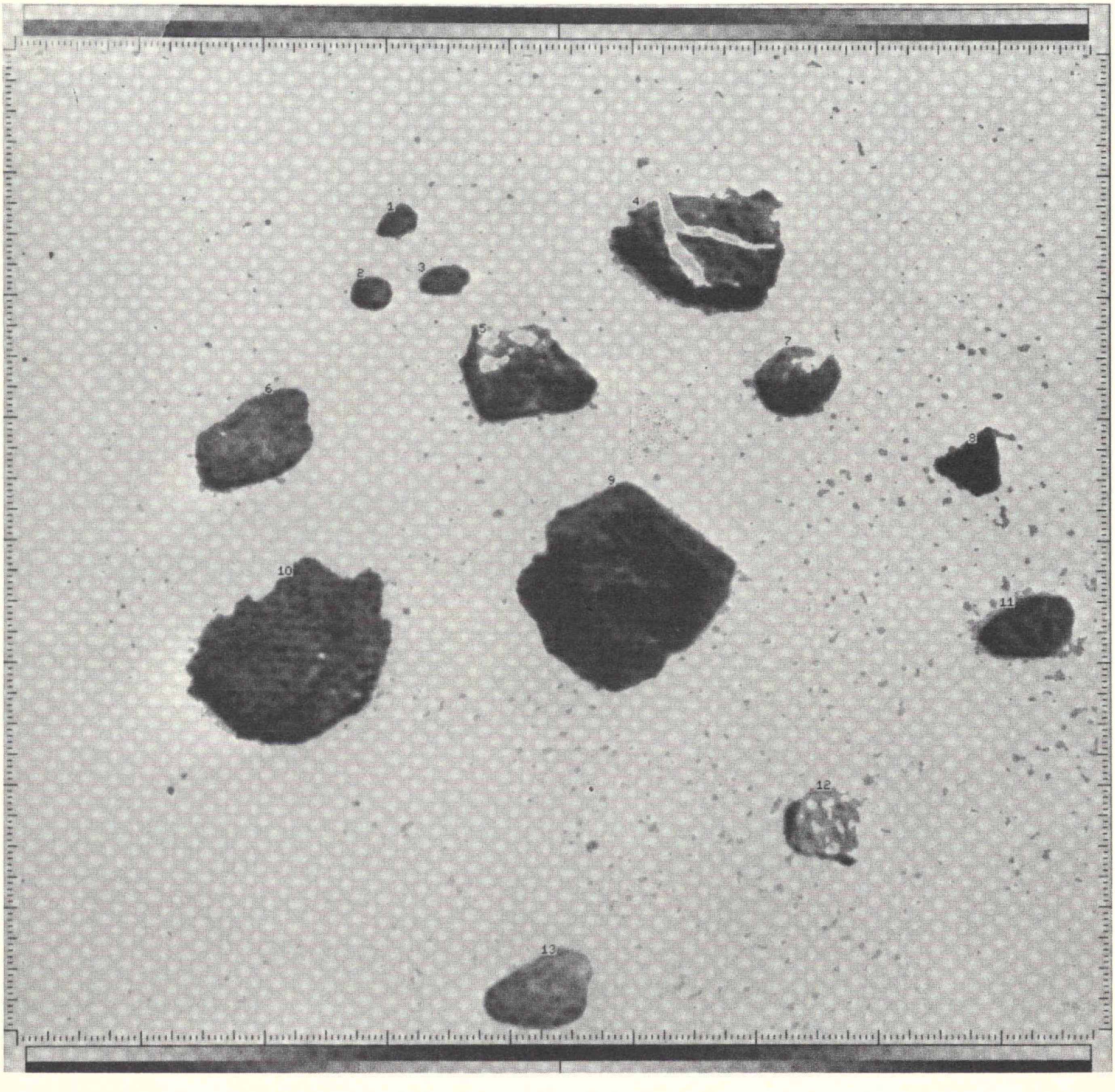


Fig. 13. The sidewalk scene after the segmentation process using the two-sigma assumption

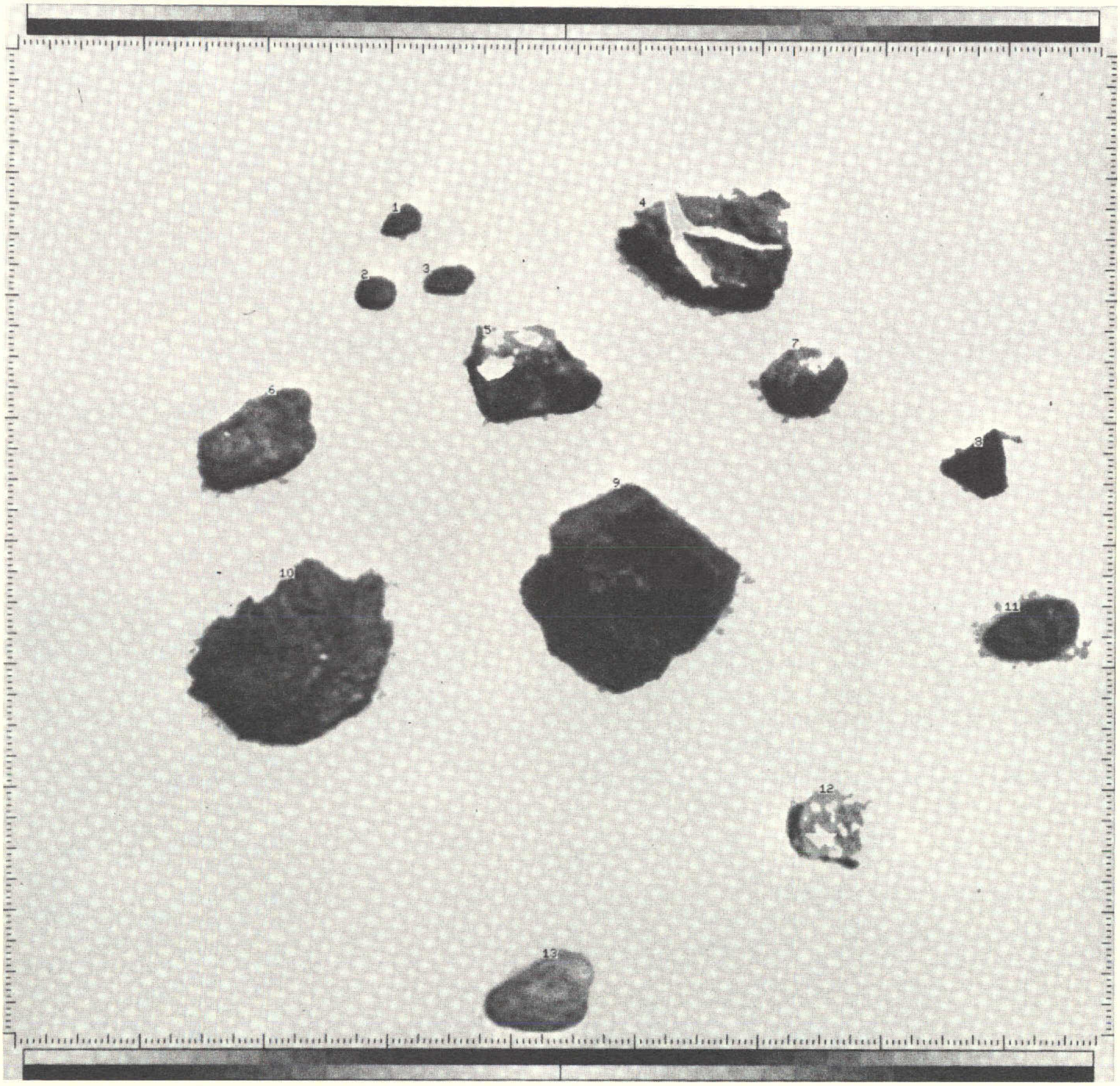


Fig. 14. The segmented sidewalk scene after the erasure of small objects

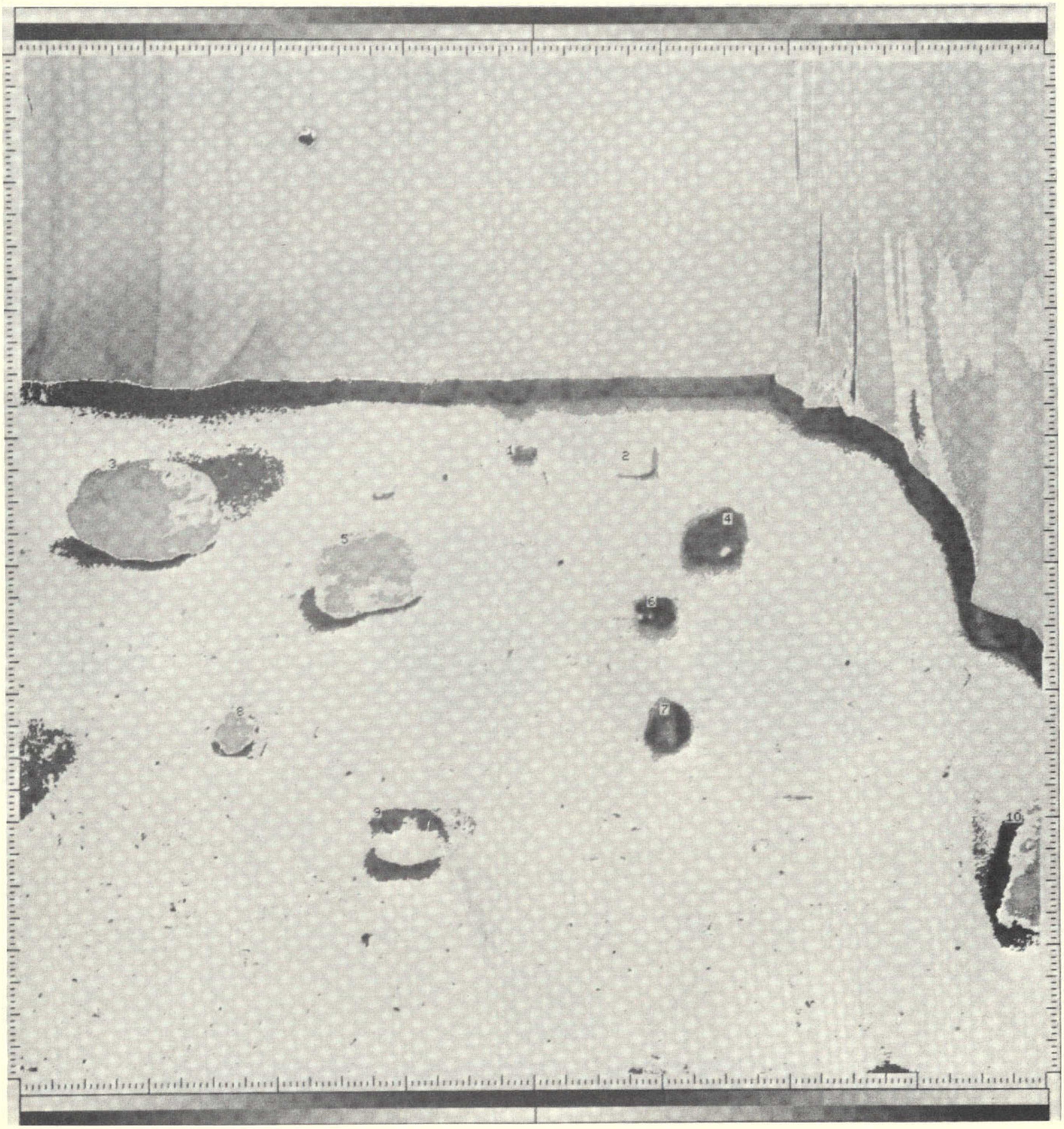


Fig. 15. The indoor laboratory scene after the segmentation process using the two-sigma assumption

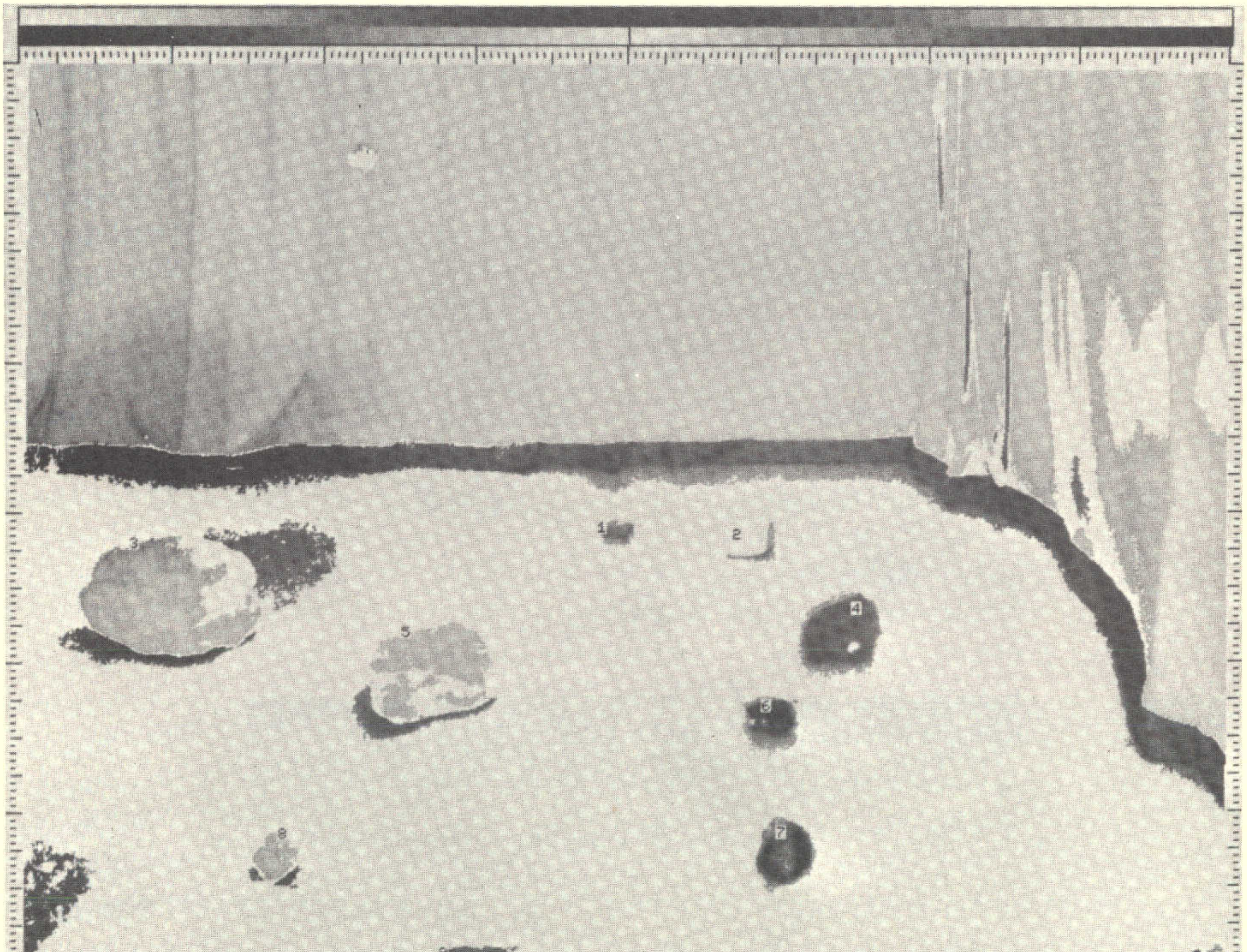
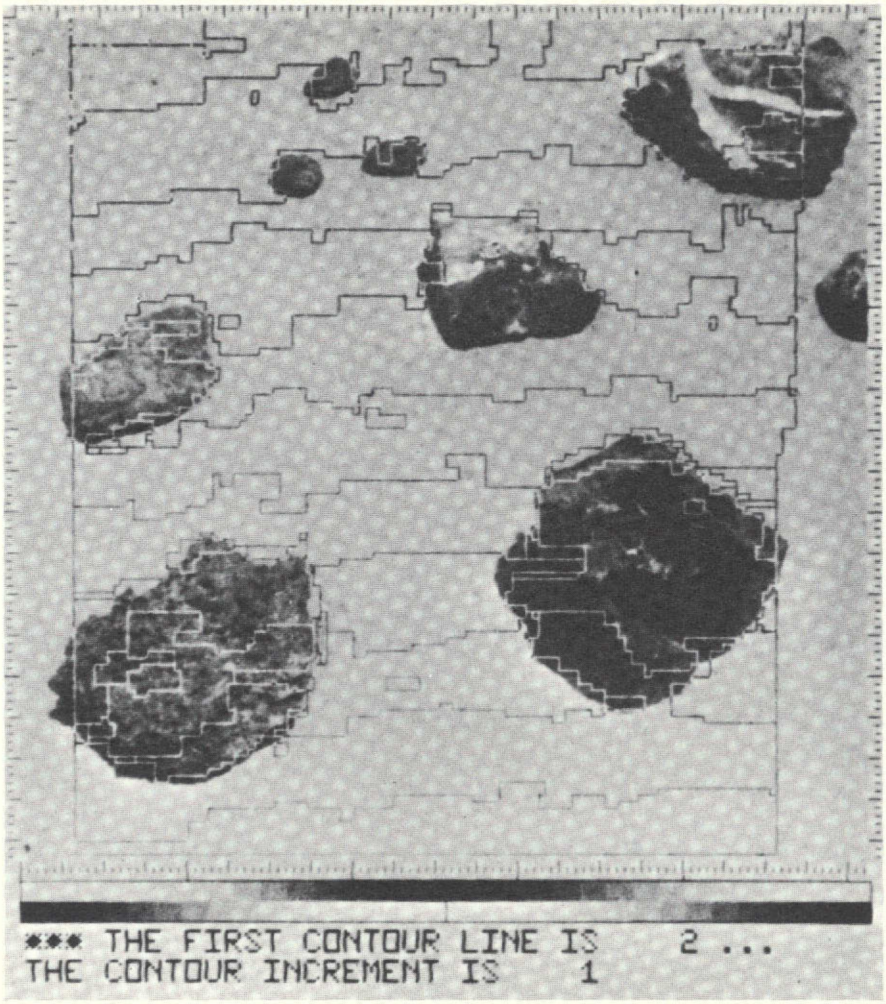


Fig. 16. The segmented indoor laboratory scene after the erasure of small objects



```
*** THE FIRST CONTOUR LINE IS    2 ...
THE CONTOUR INCREMENT IS    1
```

Fig. 17. A depth map for the sidewalk scene

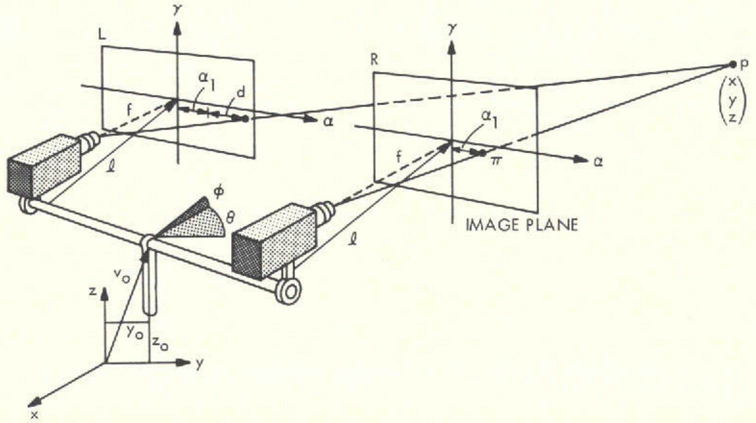


Fig. 18. Global and local coordinate systems for the picture acquisition process by the so-called retina

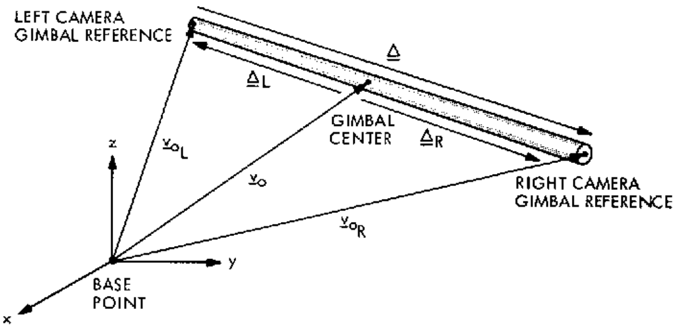


Fig. 19. Gimbal references

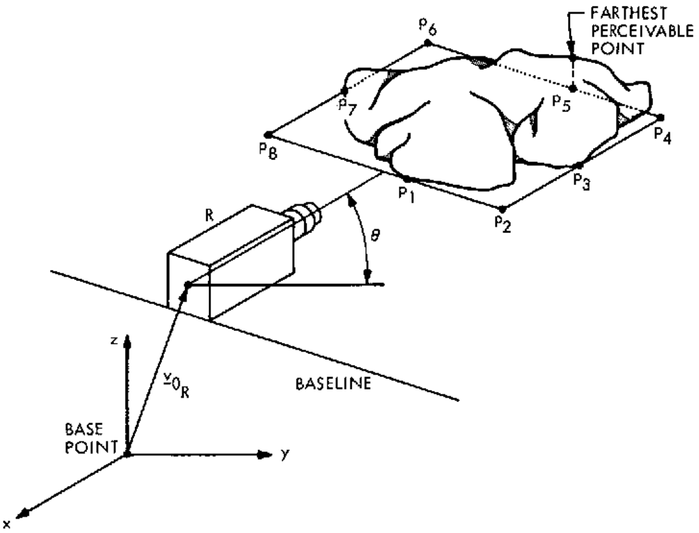


Fig. 20. Points in the global coordinate system that define the MPERP

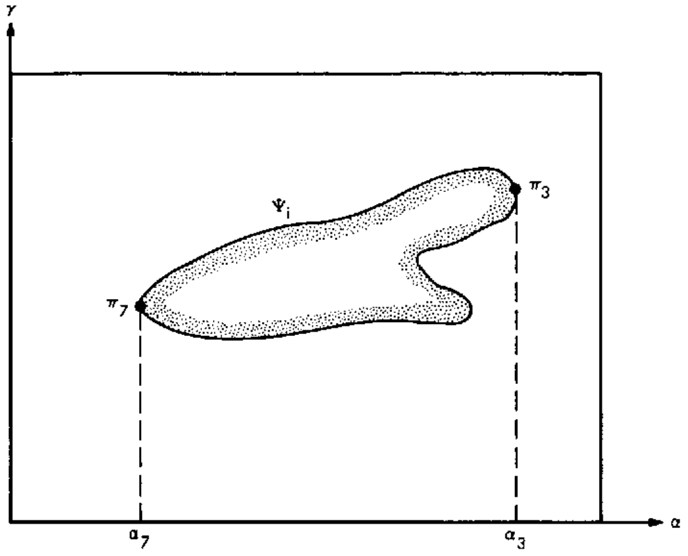


Fig. 21. The points π_3 and π_7 for a particular rock Ψ_i represented in the picture coordinate system

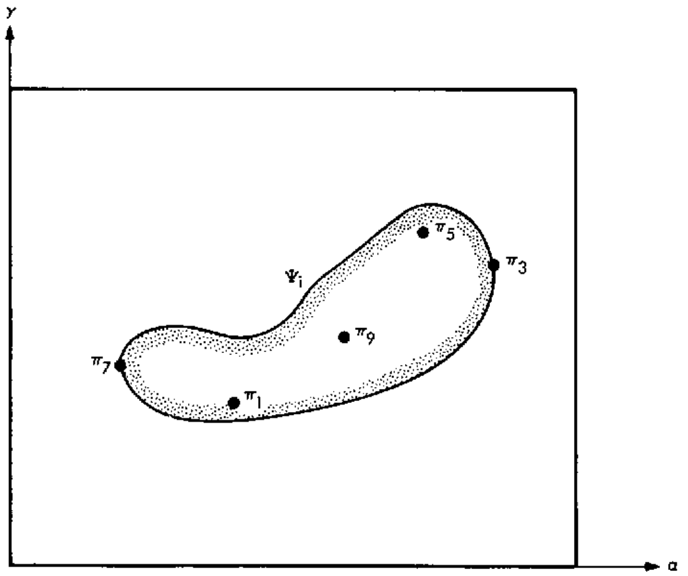


Fig. 22. A possible placement in the picture coordinate system of the five points used to construct the MPERP

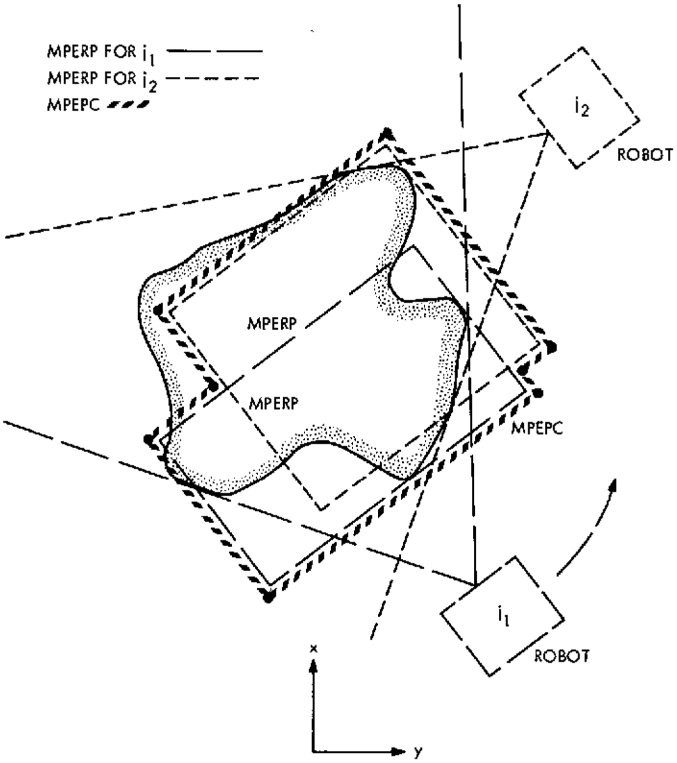


Fig. 23. The surface cross section of a rock in the x-y plane with the MPERP evaluated at two observation points and the resulting MPEPC

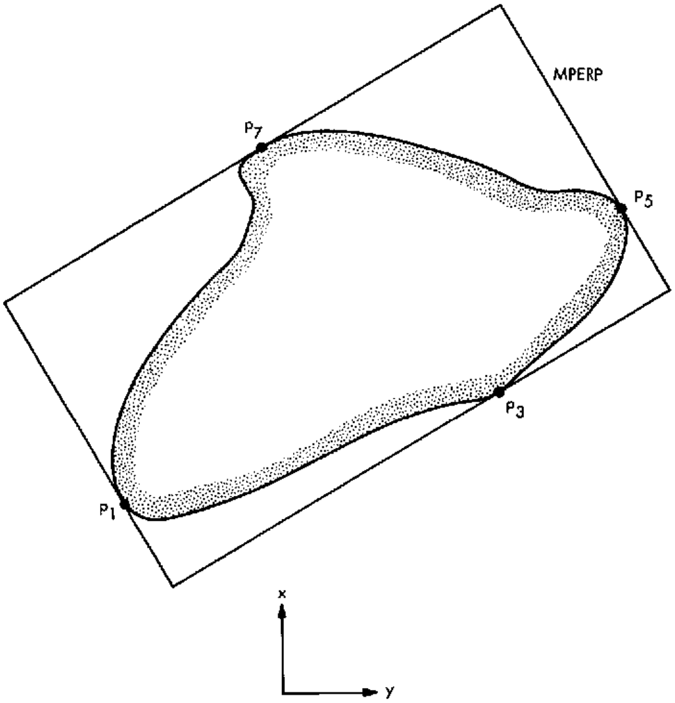


Fig. 24. The surface cross section (in the x-y plane) and the desirable MPERP for the purpose of manipulation

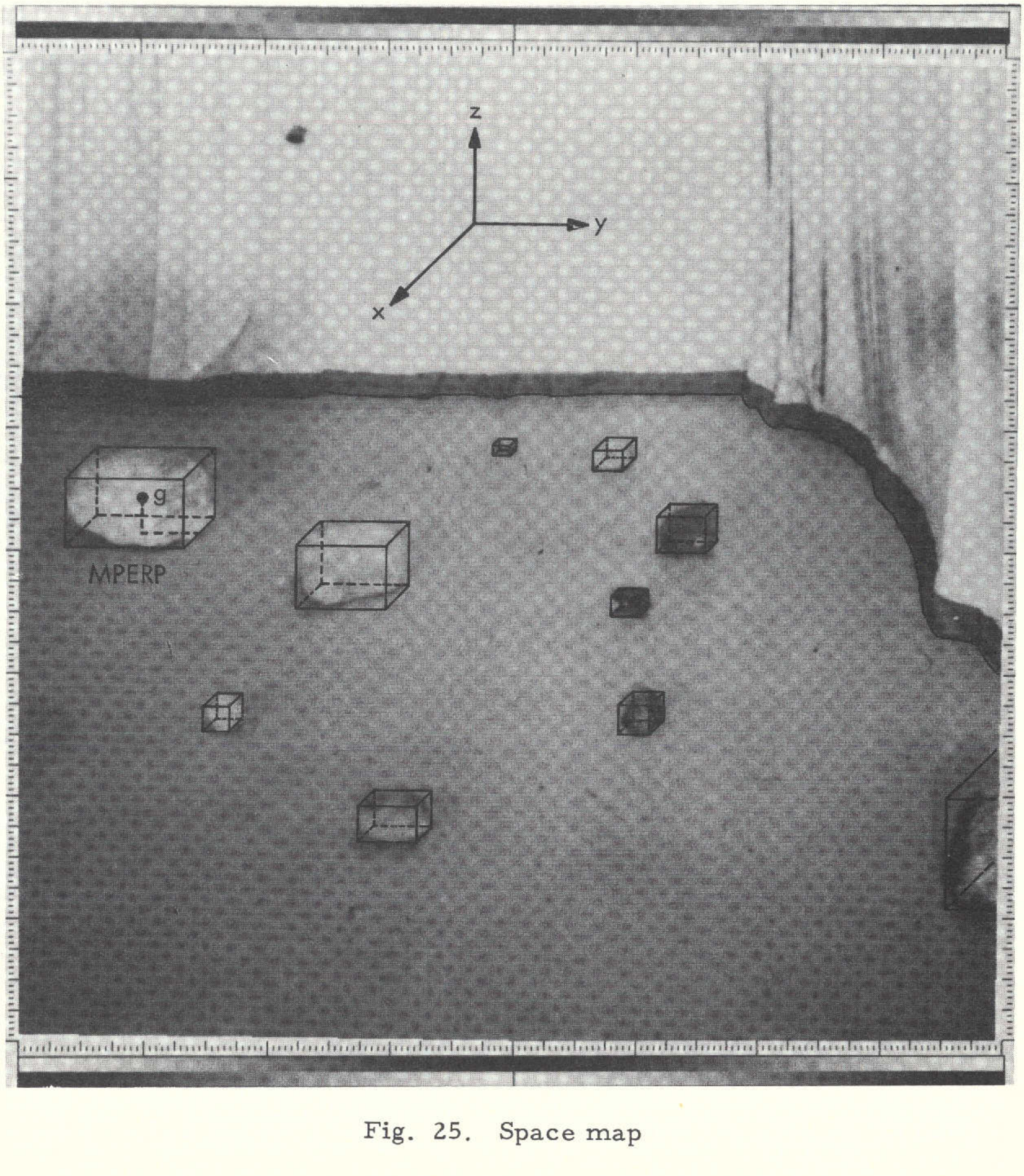


Fig. 25. Space map

NASA — JPL — Coml., L.A., Calif.