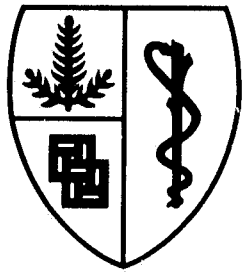
(NASA-CR-168731) MOBILITY AID FOR THE BLIND N82-21894
Final Report, 1 Dec. 1980 - 30 Nov. 1981
(Stanford Univ.) 40 p HC A03/MF A01
CSCL 05H Unclas
G3/54 09430

BIOMEDICAL TECHNOLOGY TRANSFER

Applications of NASA Science and Technology



Submitted by
STANFORD UNIVERSITY SCHOOL OF MEDICINE
CARDIOLOGY DIVISION

Prepared for
National Aeronautics and Space Administration
Technology Utilization Division
Washington, D.C. 20546

CARDIOLOGY DIVISION
Biomedical Technology Transfer

STANFORD UNIVERSITY SCHOOL OF MEDICINE

Final Report

MOBILITY AID FOR THE BLIND

December 1, 1980 - November 30, 1981

Prepared for Henry Lum, Ph.D.
Chief, Project Technology Branch
NASA-Ames Research Center
Moffett Field, California 94035

Under Cooperative Agreement NCC 2-113

Stanford University Biomedical Applications Team
Biomedical Technology Transfer Program
730 Welch Road, Suite 214
Palo Alto, California 94304

January 1982

Preface

This final report is being submitted by the Stanford University Biomedical Applications Team (SUBAT) under NASA Cooperative Agreement NCC 2-113. It presents an account of the activities and results of SUBAT's Mobility Aid for the Blind technology transfer project. The report has been prepared for NASA's Technical Monitor Henry Zum, Ph.D., Chief of the Project Technology Branch at NASA-Ames Research Center.

Although Cooperative Agreement NCC 2-113 defines its period of performance as 1 December 1980 through 30 November 1981, SUBAT did not receive its Notification of Award from NASA until 31 January 1981. The report, however, covers the entire period of performance and includes work done by SUBAT in December 1980 and January 1981 in anticipation of award.

The work reported herein was done under the general supervision of SUBAT Executive Director Donald C. Harrison, M.D. by consultants J. M. Tenebaum and Robert J. Debs, collaborators Michael Deering, Carter Collins, and Tim Healy, SUBAT Director Gary L. Steinman, and SUBAT Secretary Margo Bellamy. The programming assistance of Lynn Quam, the secretarial assistance of Marion Hazen, and the generousity of SRI International in providing use of their facilities are greatfully acknowledged.

Introduction

Purpose of report. During 1981, the Stanford University Biomedical Applications Team (SUBAT) collaborated with the Smith-Kettlewell Institute for the Visual Sciences (SKIVS) and the Project Technology Branch of NASA-Ames Research Center (ARC) to carry out the first stage of a technology transfer project entitled Mobility Aid for the Blind. This report presents SUBAT's final accounting of the activities and results of this collaborative effort.

Overview. SKIVS has been engaged for several years in a project to develop an effective mobility aid for blind pedestrians (i) which acquires consecutive images of the scenes before a moving pedestrian, (ii) which locates and identifies the pedestrian's path and potential obstacles in the path, (iii) which presents path and obstacle information to the pedestrian, and (iv) which operates in real-time. The mobility aid has three principal components: an image acquisition system, an image interpretation system, and an information presentation system. The image acquisition system consists of a miniature, solid-state TV camera which transforms the scene before the blind pedestrian into an image which can be received by the image interpretation system. The image interpretation system is implemented on a microprocessor which has been programmed to execute real-time feature extraction and scene analysis algorithms for locating and identifying the pedestrian's path and potential obstacles. Identity and location information is presented to the pedestrian by means of tactile coding and machine-generated speech.

Objectives. The objective of the technology transfer project which is imbedded in SKIVS' mobility aid program is to develop, implement, and transfer the required feature extraction and scene analysis software. The ultimate goal of the project is to overcome limitations in the capacity of SKIVS' most recent prototype mobility aid to "understand" typical urban sidewalk scenes in real-time. The present study has been undertaken to determine whether this goal can be achieved with existing image interpretation algorithms and, if so, what must be done to implement such algorithms for real-time execution on the mobility aid.

Changes in scope of study. A sequence of events beyond SUBAT's control has forced SUBAT to modify its approach to achieving study goals, to reduce the scope of the study, and to extend the study schedule beyond submission of this final report. The three most serious events were (i) a two-month delay in NASA's issuance of a Notification of Award for this study which reduced SUBAT's period of performance from one year to ten months, (ii) NASA's termination of SUBAT which forced the revision of Mobility Aid for the Blind project plans to accomodate SUBAT's demise, and (iii) NASA's failure to provide agreed upon image digitization and computer processing services which effectively eliminated the possibility of conducting image interpretation experiments. The impact of these events will be dicussed in later sections.

Background

Computer vision. The present image interpretation problem falls within the scope of computer vision, a field of interest in which artificial intelligence techniques are used to endow computers with the capacity to perceive and understand images. In practice, image interpretation schemes can perform with a degree of logical sophistication that cannot be achieved with conventional computer programming techniques. Such performance, however, typically is confined to very strictly contrained image domains for a given image interpretation system.

For the purposes of this report, the process of image interpretation can be viewed as involving a model of an image domain and two basic operations: feature extraction and semantic analysis. Feature extraction consists of recovering certain fundamental characteristics of an image or sequence of images including, for example, edge locations and orientations, surface textures and colors, light intensity levels, and contrast. The delineation of such characteristics constitute a computer's description of an image which must be interpreted. An image domain model is a detailed dynamic description of the visual environment of interest. It includes as much information about the possible status of the visual environment as might be of interest to the user, and it abstracts this information in terms of the characteristics which might be recovered by feature extraction from images of the environment. Semantic analysis is the process by which the characteristics of a given image or image sequence are related to the image domain model to produce an interpretation about the status of the visual environment.

Computer vision in the mobility aid. After several years of research and development work, SRI/VS has produced a computer vision mobility aid for blind pedestrians. The main goal in this work has been to endow the aid with the capacity to locate sidewalk position relative to the user and to warn the user of objects which might block his path or present a hazard. The current system, which is implemented on a 16 bit microprocessor, achieves this goal under very limited conditions. Severe constraints are imposed by the need to operate in real-time. The constraints restrict the image domain, the algorithmic approach to feature extraction and semantic analysis, and the level of mobility aid performance as expressed by obstacle detection and recognition rates for various classes of objects.

The image domain is restricted to clean, sun-lit, mostly shadow-free sidewalks. Certain initial conditions, such as camera orientation and sun position, must be supplied externally. Performance specifications are expressed in probabilistic terms to allow (i) the occassional occurrence of warnings about non-existent objects (false positives) and (ii) the occurrence of detection failures (false negatives), especially in respect to objects whose size falls below the camera's Nyquist sampling rate, objects whose contrast with respect to the background is low, or objects whose images are viewed against a highly textured

background. The need to deal with essentially ideal images and probabilistic performance specifications is typical of state-of-the-art efforts in other real-time applications of computer vision: e.g., industrial parts recognition[1], blood cell counting, and automated navigation.

The mobility aid relies on a crude but fast feature extraction algorithm that operates at one video frame per second to detect and link edges. More sophisticated feature extraction operations cannot be done at the present time. Semantic analysis is amortized over several frames so that predictions from previously analyzed frames can be used to facilitate more rapid interpretation of the current frame. The overall approach is similar to approaches employed by Brooks[2] and Shapiro[3]. The amortization of semantic analysis over multiple frames, in particular, is an approach which has been used by Williams[4], Tsuji[5], Roach[6], Nevatia[7], and Lavin[8].

The SRIVS mobility aid can guide a blind pedestrian successfully down the center of an urban sidewalk under these constraints. If any of the constraints are violated, however, the system usually fails. Small objects are never detected, for example, and shadows usually confuse the system. The constraints reflect ideal but unrealistic circumstances for mobility aid use. Practical use requires the development of feature extraction and semantic analysis software that can operate under more realistic conditions.

Overcoming limitations. The limitations of SRIVS' mobility aid and most other real-time computer vision systems arise from the need to carry out feature extraction operations at video frame rates on cost- and size-limited computational resources. This requirement currently forces such systems to rely on crude feature extraction algorithms which permit only simple edge detection and linking and on equally simplified descriptions of image domains. Reliable recognition, on the other hand, requires much more sophisticated initial feature extraction as well as richer descriptions of image domains. For example, if distance from viewer, three-dimensional shape, relative orientation, surface texture, and color were captured by the mobility aid, it would be possible to distinguish shadows or flat objects on a sidewalk from real obstacles and to recognize three-dimensional objects regardless of viewpoint. In view of the fact that VLSI technology is revolutionizing the capabilities of small, inexpensive computers, it is appropriate to consider incorporating more elaborate software in the mobility aid so as to achieve more practical operational capabilities.

A variety of powerful computer vision techniques are potentially applicable to the mobility aid. For example, Barrow and Tenenbaum[9] review techniques for recovering surface descriptions based on stereo correspondance, optic flow, texture gradient, and other basic cues. Other potentially applicable techniques include (i) surface shape recovery from photometric shading[10], texture[11,12], stereopsis[13,14], and motion flow[7,15];

(ii) detection of long, straight lines against complex backgrounds by means of the Hough transform; and (iii) image partitioning by means of the recursive top-down algorithm. Many of these techniques were developed in the context of NASA-sponsored robotics and remote sensing research.

Approach

Initial approach. In general terms, the initial approach to the first stage of this technology transfer project consisted of (i) reviewing existing image interpretation techniques, (ii) implementing potentially applicable techniques on a suitable computer system, (iii) developing an urban sidewalk image database, (iv) conducting image interpretation experiments to determine which techniques produce the best results on the image database, and (v) thereby identifying the image interpretation resources and needs in connection with later stages of the project.

Unfortunately, previously described events produced unanticipated time and resource constraints which effectively precluded carrying out any experiments. It became necessary to make mid-project revisions of approach and to stretch resources by soliciting services and materials from concerned individuals and institutions.

Revised approach. A new approach was devised which takes advantage of an almost universal desire among computer vision experts to extend the application of their technology to new visual environments and to obtain a consistent basis for comparing image interpretation techniques. This approach involves (i) compilation, digitization, and distribution of an urban sidewalk image database to a variety of computer vision experts, (ii) conduct of voluntary image interpretation experiments on the database with each expert using his own facilities and software, and (iii) assessment of the experimental results by the Mobility Aid for the Blind project team to determine (a) the extent to which available techniques can overcome present mobility aid limitations and (b) which limitations of the mobility aid require additional research to surmount.

Interest among computer vision experts in the Mobility Aid for the Blind project has been stimulated by two incentives. Firstly, the image database has been compiled in such a way that it can serve as a universal standard for objective comparison of alternative image interpretation algorithms. The field of computer vision is in need of such a standard, and this project's database will be of lasting benefit to the research community in that role. Secondly, modest honoraria were offered for the most effective techniques that are revealed by voluntary experimentation.

Methods and Results

Overview. The essential accomplishment of the project team to date has been the compilation, digitization, and distribution of an urban sidewalk image database which many of the foremost researchers in computer vision have agreed to process in their own laboratories. A retrospective examination of this database revealed a number of photometric quality problems which are being corrected at the present time. Arrangements have been made to complete the present study, and a voluntary follow-up report will be prepared under the supervision of J. M. Tenenbaum. The report will convey an analysis of the results of experiments conducted on the project's image database by the various researchers who have agreed to participate in the project.

Testing protocols, performance specifications, and scene variables. The prototype mobility aid was tested under a protocol which was defined in connection with the MCP-sponsored mobility aid development project. A summary of the protocol and applicable performance specifications are included as Appendix A to this report.

An analysis of the test results revealed instances of scenes which caused the mobility aid to fail and identified the scene variables which were responsible for the failures: (i) size and type of shadows on the sidewalk, (ii) degree of contrast between adjacent regions of an image, e.g., between sidewalk and street or grass, (iii) the number of objects which clutter the scene, and (iv) the incidence of "stepdowns" or holes in the sidewalk which is being viewed. These variables were incorporated into the design of the image database to provide for an effective assessment of available image interpretation techniques.

Less formal reports of user experience were productive, also, in establishing an approach to the assessment of mobility aid performance. It became apparent (i) that it is necessary to distinguish classes of objects that differ in relative importance to a user's safe and efficient travel and (ii) that obstacle detection and recognition rates must be interpreted in the context of such a classification scheme. Basically, there are several key items such as sidewalk edges, potholes, broken sidewalk slabs, and large obstacles which must be detected. Also, there are secondary items such as small off-path objects and distinguishing features of large objects (e.g., car vs truck vs bolder) whose detection is desireable but not essential. Finally, there are items such as shadows, fallen leaves, and sidewalk stains which must be ignored or suppressed by the mobility aid.

Compilation of database. The image database of representative urban sidewalk scenes was recorded photographically using a 35 mm SLR still camera and a 16 mm movie camera. Each scene was recorded as a stereo pair of 35 mm positive color slides using a single tripod-mounted camera at head height. A single camera was used to obtain sequential exposures of the two halves

of each stereo pair, and a horizontal rail mount was devised along which the camera could be translated between exposures to achieve stereo separation. This approach facilitated use of a single lens for both halves of a stereo exposure and, therefore, minimized variations in lens distortion effects between halves. A stereo separation of 2.5 inches was chosen to simulate the human eye system. Most slides were taken with a sixty degree field of view to simulate normal human peripheral vision, and a few slides were taken with a four degree field of view to simulate human foveal vision. The narrower field images appear as close-ups of texture and sidewalk cracks.

Selected scenes were recorded as 16 mm movies to obtain image sequences of a sort that would be seen by a sighted pedestrian. No attempt was made to produce stereo movies.

Database. Forty stereo slides were taken of Berkley and San Fransisco sidewalk scenes. Twelve of these slides were selected for use in the database, and eleven of these slides are reproduced as photocopies of prints in Appendix B.

The twelve slides were digitized so that they could be distributed in machine readable form. Since NASA-Ames Research Center could not digitize the slides as had been previously agreed, the project team arranged for no-cost digitization to be done at SRI International. Software was written to speed up the digitization process by permitting the 35 mm film strips (unmounted slides) to be wrapped around a drum, digitized en masse, and partitioned into individual frames by means of software manipulation. This technique produces uniformly digitized images and simplifies the registration of stereo pairs. Since there is no universal file format for computer images, a self-descriptive format which is compatible with most vision systems was devised. The scanning procedure and file format are described in Appendix C.

Problems and corrective measures. Although the resulting digitized imagery appeared acceptable on a raster display, its overall quality was inadequate for use as a standard database. Two specific problems became apparent: (i) The dynamic (light intensity) range of the images exceeded the linear range of the films density curve because (a) fast film was used, (b) the film was slightly underexposed, and (c) high contrast development techniques were used. As a result, image detail in regions of shadow and bright illumination was compromised, and the data became unsuitable for interpretation by photometric techniques. (ii) The stereo baseline was too small to produce sufficient range resolution. Given the inherent spatial resolution of the digitized images, for example, it is not possible to detect a sidewalk curb at a distance of ten feet. An analysis is provided in Appendix D.

A new set of slides is being taken to correct these problems in the database. ASA-25 film and custom developing services are being used to obtain better dynamic range and image quality. The

stereo baseline is being extended to shoulder width (18-24 inches) to improve range resolution. Ultimately, the image interpretation software in the mobility aid will be written to accept stereo image pairs with a stereo baseline of 2.5 inches. For the present, however, the extended stereo baseline is required to facilitate the use of existing software for experimental purposes.

Solicitation of project participants. The participation of computer vision researchers in the Mobility Aid for the Blind project was solicited by mail and by informal contact at professional meetings. A review of the computer vision technical literature revealed several papers which described image interpretation techniques, algorithms, and systems of potential relevance to the mobility aid problem. Each author was sent color prints of the database images and a letter of invitation to participate in the present study. The letter and its distribution list are attached as Appendix E to this report. Additional informal contacts were made by the project team at the 1981 International Joint Conference on Artificial Intelligence (Deering) and the NSF Workshop on Human and Machine Perception (Tenenbaum). The response to the project team's solicitation efforts exceeded expectations. A list of the eighteen researchers who have indicated a willingness to participate in the present study is provided in Appendix F. The list includes many of the foremost experts in the field of computer vision. Furthermore, the specific interests of these experts span most of the mobility aid's problem areas.

Presentations and Publications. Michael Deering presented the paper "Real-Time Natural Scene Analysis for a Blind Prosthesis" at the 1981 International Joint Conference on Artificial Intelligence in Seattle, 24-28 August 1981. The paper is included in this report as Appendix G. Also, Jay Tenenbaum delivered the Keynote Address at the NSF Workshop on Human and Machine Perception in Denver, 12-14 August 1981.

Plans. The revised image database will be distributed to the voluntary project participants who will conduct image interpretation experiments. Experimental results will be reported to the Mobility Aid for the Blind Project Team and analyzed. The analysis will be reported to NASA-Ames Research Center (Lum) in a voluntary follow-up report to be prepared by Tenenbaum and, if warranted, in the open literature or at a professional conference. An appropriate agency, probably IEEE, will be selected as permanent repository of the image database. Finally, the project team will develop proposals to obtain funding for later stages of the technology transfer project. Specific plans include proposing (i) continuation of mobility aid development to NSF and ARPA and (ii) VLSI implementation of image interpretation algorithms to NASA. The latter activity will be proposed in the context of robotics applications for orbiting space platforms.

Conclusions

The first stage of the Mobility Aid for the Blind technology

transfer project has not been completed as planned due to a sequence of events beyond SUBAT's control. However, the project team has provided for the eventual completion of proposed work and for the submission of a follow-up report once the work has been completed. Also, the project team is prepared to draw several conclusion from the work accomplished to date:

1. The Mobility Aid for the Blind Mobility project is perceived as being worthwhile by the computer vision research community. Many computer vision experts have agreed to participate in the project, and feedback concerning the potential use of the urban sidewalk database as a standard testbed for computer vision technology has been positive.

2. It is anticipated that the twelve images in the urban sidewalk database will exert substantial leverage on the further development of computer vision and that the database will be of lasting benefit to the computer vision research community. The wide distribution of the "industrial parts bin" image database several years ago had such an impact. The urban sidewalk database is more complex and, for the first time, permits both (i) comparitive assessment of a variety of algorithms and approaches on a single database and (ii) the assessment of simultaneously applied complimentary approaches and integrated approaches on a complex standard database.

3. The methodology of distributing images rather than distributing algorithms should provide many of the scientific results that are supposed to be produced by the more ambitious ARPA Image Understanding Testbed Project (SRI) at a fraction of that project's cost.

4. The use of the mobility aid database as a standard database will focus the research community's interest on the mobility aid problem for years to come. The results of such focused efforts should be a substantial improvement in the performance, cost, and design of mobility aids and other vision aids for the blind.

5. It is anticipated that no one algorithm will prove to be superior in image interpretation performance on urban sidewalk scenes. Different algorithms will be more successful at recognizing different classes of obstacles in the blind pedestrian's path. Ultimately, the mobility aid will incorporate several different algorithms in order to achieve adequate performance over a wide class of obstacles and conditions in the visual environment. The present task is to ascertain which existing algorithms and techniques should be incorporated into the mobility aid and and what new approaches must be developed and incorporated.

References

1. Perkins, W.: A model based vision system for industrial parts. IEEE Trans. Comput., 27:126-43, 1978.

2. Brooks, R.; Greiner, R.; Binford, T.: The ACRONYM model-based vision system. Proc. IJCAI-79, Tokyo, Japan, August 1979, 105-13.

3. Shapiro, L.; Moriarty, J.; Mulgaonkar, P.; Haralick, P.: Sticks, plates, and blobs: A three-dimensional object representation for scene analysis. First Annual National Conference on Artificial Intelligence, August 1980, 28-30.

4. Williams, T.: Depth from camera motion in a real world scene. IEEE Trans Pattern Anal and Mach Intell 2(6):511-6, November 1980.

5. Tsuji, S.; Osada, M.; Yachida, M.: Tracking and segmentation of moving objects in dynamic line images. IEEE Trans Pattern Analysis and Machine Intelligence 2(6):516-22, November 1980.

6. Roach, J.; Aggarwal, J.: Determining the three-dimensional motion and model of objects from a sequence of images. University of Texas at Austin Laboratory for Image and Signal Analysis Research Report TR-80-2, June 1980.

7. Nevatia, R.: Depth measurement by motion stereo. Comp Graphics and Image Processing 5:203-14, 1976.

8. Lavin, M.: Analysis of scenes from a moving viewpoint. In Artificial Intelligence: An MIT Perspective (P. Winston and R. Brown, eds.), 187-208.

9. Barrow, H.G.; Tenenbaum, J.M.: Computational vision. Proc IEEE 69(5):572-95, May 1981.

10. Marr, D.: Visual information processing: The structure and creation of visual representations. Proc IJCAI-79, Tokyo, Japan, August 1979, 1108-26.

11. Witkin, A.: A statistical technique for recovering surface orientation from texture in natural imagery. First Annual Conference on Artificial Intelligence, August 1980, 1-3.

12. Kender, J.: Shape from texture: An aggregation transformation that maps a class of textures into surface orientations. Proc IJCAI-79, Tokyo, Japan, August 1979, 475-80.

13. Gennery, D.: A stereo vision system for an automated vehicle. Proc IJCAI-77, Cambridge, MA, August 1977, 576-82.

14. Moravec, H.: Visual mapping by a robot rover. Proc IJCAI-79, Tokyo, Japan, August 1979, 598-600.

15. Nakayama, K.: Geometric and physiological aspects of depth perception. SPIE Vol. 120, Three-Dimensional Imaging, 1977.

APPENDIX A

EXCERPT OF 7 AUGUST 1980 LETTER FROM CARTER C. COLLINS TO GARY L. STEINMAN DESCRIBING MOBILITY AID TESTING PROTOCOL AND PERFORMANCE SPECIFICATIONS

In reference to the NASA proporal, we would tentatively formulate performance specifications for the electronic mobility aid visual processing algorithms as follows:

Class I: Detection of large, salient features in the environment such as poles (as indicated in our original grant request to NSF) and, in general, those obstacles relatively easy for the algorithms to detect. These will include features over 1' in height or depth including boxes, low walls, parked cars, bushes, etc. We would intend to supply sample photographs of sidewalk scenes with such items clearly labelled. Class I objects should be detectable at the 98% correct level.

Class II objects would include such items as the edge of the sidewalk or path, edges of a lawn or wall, crosswalk markings and other irregular path delineators. Detection and recognition accuracy for Class II objects could be 95% or better.

Class III objects will include items such as curbs (step down), broken sidewalk slabs, rocks, small objects, litter, and large chuck holes. In general Class III objects will be 4 inches or larger in vertical extent. The detection and recognition accuracy for Class III obstacles can vary over a range from 90% for those obstacles which are at present most difficult to virtually 98% for those which we can easily handle with our present algorithms. With further analysis we will be able to tighten the specifications and break them into subcategories, each with its own quoted accuracy of detection.

Class IV objects should include items down to 1/2" in height, such as a fractured sidewalk, broken tiles, litter and other small items on the sidewalk. The major use of Class IV is to reject items such as leaves and shadows which would not constitute an impediment or hazard to the blind pedestrian.

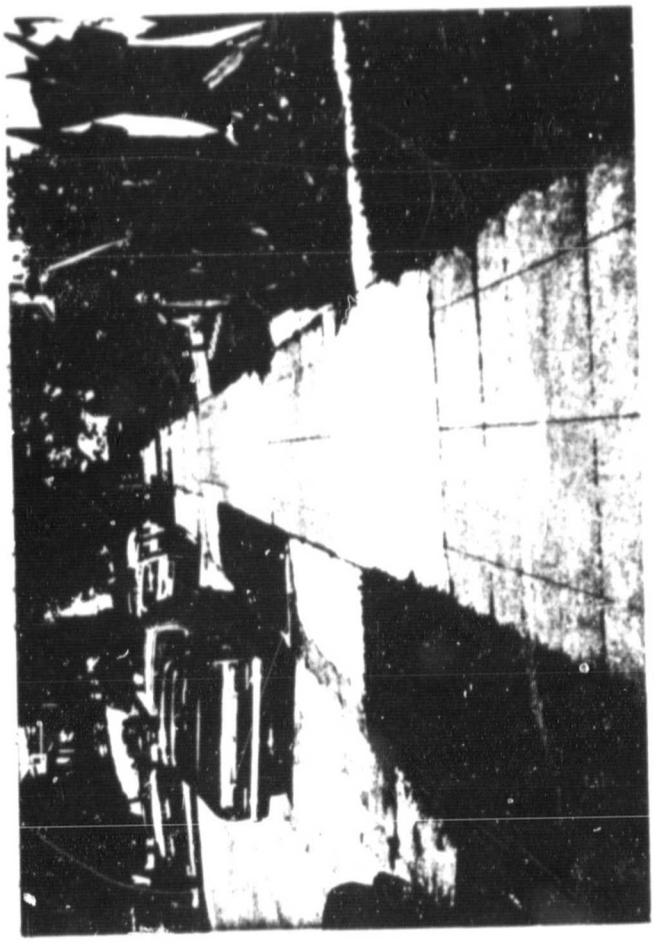
Our current microprocessor-based system should be able to handle the first 3 categories at the stated accuracies under ideal conditions which perhaps obtain some 70% of the time. However, the new algorithms which we seek to utilize or develop under NASA funding should handle all 4 classes under 90 to 95% of all lighting and weather situations (except extreme darkness). Such all-weather performance will be required in order to achieve acceptance by the blind community.

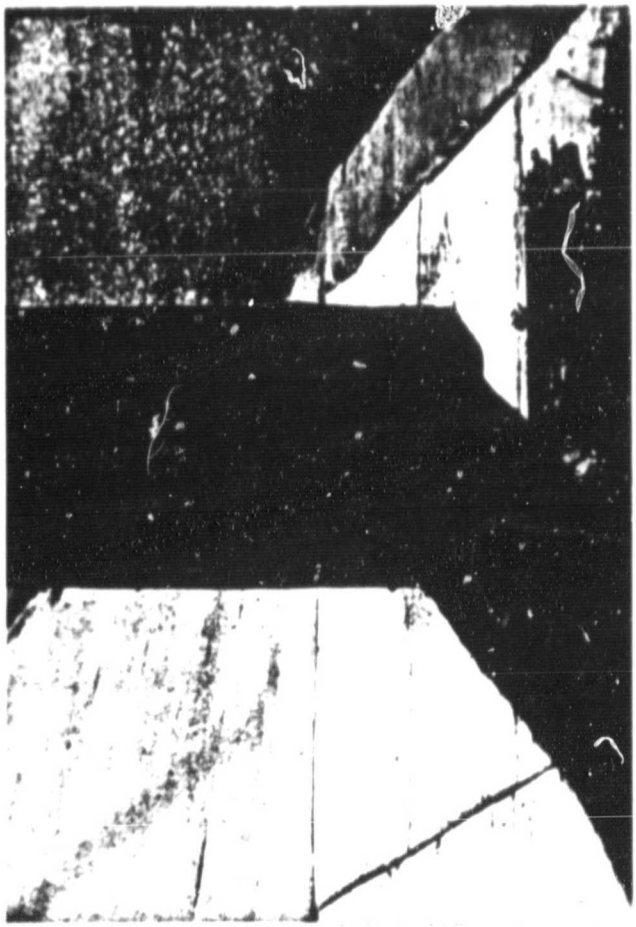
For our currently NSF funded project (Phase I), we plan to evaluate the apparatus with blind users. An investigator, utilizing a visual scoring rechnique, will tabulate the classes of obstacles encountered by button presses which will be recorded on a portable tape recorder. Results will be summarized and played back through a microcomputer to collate and analyze the performance and statistical data in all desired categories.

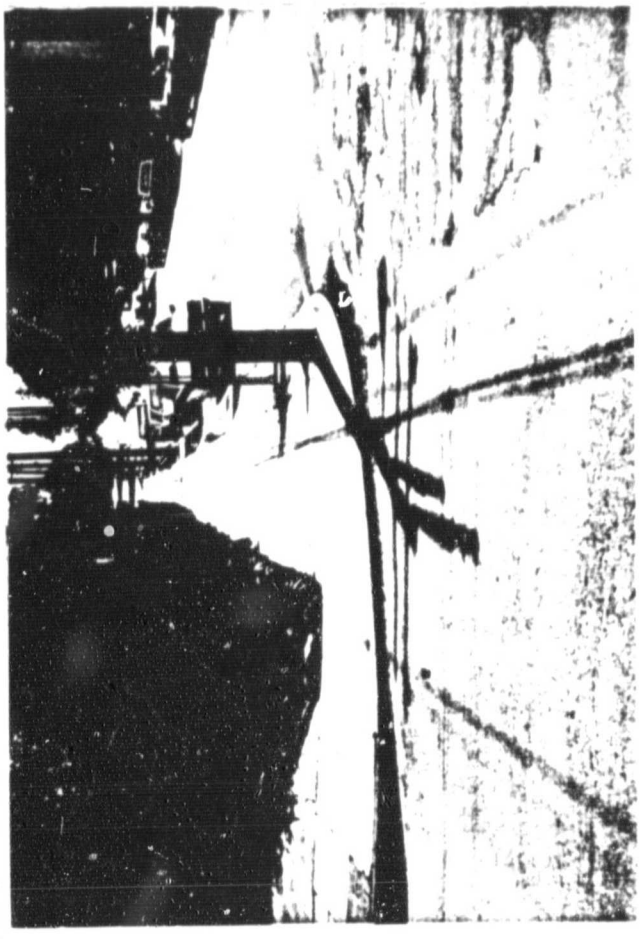
APPENDIX B

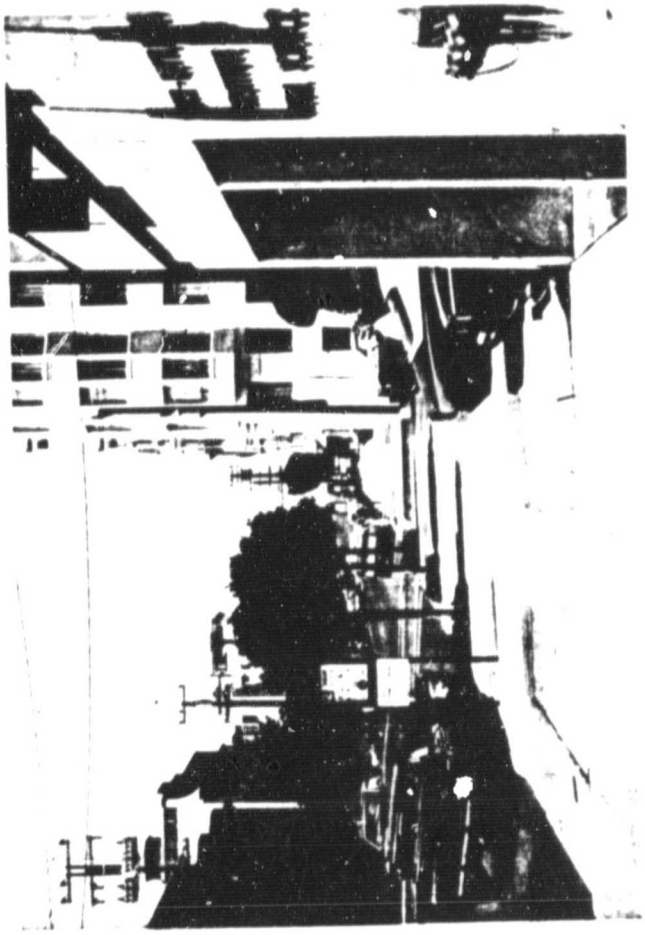
PHOTOCOPIES OF

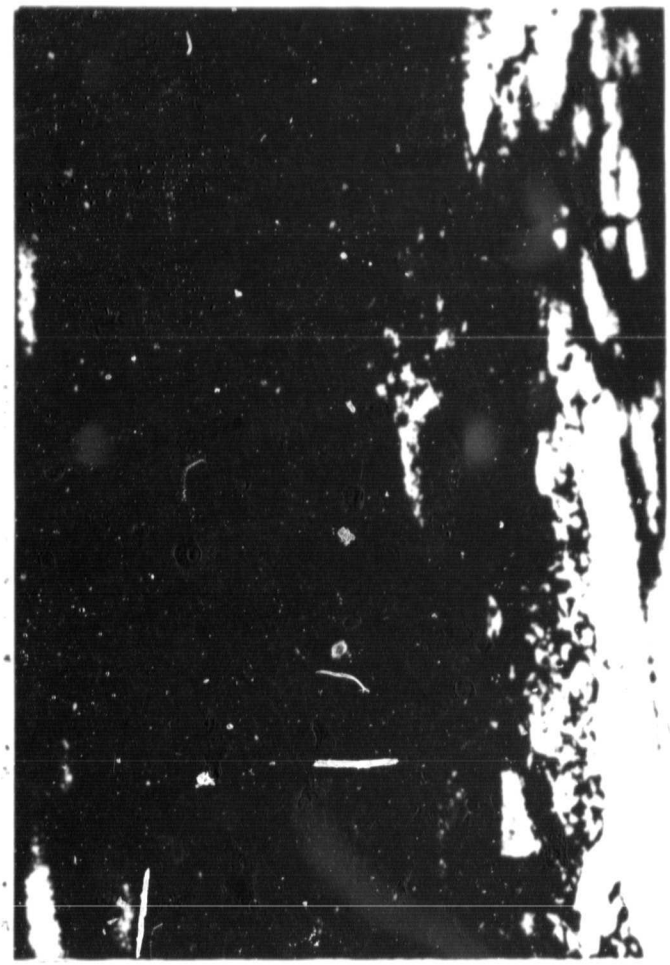
COLOR IMAGES

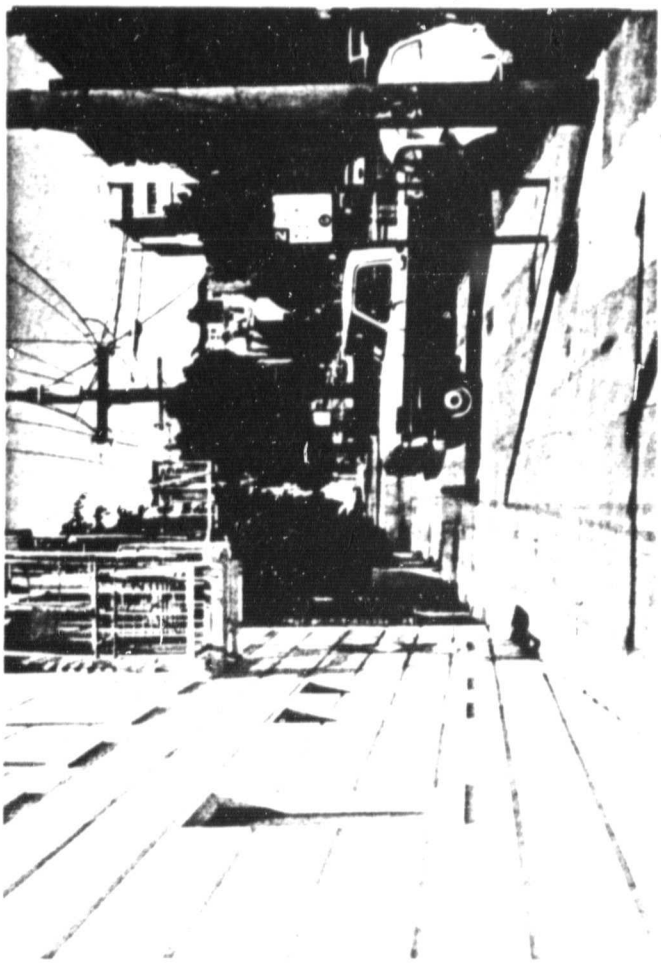


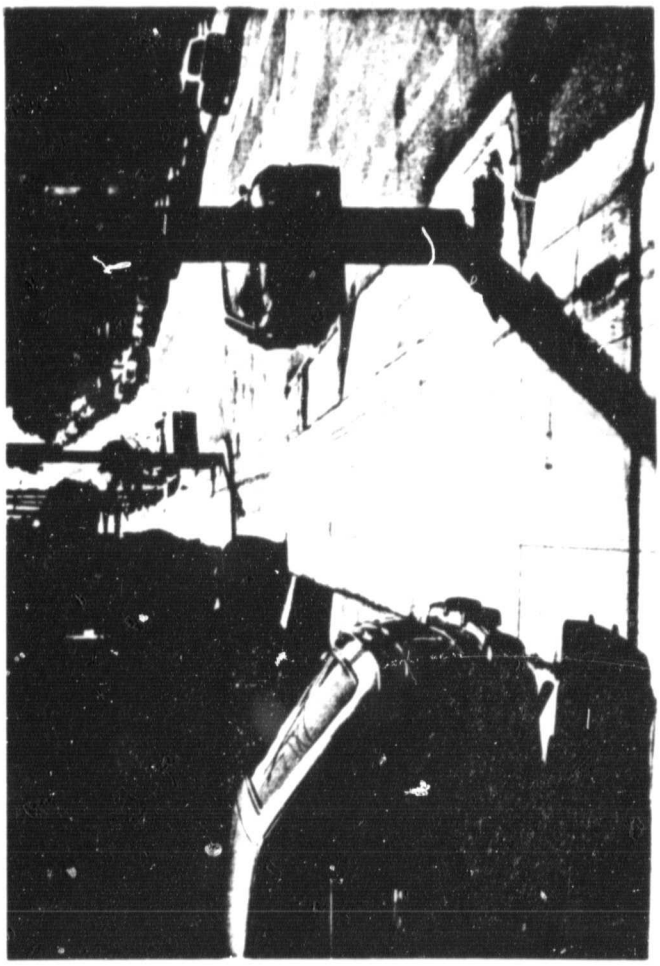


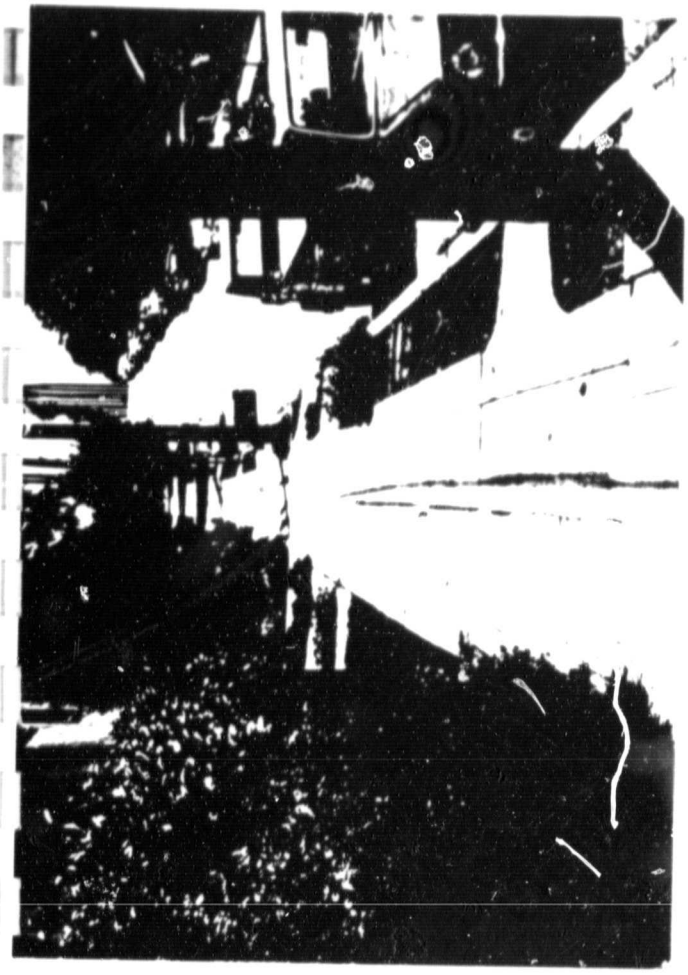


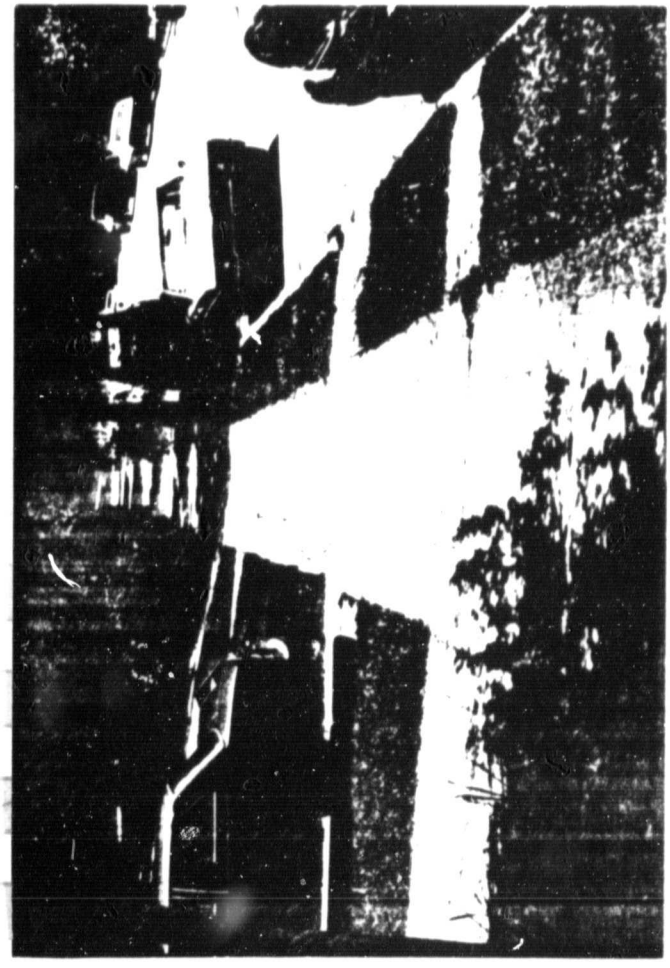












APPENDIX C
IMAGE DIGITIZATION AND FILE FORMAT

Image Digitization Procedure

35 mm images were digitized using an Optronics C-3200 Color Scanner at SPI International. Each strip was scanned at 50 microns per pixel resolution four times using clear, red, green, and blue color filters. Digitized pixel value is related to film density and film transmission by the following formulas:

$$\text{pixel value} = 255/3*(3-\text{film density})$$

$$\text{pixel value} = 255/3*(3-\text{logbase10(transmission)}),$$

where $-3 \leq \text{density} \leq 0$.

The boundaries of the individual frames were located using an image scrolling program (SCROLL). Each frame in all four color bands was extracted into images which are 700 pixels wide by 512 pixels high.

Image File Format

The digitized images are stored in files which consist of an 8-byte header followed by scan lines in left-to-right, top-to-bottom order. Each header has exactly the same information:

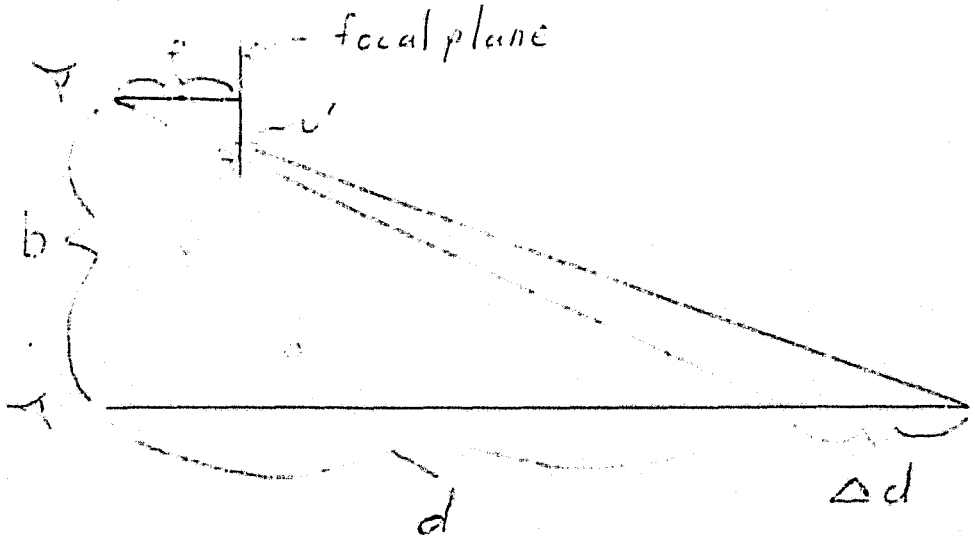
bytes 0,1	constant 1100 decimal
bytes 2,3	number of pixels per line = 700
bytes 4,5	number of lines per image = 512
bytes 6,7	number of bits per pixel = 8

A total of 88 files were produced: four each for the twenty-two 35 mm nos. The files are named as follows: FRAMEA.W, FRAMEA.R, FRAMEA.G, FRAMEA.B, ..., FRAMEV.W, FRAMEV.R, FRAMEV.G, and FRAMEV.B. The suffix to the string "FRAME" indicates one of 22 alphabetical letters in the range A through V; and the file name extensions .W, .R, .G, and .B refer to the clear (white), red, green, and blue color filters, respectively.

Effects of Viewpoint Separation on Stereopis

Stereo imagery is characterized by several parameters. These include the linear separation between the two cameras, termed the stereo baseline, the field of view and resolution of the cameras, and the distance to the three dimensional objects to be viewed. The stereo resolution of such a system can be defined in terms of the minimum depth separation between two objects that is large enough to be detectable by the system. Within the sidewalk environment the ability of a system to detect certain impediments depends upon the stereo resolution. For example, detecting step downs, holes, and broken sidewalk slabs depends upon the system being able to detect depth discontinuaties in the normally flat sidewalk surface. The difference in distance from the cameras to the edge of a hole and to the bottom of a hole must be greater than the system's stereo resolution if the hole is to be detected. For a desired depth resolution Δd, the minimum required stereo baseline b can be determined in terms of the other parameters.

Let the field of view of the camera be denoted by the angle φ. The monocular resolution of a camera is defined in terms of the number of pixels in the image. The camera can only detect features two pixels or larger in size (this is the Nyquist sample size.) Given a camera image of n by n pixels, for two points to be distinguished as separate they must lay further apart than $s = 4*f*\tan(\varphi)/n$ (where f is the focal length of the camera and s is the distance on the focal plane, both in the same units, inches or centimeters.) Δd will denote the minimum incrimental distance beyond a point at distance d from the cameras that can be resolved. The figure displays the camera geometry of objects at distances d and $d + \Delta d$.



The point at distance d from the right camera will appear at a location u on the image plane of the left camera given by the equation $u = b*f/d$. The point at distance $d + \Delta d$ will appear at $u' = b*f/(d+\Delta d)$. In the image plane of the right camera there will be no difference between the projection of the two points. Thus for the system to distinguish between an object laying at a distance d and one laying at a distance $d + \Delta d$, the quantity $u - u'$ must be greater than the detectable point separation distance s from above. This leads to the equation:

$$u - u' = \frac{b \cdot f \cdot \Delta d}{d \cdot (d+\Delta d)} > \frac{4 \cdot f \cdot \tan(\varphi)}{n}$$

solving for b:

$$b > \frac{4 \cdot d \cdot \tan(\varphi) \cdot (d+\Delta d)}{n \cdot \Delta d}$$

This equation gives the the minimum camera baseline required for a desired stereo resolution. From this it can be seen that higher stereo resolution is obtainable at the cost of a longer base line, higher resolution, narrower field of view, or lesser distance to the objects of interest. For example, given a field of view of 60 degrees, a camera resolution of 512 by 512 pixels, and an object distance of 15 feet, a baseline of nearly 21 inches would be required to obtain a stereo resolution of two feet. Because of perspective effects, a two foot stereo discontinuity is produced by objects eight inches or taller on the sidewalk, and thus such a baseline suffice to detect most sidewalk curbs.

APPENDIX E

LETTER TO

AI EXPERTS

FAIRCHILD
A Schlumberger Company

Technology Group
4001 Miranda Avenue
Palo Alto, California 94304
Telephone 415/493-3100
TWX 910/379-6435

July 6, 1981

Dear Colleague(s),

We seek your collaboration in a community-wide study to identify useful vision algorithms for incorporation in a blind mobility aid. This effort, sponsored by NASA (through the Biomedical Applications Team at Stanford University), is being undertaken to asses whether a useful prosthesis can be constructed with state of the art techniques and to ascertain where additional research is most needed.

Specifically, we are asking participants to evaluate empirically the performance of their most relevant, operational algorithms on a database of representative images.

Urban street scenes have been chosen as the experimental domain (see enclosed sample photos). A useful mobility aid must be able to identify clear navigable paths (e.g., the sidewalk) and detect the presence of significant obstacles (e.g., curbs, chuck holes, telephone poles, garbage cans, buildings and vehicles, as opposed to shadows). Semantic labeling of obstacles is desireable but not essential.

The enclosed paper provides further background on the difficult visual processing problems that arise in building a mobility aid, and describes a first attempt at an integrated solution. We are interested in better algorithms for accomplishing various stages of processing within the context of that system, as well as algorithms motivated by research in computational vision, that could overcome fundamental limitations. Within this broad charter, many types of algorithms are potentially relevant: for example, correlation tracking (e.g., for following the center line of the sidewalk), segmentation (e.g., for extracting long lines, smooth curves, and homogeneous regions, associated with the sidewalk and major obstacles), interpretation of pictorial features as physical scene events (e.g., as shadow or occlusion boundaries), recovery of intrinsic surface characteristics (e.g., range and orientation mapping using shading, texture gradient, stereo, motion stereo, etc.), extraction of prominent three dimensional surfaces (e.g., the ground plane, vertical surfaces) and volumes (e.g., cylinder representations of obstacles such as telephone poles and cars), semantic or schema based interpretation of pictorial or 3-D features, establishing symbolic correspondence between features in successive views and so forth.

We have compiled a modest image data base for use in evaluation, some samples of which are enclosed. All scenes are available in color and stereo. In some cases, foveal views of interesting features and time sequences of imagery (in the form of 16mm movie film shot while walking along the sidewalk) are also available. This data will be available both in hard copy (as color prints or positive transparencies and digitized form (via tape or ARPANET). The preferred form of output is graphical overlays (e.g., delineations of object boundaries) superimposed on the original image. Other forms of output (e.g., numeric arrays) may also be provided to permit more detailed evaluation of algorithm performance.

We hope you will be able to participate in this worthwhile cause. In so doing, you will also be contributing to the quality of computer vision research by helping establish a precedent for competitive algorithm evaluation on standard data sets. As an added incentive, a small honorarium ($150) will be awarded for each of the 20 most promising algorithms. More substantial consulting funds may also be available for refinement and/or further evaluation.

Please return the enclosed qustionaire as soon as possible. We plan to begin distributing imagery immediately on receipt, and would like to receive your results for compilation by the end of the summer. The results will be disseminated as a report and, if appropriate, at some suitable conference.

Sincerely,

Jay M. Tenenbaum
Fairchild AI Research Laboratory

Michael Deering
University of California, Berkeley and
Smith Kettelwell Institute of
Vision Research

P.S., Recipients of this letter and some algorithms of interest are listed on the next page. Suggestions for other participants and algorithms relevant to this study would be greatly appreciated.

Partial list of potential participants and potentially relevant algorithms:

Harlyn Baker	(edge-based stereo)
Dana Ballard	(Hough techniques)
Ruzena Bajscy	(texture gradient)
Steve Barnard	(vertical surface finder)
Tom Binford	(edge detection and interpretation)
Rod Brooks	(ribbon finder, model-based object finding)
Dave Burr	(image registration)
Bob Cunningham	(motion tracking)
Larry Davis	(segmentation, texture algorithms)
Martin Fischler	(linear features, analysis of range data)
Don Gennery	(stereo ground plane finder)
Eric Grimson	(stereo)
Marsha Jo Hannah	(bootstrap stereo)
Bob Harralick	(facet model)
Ellen Hildreth	(Primal sketch)
Berthold Horn	(shape from shading + contour)
Takeo Kanade	(range finder, texture)
John Kender	(texture)
Martin Levine	(segmentation algorithms)
David Lowe	(surface interpretation)
Worthy Martin	(motion)
Dave Milgram	(interpretation of range data)
Hans Moravec	(stereo, motion, obstacle avoidance)
M. Nagao	(segmentation)
H-H Nagel	(analysis of image sequences)
Ram Nevatia	(edge detection, motion stereo)
Yu-ichi Ohta	(region analysis)
Sandy Pentlin	(shape from shading)
Walt Perkins	(concurves)
Slava Prazdny	(range, orientation from motion)
Keith Price	(region extraction)
Lynn Quam	(correlation tracker)
Raj Reddy	(region extraction)
Ed Riseman	(edge and region extraction techniques)
Azriel Rosenfeld	(relaxation enhancement)
Jay Tenenbaum	(interpretation-guided segmentation)
Shimon Ullman	(shape from motion)
Jon Webb	(motion)
Tom Williams	(analysis of image sequences)
Andrew Witkin	(occluding edges, shape from texture)
Steve Zucker	(segmentation algorithms)

QUESTIONAIRE

Name:

Address:

Description of algorithms:
(please answer for each algorithm you plan to evaluate, using additional pages as required)

Type of algorithm (edge detection, stereo matching etc.)

Input (image, line drawing, range array, cylinder model etc.)

Output (image overlay, edge or region data structure, orientation array etc.)

Principles of operation (brief description)

Envisaged role in a mobility aid.

Implementation Details
(hardware, software, memory requirements, execution speed)

What types of experimental imagery will you need from us?
(Type of scene-refer to numbers on back of sample photos; color, stereo, field of view; resolution and format of digitiz data, if desired, else size and nature e.g., transparency, glo of hard copy data.)

In what form will you supply output for evaluation
(photograph of displayed results overlaid on original image, printout, tape dump of data structures etc.)

note: Transfer of imagery and results in machine readable form is preferred. Digitized imagery will be provided as files, one record per row of the image array. The same format can be used to return results as arrays of labeled pixels. Such arrays provide a uniform means for representing results at many levels of processing (e.g., range values, edge or region labels, semantic labels and so forth.) They are easy to display and compare, and are readily transformed into other data structures. Photos of graphical output are also acceptable, and would be appreciated, in any case, to permit an initial, qualitative evaluation.

Comments
(suggestions for other participants, general comments on running this study)

APPENDIX F

VOLUNTARY PROJECT PARTICIPANTS

Experimental Algorithms: Monthly Aid for the Blind

Source	Algorithm
Donald Gennery, JPL	Stereo processing algorithm for rapid identification of ground planes
Takashi Matsuyama Kyoto University	Segmentation algorithm for partitioning image into regions for further processing
Harlan Baker, Stanford University AI Laboratories	Stereo (correlator) processing algorithm for depth perception and 3-D image extraction
R. Nevatia, USC	Linear-feature (edge) extraction algorithm
Dr. Nagel, Hamburg, Germany	Motion interpretation algorithm
William Thompson, University of Minnesota	Frame-to-frame region-matching algorithm for motion interpretation, object identification, and determination of location and speed.
Keith Price, USC	Segmentation and region-matching algorithm
Martin Levine, McGill University	Segmentation algorithm
Linda Shapiro, Virginia Polytechnic Institute	Edge extraction and region-matching algorithm
Ed Riseman, University of Massachusetts	Motion interpretation algorithm
T. Williams, Digital Equipment Corporation	Motion interpretation algorithm for ground plane identification
Dana Ballard, University of Rochester	Motion interpretation algorithm

Experimental Algorithms: Monthly Aid for the Blind

Source	Algorithm
Azriel Rosenfeld and Larry Davis, University of Maryland	Unspecified
R.M. Haralick, Virginia Polytechnic Institute	Unspecified
Kanade, Carnegie - Mellon University	Unspecified
M. Bradley, MIT	Unspecified

APPENDIX G

Real-Time Natural Scene Analysis for a Blind Prosthesis

Michael Deering

Computer Science Division, Department of EECS
University of California, Berkeley
Berkeley, California 94720

Carter Collins

Smith-Kettlewell Institute of Visual Sciences
San Francisco, California 94115

• This work was supported by NSF grant no. PFR-7906299 from the Science and Technology to Aid the Handicapped Program.

ABSTRACT

A real-time computer vision system designed for the limited environment of city sidewalks is presented. This system is part of a prototype mobility aid for the blind. The overall device endeavors to keep blind pedestrians on a safe path down the sidewalk, and also warn of upcoming obstacles. The scene analysis algorithm uses semantic models of the environment to interpret edges in the multi-frame image data as borders of various objects, as well as to assign distance estimates to these objects. The input is a 64 by 64 by 6 bit gray-scale image taken from the vantage point of the shoulder of a pedestrian once a second. Along with each image, the three dimensional transformation of the camera location since the previous frame is assumed to be provided by hardware. After an initial segmentation into edge lines represented as arcs of circles, predictions of edges (generated by analysis of previous frames) are used to identify edges in the current frame. Edges not identified by this process are incorporated into the portion of the three dimensional world model that they are the most consistent with. The induced three dimensional world model of objects can then be used to provide mobility information to the blind user. The emphasis throughout the system has been on efficiency. The design trade-offs and techniques used to obtain high processing rates are discussed. Most of the vision system is currently running in real-time on a 16 bit micro-processor. Field trials of the complete prototype device will begin soon.

I Introduction

An effort to produce an optically based electronic mobility aid for blind pedestrians has led to the development of a natural scene analysis program for the typical scenes encountered by a pedestrian. The restriction to the semanticly rich domain of city sidewalks has allowed the visual processing to be performed in real time on a 16 bit microprocessor. The nature of the task is such that perfect object detection and recognition are not required, but rather the

probabilistic detection of potential obstacles. The main goal of the system is to determine approximately where the sidewalk is (with respect to the user), and secondarily to warn of objects blocking the path ahead. This task must be performed in real-time on real-world sidewalks.

Most real-time vision systems to date have dealt with very constrained image domains, due to the enormous computational requirements inherent in visual processing. These include industrial parts recognition [1], blood cell counting, and automatic navigation. Faster hardware and more efficient software techniques will gradually allow more complex domains to be handled at high speeds. Our approach has been to start with very fast segmentation, and amortize semantic processing over several frames, utilizing predictions from models of previously recognized objects to guide the parse of the current image. At the high end, our system is similar to many semantically oriented systems, such as [2] and [3]. Finally our use of multi-frame data is similar to many aspects of [4] [5][6][7][8].

II Constraints

In order to achieve real-time processing, we have had to impose several constraints upon the operations of the system:

1. Input is restricted to clean, sun-lit, mostly shadow-free sidewalks.
2. Certain initial starting conditions will be supplied to the system from the outside (which way the camera is pointing, where the sun is, etc.)
3. False positives are allowed (it is OK to occasionally warn about non-existent obstacles)
4. We must accept that within our resolution and processing time certain classes of objects are undetectable. These include objects whose width falls below the Nyquist sampling rate of the camera (mainly skinny poles), and objects with very low contrast with respect to the background, or those against a wildly changing background. At the same time, we wished to construct the program modulely, allowing knowledge about objects and scenes to be separated from the control structure (but without sacrificing efficiency.)

III Overview of the Program

The input scenes are successive 64 by 64 by 6 bit gray wide angle images taken from the vantage point of the shoulder of a pedestrian, at a rate of one or two frames per second. This relatively low resolution is the highest possible under the hardware and timing constraints. The overall organization of the program is: After video acquisition of the input scene, digitization and noise-removal, the information processed in three passes as follows:

1. segment the picture into linked chains of edges,
2. fit curves to these chains and put the mathematical description of the curves into an associative data-base, and
3. match these curves against several data bases (the world model) which include curve predictions from previous scenes.

The results of these matches identify semantically the objects belonging to the curves. Knowing whether an object is horizontal or vertical allow one to project the curves out into three dimensional space to determine their direction and range. Further heuristics are employed that utilize location information

from previous frames to make an independent motion stereo based estimate of the object ranges. At this stage the program should know where the sidewalk and any close obstacles are located, and can proceed to output this information. (Obstacles are the common sorts of large physical objects that one might encounter on a city sidewalk: phone poles, lamp posts, fire hydrants, trash cans, sign posts, automobiles (off to one side), parking meters, trees, bushes, etc.)

IV Coordinate System

The coordinate system used for the three dimensional outside world is centered on the focal point of the camera as it moves through space. The Z-axis is oriented in the direction that the camera is pointing, the Y-axis points straight up, and the X-axis points to the right of the camera. Thus the three dimensional location of all objects is always determined relative to the pedestrian, and are re-computed each frame. Points in the world are mapped to points in the image plane through the usual projective geometrical equations.

One of the fundamental problems of computer vision is that this projection of the three dimensional world (X,Y,Z) to the image plane (x,y) cannot be reversed without additional data of some kind. One of the main goals of the semantic phase of our system is to provide this additional data via semantic knowledge about the probable locations, orientations, and relationships between typical objects encountered within the sidewalk environment. This additional information usually is in the form of a hypothesis on the value of one of X,Y, or Z. Given this value, along with the image plane feature location (x,y), the remaining two three-dimensional coordinates can be found by suitable manipulation of the projection equations.

The motion of the camera in the world between frames will cause the projections of edges of three dimensional objects onto the image plane to change. The general case of the camera transformation involves six parameters, $(\Delta X, \Delta Y, \Delta Z, \vartheta, \varphi, \rho)$. For a camera mounted upon the shoulder of a pedestrian it can be safely assumed that ΔY and ρ are approximately zero, as there is little torsional rotation, and the height of a particular pedestrian remains roughly constant. (It should be noted that the mechanics of the human visual system goes to a great deal of effort to keep ρ near 0, a ρ rotation of up to six degrees of the head is countered by an opposite rotation of the eyeball in the socket. The shape of the horoptor in many animals indicate that the height of the animal is taken to be a constant for some visual processing.) The equations to perform this transformation can be combined with the image plane projection equations to obtain the location within the current image plane of a world point from a previous frame.

V The World Model (the Sidewalk World)

Our world model is the typical sidewalk environment as encountered by a pedestrian. Most modern sidewalks are constructed of slabs of white concrete, and are three to twelve feet wide. Many run in straight lines for an entire block before ending in a corner, while others may be curved. For simplicity it is assumed that all sidewalks encountered by the vision system are straight for thirty feet beyond the camera unless a corner is ahead. Sidewalks mainly differ in their width and the presence (or absence) of a grass border on their street side. These variations are modeled by a few simple parameters.

Within the 64 x 64 image the borders of the sidewalk on either side will often appear as high contrast lines. These lines will be in the lower half of the image,

at a highly inclined angle. Various objects bordering on the sidewalk sometimes are of similar optical intensity, reducing the contrast of the sidewalk edges. A large variety of objects can border city and suburban sidewalks. These include: bushes, shrubs, trees, grass, dirt, driveways, walkways, walls of all types, doors, and windows. On the street side usually one finds pavement and automobiles. These objects occur at fairly predictable locations, and many times with good contrast compared to the white sidewalk.

Most objects located upon the sidewalk proper have three fortunate properties: they do not move, have stereotypical locations with respect to the edge of the sidewalk, and usually do not block the path. Objects in this class include: phone poles, lamp posts, fire hydrants, traffic signs, parking meters, trees, mail boxes, phone booths, most trash cans, and bushes. Many of these objects also have the property of being rectilinear, and approximatable as cylinders or boxes (and thus produce good, high contrast edges.)

Unfortunately other objects are more unpredictable. These include: paper boxes, bags, newspapers, trash, garbage cans, parked bicycles, and badly parked cars. Such obstacles can appear anywhere on the sidewalk, and are not always very rectilinear. However, they do have some properties that facilitate their detection. Many are short, so their borders are within the two edges of the sidewalk. They also rest directly upon the ground, enabling their distance to be accurately determined and verified over several frames.

Finally there is the class of moving objects which are the hardest to handle, as their shape and location may change drastically from frame to frame. This class includes: pedestrians, dogs, bicycles, occasionally a car crossing the sidewalk, and wind-blown trash. Fortunately, most mobile obstacles are alive or controlled by humans, and will try not to collide with a pedestrian.

VI Low Level Processing

Initial segmentation of digital images is now a fairly well developed art, but no general technique is know to be (close to) optimal for the large class of natural images, and the speed of various algorithms can differ by orders of magnitude. The necessity for real-time performance is the most severe constraint imposed upon our system. Many promising segmentation techniques had to be rejected out of hand on efficiency grounds.

The speed of the initial segmentation algorithm dominates the performance of the overall system, as the semantic phase is usually many times faster [1]. Thus a poor quality (but fast) segmentation algorithm may be preferable to higher quality (but slower) algorithm if one can make the semantic phase work a little harder. This is the case in our system. Our segmentation algorithm only directly compares two pixels at a time, and thus is sensitive to noise, but runs at a very high speed. In some sense every module in the system after the pixel comparison has some component who's job is to help correct for the defects introduced by the initial fast segmentation.

The segmentation algorithm used is an edge following algorithm that differs from the usual (such as described in [9]), in that we follow several edges simultaneously. Most edge followers grow an edge line point by point serially from one end of the line. We instead grow many edge lines in parallel by adding to both ends of many edge lines simultaneously. The advantages of this method corresponds roughly to those gained by a breadth first versus a depth first search, in that there is more global information available when one is forced to make local decisions. This allows edge thinning to take place at the same time as edge following, and contributes to the speed of the algorithm.

In more detail, edge points in the input are found using a pseudo-random scan [10]. In the area around each point a search is made for existing edge lines and other isolated edge points. Based on complex decision rules, an existing edge line may be extended to the new point, a new edge line may be created between the new point and another point, or two edge lines may be joined through the new point. (These rules are similar to those found in [11], but with less complex weightings.) The decision rules mentioned above help thin the edges, mainly by forcing edge lines to be essentially continuous. The final result of pass 1 is the collection of edge lines obtained after all the edge points have been processed.

The edge follower tends to be conservative, as it knows that pass 3 will connect broken lines. This is possible as pass 3 has access to more global information about the objects within the scene, and thus may have reason to believe that three roughly collinear line segments may in fact be the edge of one object. Pass 1 does not have enough information to decide if a gap between two line segments is due to noise breaking up a single edge, or is really an occluding object or a gap between two objects.

In our system, noise in a single pixel many times can lead to the break-up of a potential edge line into two pieces. The defects in this edge segmentation can be modeled as higher level noise. That is, as broken edges, missing edges, and misoriented edges. Semantic rules about line segments can take these defects into account, and sometimes even make use of certain properties of the "noise". For example, the breakup of a edge line corresponding to the edge of an object in the scene may be caused by a surface discoloration near the object edge. If this is the case, then the "noise" will be serially correlated from frame to frame. Thus the broken edge will be broken in the same way in several frames, and the relative location of the break can be (and is) used as a feature of that edge (which can help to re-identify it in successive frames.)

VII The Fitting of Edges to Curves

Pass 2 gathers information about each edge line, summarizes its attributes, and sorts it to permit quick searches. Pass 1 sorts the edge lines by x-y location and computes their length, Pass 2 computes their "curvature" and angle of inclination from the horizontal, and sorts them by angle. Edge lines that are too convoluted are broken up into smaller (and simpler) segments. This covers the few cases in which the conservative edge follower described above is not sufficiently conservative. This can occur when two straight line segments intersect at a shallow angle. Pass 1 cannot distinguish this case from a single shallowly curved line segment. Pass 2's curvature statistics are needed to resolve this case.

We devised a fast algorithm to fit lines to arcs circles. The main point for our application was to within a milli-second classify a given line segment as a roughly straight line, a gently curving line, or noise. It is possible that in the future we may make use of finer details of the curves, but in an environment of rotating three dimensional objects absolute curvature is of little use, and more general information on object surface orientation and distance is needed (such as discussed in [12].)

VIII Representation of Objects

Our three dimensional representation necessarily emphasis the edges of objects, given the nature of our initial segmentation. The representation

roughly resembles a collection of three dimensional edges of the object, but the locations of the edge lines relative to each other is not fixed as in 3D wire frame models in computer graphics, but rather are allowed to vary as needed. As most objects in the sidewalk world are somewhat rectilinear, many are represented by planes parallel to the X,Y or Z axis (a similar representation was used in [4].) For example, a phone pole can be fairly well approximated by a rectangle facing the observer. The sidewalk proper is approximated by a rectangular slab who's position and width is updated every frame with new data.

To achieve high processing speeds, some of our representation is procedural rather than semantic. But as we have built up the number of objects that we handle, a number of common subroutines have emerged, allowing new objects to be added and represented fairly easily. High processing speeds verses separation of control structures and knowledge are not necessarily incompatible, but to obtain both one must have intervening software step that transforms high level abstractions into a form combinable with control structures. It also helps to have a very flexible control structure. Our system puts both edge data and object building procedures into associative data-bases, thus allowing the flow of control to be determined by the data. In retrospect, most of the object handling procedures could have been generated by machine from static descriptors rather than by hand, and we may go to such a system in the future.

IX Representation of Visual Knowledge

With the knowledge of how to recognize and represent objects handled by the object representation, the remaining visual knowledge of interest is that that tells you *where* an object is (it's distance and direction.) A number of hueristics of varying degrees of generality exist to do this job, with varying degrees of accuracy and constraints. These are:

1. If the object is know to be resting on (or very near) the ground plane, then Y is known to be -UserHeigth, and X and Z can be obtained by back projection.
2. If the object has a known distance from the edge of the sidewalk (for example, phone poles are usually one foot in), and all one has is a piece of an edge of the object (usually not the ground plane intersection), then one can obtain the objects (X,Y,Z) location as follows: take the image plane (x,y) location of the edge piece, project it as a line through the origin (the focal point of the camera) into (X,Y,Z) space, and intersect this line with the plane which is the constant distance in from the sidewalk. This intersection X and Z will be the object's location on the ground plane. (The equation of the plane parallel to the sidewalk is obtainable because the equation of the sidewalk edge is assumed to be known.) Even if the constant distance in from the sidewalk edge is incorrectly guessed, which can lead to distance error on the order of 50% or more, at least some distance information has been provided, and one can make simple decisions like "will I run into it in two seconds or twenty seconds". In future frames, motion stereo can tighten up this distance estimate (and correct the constant distance term). By the time an object initially sighted twenty or more feet away comes to within five feet of the pedestrian the location of the object will have appeared in thirty frames of solid data, hopefully pinpointing it's location to within a foot.

3. Once the effects of camera tilt and pan have been subtracted, the differences in locations of a feature in successive frames can be used to determine its range and distance by working backwards from the projection and camera transformation equations. We employ motion stereo as a secondary distance cue that is used to check up on our primary cues 1 and 2 above.
4. There are some distance heuristics that are only of use for determining the equation of the sidewalk's borders. These include making use of the known constant width of the sidewalk.
5. Finially, the location of an image feature relative to the current interpretation of scene can be used to obtain a probable distance estimate. For example, edges near the vanishing point of the sidewalk are probably (but not necessarily) far away. Edges way off to one side and in the sky most likely belong to upper stories of buildings or the background, and may be safely ignored. (One misses overhangs this way, but overhangs in general are very hard to recognize, as many are of very low contrast to begin with.)

X Semantic Analysis

The semantic phase endeavors to build a three dimensional model of the outside world that it is moving through, such that edges in the image frames correspond to edges of objects in three dimensional model. The various distance hueristics listed above are employed both to initially place objects as well as to verify their location/identity over several frames. (An object whose distance varies wildly from frame to frame may be mis-identified.) Further constraints exist that simplify the semantic analysis task. These are:

1. Most objects in the sidewalk world rest on the ground plane (though we do not assume that their point of contact is visible.)
2. Most objects can be roughly approximated by planes parallel to the x, y, or z axis (as in [4].)
3. Location accuracy need only be enough to avoid objects *most* of the time. For example, distances to objects need only be computed with an accuracy of ±20% when objects are closer than 8 feet, and ±40% when objects are further away.
4. The camera transformation will be correctly supplied most of the time (by hardware) to within 1% angular accuracy and 10% translation accuracy.

In order to speed the identification of edges in a new frame, predictions of edge locations from previous frames are used. Much of the speed of the semantic pass is due to the essentially hardware solution of the successive frame registration problem. Most re-occurring edges can have their location in successive frames determineed to with a few pixels be using the camera transformation. The overall effect is a sort of "boot-strapping" re-identification of scene features, similar to that described in [13]. (It may be possible that in the future we can dispense with the special camera motion tracking hardware and recover this information incrementally from optical flow.)

In more detail, the semantic phase is broken up into several subparts. These are:

1. Edge lines from the previous frame are first transformed by the camera transformation and then matched against edge lines in the current frame, current lines that appear to be direct descendents of previous lines are removed from the current data base as "explained". The order in which old object's edges are searched for is determined by their relative semantic

importance and data quality. Thus the edges of the sidewalk are usually searched for first, followed by the other objects roughly ranked by their number of (visible) edges.

2. The matches made in 1 induce new information that can be used to construct an updated three dimensional model of the objects that the edges belong to. These models can then make claims for gaps in their edge outlines.
3. The claims made in 2, along with various generic claims for new objects are matched against the remaining data base of edge lines. Residual lines will be claimed as background noise.
4. New object edges obtained in 3 allow for further updating of the three dimensional models, which at this point can be used by the blind navigation system. Predictions and search scheduling for the next frame are made at this point. Objects who's existence is no longer supported by the edge line evidence are deleted in favor of more consistent interpretations.

Thus at any one point in time the world model data base of the system contains models of several objects (phone poles, bushes, automobiles, etc.) that are moving by, as well as a model of the sidewalk proper.

XI Experimental Results

Figure 1 is a sample image taken from an 8mm movie of a sidewalk. One of our test sequences consists of 30 digitized images from this movie. The forward motion between each frame was one and a half feet. On our half speed XL68,000, passes 1 and 2 can process this movie at the rate of one second per frame. The semantic processing of pass three takes an additional half second per frame. When applied to this movie, the system correctly discovered and tracked the sidewalk edges, as well as edges of several objects off to the right of the sidewalk. No objects were found to be blocking the sidewalk. Figure 2 displays a digitized image from the middle of this sequence with the wire-frame model of pass 3 superimposed. (A similar fit is made for each frame of the movie.) A computer animated reconstruction of the outside environment given the world model produced by pass 3 is seen in figure 3. (The detail on the parked car is simulated.) Figures 4, 5 and 6 display three other digitized scenes from earlier in the movie, with the wire frame model produced by that system for that frame superimposed. We expect to be running field trials of the entire system in a portable cart trailing behind a blind subject shortly.

XII Incorporation into a Blind Aid

The overall functioning of this system as a blind aid is part of the lineage of a large number of previous tactile blind aids devices designed over the last twenty years [14][15][16]. The computer vision component and the blind interface component have been separated out from each other via the following reasoning:

1. Assuming a perfect computer vision system that knows where every object of interest is located, how could one best communicate this information to a blind pedestrian? What sort of user interfaces and interactions will allow the user to make rapid, accurate use of the information provided?

Fig. 1 Original Image

Fig. 2 Image+Wire Frame Model

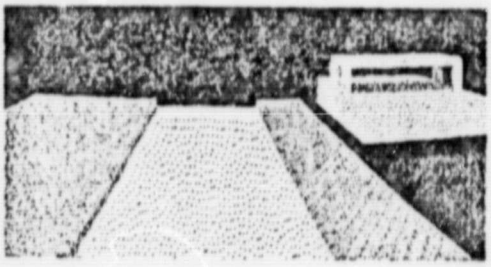


Fig. 3 Computer Reconstruction

Fig. 4 Frame #3+Wire Frame Model

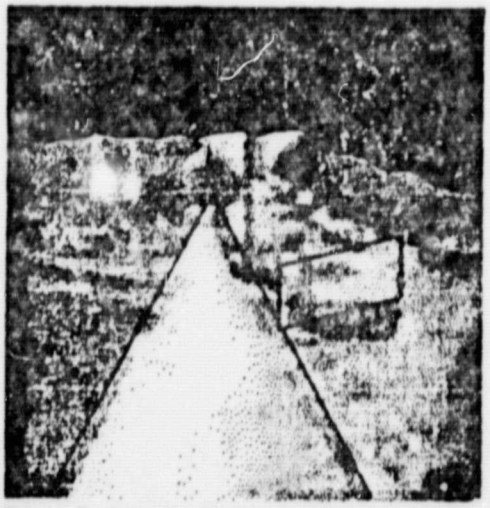


Fig. 5 Frame #4+Wire Frame Model

Fig. 6 Frame #5+Wire Frame Model

2. How does one build a portable (wearable!) computer vision system that will locate at least the majority of the objects of interest?

Our solution to 1 has been presented in the previous portions of this paper. Our solution to 2 is to use two output channels - stereo synthetic speech for cognitive (high level) information, and a linear array of 16 tactile elements as a pointing device. The stereo speech unit is a combination of a normal speech synthesis system with a audio processing unit that can "throw" the computer's speech, allowing it to appear to come from a particular direction and distance. (For example, a phone pole might seem to announce "phone pole".) The tactile array is a skin tapping device worn as a head-band, each element corresponds to a particular angular direction, and the frequency of taps of an element corresponds inversely to the distance of the feature being pointed to.

Currently we intention to have the speech unit make major announcements (the blind don't want it babbling all the time, they listen to sound shadows and street sounds.) The tactile output will be used for communicating more mundane information, such as "you're veering off to the left of the sidewalk, veer a bit to the right", by "flashing" the edge of the sidewalk on the appropriate side of the tactile display, should the user veer toward it. In any case, one of the main reasons for putting the whole system on a micro processor, rather than simulating it on a mainframe, was to have the capability of expermenting in the real world with various blind interface systems. Also, despite years of experience in testing blind aids, it is very hard to tell how the blind will react to a particular device without letting them make extensive use of it under real world conditions.

XIII Perspectives on Future Directions

Within the hardware and timing constraints imposed, we feel that the current system performs well, and cannot be much improved upon. However, for use in a robust blind aid, the system has several limitations which must be overcome. These include the low sensitivity to low contrast edges in shadowed or complex scenes, and the low resolution (~1 degree/pixel.) More important is the limitation to processing of edge data only, at the exclusion of surface data. It would be nice to have more information about the sidewalk than just where it's edges are, such as how flat it is, are there any broken sidewalk slabs or holes? Also, the current system must treat any high contrast edge on the sidewalk as a potential object edge, even though most are only flat shadows or stains.

Various surface processing techniques proposed in the literature can solve many of these limitations. Texture gradients [17] should provide a fairly robust broad classification of the scene into flat and upright surfaces. Optical flow can provide approximate distance estimates. Luminance gradients could indicate surface curvature, which could be used in identifying (and segmenting) phone poles, walls, automobiles, etc. [12]. Finally, stereo gradients can not only determine general distance estimates, but for the special case of the almost flat sidewalk, they can be tunned to spot vertical deviations as small as half an inch (such as un-even sidewalk tiles that one might trip upon). More importantly, stereo can determine that a dark patch is flat, and can be safely ignored. This would be similar to the system described in [18]. However, for this specialized stereo algorithm to work, it must have a very good estimate as to where the flat sidewalk is in the first place. This is where the other surface processing techniques enter the loop. Such a system could provide very robust performance under even extreme conditions (such as wet (and reflective!) sidewalks in the

rain), but at the expense of special purpose hardware.

Currently we are in the initial stages of designing a surface processing oriented version of our system system along the lines described above, which is to be implemented in VLSI. This system will be characterize by higher resolution (with separate foveal and peripheral resolution), higher frame rates (approaching 30 frames a second), stereo processing, and extensive use of surface processing techniques. It will differ from other vision systems in that it will be optimized for the sidewalk environment. For example, the stereo section will not have to deal with the stereo frame registration problem in its general form, but for the much simplier case of extracting the (mostly flat) ground plane. Evidence indicates that the vision system of many animals (including man's) has a built in special case solution for extracting the ground plane, which is similar to our proposed technique.

REFERENCES

[1] Perkins, W. A. "A model based vision system for industrial parts." *IEEE Trans. Comput.* vol. C-27 (1978) 126-143.

[2] Brooks, R., Greiner, R. and Binford, T. "The ACRONYM Model-Based Vision System." In *Proc. IJCAI-79.* Tokyo, August, 1979, 105-113.

[3] Saphiro, L., Moriarty, J., Mulgaonkar, P. and Haralick, R. "Sticks, Plates, and Blobs: A Three-Dimensional Object Representation for Scene Analysis." In *Proc. First NCAI.* Stanford,CA, August, 1980, 28-30.

[4] Williams, T. "Depth from Camera Motion in a Real World Scene." *IEEE Trans. Pattern Analysis and Machine Intelligence.* vol. PAMI-2, no. 6 (1980) 511-516.

[5] Tsuji, S., Osada, M. and Yachida, M. "Tracking and Segmentation of Moving Objects in Dynamic Line Images." *IEEE Trans. Pattern Analysis and Machine Intelligence.* vol. PAMI-2, no. 6 (1980) 516-522.

[6] Roach, J. and Aggarwal, J. "Determining the Three-Dimensional Motion and Model of Objects from a Sequence of Images", University of Texas at Austin Laboratory for Image and Signal Analysis research report TR-80-2.

[7] R. Nevatia, "Depth Measurement by Motion Stereo," *Computer Graphics and Image Processing*, vol. 5, pp. 203-214, 1976.

[8] M. Lavin, "Analysis of Scenes from a Moving Viewpoint," in *Artificial Intelligence: An MIT Perspective*, P. Winston and R. Brown, eds., pp. 187-208 (The MIT Press, Cambridge, Massachusetts, 1979.)

[9] McKee, J. and Aggarwal, J. "Finding the edges of the surfaces of three-dimensional curved objects by computer." *Pattern Recognition.* vol. 7 (1975) 25-52.

[10] Kavasznay, L., and Joseph, H. "Image processing." *Proc. IRE.* no. 43 (1955) 56-57.

[11] Prager, J. "Extracting and Labeling Boundary Segments in Natural Scenes." *IEEE Trans. Pattern Analysis and Machine Intelligence.* vol. PAMI-2, no. 1 (1980) 16-27.

[12] Tenenbaum, J. and Barrow, H. "Recovering Intrinsic Scene Characteristics from Images." In *Computer Vision Systems*, Hanson, A. and Riseman, E. eds., 3-26 (Academic Press, New York, NY, 1978).

[13] Hannah, M. "Bootstrap Stereo." *Proc. First NCAI.* Stanford,CA, August, 1980, 38-40.

[14] Collins, C. "Tactile television: Mechanical and Electrical Image Projection." *IEEE Trans. on Man-Machine System.* no. 11 (1970) 65-71.

[15] Collins, C. and Madey, J. "Tactile sensory replacement." *Proc. San Diego Biomedical Symposium.* Vol. 13 (1974) 15-26.

[16] Collins, C., Scadden, L. and Alden, A. "Mobility studies with a tactile imaging device." *Proc. Conf. on Systems and Devices for the Disabled,* Seattle, (1977) 170-174.

[17] Witkin, A. "A Statistical Technique for Recovering Surface Orientation from Texture in Natural Imagery." In *Proc. First NCAI.* Stanford,CA, August, 1980, 1-3.

[18] Gennery, D. "A Stereo Vision System for an Autonomous Vehicle." In *Proc. IJCAI-77,* Cambridge, MA, August, 1977, 576-582.