TEC-0069

Passive Recovery of Scene Geometry for an Unmanned Ground Vehicle

Robert C. Bolles
Martin A. Fischler

SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025-3493

September 1995

Approved for public release; distribution is unlimited.

U.S. Army Corps of Engineers
Topographic Engineering Center
7701 Telegraph Road
Alexandria, Virginia 22315-3864

US Army Corps of Engineers
Topographic Engineering Center

TEC

REPORT DOCUMENTATION PAGE			*Form Approved* *OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 1995	Final Technical Nov. 1991 - Dec. 1994

4. TITLE AND SUBTITLE	5. FUNDING NUMBERS
Passive Recovery of Scene Geometry for an Unmanned Ground Vehicle	DACA76-92-C-0003
6. AUTHOR(S) Robert C. Bolles Martin A. Fischler	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) SRI International 333 Ravenswood Avenue Menlo Park, CA 94025-3493	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Topographic Engineering Center 7701 Telegraph Road Alexandria, VA 22315-3864	19. SPONSORING / MONITORING AGENCY REPORT NUMBER TEC-0069

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT	12b. DISTRIBUTION CODE
Approved for public release; distribution is unlimited.	

13. ABSTRACT *(Maximum 200 words)*

The goal of this project was to develop techniques for constructing three-dimensional descriptions of outdoor scenes to support the navigational needs of an Unmanned Ground Vehicle (UGV), operating both during the day and at night. In this report we describe our progress in four areas:

Stereo Evaluation - We performed an in-depth evaluation of three representative stereo techniques by analyzing the results on 49 stereo pairs.

Scene Sketch - We developed a high-level representation of an outdoor scene, which we call a "scene sketch", that can describe the semantic, as well as geometric properties of a three-dimensional scene.

Spatiotemporal Filtering - We developed a technique for increasing the resolution and robustness of passive range sensors by integrating stereo and motion analysis.

FLIR Stereo - We demonstrated the effectiveness of stereo analysis applied to infrared data, which makes passive night driving possible.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Stereo analysis, motion analysis, unmanned ground vehicle, FLIR, semantic scene sketch	53
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

TABLE OF CONTENTS

PREFACE

This report was prepared under Contract DACA76-92-C-0003 for the U.S. Army Topographic Engineering Center at Alexandria, Virginia, by SRI International in Menlo Park, California. The following SRI researchers have contributed to the work described: H.H. Baker, R.C. Bolles, M.A. Fischler, P. Fua, M.J. Hannah, J. Herson, L. Iverson, and J. Woodfill.

The Contracting Officer's Representatives have been Ms. Linda Graff, Ms. Lauretta Williams, and Mr. Thomas Hay.

Passive Recovery of Scene Geometry for an Unmanned Ground Vehicle

I. Introduction

SRI performed this project as a member of ARPA's Unmanned Ground Vehicle (UGV) team of co-contractors. Martin Marietta Corporation of Denver, Colorado, was the integrating contractor and SRI was one of 40 or more co-contractors providing technology to Martin Marietta for inclusion on its vehicles.

SRI took an active role in the "tiger team" that designed the initial architecture for the vehicle sensing and control systems. Throughout the project, we worked closely with researchers from the Jet Propulsion Laboratory (JPL) and Teleos, Inc. These three groups formed the "stereo team" or "navigation sensing team," which designed, developed, evaluated, and delivered a series of stereo ranging systems to Martin Marietta for inclusion in its integrated demonstrations and technology demonstrations.

Stereo analysis is a critical technology for the UGV because it provides a passive way of measuring distances to objects in front of the vehicle. A dense array of these distances forms the raw data from which navigable paths and obstacles can be located. The military is particularly interested in passive ranging techniques because they do not involve the projection, onto the scene, of electromagnetic energy that might be detected by an enemy.

Stereo analysis, when applied to infrared data, can provide passive ranging for night operations. This extension is crucial because the military operates 24 hours a day.

During the project, SRI concentrated on the following four areas:

Stereo Evaluation – We performed an in-depth evaluation of three representative stereo techniques by analyzing their results on 45 stereo pairs.

Scene Sketch – We developed a high-level representation of an outdoor scene, which we call a "scene sketch," that can describe the semantic, as well as geometric, properties of a three-dimensional scene.

Spatiotemporal Filtering – We developed a technique for increasing the resolution and robustness of passive range sensors by integrating stereo and motion analysis.

Forward-Looking Infrared (FLIR) Stereo – We took the first steps toward demonstrating the effectiveness of stereo analysis applied to infrared data, which can support night operations.

In this report we describe our progress in these areas and briefly outline our plans for the future.

II. Stereo Evaluation

Stereo analysis, which for a long time had been viewed as a technique that was interesting, but too costly to be practical, has recently emerged as a viable tool for realtime applications, such as vehicle navigation. This has happened for two reasons. First, advances in hardware have made it practical to compute stereo matches in real time. And second, advances in algorithm development have made it possible to correctly match large portions of outdoor scenes.

Our first task in this project was to perform a qualitative evaluation of current stereo techniques to see if they were capable of providing sufficient range information to support outdoor vehicle navigation. Our goals for this evaluation were (1) to make an initial estimate of the effectiveness of current stereo techniques applied to UGV tasks, (2) to identify key problems for future research, and (3) to debug the evaluation process.

For the evaluation, we decided to examine the effectiveness of a small number of techniques applied to a large number of examples. Since we did not have the resources to perform a complete analysis of all techniques, we felt that it would be more instructive to examine a few techniques thoroughly than to evaluate many of them.

One of the guidelines we adopted for this evaluation was to develop and maintain an atmosphere of cooperation and constructive criticism among the researchers participating in it. Without this, we would not be able to focus on our ultimate goal of producing a sequence of increasingly capable stereo systems. To help establish a cooperative atmosphere, we decided to concentrate on the positive aspects of each technique and emphasize potential extensions, realizing that existing techniques were developed for different domains and different applications. We also decided to share all the raw results with the participants so they could duplicate our analysis or develop their own.

For the evaluation, SRI collected imagery from five groups, JPL, INRIA (in France), SRI, Carnegie Mellon University (CMU), and Teleos (hence the name "JISCT" for the first evaluation phase); selected 49 image pairs for analysis; converted them into a standard format; distributed the data set to the five groups for processing, along with an extensive set of instructions; collected the results; characterized them; and finally distributed the results and the associated report to the participants.

We intentionally asked each group to process a large number of pairs (10 training pairs and 45 test pairs ... 6 pairs were in both the training and test sets), because we wanted to force each group to establish a standard algorithm that was automatically applied. As a result of this approach, there are now three or four groups around the world that can readily apply end-to-end stereo techniques to new data and compare their results. In the future, we hope to expand this community to 10 or more groups. This process has opened up a new form of interaction within the computer vision community that we feel will help stimulate advances and reduce redundant development.

In the instructions to the participants, we asked each group to produce several results for each match point in addition to its computed disparity. For each point, we asked for an x and a y disparity, an estimate of the precision associated with each reported disparity, an estimate of the confidence associated with each match, and an annotation for each unmatched point, indicating why the technique could not find a match. Possible explanations for a no match included, "area too bland," "multiple choices," and "inconsistent with neighbors." Although none of the groups produced all this additional information (they all produced some of it), we felt that it was important to begin the process with the goal of producing this auxiliary information, which will be invaluable for the higher-level routines using the stereo results. We foresee a time in the not-too-distant future when the calling routine will use the precisions, confidences, and annotations to actively control the sensor parameters for the next data acquisition step. For example, if the current stereo results contain a large region with no disparities and the image regions are quite dark, the controlling routine could open the irises or increase the integration time to reexamine these dark regions.

To assist in the analysis of the results, SRI developed two sets of routines, one to gather statistics and one to display the disparities in a variety of ways. Since we did not have ground truth for the distributed imagery, we could not compare the computed disparities with objective values. However, we were able to gather statistics on two of the three types of mistakes in which we were interested by outlining selected regions in the imagery and counting the occurrence of results/no-results within these regions. We made a distinction between the following three types of mistakes:

1. False Negatives: No disparities computed for points that should have results.

2. False Positives in Unmatchable Regions: Disparities reported for points that don't have matches in the second image, for example, points occluded in one image or points out of the field of view of one of the images.

3. False Positives in Matchable Regions: Incorrect disparities reported for matchable points.

By interactively outlining regions of occluded points, regions of points out of the field of view of the second image, and regions of points in the sky, we were able to directly measure statistics for the first two types of mistakes. In addition, we outlined regions corresponding to expected problems, such as dark shadows, foliage, and bland areas. In this way we could gather statistics on the behavior of the algorithms on these special problems.

The results of the first-phase evaluation can be summarized as follows:

- We were surprised by the completeness of the results. Even though the data set contained a wide range of imagery, including some sequences designed to stretch the analysis along specific dimensions, such as noise tolerance and disparity range, the stereo systems computed disparities for 64 percent of the matchable points. On eight image pairs selected to be the most appropriate for UGV applications, the techniques computed disparities for as much as 87 percent of the points. Although the missing points (and mistakes in the reported matches) could cause problems for vehicle navigation, this level of completeness is an indication that there is a solid basis for building a passive ranging system for an outdoor vehicle.

- For the UGV-related imagery the number of gross errors was relatively small, ranging from a few "spike" errors to small regions of mistakes. We estimate that these results contained gross errors of somewhere between 1 and 5 percent. Many of these errors would have to be eliminated for the data to be used directly for planning navigable routes.

- The stereo systems made different mistakes, most of which could be explained by their correlation patch size, search technique, or match verification technique. However, since they made different mistakes, there is a possibility of combining them in a way to check each other and fill in missing data.

- All the stereo systems could be improved significantly with a relatively small amount of effort. This was the first test of this type, requiring the analysis of a large data set, and it uncovered some weaknesses in the different stereo systems that can be corrected. One area to be considered is the development of pre-analysis techniques to automatically set key parameters, such as patch size and search areas (as Teleos did). The filtering of results could also be improved, eliminating matches that differ significantly from their neighbors (as SRI did).

- There were a few surprises, such as Teleos's successful solution to one set of image pairs from CMU that includes a carpet with a repetitive pattern on it. Teleos's large patches were able to detect large regions of subtle differences, which allowed recovery of the correct disparities.

Additional information about the JISCT evaluation and its results can be found in Appendix A [3].

III. Scene Sketch

The primary purpose of the UGV's sensor system is to generate three-dimensional structural descriptions of outdoor scenes to support vehicle navigation. Ideally, these descriptions would contain geometric information that describes the location, size, and shape of objects and semantic information that specifies the material types and semantic categories of key objects. In addition, this process should be a dynamic one that continuously updates its scene description. In practice, however, current scene -modeling techniques typically concentrate on geometric recovery and analyze each snapshot of a scene independently. This approach has two major limitations. First, the navigation system cannot use semantic information to plan its paths. For example, since it cannot distinguish between a rock and a bush, it has to plan a path around all detected "bumps." On the other hand, such a system would be quite happy to plan a path across a lake because it is flat! Second, this snap-shot-based system cannot use the results of previous analyses to focus its current processing on key areas, since it's required to build a complete map from each snapshot.

In this project, we developed initial techniques that overcome both of these limitations. We developed techniques for semantically labeling scene elements and techniques for incrementally refining descriptions over time. For semantic labeling, we are developing a set of specialists, each capable of identifying one class of objects, and an integrated framework for combining these results into a single sketch of the scene. Our specialists include a technique for identifying vegetation; one for locating ridges; and one for identifying the horizon. We call the integrated model a sketch because it is not designed to be a precise, unique partitioning of the scene into labeled objects. Rather it is a qualitative description of the key items in the scene that are important for navigation. Figure 1 shows an example of this type of sketch.

We expect a sketch to include such information as

Unmapped Areas: Regions of points that have not been reliably measured will be assigned labels indicating the reasons for their failures. Possible reasons include the following:

- They are out of the field of view of one of the cameras.
- They are occluded by another object in the scene.
- The matcher was unable to find consistent matches for them.
- The matcher found consistent matches for individual points, but the pattern of matches indicates a problem in the region.

Material Types: Regions will be assigned material types, such as dirt, vegetation, water, and sky, depending on their spectral distribution, texture, and shape.

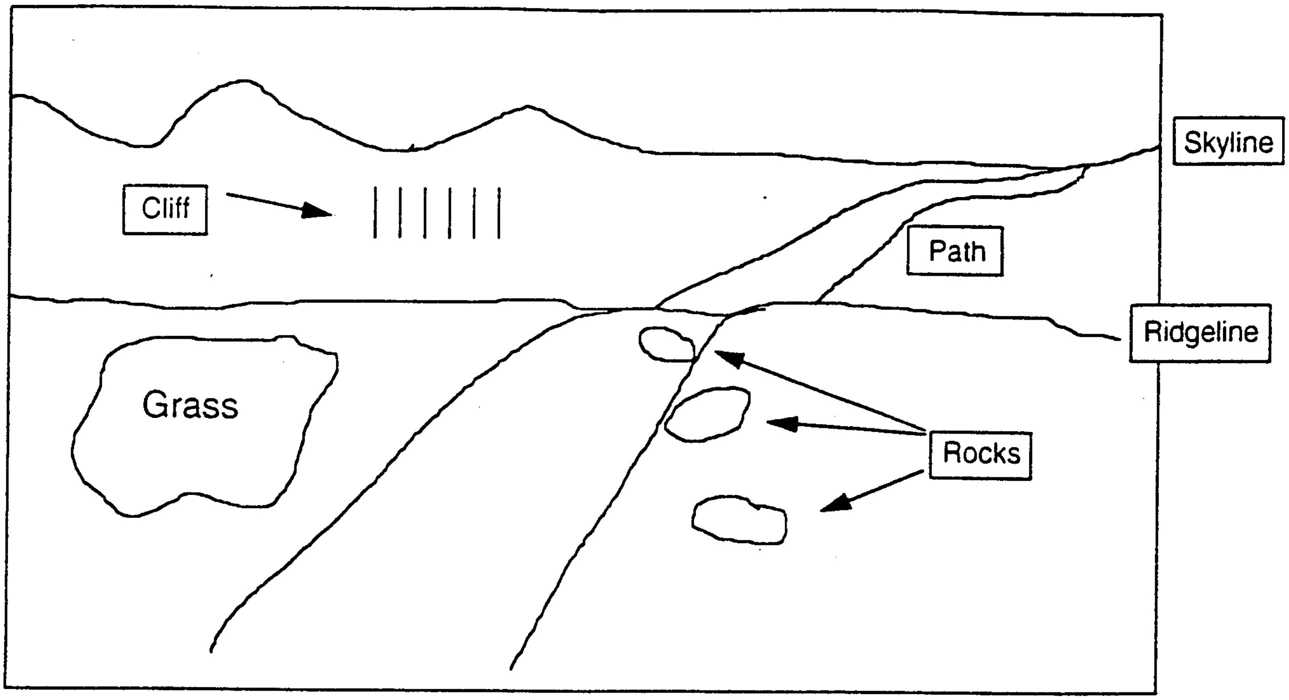


Figure 1: Local semantic scene sketch.

Semantic Labels: Recognized objects will be given labels, such as rock, bush, or road.

Higher-Level Geometric Descriptions: Key primitives, such as ridge lines, will be explicitly labeled.

We have developed an initial set of techniques to detect and classify some of these categories, including sky, vegetation, ridges, and problem areas. In the future, we plan to implement additional specialists and integrate them into our TraX System [2], which provides a method for constructing and transforming two-dimensional and three-dimensional representations over time.

IV. Spatiotemporal Filtering

One problem with current correlation-based stereo systems is that they apply smoothing or spatial aggregation operations that reduce the spatial and depth resolutions of their results. They use these operations to minimize matching mistakes, but they also have the side effect of reducing the resolution, which in turn increases the size of the smallest detectable obstacles.

To avoid this loss of resolution, we have developed an alternative technique for eliminating mistakes. Our approach filters out mistakes by checking the consistencies of multiple in(ter)dependent matches. The idea behind this strategy is the same as that behind the consistency checking techniques used in Hannah's left-to-right and right-to-left doublechecking [6], and INRIA's trinocular filtering [1].

Figure 2 shows the basic approach. Each arrow represents an independent match from one image to another. The system performs conventional stereo disparity estimation from left to right, and optical flow estimation from present to past. The disparity estimates are corroborated by performing an additional right to left stereo match and verifying that the left to right result is the inverse of the right to left one (the "left-right check"), as in [6] and [5]. Optical flow estimates are similarly corroborated by estimating the past to present flow field (the "forward-back check"). If these checks fail, the matches are marked as invalid. In addition, since an optical flow estimate is available for each pixel in both cameras, the spatiotemporal transitivity of two stereo estimates and two optical flow estimates can be checked to ensure that the four-sided "loop" of matches is consistent (the "spatiotemporal check"). If the spatiotemporal check fails, the confidence in the individual matches is decreased. Validity information is associated with what we term "pixel features," instead of "pixels" or "features," because the information is associated with individual pixels that are tracked over time.

This spatiotemporal matching strategy provides a natural way of integrating local depth images over time. As shown in Figure 3, our model of the scene is image-centered. Each pixel-feature has associated stereo disparity estimates, optical flow estimates and auxiliary information describing the numbers and types of consistency checks passed by the pixel-feature. An advantage of this type of scene modeling is that it avoids the need to explicitly compute the six degree-of-freedom transforms that relate one local three-dimensional model to another, as shown in Figure 4. Computing these transforms can be tricky when, for example, there is motion in the scene, as well as computationally expensive (e.g., see [7]). Our approach to integration differs from most other approaches, however, in that it does not provide a persistent *global* model of the scene. Rather it provides an image-centered model of the currently visible surfaces of the scene objects. As objects leave the field of view, they are lost. The information produced by our modeling system could be provided as data to a more conventional global modeling system.

A more complete description of our spatiotemporal consistency checking technique can be found in the reference in Appendix B [4].

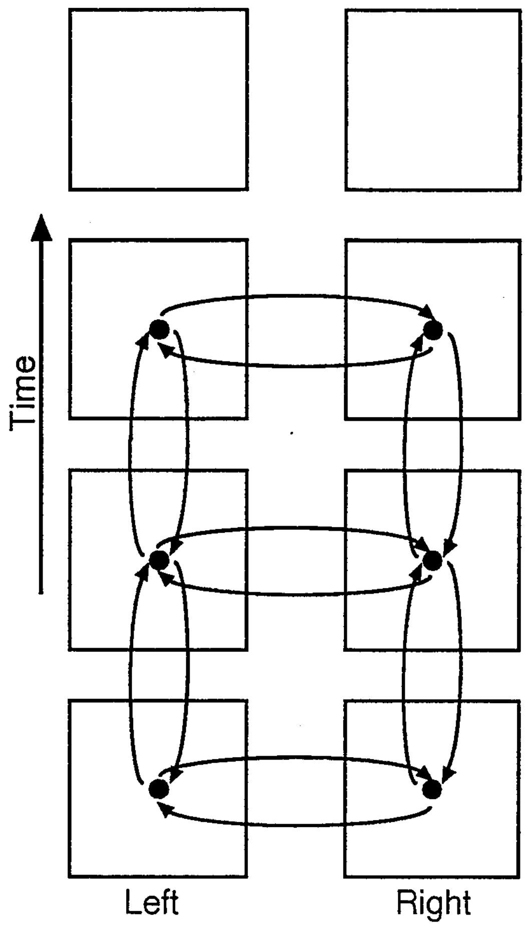


Figure 2: Spatiotemporal multiple-match consistency checking.

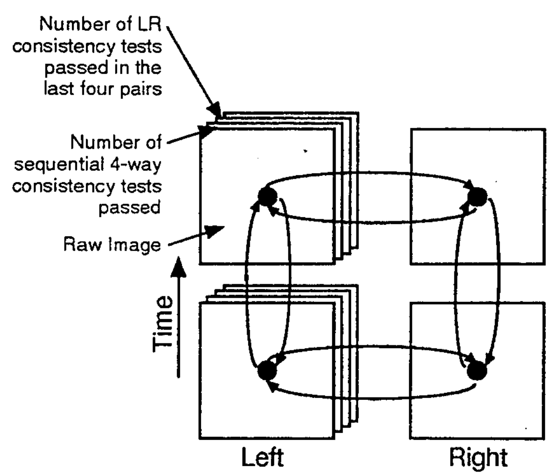


Figure 3: Image-centered modeling of the scene from a sequence of stereo pairs.

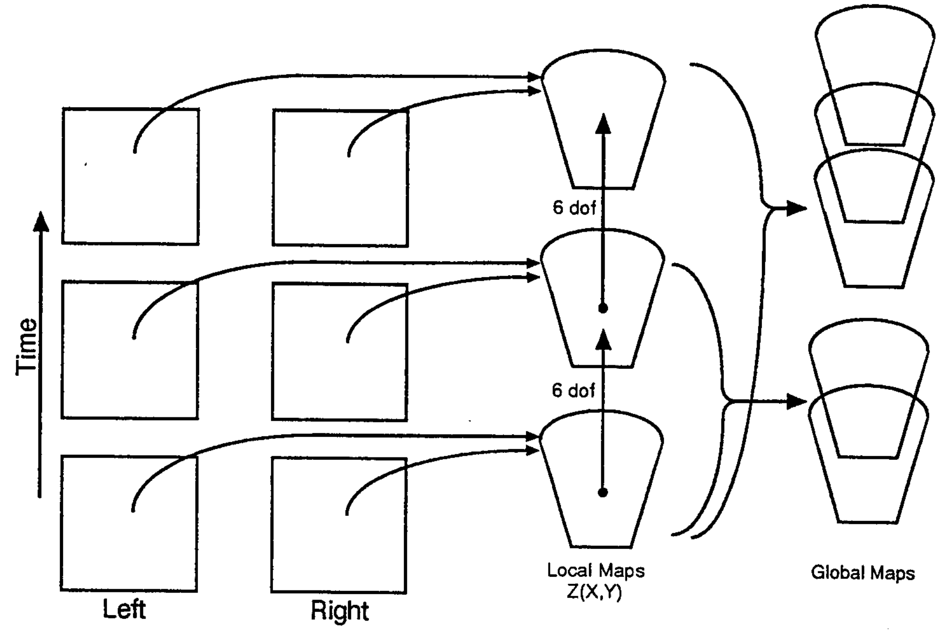


Figure 4: Global map formation from a sequence of stereo pairs.

V. FLIR Stereo

In this project, we have taken the first steps toward demonstrating the effectiveness of FLIR stereo for supporting night driving. We demonstrated that our conventional stereo techniques work well when applied to FLIR data. We installed a pair of infrared sensors on JPL's HMMWV vehicle and demonstrated real-time night ranging, and we gathered a 24-hour time-lapse sequence of infrared images to verify that there is sufficient contrast and texture throughout the day and night for stereo matching.

For our infrared experiments, we have used a pair of sensors, called Radiance I Cameras, produced by AMBER Corporation of Goleta, California. They are indium-antimonide sensors that measure radiation in the mid-IR range of 3 to 5 microns. The sensing chips have a resolution of 256 by 256 pixels and produce 12-bit intensity values. They provide analog output, which we have used primarily, and digital output, which we have implemented interfaces for and are currently in the process of integrating into our experimental system.

The digital interfaces are important for several reasons. First, they provide the purest form of data, because they avoid the corruption inherent in converting the data to a video format, such as NTSC, before digitizing it. Second, they provide one digital value per sensor element. Third, the digital interfaces provide 12-bit intensity values versus the 8 or fewer bits per pixel typically available from digitized video signals. Fourth, they provide a way to synchronize two of the sensors for stereo data acquisition. Since the AMBER sensors cannot be GENLOCKed together (like most conventional cameras), but can be started by a "start pulse," we can synchronize a pair by starting them together, and then reading the digital output. Since the digital output encodes one value per element in the 256 by 256 sensor arrays, the relative positions of the pixels within one array and from one array to another are fixed, as long as the sensors are rigidly mounted relative to each other.

In the future, we plan to complete the integration of the digital interfaces into our system, explore 12-bit correlation techniques, and then transfer the system to Martin Marietta for night-time demonstrations on its vehicles.

VI. Summary

In this project, we have made significant progress in achieving our goal of advancing the state of the art in passive scene modeling for UGV applications by

1. Participating on the tiger team that designed the sensing and control architecture for the vehicle.

2. Evaluating the strengths and weaknesses of stereo analysis applied to UGV tasks.

3. Identifying key areas for future research in stereo analysis and, more generally, in computer vision to support vehicle navigation.

4. Developing a scene-sketching framework for describing the semantic, as well as geometric, properties of a scene, which will significantly increase the information provided to a navigation path planner.

5. Developing a technique for increasing the resolution and robustness of passive range images by integrating stereo and motion analysis.

6. Developing FLIR stereo techniques to support night driving.

In the future, we plan to (1) develop additional specialists for detecting and classifying key navigational features, such as rocks, bushes, and ravines, (2) continue our development of FLIR techniques to support night driving, and (3) explore strategies for employing multiple stereo systems to provide both the wide field of view required to drive safely around obstacles and the high-resolution results required to detect small, but dangerous, obstacles.

References

[1] Ayache, N., and F. Lustman, "Fast and Reliable Passive Trinocular Stereovision," *Int'l Conf. on Computer Vision,* June 1987.

[2] R.C. Bolles, and A.F. Bobick, "Exploiting Temporal Coherence in Scene Analysis for Autonomous Navigation," *Proc. IEEE Int'l Conf. on Robotics and Automation,* Scottsdale, Arizona, pp. 990-996, May 1989.

[3] R.C. Bolles, H.H. Baker, M.J. Hannah, "The JISCT Stereo Evaluation," *Proc. Image Understanding Workshop,* Washington, D.C., pp. 263-274, April 1993.

[4] Bolles, R.C., and J. Woodfill, "Spatiotemporal Consistency Checking of Passive Range Data," *Proc. International Symposium of Robotics Research*, Pittsburgh, PA, October 1993 *(see Appendix A).*

[5] Fua, P.V., "A Parallel Stereo Algorithm that Produces Dense Depth Maps and Preserves Image Features," *Machine Vision and Applications*, 1991.

[6] Hannah, M.J., "A System for Digital Stereo Image Matching," *Photogrammetric Engineering and Remote Sensing,* Vol. 55, No. 12, pp. 1765-1770, December 1989.

[7] Szeliski, R., "Bayesian Modeling of Uncertainty in Low-Level Vision," *Int'l Jnl of Computer Vision,* Vol. 5, No. 3, pp. 271-301, December 1990.

Appendix A

"The JISCT Stereo Evaluation"
R.C. Bolles, H.H. Baker, and M.J. Hannah

in the Proceedings of the Image Understanding Workshop
Washington, D.C., April 1993

The JISCT Stereo Evaluation*

Robert C. Bolles, H. Harlyn Baker, and Marsha Jo Hannah

Artificial Intelligence Center, SRI International
333 Ravenswood Ave., Menlo Park, CA 94025
(bolles@ai.sri.com baker@ai.sri.com hannah@ai.sri.com)

Abstract

The results of the "JISCT" Stereo Evaluation (named after the five groups contributing imagery: JPL, INRIA (in France), SRI, CMU, and Teleos) are presented. The goals of this evaluation, which was the first phase of a multiphase evaluation process, were (1) to get an initial estimate of the effectiveness of current stereo techniques applied to Unmanned Ground Vehicle (UGV) tasks, (2) to identify key problems for future research, and (3) to debug the evaluation process so that it can be repeated with a larger group of participants. SRI collected 49 pairs of images, distributed them to the five participants, and received complete results from three groups — INRIA, SRI, and Teleos. SRI compared the results by interactively analyzing them and automatically gathering statistics.

We were surprised by the completeness of everyone's results. On the eight image pairs that we thought were the most representative of UGV tasks, the techniques computed disparities for as much as 87% of the points with only a few "spike" errors and some scattered regions of points without matches. Although the missing points (and mistakes in the reported matches) could cause problems for vehicle navigation, this level of completeness is an indication that there is a solid basis for building a passive ranging system for an outdoor vehicle. On the other hand, none of these techniques have "solved the stereo problem" — we selected a number of important areas for future research, including filtering out gross errors and handling the wide dynamic range of intensities common in outdoor imagery.

1 Introduction

Stereo analysis, which for a long time had been viewed as an interesting, but too-costly-to-be-practical technique, has emerged as a viable tool for realtime applications such as vehicle navigation. This has happened for two reasons. First, advances in hardware have made it practical to compute stereo matches "in real time." And second, advances in algorithm development have made it possible to correctly match large portions of outdoor scenes.

An important next step in the development and use of practical stereo systems is the characterization of their capabilities. Potential users, such as system integrators and automatic task planners, need to know their computational requirements, their speeds, their precision, their mistakes, and so forth, in order to model their behavior and reason about their use. With this in mind, SRI, JPL, and Teleos began a multiphase evaluation process last year within the Unmanned Ground Vehicle (UGV) Project. The first phase of that evaluation has been completed, and the second phase has begun. This paper describes the results of the first phase.

The overall plan for our complete evaluation process is to pursue a three-pronged approach, including analytic models, qualitative "behavioral" models, and statistical performance models. The analytic models would be used to estimate such things as the expected depth precision computable with a specific camera configuration. The qualitative models would be used to identify key problems for future research, for example, detection of holes, analysis of shadowed regions, and measurement of bland areas. The statistical models would be used to produce quantitative estimates of such key factors as the smallest obstacle detectable at a specified distance. SRI has taken the lead in the qualitative evaluation; JPL has taken the lead in the quantitative analysis.

For the qualitative analysis, we decided to start by examining a small number of techniques in order to debug the process, and then expand the evaluation to include a much larger set of participants. The goals of the first phase were to get an initial estimate of the effectiveness of current stereo techniques applied to UGV tasks and, from this, to identify key problems for future research.

One of the high-level guidelines we adopted was to develop and maintain an atmosphere of cooperation and constructive criticism among the researchers participating in the evaluation. Without this we would not be

*Supported by Advanced Research Projects Agency Contract DACA76-92-C-0003.

able to focus on our ultimate goal of producing a sequence of increasingly capable stereo systems. To help establish a cooperative atmosphere, we decided to concentrate on the positive aspects of each algorithm and highlight ways to strengthen existing techniques, realizing that they were developed for different domains and different applications. We also decided to share all the raw results with the participants so they could duplicate our analysis or develop their own.

For the first phase of the qualitative evaluation, SRI collected imagery from five groups. JPL, INRIA (in France), SRI, CMU, and Teleos (hence the name "JISCT" for the first evaluation phase); selected 49 pairs for analysis; converted them into a standard format; distributed the dataset to the five groups for processing, along with an extensive set of instructions; collected the results; characterized them; and finally distributed the results and the associated report to the participants.

We intentionally asked each group to process a large number of pairs (10 training pairs and 45 "test" pairs ... 6 pairs were in both the training and test sets; we made an administrative mistake on one of the test pairs, reducing the total to 44), because we wanted to force them to establish a standard algorithm that was automatically applied. As a result of this, there are now four groups around the world that can readily apply end-to-end stereo techniques to new data and compare their results. As part of the second phase we hope to expand this community to 10 or more groups. This process is opening up a new form of interaction within the computer vision community that we feel will help stimulate advances and reduce redundant development.

In the instructions to the participants, we asked each group to produce several results for each matched point in addition to its computed disparity. For each point we asked for an x and a y disparity, an estimate of the precision associated with each reported disparity, an estimate of the confidence associated with each match, and an annotation for each unmatched point, indicating why the technique could not find a match. Possible explanations for no match included "area too bland," "multiple choices," and "inconsistent with neighbors." Although none of the groups produced all this additional information (they all produced some of it), we felt that it was important to begin the process with the goal of producing this auxiliary information, which will be invaluable for the higher-level routines using the stereo results. We foresee a time in the not too distant future when the calling routine will use the precisions, confidences, and annotations to actively control the sensor parameters for the next data acquisition step. For example, if the current stereo results contain a large region of points without disparities and the image region is quite dark, the controlling routine could open the irises or increase the integration time to reexamine these dark regions.

Four groups returned results and write-ups to SRI — Teleos, SRI, and two from INRIA. One of the INRIA sets was from a technique that locates linear features and then matches these features. Since this technique reports only disparities along the matched edges, it was not possible to directly compare its results to the others. Therefore, we concentrated our analysis on the three correlation-based algorithms.

Each participating group analyzed its own results. In addition, Harlyn Baker and Marsha Jo Hannah of SRI analyzed the results from all the groups on all 44 pairs and wrote short reviews of them. In the full report [Bolles, Baker, & Hannah], their comments are included as appendices. These comments, plus the automatically compiled statistics, form the core of this evaluation.

Initially, we were a little reluctant to compute and publish statistics that may be taken out of context. On the other hand, statistics, if reported with sufficient caveats, can provide a convenient basis for comparing techniques. In this paper, we summarize the qualitative results and quantitative statistics. The validities of both are limited by the dataset, which implicitly defines the range of data for which the conclusions directly apply, and by the analyzers, who naturally focused on issues they were most interested in.

This paper is organized as follows. In Section 2, we briefly describe the key strategies and parameters of the three principal techniques, highlighting their similarities and differences. In Section 3, we describe our experimental procedure. In Section 4, we present the automatically gathered statistics, which we refer to as the believe-everything-they-tell-you statistics because they are based on the number of "reported" disparities in specified regions of the test data, not on the number of "correct" disparities. In Section 5, we summarize our qualitative analysis and briefly discuss open issues for future research. In Section 6, we conclude with an evaluation of the JISCT evaluation and make some suggestions for the next step in the evaluation process.

2 Technique Summaries

We evaluated three techniques, whose key aspects are highlighted below.

2.1 INRIA

This technique was originally implemented as part of a European space project to produce three-dimensional models of scenes containing rocks and sand. It is implemented in C on a Sun. A similar technique is implemented on a Connection Machine (by Pascal Fua) at SRI. Key aspects are

- The algorithm computes a disparity for every pixel in an image by matching patches (usually 11x11 pixels) at one or two image resolutions, independently.

The basic algorithm "INRIA-1" matches only at one resolution.

- The technique uses an approximation to normalized correlation, referred to as C5, because it can be implemented efficiently using a sliding computation of the basic sums.
- The algorithm searches only along epipolar lines, which are assumed to be horizontal.
- The algorithm expects a range of disparities to be specified for each image pair to be analyzed.
- The technique verifies all matches by independently matching patches from the left image in the right image and patches from the right image in the left image. If the match for a patch from the left image is not mapped back to within a pixel of its location in the left image, the point is not assigned a disparity.
- The technique computes a subpixel location for each match by fitting a second-order curve to the correlation values surrounding the best match.
- After computing disparities for as many pixels in the left image as possible, the algorithm filters out isolated matches by morphologically shrinking the regions of matches. It typically shrinks the regions three times, grows the result three times, and then ANDs this result with the original image of results. This process can erase regions as large as 6x6 pixels.
- The algorithm computes a confidence value for each disparity by differencing the heights of the two highest matching peaks.
- The technique estimates the precision of a disparity value by fitting a Gaussian to the matching peak, using its standard deviation as the precision measure.
- The technique does not attempt matches near the edges of an image.
- The second set of results provided for this evaluation often was produced by matching at two image resolutions and picking the highest resolution for which there was a valid match.

2.2 SRI

This stereo system has evolved over 20 years, beginning with early Martian Rover research, migrating into the aerial mapping domain, and now coming back to ground-level analysis. Its goal has been to produce a set of high-quality matches from a wide range of (possibly uncalibrated) imagery. The algorithm is a multistage process that uses one matching technique to get a few solid matches at high-information points, and then uses these matches to guide another matching technique, whose results become anchors for yet another technique, etc, with culling of mistakes occurring at many levels. At each stage, the algorithm acquires more supporting matches to suggest limits for the disparity search, so the algorithm can attempt to match points that have less "interesting" information, using less hierarchy. For this evaluation, code was added to produce "dense" matches; this included stages that grow regions of matches around previously matched points, and fill in a regular grid of matches. In total, the standard algorithm for this evaluation involved seven stages of matching and three filtering steps. The algorithm is implemented in C on a Sun; speed has not been a priority.

Some key aspects are

- The algorithm applies a version of hierarchical matching for each point that it analyzes. At the early stages of the process, it uses all available image resolutions, starting at the coarsest, using the match found at that level to predict the location of the match at the next finer level, then refining it, and so forth. At the final stage, where the dense grid of points is computed, the algorithm uses only one or two levels.
- At each image resolution (level), the algorithm does a two-dimensional search near the epipolar line and then hill-climbs around the best match. The epipolar lines can be at any angle in the second image, and if there is no camera model (due to bad matches at early stages, or because the camera isn't modelable by a pinhole camera), the algorithms search over areas—(dx,dy) boxes—defined by surrounding matches.
- The algorithm uses normalized cross correlation (correcting for a linear intensity change from image to image) on 11x11 patches typically. Later stages, such as the region-growing step, can use smaller patches. The final match includes a subpixel estimate of the disparity, computed by fitting two parabolas to the nearby correlation values.
- Each match from one image to another is verified by applying the same technique to match back into the original image. If the return match is not within a pixel of the original point, the match is discarded as unreliable.
- The algorithm applies several other "filters" to weed out mistakes, including a threshold on interest value, thresholds on relative and absolute correlation values, tests for matches outside an image, and tests for unusual disparity values within a region of the image.
- Later stages of the algorithm use previously computed disparities in the neighborhood of a new point

to be matched, to specify the range of disparities to be considered. The neighborhoods are typically large, beginning at 1/4th of the image area, and gradually reducing to 1/64th of the image for this experiment. This technique assumes that the scene is composed of relatively large continuous surfaces.

- Since a confidence for each match was requested for this experiment, one was supplied by computing the ratio of the correlation value to the autocorrelation threshold.

2.3 TELEOS

This technique has been designed for efficient implementation and recently has been geared toward active vision in which the basic stereo process matches 100 to 200 selected points in a 1/30th of a second. It is implemented on a combination of two special boards and a Datacube system. For this evaluation, however, the hardware was not available and so a Lisp version of the algorithm (running on a Lisp Machine) was used. Some key aspects are

- The algorithm uses large correlation windows (ranging from 24x24 to 96x96 pixels).
- The algorithm computes binary correlation values from the Laplacian of Gaussian of the original images.
- The algorithm analyzes the data only at one resolution. It automatically selects the size of the convolution operators by analyzing the peak shapes of matches at 25 points in each new image pair. It selects the smallest window size that produces a significant difference between the heights of the top two highest peaks.
- At each point in the image, the algorithm starts with the disparity computed for the neighboring pixel and tries to locate a match at a similar disparity. A serpentine search, which analyzes the first row from left to right, the second row from right to left, and so forth, is used in order to reduce the computation time on the Lisp Machine.
- The algorithm searches off the epipolar line for the best match.
- The algorithm also examines the effect of skewing the patch being matched. It analyzes skews ranging from -.5 pixels per line to +.5 pixels per line. This analysis is applied only at the end of the search when the best match has been selected.
- The algorithm estimates a subpixel disparity value by fitting a quadratic function to the best peak.
- The algorithm does not try to match points near the edges of an image.

3 Experimental Procedure

The goal of this initial evaluation was to produce a qualitative characterization of the capabilities of current stereo techniques applied to UGV tasks. The intent, as stated in the instructions distributed to each participant, was to produce a description such as the following:

> On the 44 image pairs in the database our techniques correctly measured disparities to 65% of the points on the ground and 40% of the points on obstacles, such as trees, bushes, and rocks. The top five problems for our techniques were dynamic range, holes, bland areas, repeated structure, and poor range resolution. We estimate that these problems occur in the UGV scenarios with frequencies of ...

The idea was to produce a characterization that would focus future work on key UGV problems.

Our basic approach to developing this type of characterization was to apply the techniques to a large dataset, visually display the results in ways to highlight unusual events, gather basic statistics, and where possible, summarize our observations in descriptions that link observed behaviors to aspects of the techniques.

To start the process, SRI compiled a database of 49 image pairs from JPL, INRIA, SRI, CMU, and Teleos. We converted the images into a standard format and then distributed them to the five contributing groups for analysis. The groups were instructed to use 10 pairs as a training set, "freeze" their algorithm, and then process the whole set of 45 pairs. Results and commentary from four stereo systems were returned to SRI — Teleos, SRI, and two from INRIA. One of the INRIA sets, using edge-based feature analysis, could not easily be compared with the others. We concentrated our analysis on the three correlation-based system results.

To assist in the analysis of the results, SRI developed two sets of routines, one to gather statistics and one to display the disparities in a variety of ways. Since we did not have ground truth for the distributed imagery, we were not able to compare the computed disparities with objective values. However, we were able to gather statistics on two of the three types of mistakes that we are interested in by outlining special regions in the imagery and counting the occurrence of results within these regions.

We made a distinction between the following three types of mistakes:

False Negatives: No disparities computed for points that should have results.

False Positives in Unmatchable Regions: Disparities reported for points that don't have matches in the second image, for example, points occluded in one image or points out of the field of view of one of the images.

False Positives in Matchable Regions: Incorrect disparities reported for matchable points.

By interactively outlining regions of occluded points, regions of points out of the field of view of the second image, and regions of points in the sky, we were able to directly measure statistics for the first two types of mistakes. In addition, we outlined regions corresponding to expected problems, such as dark shadows, foliage, and bland areas. In this way we could gather statistics on the behavior of the algorithms on these special problems.

As part of the initial instructions we asked each group to extend its algorithm to produce an image of annotations that summarizes the result of the analysis, pixel by pixel. At each pixel we asked for a code from the following list:

```
 0:  no match attempted
 1:  matched fine

   NO MATCH BECAUSE
 2:  too bland, no information to key on
 3:  low match value (e.g., correlation value)
 4:  multiple choices (ie, repeated structure)
 5:  back-match inconsistency
 6:  point out of camera's field of view
 7:  point occluded by an object in the scene
 8:  point too far off the epipolar line
 9:  point inconsistent with neighbors
10:  other
```

The reason for requesting these codes is to encourage future algorithms to provide this additional information, which can be used by the higher-level vision techniques to decide what should be done next. For example, if no results are reported for a region directly ahead of the vehicle and the region is too bland and very dark, one option might be to open the irises on the cameras (or increase the integration time) in order to see into the dark area.

INRIA reported codes of 1 and 10; SRI reported all codes except for 4 and 7; and Teleos reported codes of 0, 1, 2, and 3. Therefore, we were able to count the number of matches attempted in each region and the number of disparities reported.

To estimate the frequency of incorrectly reported disparities (the third type of mistake), we either compared them to interactively selected values or located an aberration in the local pattern of disparities when they were displayed on the screen. We experimented with a variety of display techniques, including displaying the disparities as color-coded dots in stereo, heights above a three-dimensional "ground" plane, and disparity-displaced vertical lines. We are continuing to look for better ways to display three-dimensional results, because most current techniques encourage the human eye to "smooth over" differences, making the results look better than they actually are.

4 Statistics Summary

The statistics that we refer to as believe-everything-they-tell-you statistics are based on the number of reported disparities in specified regions of the test data. These statistics do *not* distinguish between "correct" and "incorrect" disparity values, just reported values and unreported values. They do, however, provide enough information to estimate three important quantities, the number of false negatives (matchable points that were not assigned a disparity), the number of false positives occurring in unmatchable regions, and the number of matchable pixels that were assigned disparities.

To help focus attention of key areas of the test data, we interactively outlined regions in the left images of 20 of the 44 image pairs (see Figure 1 and Figure 2). One of the most important regions is what we called "matchable-data." It eliminates several types of points that do not have matches in the right image, including null bands that do not contain grayscale data (but are included in the images to fill them out to a standard size, such as 512 by 512 pixels) and pixels that are out of the field of view of the right camera. In the 20 images we examined, the percentage of unmatchable points ranged from 4.3% to 46.0% and averaged 12.3%.

The statistics were gathered by a program that counted the number of disparities (dx disparities) reported in the specified region (or the whole image, if that was appropriate).

Figure 3 shows the results on all 44 image pairs. Note that

- The dataset contains a wide variety of imagery; some of it is realistic (containing dirt roads and cross-country scenes) and some is designed to test the algorithms along one dimension, such as baseline and noise tolerance. Some of the imagery is even trick imagery (the shoe images from CMU).

- The numbers in parentheses after each group's name (along the top of the table) indicate the number of test pairs in the dataset from that group.

- The INRIA-2 results are in parentheses because different parameter settings were used for different image pairs. However, the usual change was for the technique to match at two spatial resolutions instead of just one, and then combine the results. If a second set of parameters was not tried for a pair, we left the entry blank and used the INRIA-1 results in our computation of INRIA-2's average.

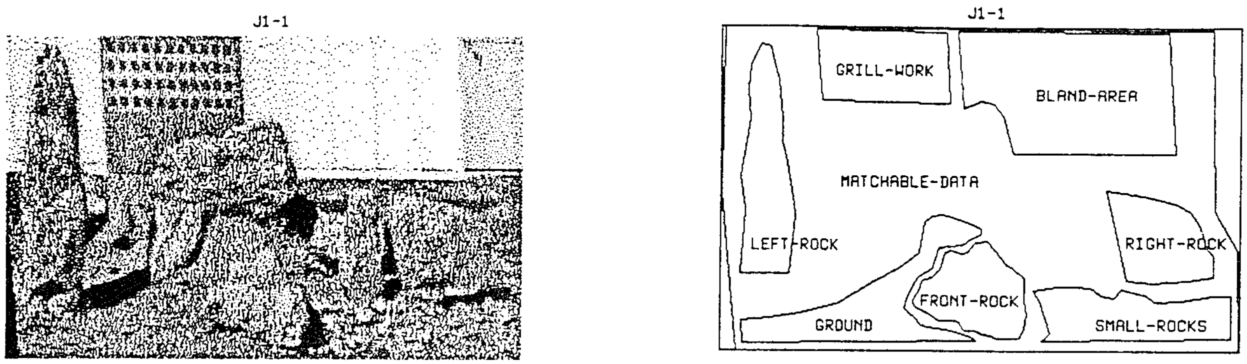


Figure 1: Interactively outlined, special-interest regions for the J1 image pair from INRIA.

- If we did not outline a "matchable-data" region for a pair, we used the full-image statistics in our computations. This reduces the effectiveness totals somewhat (possibly by as much as 7%).

Given the diversity of the data, we were pleased with the completeness of the results.

In order to examine the behavior of the techniques on typical UGV imagery, we selected the eight images from the dataset that were the most appropriate for UGV tasks and collected statistics on that subset. Figure 4 shows the results on these data. The INRIA-2, SRI-2, and Teleos-1 techniques performed well, computing disparities for 86 or 87% of the matchable points. Note, however, that these images did not contain difficult obstacles, such as holes, ditches, and small rocks—the obstacles were large rocks, bushes, and trees.

Figure 5 shows the results on the 17 large obstacles in the dataset. The techniques did an excellent job of detecting these objects, which stick up above the ground — they only had a little trouble in shadowed regions on them.

With respect to shadows, the techniques had a significantly harder time computing disparities for points in shadowed regions than in sun-lit regions. Figure 6 shows the results for points in shadows.

The techniques also had trouble with bland regions, as expected. Figure 7 shows the results on these areas. The techniques typically computed results around the edges of the regions — the larger the correlation windows, the more points were computed, because correlation windows naturally extend matches into the interior of bland regions by about half their diameter.

There are several potentially important problem areas that were not covered in this initial dataset, including holes, sand, small- to medium-sized rocks and bushes, reflective surfaces (water or windows), and moving objects. One of our goals for the second phase of this evaluation is to include examples of these problems.

5 Qualitative Analysis

We were surprised by the completeness of everyone's results. Even though the dataset contained a wide range of imagery, including some sequences designed to stretch the analysis along specific dimensions, such as noise tolerance and disparity range, the techniques computed disparities for 64% of the matchable points. On the eight image pairs that we selected as the most appropriate for UGV applications, the techniques computed disparities for as much as 87% of the points. Although the missing points (and mistakes in the reported matches) could cause problems for vehicle navigation, this level of completeness is an indication that there is a solid basis for building a passive ranging system for an outdoor vehicle.

The number of gross errors varied considerably from image pair to image pair. For most "realistic" images the number was relatively small, ranging from a few "spike" errors to small regions of mistakes. We estimate that for these images there were between 1 and 5% gross errors in the results. In many cases, the worst errors cluster into areas that are "breaking up" for one reason or another (usually poor information plus a poor "guess" for the disparity range); if we can "fix" these areas, then the remaining "spike" errors should be amenable to culling techniques. In any case, most of these errors would have to be eliminated in order for the data to be used directly for planning navigable routes.

The techniques made different mistakes, most of which could be explained by their correlation patch size, search technique, or match verification technique. However, since they made different mistakes, there is a possibility of combining them in a way to check each other and fill in missing data.

All the techniques could be improved significantly with a relatively small amount of effort. This was the first test of this type, requiring the analysis of a large dataset, and it uncovered some weaknesses that can be corrected. One area to be considered is the development of preanalysis

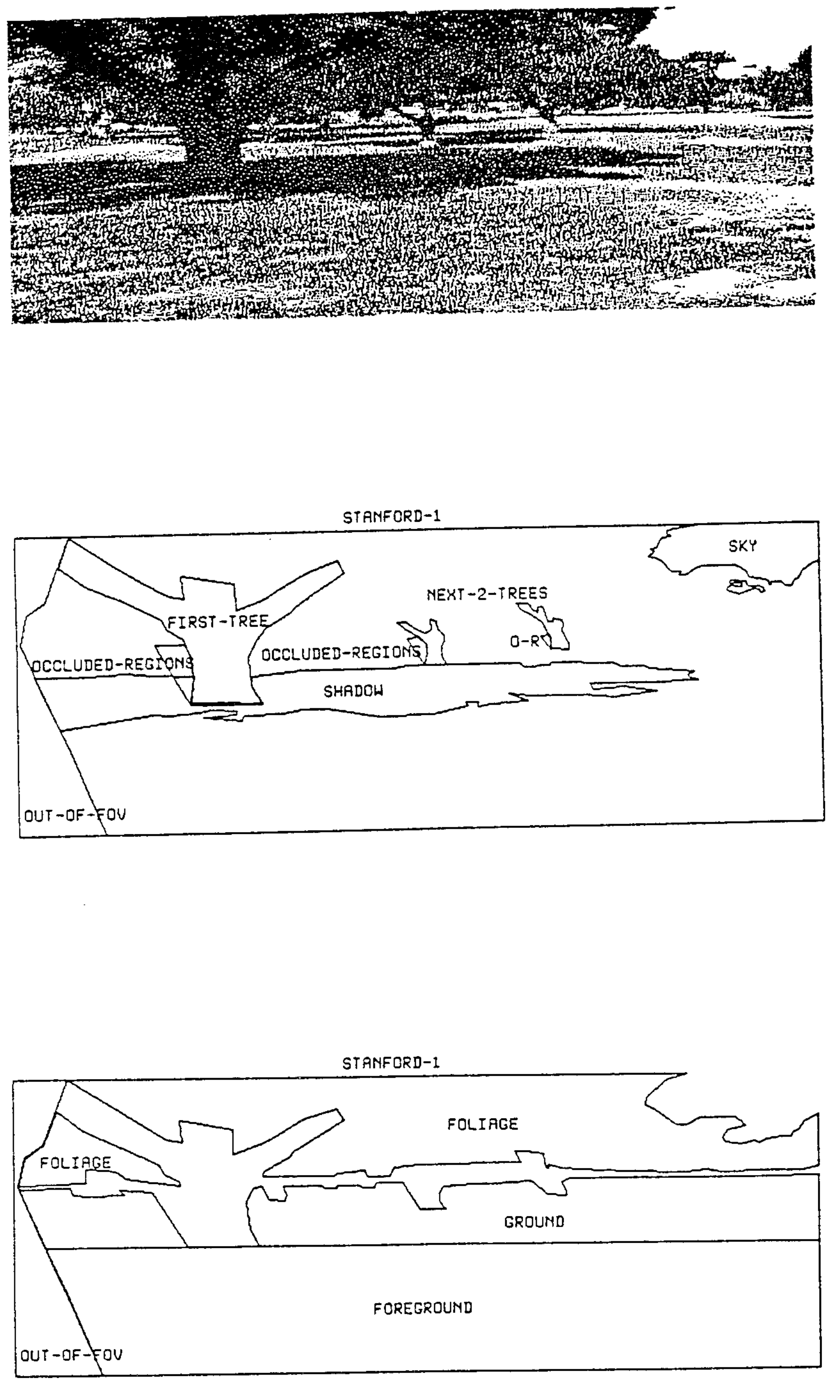


Figure 2: Special-interest regions for the STANFORD image pair from SRI.

	JPL(5)	INRIA(8)	SRI(15)	CMU(9)	Teleos(7)	Weighted Average
INRIA-1	63	66	42	89	35	57
(INRIA-2)	(92)	(75)	(60)	(70)	(50)	(67)
SRI-2	94	74	61	64	39	64
Teleos-1	95	81	45	87	77	71
Average	84	74	49	80	50	64

Figure 3: Percentage of "matchable" pixels assigned disparities on all 44 image pairs.

	Arroyo	EPI16	HMMWV1	HMMWV2	J1	Road	Rock	StanDbl	Average
INRIA-1	90	60	67	79	72	40	37	47	62
(INRIA-2)		(85)	(95)	(95)		(90)	(88)	(76)	(86)
SRI-2	91	72	94	94	73	97	94	78	87
Teleos-1	98	72	93	91	74	98	95	72	87
Average	93	68	85	88	73	78	75	66	79

Figure 4: Percentage of "matchable" pixels assigned disparities on the eight most representative pairs.

	Arroyo			HMMWV1			HMMWV2		Rock		
	Bush1	Bush2	Rock	LMound	Rock	RMound	Rock	Etc	LBush	RBush	Rock
Pixels:	56	68	10	130	18	106	26	839	174	105	23
INRIA-1	95	88	100	98	100	88	100	96	67	30	57
(INRIA-2)				(100)	(100)	(100)	(100)	(98)	(74)	(74)	(100)
SRI-2	91	82	90	99	94	96	96	95	68	64	96
Teleos-1	90	100	90	88	100	97	88	94	74	72	57
Average	92	90	93	95	98	94	95	95	70	55	70

	J1			StanDbl		Ball2	Unweighted
	RRock	FRock	LRock	1Tree	2&3T	Tennis-Ball	Average
Pixels:	70	70	98	276	37	145	
INRIA-1	100	100	100	63	86	92	85
(INRIA-2)				(92)	(100)	(94)	(95)
SRI-2	100	99	99	50	76	90	87
Teleos-1	98	89	100	94	62	86	87
Average	99	96	100	69	75	89	86

Figure 5: Percentage of "matchable" pixels on large obstacles assigned disparities.

	Stanford		StanDbl		Unweighted
	Shadow	1stTree	Shadow	1stTree	Average
Pixels:	215	140	448	276	
INRIA-1	40	71	38	63	53
(INRIA-2)	(84)	(96)	(73)	(92)	(86)
SRI-2	59	61	65	50	59
Teleos-1	1	82	29	94	52
Average	33	71	44	69	55

Figure 6: Percentage of "matchable" pixels in shadows assigned disparities.

Pixels:	iRoad2 Road 1077	J1 Bland 229	Ball2 Matchable- 3126	Ball4 Matchable- 3126	Unweighted Average
INRIA-1	12	22	56	39	32
(INRIA-2)	(86)		(72)	(62)	(61)
SRI-2	63	32	76	43	54
Teleos-1	83	10	86	84	66
Average	53	21	73	55	51

Figure 7: Percentage of "matchable" pixels in bland areas assigned disparities.

techniques to automatically set key parameters, such as patch size and search areas (as Teleos does). Another place for improvement is in the filtering of the results to eliminate matches that differ significantly from their neighbors (as SRI and INRIA do).

There were a few surprises, such as Teleos's successful solution to one set of image pairs from CMU that includes a carpet with a repetitive pattern on it. Teleos's large patches were able to detect large regions of subtle differences, which led to the correct disparities.

5.1 Technique-Oriented Summaries

No one of these algorithms has completely solved the stereo problem, although all can produce basically usable results on most reasonable imagery. Each has strengths and weaknesses—and very often an algorithm's strength on one dataset is its weakness on another!

INRIA's algorithms assume that the images are in epipolar alignment. This makes their searches more efficient, and keeps matches from wandering off of the epipolar lines (for instance, "climbing" the edges of tree trunks). However, when presented with nonepipolar imagery, INRIA-1 fell apart; INRIA-2 did better, but had a persistent problem, producing rough disparity contours, which are apparently due to the way the pyramid was handled. The low-resolution results were simply zoomed-out using pixel replication. This epipolar line constraint also limits the usefulness of INRIA's algorithms on imagery from nonpinhole cameras.

SRI's algorithm mostly disregards the epipolar constraint. Consequently, it had no particular problems handling nonepipolar imagery. However, it failed to match many of the very smooth tree edges in the EPI sequence, probably because its matches "slid" up the linear sides of the trees.

INRIA's algorithms search the entire width of the epipolar line. This helped them to do well on some datasets, but when the ground texture was ambiguous, their technique tended to return no match because of multiple choices.

SRI's algorithm depends on early matches to "set the context", so that later searches for matches can be confined to the disparities in that neighborhood. When there is enough global texture for the initial matches to give a good sampling of the disparities, this works well, enabling SRI-2 to produce ground plane matches where the others couldn't. However, when lack of foreground detail keeps SRI-2 from having the right initial matches, it fails to match, or finds random mismatches.

Teleos's algorithm uses very large windows dynamically skewed to accommodate tilted planes. This causes it to do well on some ground planes where it was able to disambiguate the pattern through minor variations, but not on others where the ground plane tilt was out of the allowed range of skewing. Of course, these large windows also cause it to have problems with any scene containing depth discontinuities—it either finds no match, or tries to blend the foreground object into the background objects, or widens the foreground object out onto the background. In addition, Teleos-1's scanning heuristic creates some rather peculiar artifacts— extending objects in opposite directions on alternate scan lines. However, its ability to "see" into low-contrast situations is very good.

The Teleos system, with its large correlation windows, also produces smaller range images, because it limits matching to areas where the full correlation patch is within the image. In an active vision system, the sensors could be reoriented to center objects of interest that may initially appear on the boundary of an image.

Both INRIA's and SRI's algorithms use fairly small windows. This removes much of the need for window skewing and warping, although on extremely tipped planes, warping would be helpful. INRIA-1, INRIA-2, and SRI-2 all do better on tilted planes if the information is slightly "fuzzy". These algorithms don't do nearly as well in the presence of man-made ambiguous patterns.

SRI's algorithm tends to leave more holes in the data—low-information places that it refuses to try to match, ambiguous places where it can't backmatch successfully, or error matches that it has detected and removed. This gives the data a "lacey" appearance, and it should probably be followed by an interpolation step, to fill in these problem areas. (The SRI technique is capable of interpolation, but it was not used in this evaluation.) SRI-2 often leaves a nice band of no-matches outlining depth discontinuities, where one doesn't really want separate objects "smoothed" together. SRI-2 also often refuses to match areas like the sky, which technically don't have a match.

None of the algorithms currently distinguishes between good image data and the "null data" areas caused by image digitization, reprojection, and so forth. This can lead to rather peculiar mismatches around these areas of null data. All of the algorithms should add the ability to accept a mask telling what parts of the image not to try to match. Better yet would be a preprocessing step to construct these masks automatically.

It was interesting to see how much better all of the algorithms did on the imagery taken by JPL than on the SRI imagery. A major factor is the unusual aspect ratio of the SRI imagery caused by digitizing individual fields, since the vehicle was moving fast enough to show a significant difference between fields. JPL's imagery was taken while the vehicle was standing still. Other differences that may have contributed include image contrast, epipolar geometry, and look angle (SRI's cameras were looking far forward, whereas JPL's were looking down a bit more). We note that the exchange of imagery can help in algorithm development by avoiding inadvertently "tuning" one's algorithm to one's particular style of imagery.

5.2 Open Research Problems

After examining the results from this dataset, we have selected the following topics for future research in the area of low-level passive range sensing:

1. Filtering out gross errors caused by erroneous matches.

2. Handling the wide dynamic range in intensities common in outdoor imagery, from dark shadowed regions up to specularities off shiny surfaces.

3. Handling the large range in adjacent disparities arising from narrow foreground obstacles.

4. Adjusting algorithm parameters automatically to properly handle different image regions, such as bland areas and texture regions.

5. Detecting multiple matches and selecting the correct one, possibly by analyzing multiple images.

6. Providing validation and confidence estimation mechanisms.

7. Detecting occlusion edges and reporting accurate depths on both sides of them.

8. Detecting and characterizing small- to medium-sized obstacles, such as rocks and bushes.

9. Detecting "negative" obstacles, such as holes and ditches.

Although the JISCT dataset did not include examples of the last two areas, they are clearly important for cross-country navigation.

6 Conclusion

As a result of this phase of our stereo evaluation, we can make a few general observations and develop a few ideas for the project's next phase.

First, the time is right for evaluation. If promising computer vision techniques, such as stereo analysis and road following, are to make the transition from the research laboratory to practical systems, their characteristics will have to be well enough documented that system engineers can understand them and predict their behavior. We view this evaluation as the first tentative step toward developing this type of characterization.

Second, evaluations of this type require a significant effort. To give an idea of what is involved in such an evaluation, SRI did the following: gathered imagery from five groups, converted it into a standard format, designed the experimental procedure, distributed the imagery to the participants, collected the results, converted them into a uniform format (correcting for a few mistakes in the original specifications), developed visualization routines, used these routines to interactively examine all the results, developed statistics gathering routines, applied these routines to the results, wrote the report, and finally distributed the report and copies of everyone's results.

Third, ideally an evaluation of this type should be performed periodically to provide estimates of the relative improvements of the techniques.

6.1 Critique of the JISCT Evaluation

Some things that were done correctly:

- We developed a cooperative attitude among the participants. This was the first time our community had tried establishing an ongoing evaluation process and we knew that we'd make mistakes. We also knew that the participants have their egos involved in their systems, and we wanted to emphasize the constructive aspects of comparing techniques.

- The experimental procedure was almost right. The idea of distributing a large number of stereo pairs, using some for a training set, freezing the "official" algorithm, and then applying it to 45 test pairs is correct. The large number of pairs virtually forced the groups to implement an automatic technique, which they could apply to any image pair. As a result, there are now four systems around the world that can be easily tested on new imagery.

- The idea of asking for precision estimates, confidence estimates, and annotations was correct. Although no group produced them all, future systems will be expected to because this information is so important for higher-level users of the results.

- The basic idea of sharing data from several groups was good because applying the algorithms to this diverse set of images brought to light several implicit and explicit assumptions and parameters in the algorithms.

- Since any evaluation of this type can only include a limited set of imagery that attempts to cover all possible dimensions, the idea of including several small controlled experiments worked well. For example, the set of images from Teleos explored the ability of the algorithms to handle increasing noise; the SRI EPI sequence tested a range of baselines.

Some things that should be changed:

- The lack of ground truth significantly limited the types of automatic "objective" evaluations possible. Ground truth is expensive, but there is no substitute for assessing quantitative issues.

- For this initial phase we built our dataset primarily from existing data. In the future we need to gather data that is more realistic and appropriate to the task. In particular, for UGV tasks, the data should be from the demonstration sites and include examples of the common "obstacles," such as ruts, bushes, rocks, ditches, and water. Future datasets should also include sequences of images and trinocular data, not just individual pairs.

- The whole process took too long (almost a year). Techniques can change faster than that. To be relevant, the results should be returned within a few months. This turnaround time is more possible now that we have been through the process once and have developed routines for analyzing the data.

- More auxiliary data (e.g., calibration information) should be supplied with the dataset. Some techniques rely on this information to reduce search and set key parameters. Also, it will generally be available in most applications.

6.2 Plans for the Next Evaluation Phase

We plan to include three types of imagery in the next dataset: demonstration-related pairs and sequences, a few image-intensified pairs, and some synthetic pairs that are less artifactual than previous ones. One of our goals for this phase is to explore more rugged off-road scenes, including deep ruts, tall grass, and ditches, so we are including several examples of each in the new dataset. The image-intensified data will provide our first look at applying our techniques to night-vision-type imagery. The synthetic data is formed from real pairs by modifying a set of computed disparities, and then forming a new right image based on these disparities. This data, although still not completely realistic, is significantly better than previous versions and provides complete ground truth.

We plan to distribute the dataset to 10 or 15 research groups for analysis. After debugging the process, we are in a position to open up the evaluation to include a wider group of participants.

Reference

Bolles, R.C., H.H. Baker, and M.J. Hannah, "The "JISCT" Stereo Evaluation," SRI International Report, January 1993.

Appendix B

"Spatiotemporal Consistency Checking of Passive Range Data"
R.C. Bolles and J. Woodfill

presented at the International Symposium on Robotics Research
Pittsburgh, Pennsylvania, October 1993

Spatiotemporal Consistency Checking Of Passive Range Data

Robert C. Bolles, SRI International, bolles@ai.sri.com
John Woodfill, Interval Research, woodfill@interval.com

Abstract

A spatiotemporal technique for consistency checking and cross-temporal integration of stereo range results is presented. This technique is designed as part of a passive ranging system whose goal is to produce range images with as high a resolution as possible in order to support the detection of as small objects as possible. The approach is to minimize the application of smoothing and spatial-aggregation operations, which reduce resolution, and to employ a set of multiple-match consistency checks over space and time to filter out mistakes. We present the basic approach, describe two implementations of it (one of which is a research-oriented system that runs on a Connection Machine and the other of which runs on a Sparc10 and provides real-time feedback for SRI's indoor robot, Flakey), present an initial characterization of the effectiveness of the technique, and conclude with ideas for future work.

1 Introduction

The ultimate goal of this research is to develop passive range sensing techniques that provide the spatial and depth resolutions required to detect small, but dangerous, navigation obstacles, such as holes and medium-sized rocks. Current feature-based techniques provide insufficiently dense results to detect these scene elements while correlation-based stereo and motion systems typically smooth over them. Current correlation systems apply smoothing or spatial aggregation operations in three places, in image preprocessing (e.g., performing Gaussian smoothing to reduce image noise), in matching (e.g., using large correlation windows to provide an ample statistical footing), and in post processing (e.g., eliminating results that differ significantly from their neighbors). These operations reduce the chance of errors, but they also dramatically reduce the resolution of the results. Our approach, on the other hand, is to minimize the use of spatial aggregation in order to maximize the resolution and to provide an alternate set of filtering techniques to detect mistakes.

We propose a class of filtering strategies based on checking the consistences of multiple in(ter)dependent matches. The idea behind this class of strategies is the same as that behind the consistency checking techniques used in Hannah's left-to-right and right-to-left doublechecking [7], and INRIA's trinocular filtering [1].

In this paper we introduce a natural extension of these techniques to include spatiotemporal consistency checking. Figure 1 shows the basic approach. Each arrow represents an independent match from one image to another. The system performs conventional stereo disparity estimation from left to right and optical flow estimation from present to past. The disparity estimates are corroborated by performing an additional right to left stereo match and verifying that the left to right result is the inverse of the right to left (the "left-right check"), as in [7] and [6]. Optical flow estimates are similarly corroborated by estimating the past to present flow field (the "forward-back check"). If these checks fail, the matches are marked as invalid. In addition, since an optical flow estimate is available for each pixel in both cameras, the spatiotemporal transitivity of two stereo estimates and two optical flow estimates can be checked to insure that the four sided "loop" of matches is consistent (the "spatiotemporal check"). If the spatiotemporal check fails, the confidence in the individual matches is decreased. Validity information is associated with what we term "pixel features," instead of "pixels" or "features," because the information is associated with individual pixels that are tracked over time.

This spatiotemporal matching strategy provides a natural way of integrating local depth images over time. As shown in Figure 2, our model of the scene is image-centered. Each pixel-feature has associated stereo disparity estimates, optical flow estimates and auxiliary information describing the numbers and types of consistency checks passed by the pixel-feature. An advantage of this type of scene modeling is that it avoids the need to explicitly compute the six degree-of-freedom transforms that relate one local 3D model to another, as shown in Figure 3. Computing these transforms can be tricky when, for example, there is motion in the scene, as well as computationally expensive (e.g., see [14]). Our approach to integration differs from most other approaches, however, in that it does not provide a persistent *global* model of the scene. Rather it provides an image-centered model of the currently visible surfaces of the scene objects. As objects leave the field of view, they are lost. The information produced by this type of modeling system could be provided as data to a more conventional global modeling system.

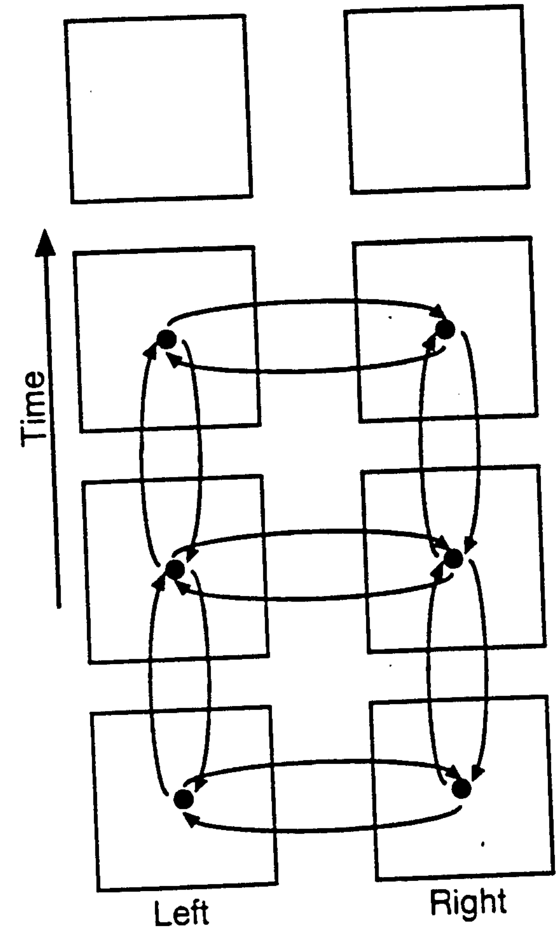


Figure 1: Spatiotemporal multiple-match consistency checking.

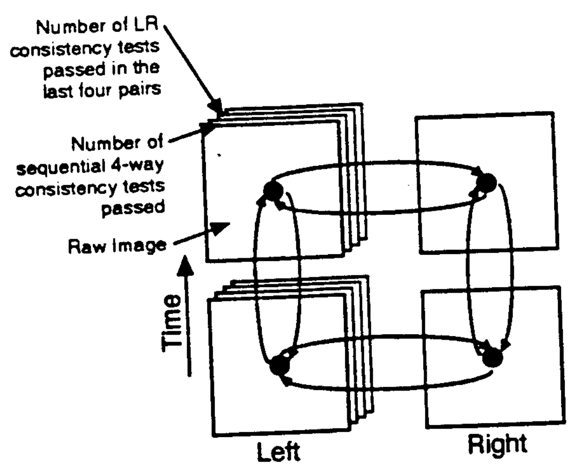


Figure 2: Image-centered modeling of the scene from a sequence of stereo pairs.

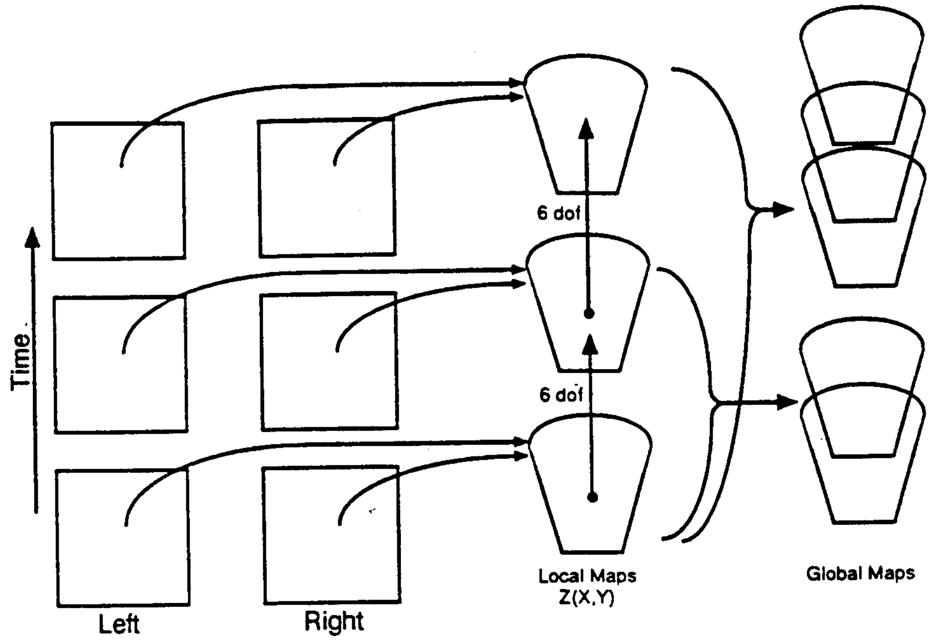


Figure 3: Global map formation from a sequence of stereo pairs.

The remainder of this paper is organized as follows. In Section 2 we describe several techniques for attaining credible stereo disparity estimates and discuss how they relate to our consistency-checking approach. In Section 3, we describe an experimental system that makes use of spatiotemporal consistency checking, and present examples of its use within a sensor system to support cross-country navigation. In Section 4, we briefly describe a second implementation of these ideas in a real-time system for SRI's indoor robot, Flakey. And finally, in Section 5, we draw some conclusions and discuss ideas for improvements and future work.

2 Credible Stereo Estimates

Producing stereo disparity estimates is conceptually simple: for each pixel in one image, find the best match (according to some metric) in the other image. However, attaining credible disparity estimates is a more complex problem, to which there are many approaches.

System Engineering: Constrain the environment and/or the sensing system to eliminate or minimize the impact of as many factors as possible. For example, restrict the lighting to be from a well-modeled source. This approach also includes techniques to reduce the search region by calibrating the cameras relative to one another. Smaller, more focused search regions reduce the computation required to find a match, and more importantly, reduce the chance of conflating similar appearing but distinct points.

Selective matching: Prefilter the images to select the best points to be matched. Only attempt matches for features that are highly distinctive. Hannah for example does not attempt matches for points in areas with low local variance [8].

Multiple cameras: Use N calibrated cameras to obtain better depth estimates. The correlation surfaces for $(N-1)$ matches can be merged at each point, and the best "global" match can be selected. To merge the results, the raw disparities are converted into a common representation. Moravec implemented a system of this type by sliding a camera to nine different positions along a bar [11]. Kanade et al. have developed a technique of this type, using an inverse disparity representation to integrate multiple results [9]. They have applied their technique to camera configurations with three or more cameras in a line. Kanade is currently developing another version of this type of system using seven cameras arranged in an L-shaped configuration.

Match evaluation: After computing a set of matches, eliminate the "incorrect" ones. This can be done in many ways. For example, compare a pixel's disparity to the disparities of its neighbors and discard it if it's significantly different from all of them. Or compare the results of two different matching methods; if they disagree at a pixel, mark it as untrustworthy.

Most stereo systems use a combination of these techniques to maximize their chances of producing dense and reliable results. In this paper we concentrate on match evaluation techniques.

A number of techniques have been used to evaluate stereo matches, some with more success than others. Probably the first technique to be tried was simply a threshold on the correlation value computed for the best match. Unfortunately, although this value is related to the validity of a match, the range of values associated with correct matches significantly overlaps the range of values for incorrect matches. This means that for any reasonable threshold there is a large number of correct matches labeled as "bad" and incorrect matches labeled as "good." This situation occurs because the correlation value is a function of several interrelated factors, such as the change in perspective from one viewpoint to another, the reflectance properties of the scene feature, the amount of noise in the images, the overall change in intensities from one image to another, and the number of nearby features that have a similar appearance.

Given this complex interrelationship of factors, many approaches to evaluating stereo matches have been tried:

Factor Analysis: Develop computational models of as many of the factors as possible, implement a matching technique that estimates the parameters of these models, and then set thresholds on these parameter values. For example, Baltsavias has implemented an iterative matching technique that can estimate such things as the surface normal of a scene feature and the gain and offset between two image windows [2].

Correlation Surface Analysis: Examine the correlation surface near the best match and compute properties such as the height of the highest peak relative to the "background" or the number of alternative matches (i.e., significant peaks) within a certain distance of the highest peak. For example, Nishihara has implemented an evaluation procedure that only accepts a match if its peak is significantly higher than the second best one and the width of the peak is greater than some threshold [3].

Object Surface Constraints: Invoke constraints derived from assumptions about the types of surfaces in the scene. For example, Pollard et al. apply a threshold of one pixel on the disparity gradient between two measured points [13]. As another example, Hannah has implemented an outlier rejection process that examines a large region around each result and marks points as inconsistent when they are more than 3 or 4 standard deviations away from the mean of the disparities in the region [3].

Multiple In(ter)dependent Matches: Perform multiple matches for each feature and compare the positions of their results. If the positions do not agree, mark the results as inconsistent. For example, Ayache and Lustman have implemented a trinocular stereo system in which two matches are made for each point, one from image1 to image2 and one from image2 to image3. Then, as shown in Figure 4, if the point in image3 is close enough to the epipolar line corresponding to the point in image1, mark the points as consistent [1].

Existing stereo systems have typically used combinations of these techniques to winnow out mistakes. In this paper we focus on the multiple-match approach and extend it to include temporal consistency for multi-camera sensors. One reason we concentrate on multiple-match techniques is that it is easy to set a threshold for "good" matches. We use a threshold of one pixel for all tests. For some of the other tests, such as correlation surface analysis and outlier rejection, it is difficult to find a principled way of setting the thresholds.

Several multiple-match techniques are possible and many of them have been incorporated into existing stereo systems to filter out mistakes. Some examples are:

Compare Multiple Depth Estimates: Given N calibrated cameras (where N is 3 or more), (1) select one camera as the pivot camera, (2) for each point in the pivot camera's image, compute $(N-1)$ depth estimates by analyzing pairs of images, one of which is from the pivot camera, (3) mark points as consistent if the depth estimates are approximately equal. Yoshida and Hirose have implemented a five-camera sensor based on this approach [17].

Epipolar-line Check: Given three calibrated cameras, there is a way to use two matches for each point to check the results, as shown in Figure 4: Match points from image1 in image2 and points from image2 in image3; If the distance between image3's point and the epipolar line corresponding to image1's point is sufficiently small, mark the point as consistent [1].

Compare Different Techniques: Apply two or more matching techniques and compare the positions of their results. For example, different search strategies and/or different correlation metrics could be used. As far as we know, no one has implemented this approach. Several stereo systems use different sized windows within a hierarchical search strategy, where one match guides the search for a higher-resolution one, but nobody has applied two completely different techniques (e.g., a correlation-based technique and an edge-based technique) and then merged their results.

Left-right Check: Perform each match twice, once from the left image to the right and once from the right image to the left (see Figure 5); if the left-to-right match and right-to-left matches are approximately inverses, mark the point as consistent (e.g., see [7] or [6]).

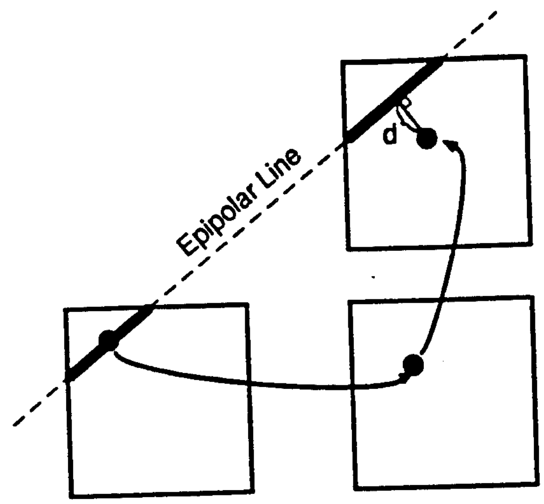


Figure 4: Trinocular epipolar-line consistency check.

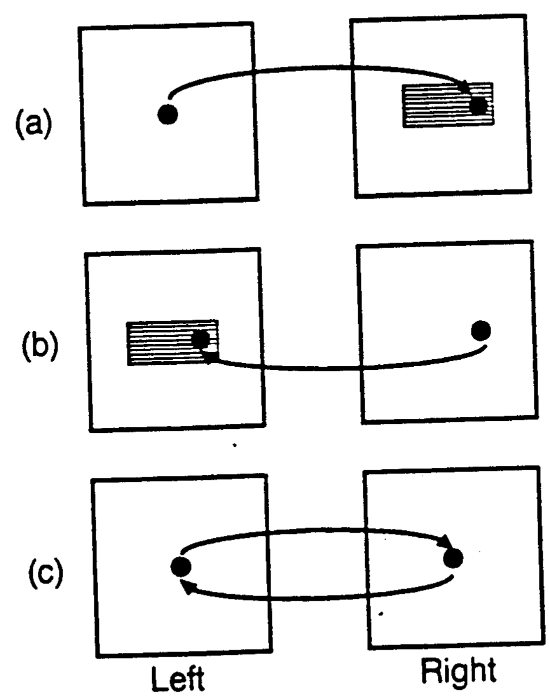


Figure 5: Left-right consistency check.

Transitivity Check: Given three uncalibrated cameras (or three crudely calibrated cameras), another way to identify possible mistakes is to perform three matches for each point, one from image1 to image2, one from image2 to image3, and one from image3 to image1 (see Figure 6). If the distance from the starting point in image1 to the final point produced by traversing the loop is small enough, mark the point as consistent. This requires more matches than the epipolar-line check, but it can be applied without knowing a precise calibration.

Object-Relative Motion Check: If the motion of the vehicle is not known (and cannot be easily computed), then pairs of images taken at different times can be used to filter out mistakes by (1) tracking points over time, (2) selecting a point in the scene as a reference point, (3) computing the x-y-z locations of all points relative to the reference point, and (4) marking points with stable relative distances as consistent. Moezzi et al. have implemented a system based on this approach [10].

Vehicle-Relative Motion Check: Given stereo images taken over time from a vehicle making a known motion (or a motion that can be computed), candidate matches can be evaluated by (1) computing a vehicle-relative x-y-z location from one image pair, (2) tracking the point into a second image pair, (3) computing another location for the point, adding in the known motion, and finally (4) checking to see if the two estimates are approximately equal. If not, the point is either a mistake or on a moving object.

We are interested in the case of imagery captured from two cameras mounted on a vehicle moving on relatively rough terrain. The cameras are relatively well aligned, but may move slightly with respect to each other and are uncalibrated. The environment is unmodelled and may contain moving objects. No additional information is available about the dynamic motion of the vehicle other than the video data itself.

We are experimenting with a version of the transitivity check that compares stereo matches over time (see Figure 7). In keeping with our goal of minimizing smoothing, we work directly with the raw intensities at full field resolution and use small correlation windows to perform both the stereo matching from left to right and optic-flow matching over time. If the two disparity maps D_c and D_p and the two flow fields M_l and M_r are viewed as maps from pixels to pixels, then for each pixel P, the spatiotemporal transitivity check measures the distance between $M_r(D_c(P))$ and $D_p(M_l(P))$. If the distance between these two points is within one pixel, P is marked as consistent.

We are still exploring ways of characterizing the effectiveness of this type of filter. As discussed in the next section, we have found that, for example, the spatiotemporal test catches a significant number of mistakes not detected by the left-right check.

Since some erroneous matches pass both the spatiotemporal test and the left-right check, we are exploring two additional techniques for consistency checking. One is a form of the vehicle-relative motion check that makes use of local estimates of depth change derived from simple assumptions about the vehicles path and measured changes in scene depth. The second involves the use of local consensus to predict and corroborate depth estimates that may fail other consistency checks.

In the next section we describe the experimental system used to explore these tests and their interactions.

3 The MIME System

Our purpose in implementing the Multiple In(ter)dependent Match Evaluation (MIME) System was to explore the idea of maximizing the resolution of range data produced by a passive sensor. Our approach has been to minimize the use of explicit or implicit smoothing operations and to recover the beneficial filtering effects of smoothing by applying a set of tests that compare the results of multiple matches.

The MIME system is implemented on a Connection Machine. As a research system, the system is designed to facilitate the comparison of different matching and filtering strategies, not for speed. As such, it has 20 or 30 top-level switches and parameters for specifying processing configurations. The switches include whether or not to discard measurements that fail left-right, forward-back, and spatiotemporal checking. The stereo and motion algorithms have parameters such as the sizes of correlation and search windows.

Imagery

Our experimental data was obtained by mounting a pair of monochrome cameras on an HMMWV vehicle, aligning the cameras manually so their optical axes were approximately parallel, recording the data on 8mm videotape as the vehicle was driven on and off road, and finally digitizing sequences of video fields from the tapes. The epipolar geometry is not known precisely, partly because the relative position of the cameras is not known precisely and partly because we have not attempted to compute lens distortions and the like. In general, we and others have found it difficult to maintain precise calibrations as the vehicle bounces along over rocks and ditches, making it desirable to have a system that is capable of working with less constrained imagery.

The overall image properties vary from one image to the next for several reasons. First, some of the data was gathered with auto-iris lenses. These lenses cover a

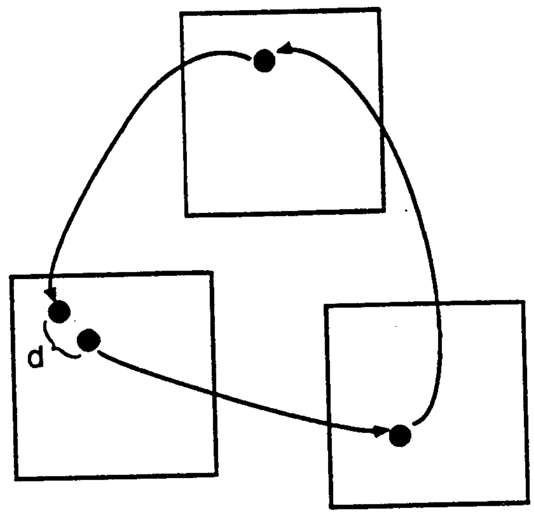


Figure 6: Trinocular consistency check for uncalibrated cameras.

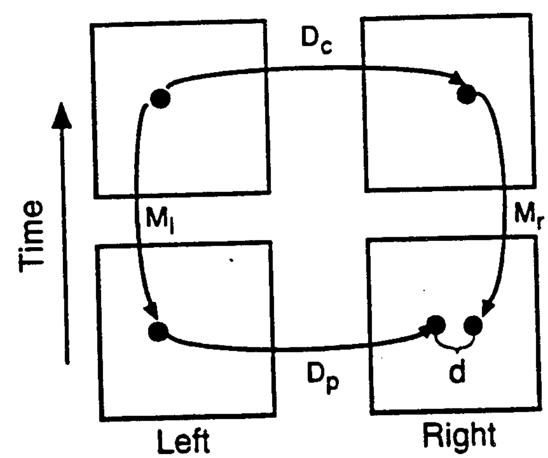


Figure 7: Spatiotemporal check.

wide dynamic range of lighting conditions by automatically adjusting their apertures, but have the problem that they are not linked together, so one aperture might be opening while the other is closing. In addition, the lens control systems tend to "hunt" for the best settings, causing intensities and depth-of-field to fluctuate continuously. A second reason for intensity differences is that there is no way to completely turn off the automatic gain control (AGC) on our COHU cameras. A third reason for intensity differences is that our method of synchronizing the two cameras involves using the output from one camera to synchronize the second one. This synchronizes the cameras, but it tends to drop the intensity of the initial camera a little because its output is doubly terminated. Lastly, dirt and smudges on the lenses produce regions that are darker than their matches.

In this kind of imagery, absolute intensity levels cannot be relied upon, forcing us to use a normalized correlation metric. The normalized metric does a good job of factoring out the gain and bias between the two correlation windows, however, factoring out gain and bias reduces the distinctiveness of a correlation window. In other words, if the intensity transforms were better constrained, a tighter requirement could be placed on potential matches, which would reduce the chances of incorrect matches. This line of reasoning is similar to the argument for using as much geometric information (e.g., epipolar constraints and a limited range of acceptable object distances) as possible to limit the search areas for matches. There are two benefits. First, the smaller the search area, the faster the search can be performed. And second, the smaller the search area, the higher the probability is of finding the correct match. Similarly in the intensity domain, the better the image-to-image intensity correspondences are known, the more directly the matches can be performed and the more likely the correct match will be found.

System Description

The system is based on four multiple-match filters: the left-right check, forward-back check, the spatiotemporal check, and the forward motion version of the vehicle-relative check. The last filter has been implemented, but not thoroughly tested.

Given a new image pair, the "complete" system performs the following sequence of operations (see Figure 7):

1. Compute two dense stereo depth maps, one mapping pixels in the left image to points on the right image, D_c^{lr}, and one mapping pixels on the right image to points on the left image, D_c^{rl}.

2. Perform the left-right check by evaluating whether the two depth maps, D_c^{lr} and D_c^{rl} are approximate inverses.

3. Compute four dense flow fields, two mapping pixels on the current left and right images to points on the previous left and right images, M_l^{cp} and M_r^{cp}, and two mapping pixels on the previous left and right images to points on the current left and right images, M_l^{pc} and M_r^{pc}.

4. Perform forward-back checks by testing whether the left flow fields, M_l^{cp} and M_l^{pc}, are inverses, and whether the right flow fields, M_r^{cp} and M_r^{pc}, are inverses.

5. Perform the spatiotemporal check, using the depth map computed for the previous pair of images D_p^{lr}, the two flow fields just computed M_l^{cp} and M_r^{cp}, and the just computed depth map D_c^{lr}. If $|M_r^{cp}(D_c^{lr}(P)) - D_p^{lr}(M_l^{cp}(P))| < \epsilon$, pixel P is labeled consistent.

6. Bring forward the image-centered information by mapping the data pertaining to pixels in the previous left and right images through the flow fields M_r and M_l. If pixel P corresponds to pixel P' in the previous image, i.e., if $M(P) = P'$, and some property R such as "valid since cycle 10" held of pixel P' in the previous image, then $R(M(P))$ can be asserted, in the present image.

7. Update the image-centered information to include the results of the current consistency checks.

Figure 8 shows an example of the type of sequence processed by the MIME system. The images have a rectangular aspect ratio because they are individual fields digitized from a videotape. They are a sequence of even fields, which are taken 1/30th of a second apart. Figure 9 shows the disparities computed and filtered from 4 image pairs. In this figure, lighter points are closer to the sensor. The dark spots on the depth map are points that failed one of the consistency checks over the processing of the four shown pairs. To emphasize the heights of objects, we transform these raw disparities into ones relative to a horizontal plane, as shown in Figure 10.

Stereo algorithm

The metric used for determining stereo correspondence is the standard normalized correlation metric [8]. Since we are interested in detecting small scene elements, all processing is done at a single scale. In the imagery we are considering, horizontal disparities fall in the range between 0 and approximately 85 pixels. Since the cameras are only approximately aligned, scan lines do not necessarily correspond to epipolar lines. Vertical disparities in our imagery generally fall in the range of -3 to 3 pixels.

The combination of single scale stereo with weakly aligned cameras results in large search windows; approximately seven scan lines of 85 disparities each, in our

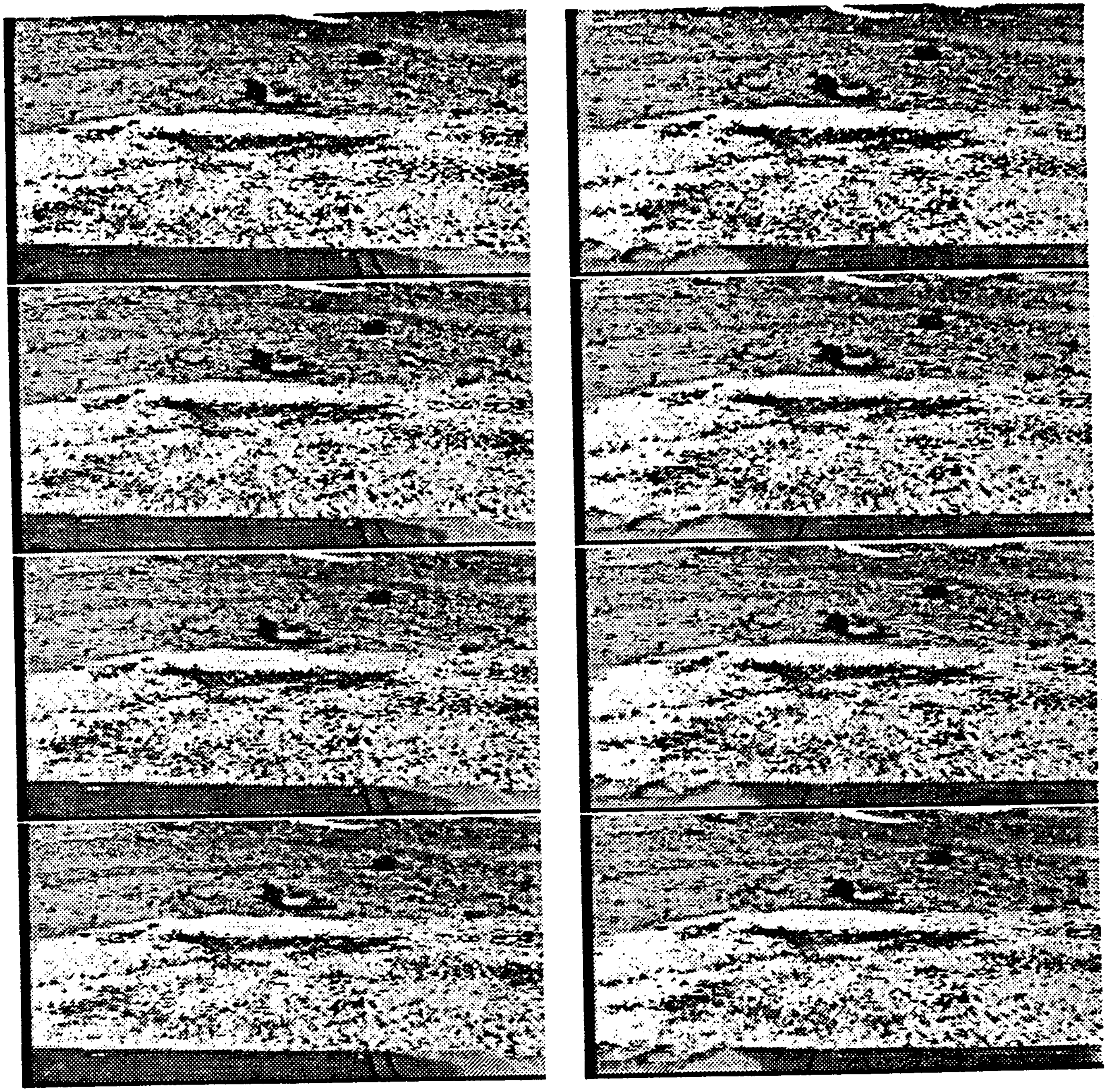


Figure 8: Sequence of stereo pairs of a pseudo-Mars scene.

Figure 9: Temporally filtered stereo disparities for a pair of images from Figure 8.

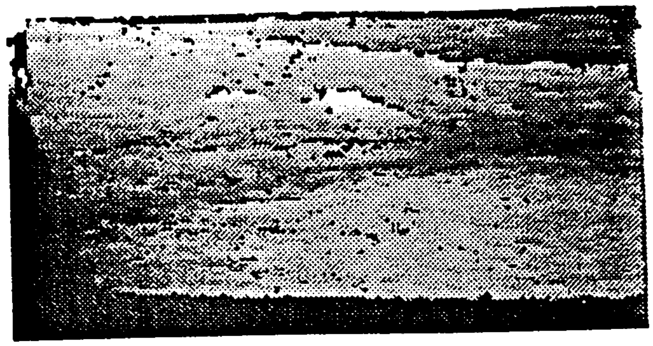


Figure 10: A skewed version of the disparities shown in Figure 9.

case. However, for any given stereo pair, vertical disparities do not vary greatly in a local area. Making use of this observation, the problem of determining vertical and horizontal disparities can be factored into two subproblems. The determination of vertical disparity is done first and makes use of a smoothing technique known as mode filtering [15, 16]. Subsequently, once the vertical disparity is known, the horizontal disparity can be determined from a single scan line. The stereo algorithm performed at each pixel P, has three steps:

1. Determine the vertical disparity of the best match in the entire search window.
2. Mode filter the vertical disparities from step 1 (i.e., select the most popular vertical disparity in the local area surrounding each pixel as the y-component of the pixel's disparity.)
3. Report a disparity for each point by choosing the x-component of the disparity computed on the scan-line corresponding to the y-component selected in step 2.

We include the mode filtering step in the stereo algorithm despite the fact that it is a type of spatial aggregation for two reasons. First, we expect the y disparities to vary relatively smoothly in the scene, except at large depth discontinuities. And second, the smoothing is not directly applied to the raw image or the results, rather it is being used to compute an intermediate result that is used to locate the final matches. Figure 11 shows the unfiltered y disparities computed for one of the pairs of images in Figure 8. Figure 12 shows the mode-filtered version of these disparities, which are used to select the row for the best match. As mentioned earlier, if the epipolar constraints are known precisely, the steps used to determine y disparities would not be necessary.

Optical-flow algorithm

The computation of optical flow is performed using sum of squared differences (SSD) correlation followed by a mode filtering step. The use of SSD correlation is justified since the images are taken from the same camera one thirtieth of a second apart, and absolute intensity levels are not expected to change drastically between frames. When there is the potential for rotational motion in the scene, it is important to keep the optical-flow correlation window as small as possible. Mode filtering makes sense because the flow field resulting from forward motion is expected to vary relatively smoothly in an image, except at depth discontinuities.

Approximate inverse check

We have introduced a slightly different left-right (forward-back) check than used by previous researchers. The first versions of this filter required that the left-to-right and right-to-left matches to be exact inverses (i.e., to the pixel). However, because of quantization effects, the commonly used version of the test allows the inverse to be within one pixel of the starting pixel, as shown in Figure 13. Our version loosens that constraint a bit more by taking into account the possibility that the right-to-left match may land on a pixel that doesn't have a valid right-to-left match. It accepts a pixel in the left image if the matching pixel in the right image *or one of its two neighbors* maps back to within one pixel of the initial point (see Figure 14). This change makes the test more symmetric. It also accepts a few more pixels in the left image as valid. Figure 15 shows the results after the left-right test has been applied. Figure 16 shows the results after both the left-right and spatiotemporal tests have been applied. Points in black indicate pixels that have been deemed inconsistent.

Additional consistency checks

Errors persist through the left-right check, the forward-back check and the spatiotemporal check over many image pairs. To catch these mistakes, we are developing two additional consistency checks. One is a version of the vehicle-relative motion check that determines local approximations of measured depth changes to detect mistakes. In the simple case of linear forward motion, all pixels corresponding to stationary elements in the scene should change depth by the same amount. If the depth of

Figure 11: Unfiltered y disparities.

Figure 12: Mode filtered y disparities.

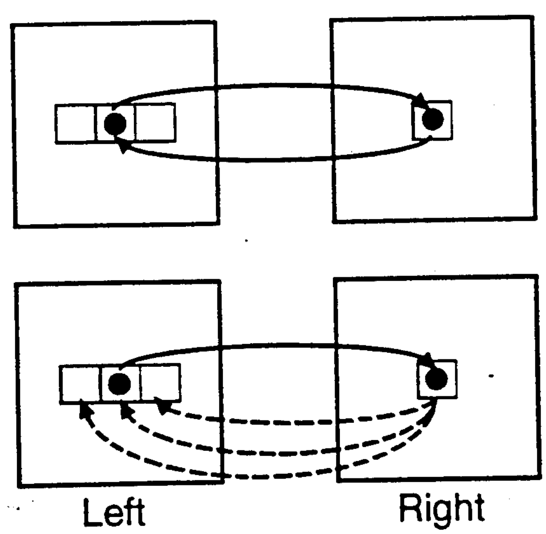


Figure 13: Normal left-right consistency check.

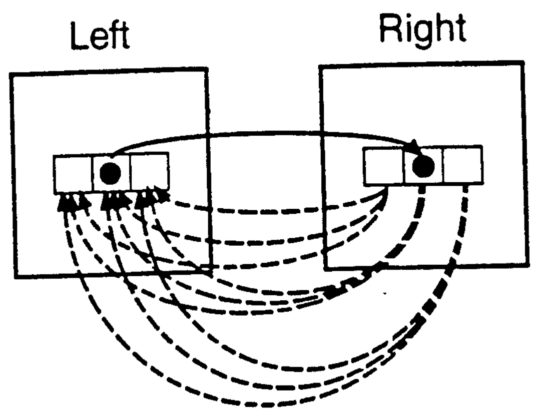


Figure 14: New left-right consistency check.

Figure 15: Disparity results after the left-right check has been applied.

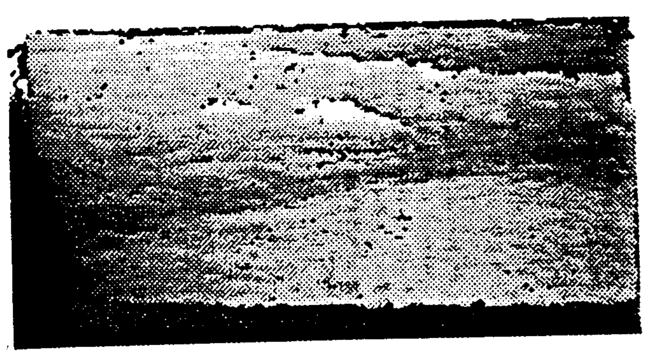


Figure 16: Disparity results after both the left-right check and the spatiotemporal check have been applied.

a pixel is mismeasured, its depth change will differ from the change measured elsewhere. For example, consider the two scene features A and B in Figure 17a, which are viewed by two one-dimensional cameras. The point A projects into aL in the left image and aR in the right image. If the matching system erroneously identifies bR as the match for aL (as shown in Figure 17a), the system produces a fictitious scene point, shown as the hollow point in the upper right corner of the diagram. If the pair of cameras is moved forward (i.e., from left to right) and the matching system persists in matching B to A, as shown in Figure 17b, then the fictitious point appears to move a distance d in the world. If d is significantly larger than the change potentially caused by inaccurate disparity measurements, then the aL-to-bR match is either a mistake or the corresponding point is moving in the scene.

The second form of additional consistency check involves the use of the results of the neighbors of an invalid point to estimate an expected value of the invalid point. This estimate can be used in two ways. Given an estimate of an invalid point's disparity, the system can search a small window about that estimate for the best match. If that match passes all the filters, it is marked as valid and included in the reported results. In our images, this process fills in 15 to 20% of the pixels initially marked as inconsistent by the left-right or spatiotemporal tests. Alternately, if the initial match for the invalid pixel agrees with this expected value, it can be labeled as consistent.

An open question about this process is the selection of the size of the search window about a suggested disparity. If the window is as large as the initial search window the same disparity will be found. At the other extreme, if the window is reduced to a single pixel, the system would report that pixel as the match and it would automatically pass all tests because, by definition, it is the best match in the region. So the question is how to reduce the size of the search region in a principled way so that it limits the search to an appropriate sized region without invalidating the evaluation procedures. We arbitrarily used regions that were 10 pixels wide in our experiments.

Experimental Results

In order to characterize the effectiveness of the various tests within the MIME system, we have applied it to several different image sequences with several different parameter and switch settings. Figure 18 shows a typical set of statistics produced by the system when both the left-right and spatiotemporal tests are applied. The parameters used for this and other sequences are characterized in Figure 19.

For this particular sequence, the vehicle was turning to the left as it approached a deep rut in a relatively flat field (see Figure 20). Figure 21 shows the region of the image in which we gathered statistics. The rest of the image is out of the field of view of the right image. The left-right test marks an average of 12% of the pixels in the left image as inconsistent.

The motion tracking procedure is virtually perfect and the forward-back check has no trouble verifying the flow vectors. The forward-back check marks fewer than 2 pixels in every 10000 as inconsistent.

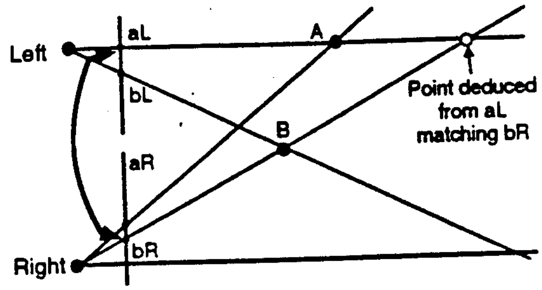
The spatiotemporal loop check marks an additional 8% of the points as possible errors, reducing the average number of "consistent" points in the left image to be 79.7%. When these tests are convolved together over 4 image pairs, the number of completely compatible points is 67.6%. When the suggestion procedure is used to fill in missing data, this number increase a few percentage points to about 72%. The number of gross errors passing all the tests is on the order of 10 to 20 pixels per image.

Persistent errors

Figure 22 illustrates a situation in which the left-right test fails to filter out a mistake. Two events conspire to produce this erroneous result. In Figure 22, aL is matched to bR, instead of aR, because something has altered A's appearance in the right image (e.g., A may be partially occluded). Similarly, bR is incorrectly matched to aL because B's appearance in the left image is different. As a result of these two mistakes, the left-right test erroneously accepts the aL-to-bR match.

Figures 23 and 24 show an example of this type of mistake. Figure 23 shows the context of the mistake. It occurs on the front edge of a deep rut. Figure 24 shows blown-up versions of the images around the mistake. The correlation window on the right in the left image is mistakenly matched to the left window in the right image, instead of the right one. This happens because the window straddles an occlusion edge between two regions at different depths, the front edge of the rut and the back of the rut. In the right image, these two subwindows have different disparities, so that there is no coherent window matching the one in the left image. As a result, the matching system finds a completely new window in the right image that looks like the one in the left. This new window is also along the edge of the rut, causing the same problem for the matching procedure when it tries to match from right to left. Unfortunately, but not too surprisingly, this right-to-left match happens to find the original window in the left image as its best match (instead of the left window in the left image). As a result, the mistake passes the left-right test.

Figure 25 shows another example of how a mistake can pass the left-right test. The X on the left of the left image is not in the field of the view of the right camera. Therefore, the best match for it is the only visible X in the right image. If the search from right to left happens



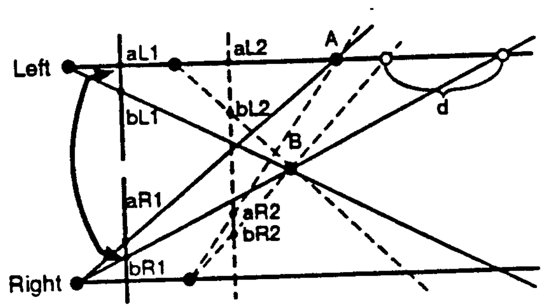


Figure 17: (a) aL-to-bR matching mistake and the deduced scene point. (b) Implied motion of the deduced scene point caused by a reoccurrence of the mistake.

Image Pair	Left-right Consistent	Forward-back Consistent	Spatiotemporal Consistent	Completely Consistent for Last Four Pairs	Completely Consistent w/ Prediction
1	90.27	–	–	–	–
2	88.83	100.00	82.36	–	–
3	87.57	99.96	80.25	–	–
4	88.11	99.97	77.39	70.04	70.04
5	89.55	99.98	77.54	66.32	66.71
6	89.83	100.00	84.43	66.68	67.55
7	88.33	99.99	82.62	67.41	69.13
8	89.92	100.00	82.05	69.82	72.30
9	89.32	100.00	85.22	75.13	78.91
10	87.19	100.00	77.76	70.03	74.94
11	86.41	99.89	77.77	66.88	72.83
12	86.43	100.00	75.37	65.16	71.28
13	87.37	99.98	75.87	62.91	68.70
14	86.81	100.00	82.82	64.72	70.74
15	86.72	99.82	78.55	64.00	69.93
16	86.32	100.00	82.29	67.52	72.82
17	86.45	100.00	77.49	70.51	74.77
18	84.76	100.00	76.89	67.15	72.84
19	85.37	100.00	78.67	67.80	74.15
Average:	87.7	99.98	79.7	67.6	71.7
Incremental:	12.3	0.02	8.0	12.1	4.1

Figure 18: Table of consistent pixels over time.

Parameter	Value
Width of stereo search window	85
Height of stereo search window	7
Width of stereo correlation window	11
Height of stereo correlation window	7
Diameter of stereo mode filter	15
Width of motion search window	19
Height of motion search window	9
Width of motion correlation window	7
Height of motion correlation window	7

Figure 19: Table of parameters.

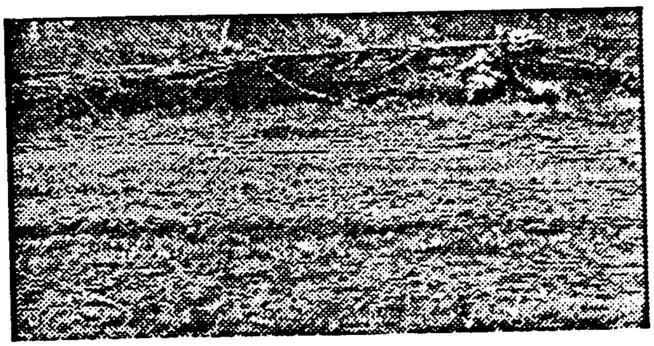


Figure 20: A scene with a deep rut crossing from left to right.

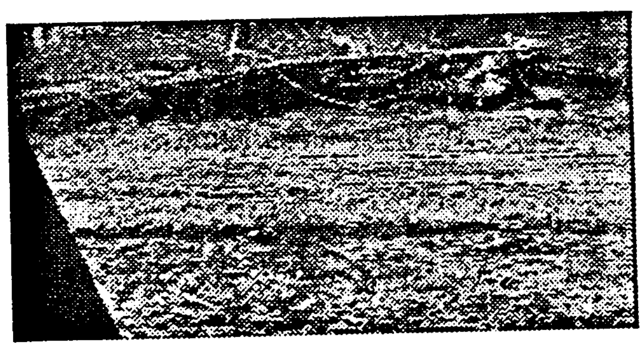


Figure 21: Region from the left image in Figure 20 from which the statistics in Figure 18 were computed.

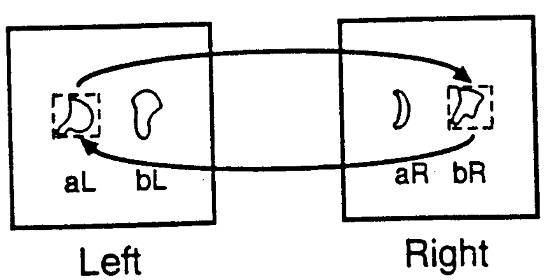


Figure 22: A pair of mistakes that conspire to pass the left-right check.

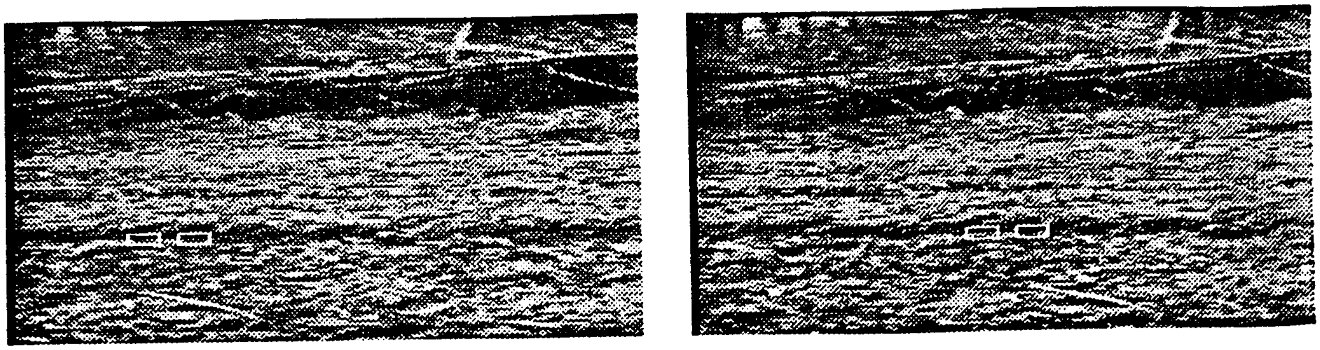


Figure 23: A pixel feature that erroneously passes the left-right check.

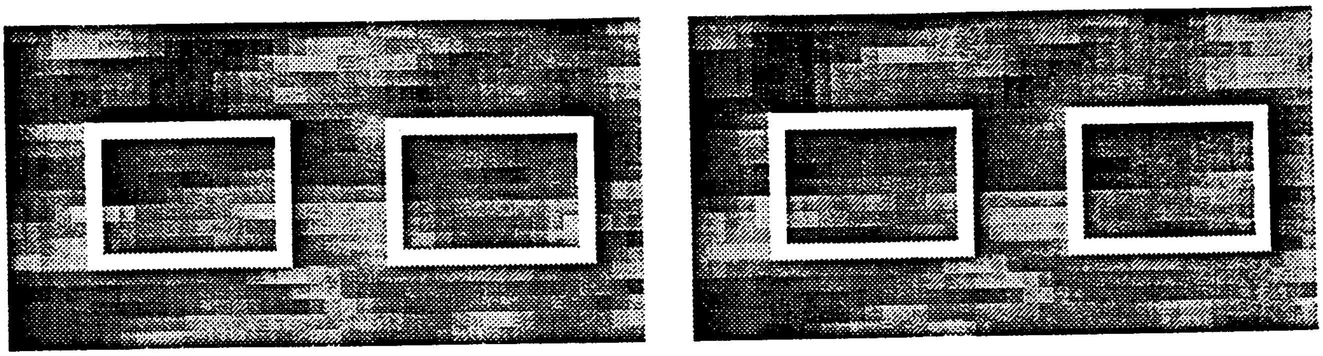


Figure 24: A blown-up version of the mistake shown in Figure 23

to prefer the left X, as indicated in the diagram, then this pair of mistakes leads to a mismatch that is not caught by the left-right check. Again, it took two events to produce the problem.

Figure 26 shows an example of how a mistake that is missed by the left-right check can also be missed by the spatiotemporal check. This example is similar to the example in Figure 22. If the cause of the mistakes in the first pair of images (e.g., partial occlusions) persists over time, the spatiotemporal test may also miss the mistake. Usually, however, the views of the scene features change enough that the test detects an inconsistency and marks the results as invalid.

4 Flakey's Stereo system

For several years, Flakey, SRI's indoor robot, has used ultrasonic sensors and a structured-light sensor to locate potential obstacles in its path. These sensors, however, have several limitations. For example, sonar cannot detect thin objects, such as table legs. And the structured-light sensor can only measures distances to points that are in a particular plane and are close to the sensor. Therefore, in order to increase both Flakey's sensing resolution and sensing range, we have implemented a streamlined version of the MIME stereo system on the on-board Sparc10 processor. The resulting system produces a 105-by-240 range image in .4sec. In addition, Flakey's control system can select horizontal stripes from this image to be recomputed at a higher rate. For example, it can compute 20 rows of the range image at 30hertz.

Flakey uses a dual lens system to project a pair of images into a single video field. We originally installed the optical 2-to-1 lens with the hope that it would minimize the overall change in intensities from one image to the next. However, the lens has such strong vignetting problems that both half images are significantly darker at the edges than they are at the middle. We tried two separate cameras, but our inability to turn off their automatic gain control has made them difficult to use, especially indoors where there are a number of specular fixtures and bright lights.

In order to run the stereo matching algorithm as fast as possible, we did the following:

- Subsampled the images from left to right, reducing 315 columns down to 105. This reduced the range of disparities from about 50 to 16.
- Simplified the correlation metric to be the sum of squared differences, which can be computed significantly faster than normalized cross correlation. The sums are computed incrementally by sliding the the search region across the image.
- Deleted the spatiotemporal consistency test, but kept the left-right check to validate matches. Therefore, each match is performed twice. In addition, we added an interest operator to flag points in the left

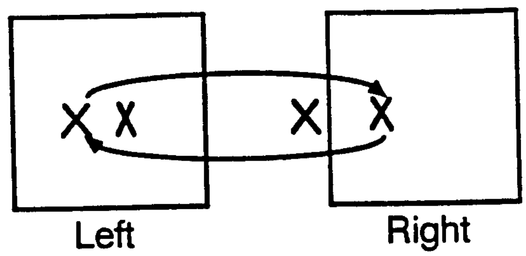


Figure 25: A pair of mistakes, one of which is caused by a feature being out of the field of view of the other image.

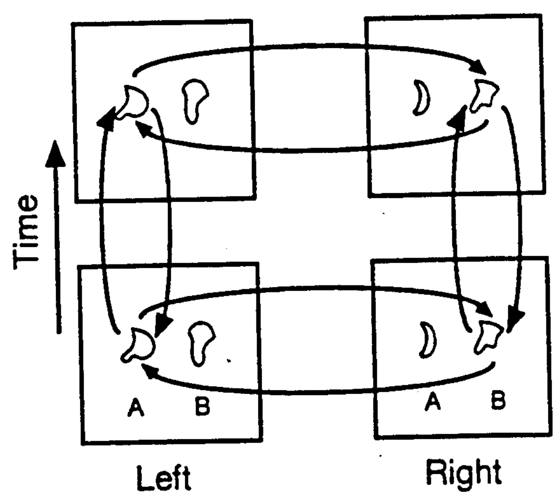


Figure 26: A mistake that passes the spatiotemporal test due to recurring errors.

image that are unlikely to produce reliable results, because they lack texture.

Flakey merges the stereo data with its sonar data, and then plans its paths in the same way it always has. It uses a two-dimensional, robot-centered map to keep track of potential obstacles and landmarks.

In the future, we plan to add behaviors to Flakey so that it can plan data-gathering maneuvers to examine unmeasured regions or closely inspect points of special interest.

5 Conclusion and Future Work

In this paper we have (1) introduced a spatiotemporal consistency check for evaluating stereo results and (2) incorporated it into a system for integrating range data over time. We have begun the process of characterizing the utility of this approach and its relationship to other similar techniques.

One observation we've made is that outdoor natural scenes contain sufficient texture to support dense correlation matching. For example, our matching technique locates matches for 70 to 80% of the pixels in most images, even though we use relatively small correlation windows. The 20 to 30% mistakes and no-data regions are caused by such things as bland areas, repeated patterns, and occlusions. Even though the mistakes represent a relatively small fraction of the results, most tasks require significantly more complete and more reliable data. For example, a navigation system cannot recommend driving over areas containing unmeasured regions or unexplained points floating above the ground. Therefore, there is a need for evaluation techniques to assign confidences to individual pixel features and for higher-level sensor control strategies to reexamine no-data or questionable regions.

The multiple-match consistency checking procedures discussed in this paper provide a form of "structural filtering" that we prefer over such techniques as thresholding correlation values. Structural filtering techniques are based on distance measurements for which it is relatively easy to determine appropriate thresholds. We view the spatiotemporal filtering technique as one of several techniques from which a stereo system can be constructed. One benefit of applying spatiotemporal techniques is that they provides a natural way to integrate range information over time, which opens up the possibility of additional temporal analysis.

In the future we plan to complete the characterization of this approach, explore higher-level explanations of the pixels marked invalid by the consistency checks (e.g., produce explanations in terms of occlusions and bland areas), and investigate techniques for combining the results of multiple "binary" consistency checks to form scenes models capable of answering such questions as "What are navigable areas in front of the vehicle?" and "Where are there preliminary indications of a possible obstacle that should be examined more closely?"

Acknowledgments

We would like to thank Harlyn Baker, Marsha Jo Hannah, Marty Fischler and Ramin Zabih for helpful discussions of these ideas. This work was supported by ARPA under contract DACA76-92-C-0003. John Woodfill was partially supported by NSF Postdoctoral grant CDA9211152.

References

[1] Ayache, N., and F. Lustman, "Fast and Reliable Passive Trinocular Stereovision," *Int'l Conf. on Computer Vision*, June 1987.

[2] Baltsavias, E.P., "Multiphoto Geometrically Constrained Matching," Institute for Geodosy and Photogrammetry, Zurick, Switzerland, December 1991.

[3] Bolles, R.C., H.H. Baker, and M.J. Hannah, "The JISCT Stereo Evaluation," SRI International Report, January 1993.

[4] Bolles, R.C., H.H. Baker, and M.J. Hannah, "The JISCT Stereo Evaluation," *Proc. ARPA Image Understanding Workshop*, Washington, D.C., pp. 263-274, April 1993.

[5] Dhond, U.R., and J.K. Aggarwal, "A Cost-Benefit Analysis of a Third Camera for Stereo Correspondence," *Int'l Jnl. of Computer Vision*, Vol. 6, No. 1, pp. 39-58, April 1991.

[6] Fua, P.V., "A Parallel Stereo Algorithm that Produces Dense Depth Maps and Preserves Image Features," *Machine Vision and Applications*, 1991.

[7] Hannah, M.J., "A System for Digital Stereo Image Matching," *Photogrammetric Engineering and Remote Sensing*, Vol. 55, No. 12, pp. 1765-1770, December 1989.

[8] Hannah, M.J., "A *Computer Matching of Areas in Stereo Images*. PhD thesis, Stanford University, July 1974.

[9] Kanade, T., M. Okutomi, T. Nakahara, "A Multiple-baseline Stereo Method," *Proc. Image Understanding Workshop*, San Diego, Ca, pp. 409-426, January 1992.

[10] Moezzi, S., S.L. Bartlett, and T.E. Weymouth, "The Camera Stability Problem and Dynamic Stereo Vision," *Proc. Computer Vision & Pattern Recognition Conf.*, pp. 109-113, 1991.

[11] Moravec, H.P., "Visual Mapping by a Robot Rover," *Proc. Int'l Joint Conf. on Artificial Intelligence*, Tokyo, Japan, pp.598-600, August 1979.

[12] Nishihara, H.K., "Practical Real-Time Imaging Stereo Matcher," Opt. Eng., 23, 5, 536-545, Sept.-Oct. 1984. Also in *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms*, edited by M.A. Fischler and O.Firschein, Morgan Kaufmann, Los Altos, 1987.

[13] Pollard, S.B., J.E.W. Mayhew, and J.P. Frisby, "PMF: A Stereo Correspondence Algorithm Using a Disparity Gradient Limit," *Perception*, Vol. 14, pp. 449-470, 1981.

[14] Szeliski, R., "Bayesian Modeling of Uncertainty in Low-Level Vision," *Int'l Jnl of Computer Vision*, Vol. 5, No. 3, pp. 271-301, December 1990.

[15] Woodfill, J., *Motion Vision and Tracking for Robots in Dynamic, Unstructured Environments*. PhD thesis, Stanford University, August 1992.

[16] Woodfill, J. and R. Zabih, "An algorithm for real-time tracking of non-rigid objects," *Proceedings of AAAI-91, Anaheim, CA.*, pages 718–723. The MIT Press, 1991.

[17] Yoshida, K. and S. Hirose, "Real-time Stereo Vision with Multiple Arrayed Camera," *IEEE Conf. on Robotics & Automation*, 1992.