SRI International

DESCRIPTION OF SRI'S BASELINE STEREO SYSTEM

Technical Note No. 342

October 1984

By: Marsha Jo Hannah, Senior Computer Scientist

Artificial Intelligence Center
Computer Science and Technology Division

This research was supported by the Defense Advanced Research Projects Agency under Contract No. MDA903-83-C-0027 (5355).

SRI International®

333 Ravenswood Ave. • Menlo Park, CA 94025
(415) 326-6200 • TWX: 910-373-2046 • Telex: 334-486

Description of SRI's Baseline Stereo System

Marsha Jo Hannah
Artificial Intelligence Center, SRI International
333 Ravenswood Ave, Menlo Park, CA 94025

Abstract

We are implementing a baseline system for automated area-based stereo compilation. This system, STSYS, operates in several passes over the data, during which it iteratively builds, checks, and refines its model of the 3-dimensional world, as represented by a pair of images. In this paper, we describe the components of STSYS and give examples of the results it produces. We find that these results agree reasonably well with those produced on the interactive DIMP system at ETL, the best available benchmark.

Introduction

Automatic techniques for the production of 3-dimensional data via stereo compilation are receiving increased interest for a variety of applications, including cartography [Panton, 1978], autonomous vehicle navigation [Hannah, 1980], and industrial automation [Nishihara & Poggio, 1983]. Conventional stereo compilation techniques, which are based on area correlation, can produce incorrect results under a variety of conditions, for example, when views are widely separated in space or time, in the vicinity of partial occlusions, in featureless or noisy areas, and in the presence of repeated patterns.

We are investigating ways to overcome these inadequacies. Our research strategy is first to implement a baseline system that performs conventional stereo compilation, then to replace pieces of the system with improved modules as we develop them. Thus, our baseline system will be the core of an ever-improving stereo system. We also intend to test the baseline system against a "challenge data base" of image areas where conventional stereo techniques fail.

As currently implemented, our system includes routines to perform the following operations automatically:

* Select "interesting" points for sparse matching
* Search 2D regions for sparse matches
* If necessary for uncalibrated imagery, compute relative camera parameters from sparse matches
* Compute epipolar lines
* Locate epipolar matches, using disparity estimates from sparse matches when available
* Evaluate matched points for local consistency and believability
* Interpolate between matched points
* Display images and results in left-right stereo, red-green stereo, or as a monocular disparity field
* Compute range data and x-y-z coordinates for matched point pairs
* Display terrain data in perspective with hidden lines removed.

We are currently exploring improved techniques for image matching, for match evaluation, and for terrain surface interpolation.

The Stereo System

Over the past several months, SRI has integrated existing pieces of stereo code into a baseline system for automated area-based stereo compilation. The system operates in several passes over the data, during which it iteratively builds, checks, and refines its model of the 3-dimensional world represented by a pair of images.

The driving program is called STSYS (STereo SYStem). It invokes a variety of modules to perform the necessary processing for stereo compilation. In theory, the modules are independent and can be replaced with improved versions at will; in practice, there are some unavoidable interdependencies of global variables that will have to be attended to.

The following sections describe the components of STSYS in the order they are normally invoked; examples of their results are included. Comments are also made as to improvements that could be made to each of the modules.

REDUCE

The basis for the image matching techniques is a hierarchy of images, as shown in Figure 1. REDUCE is the module that forms this hierarchy from the original images. In the example used for the figures, the original images are a pair of image "chips" digitized from standard 9"x9" mapping photos taken over Phoenix South Mountain Park, near Guadalupe (a suburb of Phoenix), Arizona. These images are 2048x2048 pixels in size, and cover an area that is approximately 2 kilometers square on the ground; elevations in the area range from 360 to 540 meters. The reduction hierarchy consists of a pyramid of images, each at half of the resolution of its parent; in this case REDUCE produces pairs of images that are 1024x1024, 512x512, 256x256, 128x128, 64x64, 32x32, and 16x16 pixels in size. (Figure 1 shows only the 256x256 through 16x16 image pairs.)

At present, REDUCE produces pixels in each reduced image by simple averaging of the pixels in an NxN square in the next-largest image (in the above case, N=2). It is known that this technique can produce artifacts in the data, and a more sophisticated technique of convolving the image with a Gaussian, then sub-sampling, is preferred [Burt, 1980]. Substitution of this technique will be one of the first enhancements made to STSYS.

INTEREST

The first step in the matching process is to procure a set of well-scattered, reliable matches in the image. Our approach is first to select areas in one image that contain sufficient information to produce reliable matches. To accomplish this, a statistical operator based on image variance and edge directionality is passed over the image; local peaks in the output of this operator are recorded as the preferred places to attempt the matching process.

Historically, such operators have been called *interest* operators, and the peaks in the operator output have been called *interesting points* [Moravec, 1980]. This nomenclature is somewhat misleading, as the points selected are rarely interesting to a human observer; however, these terms have been in use in the computer vision community for over 10 years. It should be noted that present interest operators are not feature detectors; the same operator run over both images of a stereo pair will not necessarily pick out the same points in the two images. In our system, the interest operator is run in only one of the images, where it selects points that are to be matched in the second image by other means. (A possible enhancement to STSYS would be to design and implement efficient interest operators that really do choose "interesting points," such as crossroads, building corners, sharp bends in rivers, etc.)

INTEREST permits the user to specify the operator to be used [Hannah, 1980], the window size over which it is calculated, and the window radius for testing local peaks. It also provides the capability to divide the image into a grid of subimages, and records the relative ranks of the interesting points within their grid cells; this permits the most interesting point(s) for each area to be matched first. Figure 2 shows the interesting points for the right image of the Phoenix pair; the numbers indicate the 1st, 2nd, 3rd, and 4th most interesting points in a 6x6 grid of cells.

Preliminary Matching

At this point in the processing, it is possible to take one of two different approaches to the matching. If nothing is known regarding the absolute camera positions and orientations (as would be the case for an amateur, handheld stereo pair), an unstructured hierarchical matching algorithm is used on the most interesting points. The results of these matches are used in seeking a solution for a simplistic relative camera model (5 angles describing the relative positions and orientations of 2 ideal cameras [Hannah, 1974]), which can then be used for the epipolar constraint in further matching. This approach uses the modules HMATCH and C2MODEL, described below. On the other hand, if the camera parameters are known (as would be the case for the highly calibrated cartographic stereo images intended for terrain mapping) matching can proceed directly with the epipolar constraints, using the module LMATCH.

HMATCH

HMATCH assumes that nothing is known about the relative orientations of the images, other than that they cover approximately the same area, at about the same scale, with no major rotation between the images. It matches each specified point (usually the most interesting point in each grid cell) using an unguided hierarchical matching technique similar to that reported in [Moravec, 1980]. This technique begins with the point in the largest image (the 2048x2048 right image of the Phoenix set), traces it back through that image's hierarchy (in our example, it repeatedly halves the co-ordinates of the point) until it reaches an image that is approximately the size of the correlation window (the 16x16 image for the 11x11 correlation windows that we used). It then uses a 2-dimensional spiral search, followed by a hill-climbing search for the maximum of the normalized cross-correlation between the image windows [Quam, 1971]. This global match is then refined back down the image hierarchy; that is, the disparity at each level (suitably magnified to account for relative image scales) is used as a starting point for a hill-climb search at the next level. The correlation window size remains constant at all levels of the hierarchy, so the match is effectively performed first over the entire image, then over increasingly local areas of the image. This technique permits the use of the overall image structure to set the context for a match; the gradually increasing detail in the imagery is then followed down through the hierarchy to the final match.

Figure 3 shows the results of this technique on a point in the Phoenix set. The image hierarchy is the same as in Figure 1, with the addition of 63x63 image chips covering the matched area in the 2048x2048, 1024x1024, and 512x512 images; these are shown in the upper right corner of each hierarchy. The matching began in the right image in the 2048x2048 chip, traced the right point through the hierarchy (approximately clockwise in the figure) to the 16x16 right image, matched it to the 16x16 left image, then refined it back through the left image hierarchy until reaching the left 2048x2048 chip.

It is instructive to look at the correlation coefficients for these matches (see Table 1). In the smaller images, the correlation is poor, since the window covers a large area of terrain with a great deal of relief. As the matching moves up the hierarchy, the correlation improves, because the window now approximates an area at a single elevation. After reaching the 512x512 images, however, the correlation begins to decline, both in absolute value and with respect to an autocorrelation-based threshold [Hannah, 1974]. This is due to noise in the images; if one examines the chip from the 2048x2048 left image, one will see several streaks across the image, representing scratches on the original photograph and/or dropped data in the digitization; close examination also reveals a grainy noise pattern. Because the degraded correlations will cause difficulties in determining which matches are the correct ones, our processing has gone only to the 512x512 images. More code will be added to STSYS in order to refine the final matches from this level down to the original 2048x2048 images.

Figure 4 shows the results of HMATCH on the most interesting point in each grid cell. Only the points thought to have been matched correctly are shown; those with poor correlation or whose matches fell outside of the image have been discarded by STSYS.

C2MODEL

If no camera calibration information is available, the module C2MODEL calculates a simplistic relative camera model from a set of matched point pairs. This is accomplished by searching for 5 angles--the azimuth and elevation of the second camera's focal point with respect to the first camera; and pan, tilt, and roll of the second camera's axes with respect to those of the first. The object of the search is to minimize the error between the matched point in the second image and the epipolar line produced when the point in the first image is projected through the hypothesized cameras. The search proceeds by a linearization of the equations and their analytic derivatives [Gennery, 1980]. Once a solution is found, the reliability of the matched points is assessed. Points that appear to contribute too much error to the solution are removed from the calculation, and the solution is redone. Either this process reaches a successful conclusion when the point set is found to be consistent, or it reports failure if too many of the point pairs are rejected.

The resulting camera model is quite crude, as it must depend on a guess as to the focal lengths of the cameras and the length of the baseline between the cameras. Also, it assumes that we are using ideal cameras, thus totally ignoring the internal calibration of the cameras. It is, however, suitable for approximating the epipolar constraint to simplify further matching.

LMATCH

If the camera parameters are given (or once the crude ones have been derived), matching can proceed somewhat more efficiently. The camera parameters define the manner in which a point in the first image projects to a line in the second image--the epipolar constraint. This constraint can be used to cut the search from 2 dimensions (all over the image) to 1 dimension (back and forth along the epipolar line).

LMATCH proceeds very much like HMATCH, except that the search for a match is confined to the vicinity of the epipolar line. Because we assume that there is no outside information to indicate where these preliminary matches lie along the line, we again use the hierarchical technique to search out and refine the match. If relative camera parameters have been derived, LMATCH is used on the second most interesting point per grid cell, plus any points that C2MODEL indicated were unreliable; the results of this mode are shown in Figure 5. If the true camera parameters have been supplied, LMATCH is used on the two most interesting points in each grid cell; these results are shown in Figure 6.

Anchored Matching

Once several reliable matches have been found, they can be used as "anchor" points for further matching. Our basic technique for this again uses the grid cells in the image. A given point will lie in some grid cell; the closest matched point(s) will lie in that cell or in one of the 8 neighboring cells. Under the assumption that the world is generally continuous, a point would be expected to have a disparity similar to that of its nearest neighbors. Thus, to approximate the disparity at a point, we first calculate the average of the disparities of the well-matched points in the current and neighboring cells, weighted by the inverse of the distance between the current point and the neighboring point. (As we develop more sophisticated interpolation schemes, this disparity approximation technique will be upgraded.) This approximate disparity is used along with the epipolar constraint to perform a very local search for the match to a point. Note that a point is considered to be well-matched if it has a correlation above a user-settable absolute threshold, usually 0.5, as well as having a correlation above a variable threshold, based on the autocorrelation function around the point in the first image (see Table 1 for examples). Our definition of well-matched should also be upgraded to include distance off of epipolar lines as well as a measure of how consistent the disparity is with its neighbors.

PMATCH

At this point in our processing, we have matched the two most interesting points in each grid cell. This is still rather sparse information, so we next invoke the module PMATCH to match the balance of the interesting points. It uses the anchored match technique described above, with a generous search radius along the epipolar line, to find these matches. Figure 7 shows the results of this module. Two different marks are used for the matches, denoting whether their correlations indicate that they are well-matched.

GMATCH

We next produce matched points on a closely spaced grid. The module GMATCH also uses the anchored match technique, with a somewhat restricted search radius along the epipolar line, to calculate matches on a user-specified grid. Figure 8 shows the results of this module on a 20x20 grid, again using different marks for the different qualities of match.

Terrain Modeling

Given the dense grid of matched points and the camera calibration, it is possible to derive a digital terrain model. If external and internal camera information is available, the module SRIDTM can be used to create a reasonably accurate DTM, which can then be displayed with another program, DTMICP. (An example of DTMICP output is shown in Figure 9; it can also produce range images of the terrain or pictures of the original imagery "painted" on the terrain.) If the only camera information is C2MODEL's relative model, then the module RELDEPTH can be used to create a relative DTM. However, due to the many over-simplifications and the computational instability of the relative camera model, such relative DTMs are of very low accuracy, and their use is discouraged.

Often, a terrain model is desired that has its points more closely spaced than that provided by the stereo matching process. Sometimes, there may be areas of the images that cannot be matched, due to noise in the data, insufficient information, or changes such as moving vebicles; this will result in "holes" in the grid of terrain data, which must be filled in somehow. In either case, interpolation of the matched data points is necessary to provide information at other points. Work on this topic is reported separately [Smith, 1984].

Evaluation

Evaluation of the accuracy of STSYS is difficult, as there do not seem to exist stereo data sets with known ground truth to compare against our results. We do, however, have the results of an interactive stereo compilation algorithm called Digital Interactive Mapping Program (DIMP), produced and operated by the U.S. Army Engineer Topographic Laboratories (ETL) [Norvelle, 1981]. It should be noted, however, that ETL's results were obtained by an interactively coached process, which was run on a 5x5 grid in the 2048x2048 images and used correlation windows warped to account for the local steepness of the terrain, while ours were obtained by a fully automatic process that ran on a 20x20 grid in the 512x512 images without warping; comparing them is a little like comparing apples and oranges. (Another planned upgrade to our matching techniques is the use of warped correlation in the match refining step.)

Of our matches (both interesting points and grid points), approximately 98% agree reasonably well with the nearby ETL matches at the resolution of the 512x512 images. Of the remaining 2%, most are clearly blunders on our part, although a few appear to be the result of errors in the DIMP compilation. It is not known what fraction (if any) of the 98% represent places where our processing and the DIMP processing produced similar wrong answers.

Discussion

SRI has an operational baseline system for automated area-based stereo compilation. This system, STSYS, operates in several passes over the data, during which it iteratively builds, checks, and refines its model of the 3-dimensional world represented by a pair of images. In this

paper, we have described the components of STSYS and given examples of the results it produces. We have compared these results to those produced on the interactive DIMP system at ETL, and found that they compare favorably.

STSYS is, at present, an experimental program; no attempt has been made to optimize it for best results or fastest operation. The program is still evolving, and will not be ready for transfer to other users until its methods stabilize. Likewise, more complete documentation must wait on completion of the code.

Acknowledgements

The research reported herein was supported by the Defense Advanced Research Projects Agency under Contract MDA903-83-C-0027, which is monitored by the U.S. Army Engineer Topographic Laboratory. The views and conclusions contained in this paper are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or of the United States Government.

I would like to thank Robert Bolles, Lynn Quam, Grahame Smith, and Martin Fischler for their support on this project.

References

Burt, Peter J., 1980. "Fast, Hierarchical Correlations with Gaussian-like Kernels," University of Maryland Computer Science Center Report TR-860, January, 1980.

Gennery, Donald B., 1980. "Modelling the Environment of an Exploring Vehicle by means of Stereo Vision," Ph.D. Thesis, Stanford University Computer Science Department Report STAN-CS-80-805, June, 1980.

Hannah, Marsha Jo, 1974. "Computer Matching of Areas in Stereo Images," Ph.D. Thesis, Stanford University Computer Science Department Report STAN-CS-74-438, July, 1974.

Hannah, Marsha Jo, 1980. "Bootstrap Stereo," *Proceedings: Image Understanding Workshop,* College Park, MD, April, 1980, pp. 201-208.

Moravec, Hans P., 1980. "Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover," Ph.D. Thesis, Stanford University Computer Science Department Report STAN-CS-80-813, September, 1980.

Nishihara, H. Keith, and Tomaso Poggio, 1983. "Stereo Vision for Robotics," *Proceedings of the International Symposium of Robotics Research,* Bretton Woods, NH, September, 1983.

Norvelle, F. Raye, 1981. "Interactive Digital Correlation Techniques for Automatic Compilation of Elevation Data," U.S. Army Engineer Topographic Laboratories Report ETL-0272, October, 1981.

Panton, Dale J., 1978. "A Flexible Approach to Digital Stereo Mapping," *Photogrammetric Engineering and Remote Sensing,* Vol. 44, No. 12, pp. 1499-1512.

Quam, Lynn H., 1971. "Computer Comparison of Pictures," Ph.D. Thesis, Stanford University Computer Science Department Report STAN-CS-71-219, May, 1971.

Smith, Grahame B., 1984. "A Fast Surface Interpolation Technique," SRI International Artificial Intelligence Center Technical Memo, in preparation, June, 1984.

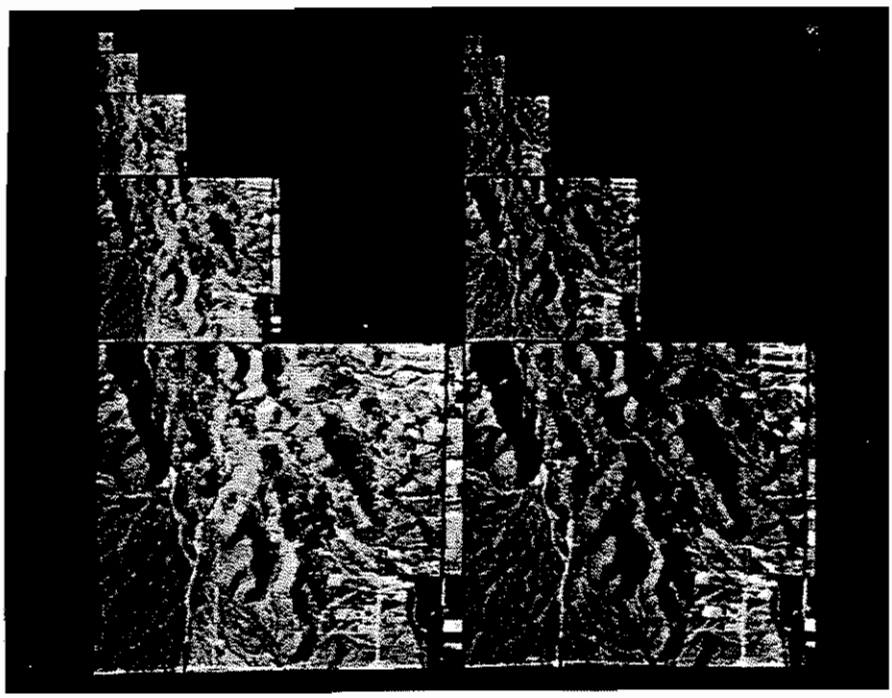

Figure 1--Reduction Image Hierarchy.

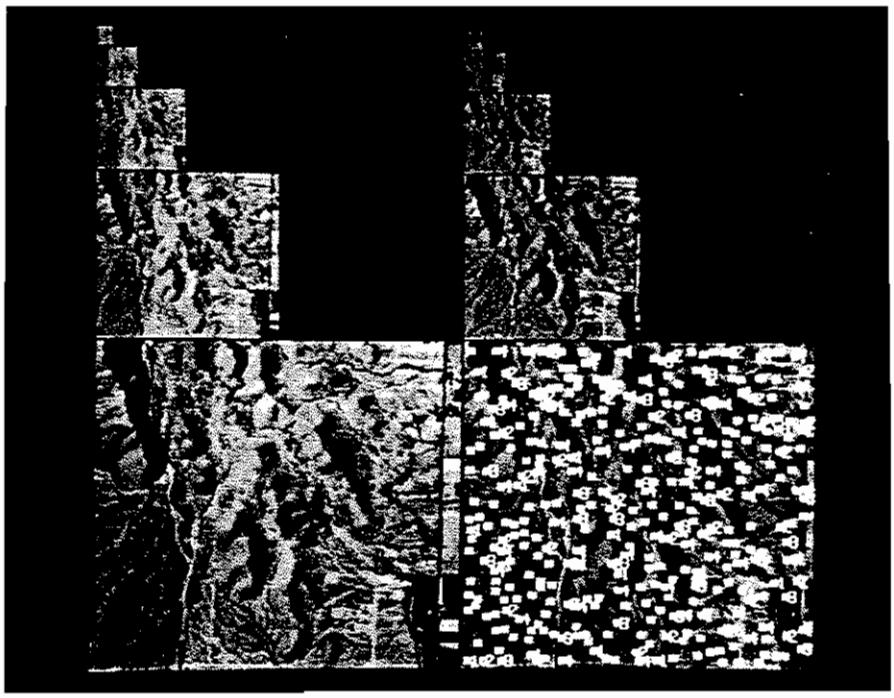

Figure 2--Interesting Points, Ranked by Grid Cell.

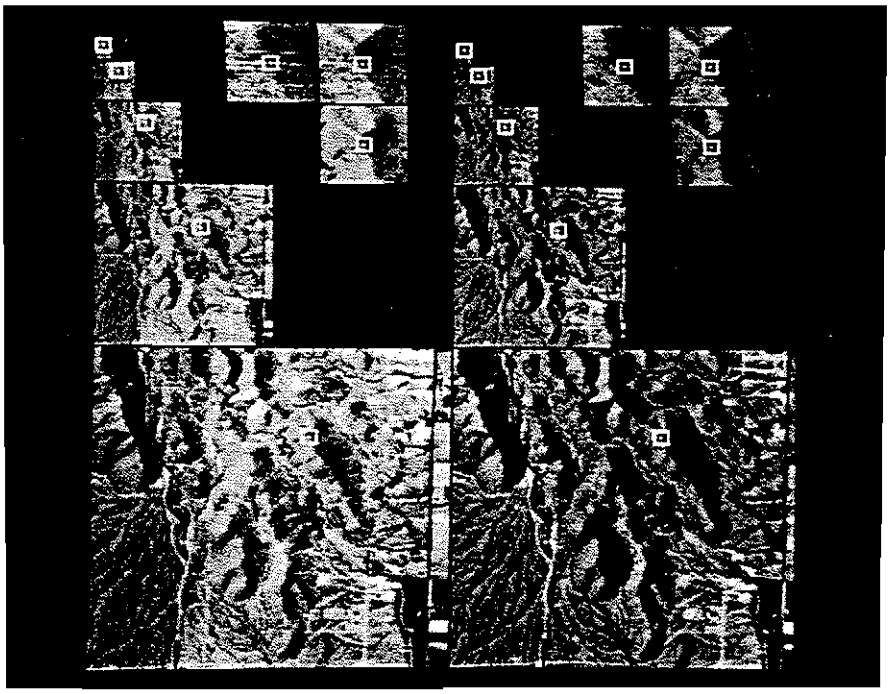

Figure 3--Hierarchical Match of an Interesting Point.

Table 1–Hierarchical Correlations for Point in Figure 3.

Image size	Point 1	Point 2	Correlation	Autocorrelation
16x16	(9,11)	(9,11)	0.140452	0.577717
32x32	(19,23)	(19,23)	0.384883	0.437053
64x64	(38,46)	(37,46)	0.738581	0.738427
128x128	(77,92)	(76,92)	0.929933	0.885289
256x256	(154,184)	(153,184)	0.954606	0.918228
512x512	(308,369)	(306,369)	0.916062	0.929428
1024x1024	(616,738)	(612,737)	0.750448	0.932947
2048x2048	(1232,1476)	(1222,1475)	0.341622	0.790917

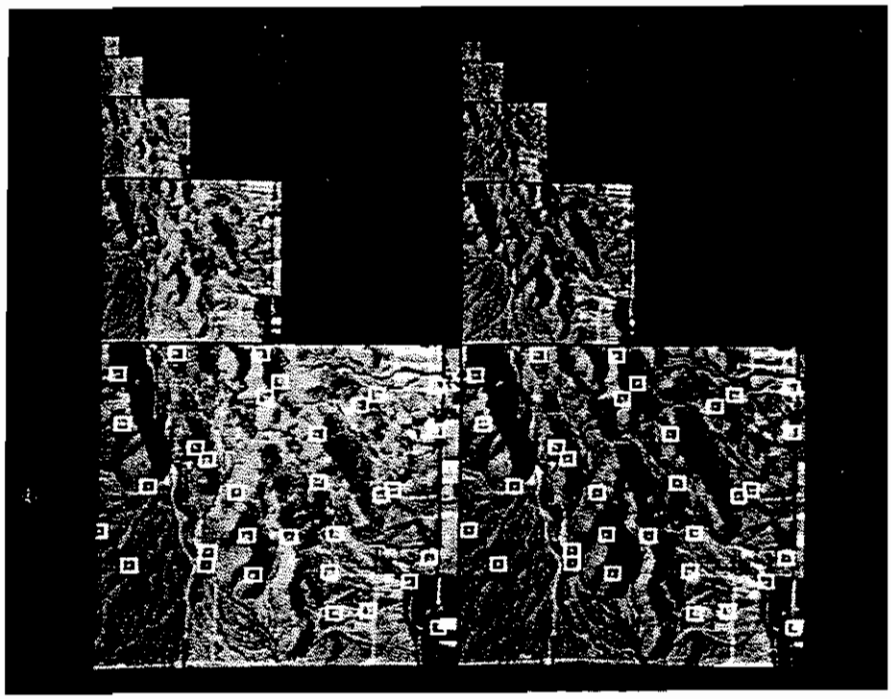

Figure 4--Results of Unstructured Hierarchical Matching of Most Interesting Point in Each Grid Cell.

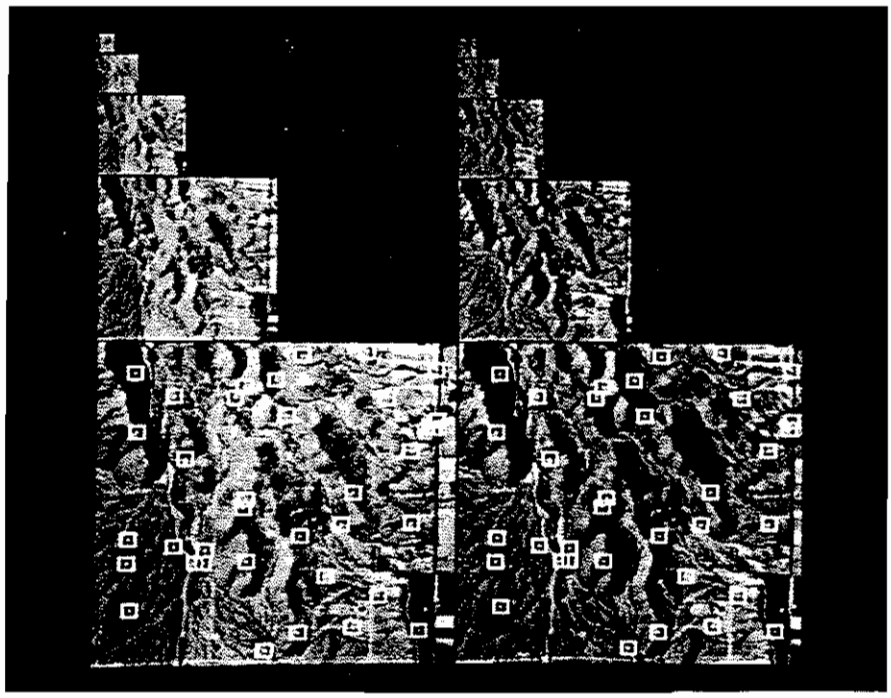

Figure 5--Results of Epipolar Hierarchical Matching of Second Most Interesting Point in Each Grid Cell.

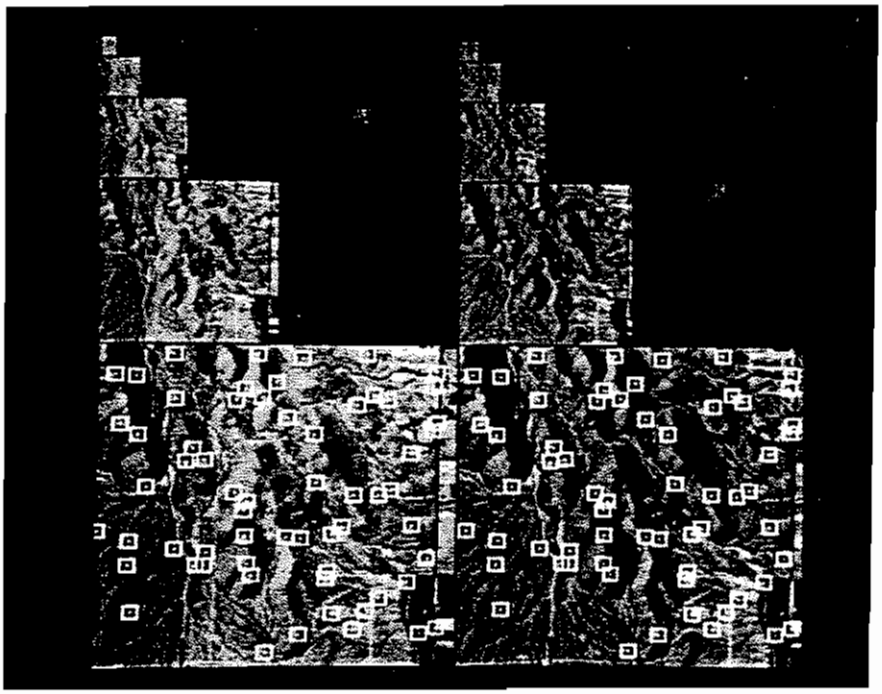

Figure 6--Results of Epipolar Hierarchical Matching of Two Most Interesting Point in Each Grid Cell.

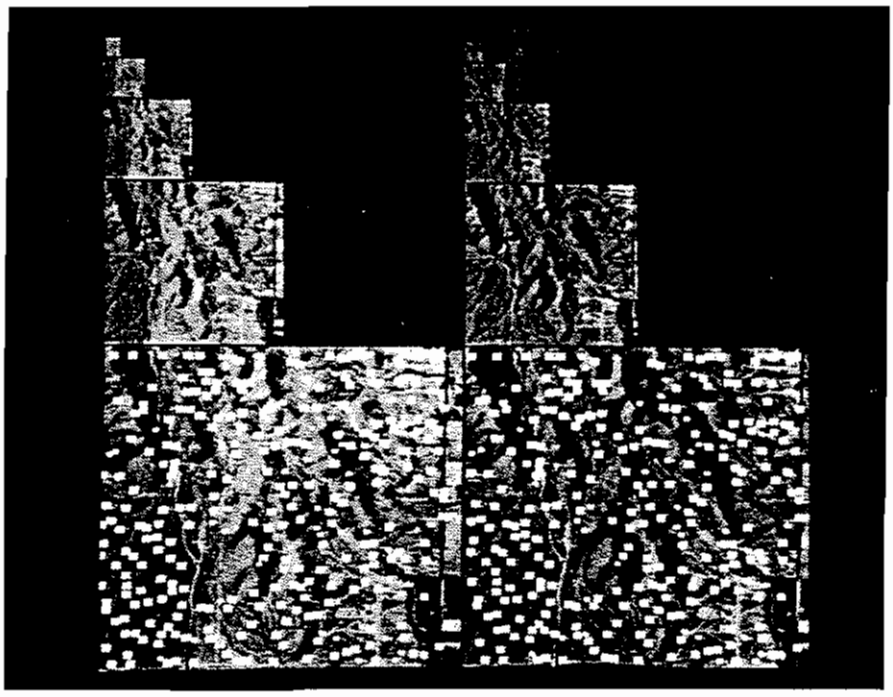

Figure 7--Results of Anchored Epipolar Matching of Remaining Interesting Points.

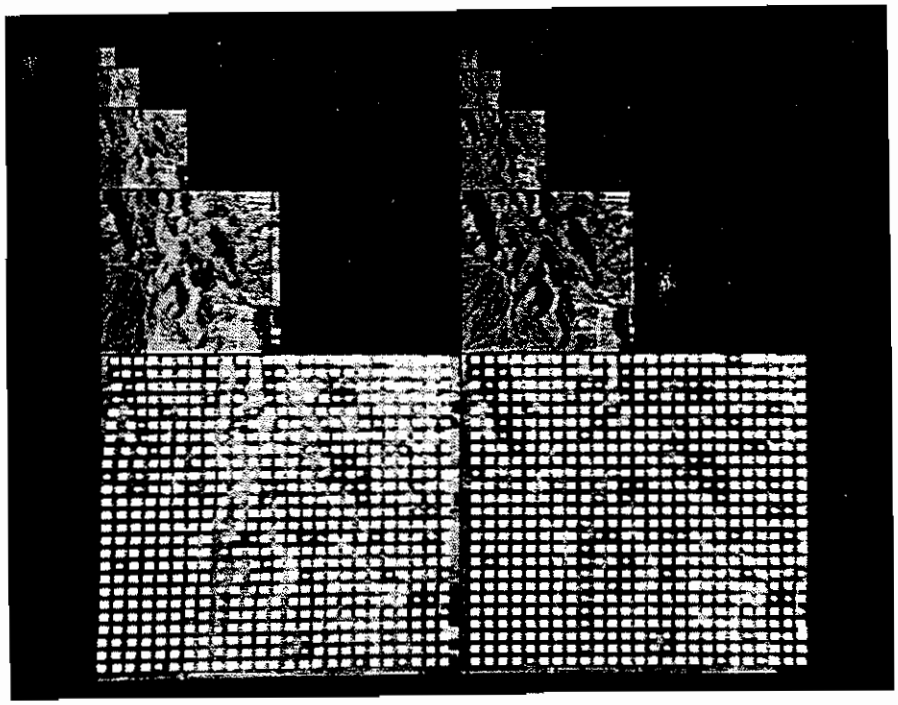

Figure 8--Results of Anchored Epipolar Matching of a Grid of Points.

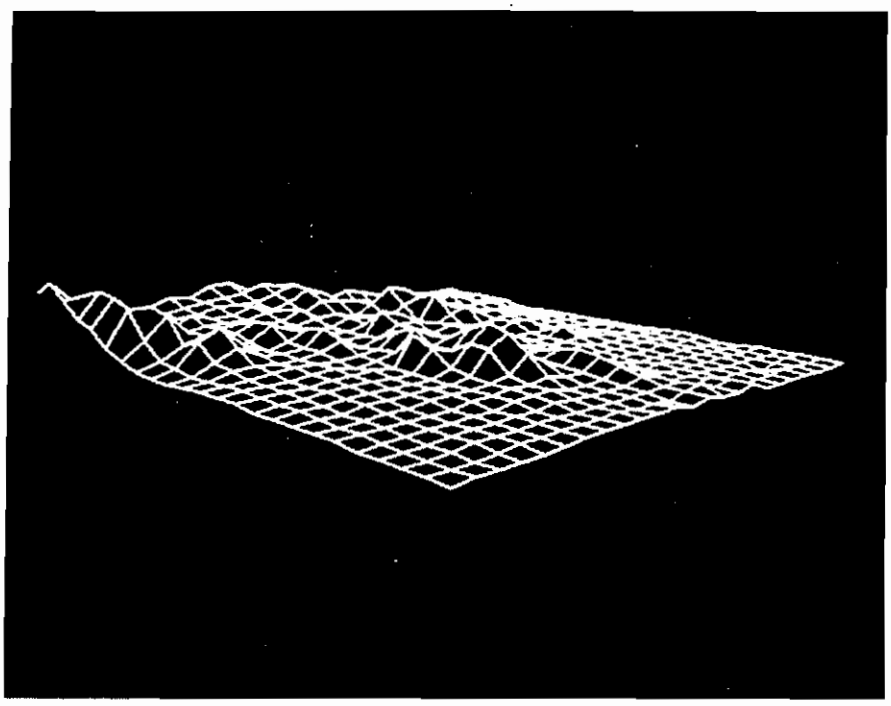

Figure 9--A View of the Resulting Digital Terrain Model.