# A System for Digital Stereo Image Matching

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

ABSTRACT: The digital stereo image processing literature is briefly reviewed, and an automatic system developed at SRI International for digital matching of points in aerial stereo imagery is described. Our system uses area-based correlation, but couples this very basic measure of match with a variety of novel search techniques to develop a disparity model for a given stereo image pair. The techniques used are hierarchical in nature, incorporate iterative refinement, and use a best-first strategy in the matching process; matches are then checked by a new technique that we call back-matching. Our techniques are illustrated using some of the results produced by this system when we participated in Image Matching Test A for ISPRS Working Group III/4.

## INTRODUCTION

AUTOMATIC TECHNIQUES for the production of three-dimensional (3-D) data by means of digital stereo matching are receiving increased interest for a variety of applications, including cartography (Panton, 1978), autonomous vehicle navigation (Hannah, 1980), and industrial automation (Nishihara and Poggio, 1983). The first and most difficult step in recovering 3-D information from a pair of stereo images is that of matching points from one digital image of the pair to the corresponding points in the second image. Many computational algorithms have been used in attempts to solve this problem (see Brady, (1982) or Barnard and Fischler (1982) for surveys of the field). These techniques primarily use area-based measures, such as correlation between image patches, or edge-based methods that match linear features in images.

Area matching techniques are the oldest and simplest of the stereo matching algorithms. Each image point to be matched is in fact the center of a small window of points in the first or reference image; this window is statistically compared with similarly sized windows of points in the second or target image of the stereo pair. The measure of match is either a difference metric that is minimized, such as RMS difference, or more commonly a correlation measure that is maximized, such as mean- and variance-normalized cross-correlation (Hannah, 1974). Because comparison of a given reference window to every possible target window is computationally expensive, various heuristics have been developed to limit the area that must be searched. In addition to the well-known epipolar constraint, these techniques have included extrapolation from already computed neighboring disparities (Panton, 1978), the use of image hierarchies (Moravec, 1980), and successive iterations of correlation and interpolation (Quam, 1984) in a hierarchy. Correlation works well most of the time, but encounters difficulties when the two images are taken from extremely different viewpoints, are of a scene that does not contain adequate visual texture, or are of a scene with many depth discontinuities. However, in these instances, and in the presence of image noise, correlation degrades gracefully—it usually continues to find the correct answer, but with reduced confidence measures.

Studies of human vision (Marr and Poggio, 1976) led to the development of edge-based methods, in which linear features are first extracted from the images by an edge operator (Hueckel, 1971; Hildreth, 1980), then matched using the epipolar constraint. Because the processing to extract edges throws away much of the information in the image, many heuristics have been developed to overcome the resulting match ambiguities. These include *a priori* modeling of the scene (Arnold, 1978), multiresolution coarse-to-fine strategies (Grimson, 1981), and longest-first prioritization of the order of edge matching (Baker, 1985). Because edges are usually somewhat sparse in the image, depths in areas between edges are filled in either by interpolation (Grimson, 1981) or by algorithms that match the image intensities between edges, using dynamic programming techniques (Baker and Binford, 1981; Ohta and Kanade, 1985). Most edge matching algorithms rely on the relative sparsity of edges, and thus tend to be confounded by images with densely textured areas or moderate levels of image noise, which is precisely where area-based matching excels. For this reason, edge matching should be regarded as complementary to, rather than as competing with, area-based matching. A simple experiment in the fusion of these two techniques (Baker, 1985) showed vastly improved results over either technique used alone.

In implementing a system at SRI International (SRI), we chose to base it on area-based techniques, because of their robustness and wide range of applicability over image types. These attributes were again demonstrated by our results on the test imagery provided by ISPRS Working Group III/4 as part of their Image Matching Test A.

## DESCRIPTION OF SRI'S STEREO SYSTEM

Over the past few years, SRI has integrated and improved existing pieces of stereo software into a baseline system for automated, area-based stereo compilation on aerial imagery. The system operates in several passes over the data, during which it iteratively builds and refines its model of a portion of the 3-D world represented by the disparities between a pair of images.

The first step in our matching process is to select a set of well-scattered windows in one image, such that each window contains sufficient information to produce a reliable match. To accomplish this, a statistical operator is passed over the image; this operator is a product of the image variance and the minimum of ratios of directed differences (hence edge strength) over windows of the specified size (Hannah, 1980). Local peaks in the output of this operator are recorded as the preferred places to attempt the matching process (Figure 1). The motivation behind this operator is that it penalizes windows with low information and windows whose only information is contained in strongly linear edges, because either of these situations can cause difficulties in obtaining the correct match by means of area-based correlation. The chosen windows are characterized by their center points, which are referred to as "interesting points" (Moravec, 1980). To ensure that we are working with selected windows that are well-scattered in the image, the image is divided into a grid of subimages, and the relative ranks of the best few interesting points within their grid cell are recorded; this permits the most interesting points in each area to be matched first.

Whether or not point $(x_1, y_1)$ in the first image $I_1$ is matched by point $(x_2, y_2)$ in the second image $I_2$ is determined by com-

PHOTOGRAMMETRIC ENGINEERING AND REMOTE SENSING,
Vol. 55, No. 12, December 1989, pp. 1765-1770.

0099-1112/89/5512–1765$02.25/0

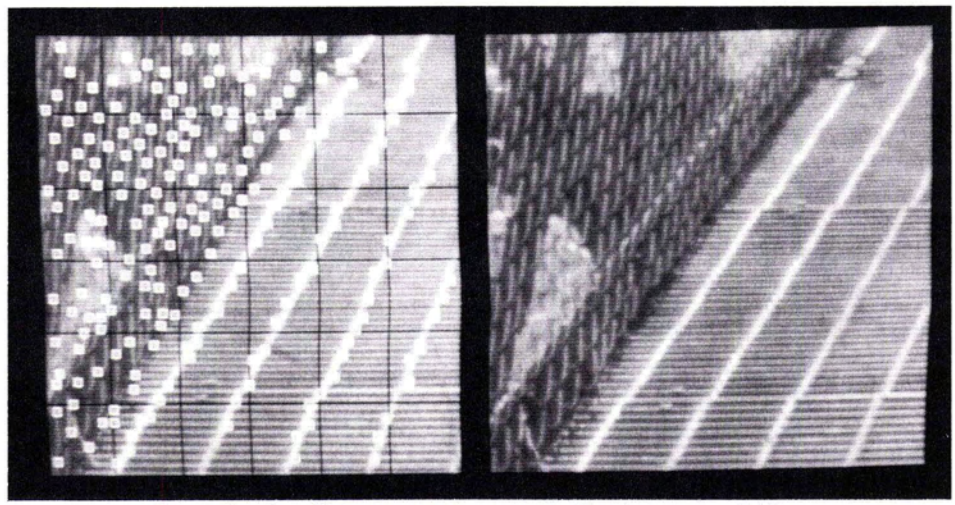

FIG. 1. Results of interest operator on ISPRS Test Image 5 — Bridge.

puting the cross-correlation, normalized by both mean and variance, over windows (typically 11 by 11 pixels) surrounding the points (Hannah, 1974). The matching point is taken to be the point in $I_2$ with highest correlation, as located by one of several search algorithms.

Our system employs several different matching algorithms. The underlying strategy is to begin with a few points that are highly likely to be matchable (based on their "interest" values, i.e., the information content of the windows surrounding the points); these are matched by very global, but very conservative, search algorithms. Each successive algorithm operates on less promising points, but uses more information from matches made at previous levels to constrain the search to smaller and smaller portions of the epipolar line, until eventually all interesting points have been processed. All of our matching algorithms use image hierarchies to some extent. Pixels in each reduced image of the hierarchy are produced by convolving the parent image with Gaussian, then sampling (Burt, 1980); images are almost always reduced in size by a factor of 2 at each step of the hierarchy (Figure 2).

The first matching algorithm, unconstrained hierarchical matching, 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 nor any significant time-lapse changes. Each specified point (usually the most interesting point in each grid cell) is matched using an unguided hierarchical matching technique (Moravec, 1980). This technique begins with a point in the largest image (that is, the highest-resolution image, in this case, the 240- by 240-pixel left image for each of the test sets) and numerically traces that point back up through that image's hierarchy (moving toward smaller, hence lower-resolution, images) by repeatedly scaling down the coordinates of the point until it reaches an image that is approximately the size of the correlation window. It then uses a two-dimensional spiral search, followed by a hill-climbing search for the maximum of the correlation between the image windows (Quam, 1971). This global match is then refined back down the image hierarchy (moving toward the larger, hence higher-resolution, images); that is, the disparity at each level (suitably magnified to account for relative image scales) is used as a starting point for a hill-climbing search at the next level (Figure 2). 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 result of applying unconstrained hierarchical matching to the most interesting point in each grid cell.

In this matching technique, as in all the others we use, matches must pass fairly strict tests in order to be considered correct, and only the successful matches are recorded for further use. At any level in the hierarchy, matches with very poor correlation (compared either against an absolute threshold or with respect to an autocorrelation-based threshold (Hannah, 1974)) are discarded, as are matches that fall outside the image.

Each match must also be confirmed by a technique that we call back-matching. Having found that point $(x_1, y_1)$ in the first image $I_1$ is best matched by $(x_2, y_2)$ in the second image $I_2$, we then repeat the entire matching algorithm, this time starting with $(x_2, y_2)$ in $I_2$ and searching for the point $(x'_1, y'_1)$ in $I_1$ that best matches $(x_2, y_2)$. If $(x_1, y_1)$ and $(x'_1, y'_1)$ differ by more than one pixel, the entire match is discarded as being unreliable.

The addition of back-matching to our matching algorithms has improved both the number of points for which matches could be found and the reliability of the matches accepted. Matches are more reliable because each one has been confirmed by a second, independent matching process. More matches can be found because the matching process no longer needs to rely on carefully tuned correlation thresholds to separate good matches from bad. Confirmed matches are accepted, despite low correlation values, while matches with high correlations that cannot be confirmed are rejected. Correlation thresholds are still used, but they are now set at much lower levels than was possible without back-matching, thus permitting many more matches.

Further processing makes use of camera models (i.e., relative or absolute orientation data). For imagery supplied with camera models, the given information is used. If camera models are unavailable, undecipherable, or unreliable, the system can calculate a simplistic relative camera model from the set of point pairs produced by unconstrained hierarchical matching. This is accomplished by searching for five angles that describe the relative positions and orientations of two ideal pinhole cameras (Hannah, 1974). The object of the search is to minimize the error between $(x_2, y_2)$ in $I_2$ and the epipolar line produced when $(x_1, y_1)$ in $I_1$ is projected into space, then into $I_2$ through the hypothesized pinhole cameras. The search proceeds by a lineari-

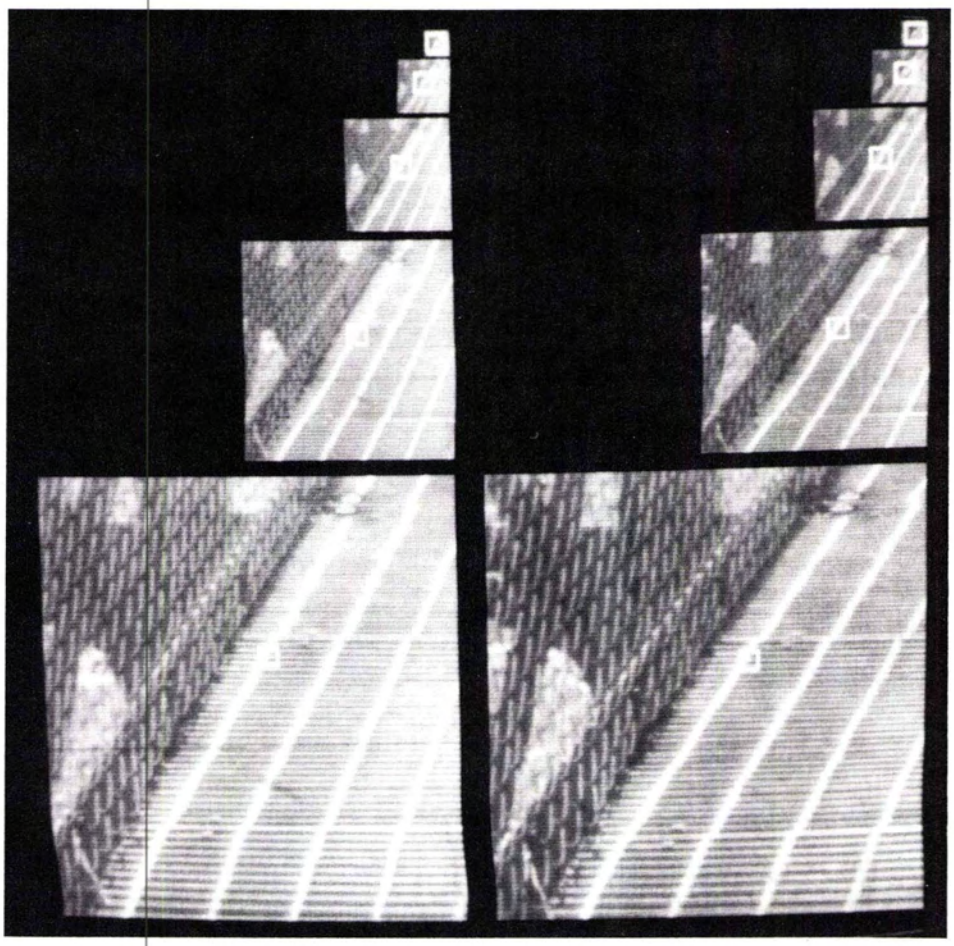

FIG. 2. Example of unconstrained hierarchical matching.

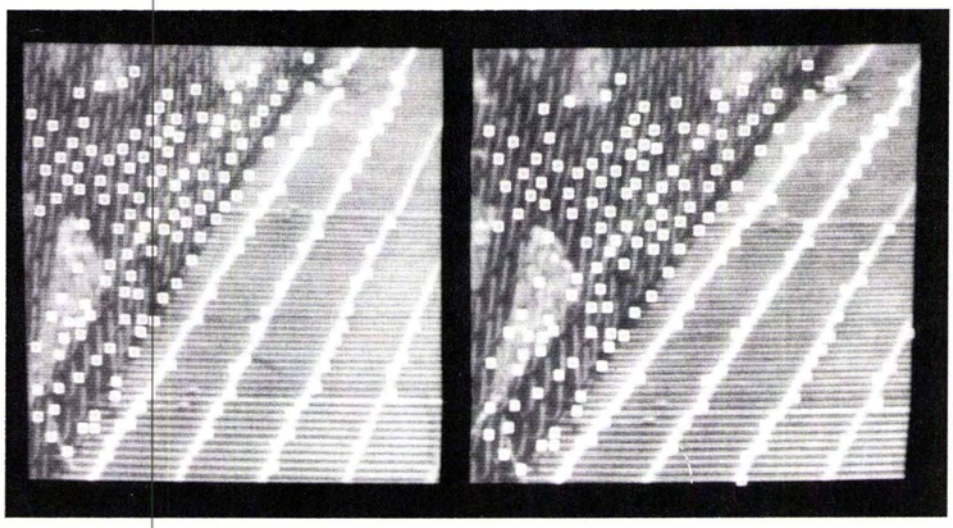

FIG. 3. Results of unconstrained hierarchical matching.

zation of the equations and their analytic derivatives (Gennery, 1980). Once a solution is found, the reliability of each matched point is assessed. Points that appear to contribute too much error to the solution are removed from the calculation, and the solution is redone. This process will reach a successful conclusion if a subset of points is found that allows convergence to a consistent model, or it will report failure if too many points are rejected.

The next technique to be applied is epipolar constrained hierarchical matching. Having the camera parameters, we now know the manner in which a point in the first image projects to a line in the second image—the epipolar constraint. This

constraint allows us to cut the search from two dimensions (all around the point) to one dimension (back and forth along the epipolar line) at each level of the hierarchy. In all other respects, epipolar constrained hierarchical matching proceeds very much like unconstrained hierarchical matching, with the additional match-evaluation criteria that matches must lie within a specified distance of the epipolar line. This technique is used on any unmatched points among the two most interesting points for each grid cell.

Once a good basis of reliable matches has been found, these matches can be used as "anchor" points for the anchored matching technique, which again uses the grid cells in the image (Figure 4). A given point will lie in some grid cell; the closest matched points should lie in that cell or in one of the eight 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 neighbors. Thus, the disparity for a point is expected to lie in the interval of the disparities of the well-matched points in the current and neighboring cells. This disparity interval is used along with the epipolar constraint to perform a very local search for the match to a point, perhaps proceeding one or two levels up the image hierarchy, to provide context for the match. All matches are required to pass the same tests as for the hierarchical matching algoithms described previously, including an anchored-matching version of the back-matching test. Figure 5 shows all of the interesting points that were matched by these techniques.

Our system can produce matched points on a regularly spaced grid (in image coordinates), if desired. This matching algorithm also uses the anchored matching technique, searching along specified portions of the epipolar line, to calculate matches for the user-specified grid of points in the first image (Figure 6). However, holes can result if a grid point does not have suitable information for matching, and again, only matches that pass all the tests are recorded. This highlights a problem with matching a grid of points – not all areas of an image have information suitable for matching, and forcing a match at such areas can lead to poor results. Matching on a grid in the image must be used with caution.

Our system also incorporates code that can use randomly spaced matches to interpolate either disparities or elevation values for points that were not matched directly. This technique can be used either to fill in holes in a grid, or to produce results on a grid that is more closely spaced than that provided by the stereo matching process. However, this technique [Smith, 1984] explicitly assumes a single, continuous surface, an assumption that is not always met in stereo imagery. For this reason, interpolation must also be used with caution.

## RESULTS ON IMAGE MATCHING TEST A

To test our system, we participated in the ISPRS Working Group III/4 Test A on Image Matching (Gülch, 1988). In this test, 12 pairs of digital images, each 240 by 240 pixels in size, were provided to various stereo researchers, representing both the photogrammetric community and computer vision researchers. The objective of the test was to assess the state of the art in image matching – how well the various procedures could handle imagery of different types and complexities, how much *a priori* information was needed to produce good results, how well the procedures could assess the quality of their results, and how precise the disparities were. Each participant's results (limited to 300 points per image) were submitted to the University of Stuttgart's Institute for Photogrammetry, where they were checked against disparities (parallaxes) determined by traditional manual photogrammetry. The imagery used in this test represented a wide variety of image subjects (from engine compartments of cars to aerial photos), image scales (from 1:20 to 1:30,000), and image quality (from very crisp to blurred or grainy). Most of this imagery does not reproduce well, so we have used only a few of the clearer images in our examples.

For each of the 12 image pairs they provided, we attempted to perform their Standard Task B – determination of the parallaxes at selected points – which is what our system does best; for a few of the images, we also performed Standard Task A – determination of the parallaxes at a grid of points. Most of the images were run with the standard parameters for the system, which had been set while processing a large, high-quality, aerial image pair unrelated to the test imagery. Because of incompatibilities in format, we did not use the camera information given with each test image pair, or any other *a priori* information; we used the raw images, without transforming to normal images or doing any other resampling.

For the most part, matching proceeded routinely, using the standard procedures and parameters. One parameter – a threshold on the "interest" value that indirectly controls the number of interesting points that the system has to work with – was changed for each data set; this was done to produce between 100 and 300 points per image, as requested for the test.

For 10 of the 12 tests, we were able to handle the image pairs without substantially altering the default parameters or proces-

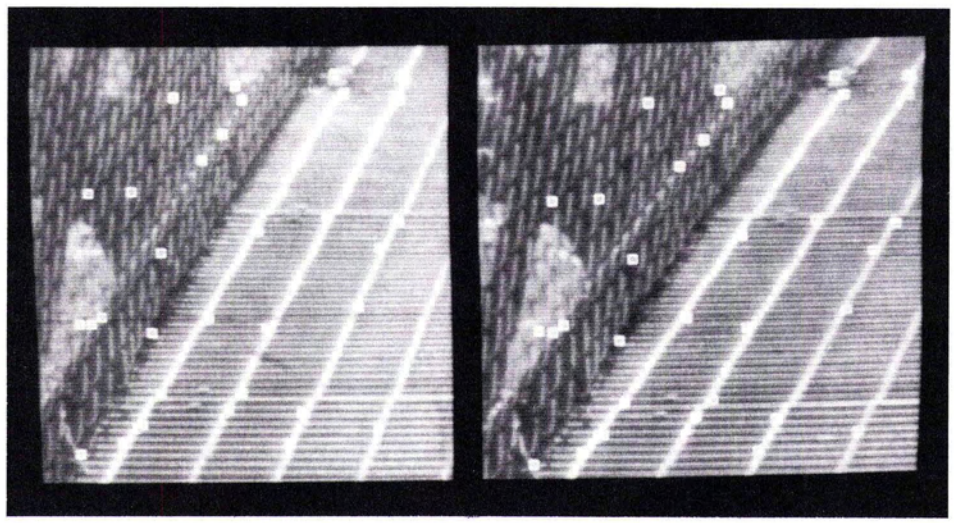

FIG. 4. Example of anchor points for matching.

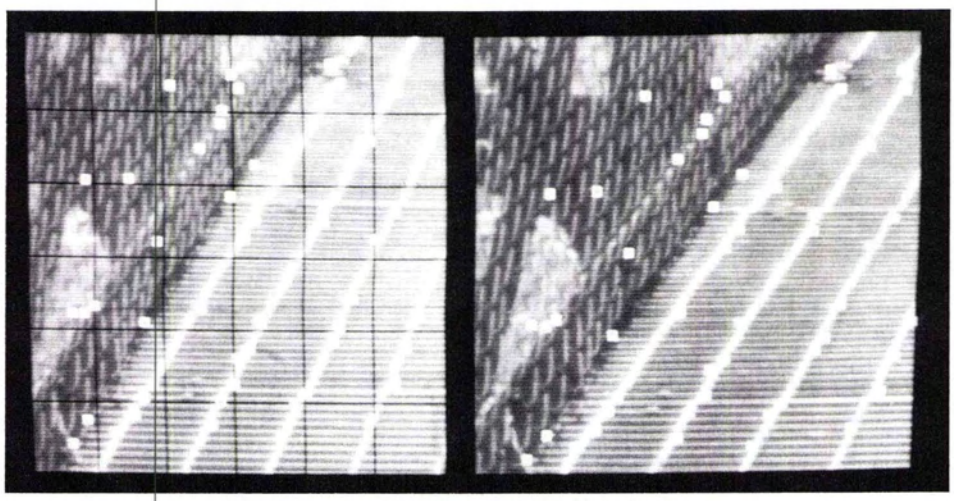

FIG. 5. Combined results of all matching algorithms.

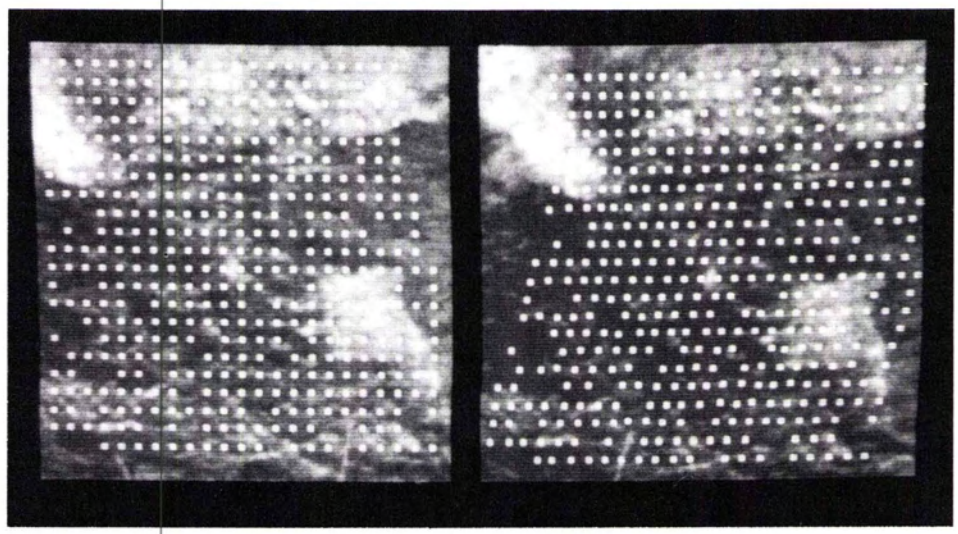

FIG. 6. Results of grid matching algorithm on ISPRS Test Image 2—Quarry.

ing sequence (Hannah, 1988). Our only problems came on the two image pairs of the Olympia dome in Munich; this is not surprising, because our algorithms were designed for use on highly textured natural terrain, not the bland faces of cultural objects with mostly linear or ambiguous features. For one of the Olympia images, we were able to obtain results (Figure 7) with a minor change in parameters and the hand-deletion of a single mismatch that was preventing us from forming a camera model. (If our relative camera model solver included the RANSAC technique (Fischler and Bolles, 1980), we believe it could have proceeded without manual intervention.) On the other Olympia image pair, the unconstrained hierarchical matching technique failed to produce any correct matches. We believe that the combination of repetitive structures, very different points of view, and the transparency of the dome caused our hierarchical matching techniques to fail. If we had elected to use the accompanying camera information, we might have been able to do some matching, although the transparency would undoubtedly have led to numerous problems.

For the most part, our results appear to be reasonably correct, even in the face of large disparity ranges within small areas of the image. Preliminary review by the test architects indicated that our algorithm was in a two-way tie for most images processed (11 out of 12), and was clearly the most accurate of the algorithms participating in the test (first in accuracy in 6 of the 12 images, and never below fifth among the 17 test participants) (Gülch, 1988). As of this writing, we have not received the final results of the committee's detailed analysis, which should be very interesting.

## SUMMARY

In this paper, we have described SRI's automatic system for stereo image matching, a system that uses area-based correlation but applies this basic technique in a variety of novel ways. Our techniques are hierarchical in nature, and use iterative refinement, as well as a best-first strategy, in the matching process, and they apply the new constraint of back-matching to verify matches. Finally, we have illustrated our techniques by presenting some of our results on the Image Matching Test A data set recently distributed by ISPRS Working Group III/4. Our results on ISPRS test imagery have again demonstrated the robustness of the correlation-based matching technique, its wide range of applicability over image types, and its accuracy.

## ACKNOWLEDGMENTS

The research reported herein was supported by the Defense Advanced Research Projects Agency under Contracts MDA903-83-C-0027, MDA903-86-C-0084, and DACA76-85-C-0004. The

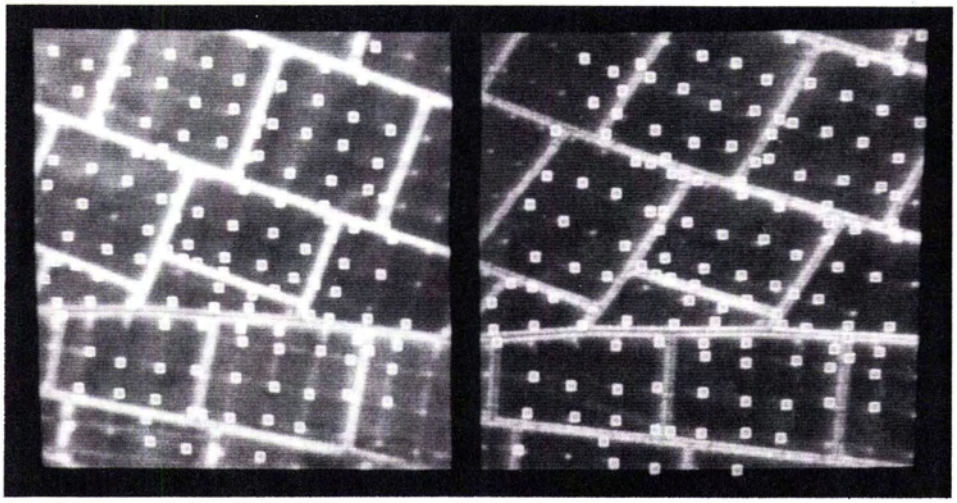

FIG. 7. Overall results on ISPRS Test Image 3 — Olympia I.

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 Harlyn Baker, Lynn Quam, Grahame Smith, and Martin Fischler for their support on this project.

## REFERENCES

Arnold, R. D., 1978. Local Context in Matching Edges for Stereo Vision, *Proceedings: Image Understanding Workshop*, Cambridge, Mass., May 1978, pp. 65–72.

Baker, H. H., 1985. Reimplementation of the Stanford Stereo Systsem and Integration Experiments with the SRI Baseline Stereo System, unpublished presentation at *Image Understanding Workshop*, Miami Beach, Florida, December 1985; also available as SRI International Artificial Intelligence Center Technical Note 431, January 1988.

Baker, H. H., and T. O. Binford, 1981. Depth from Edge and Intensity Based Stereo, *Seventh International Joint Conference on Artificial Intelligence*, Vancouver, B.C., August 1981, pp. 631–636.

Barnard, S. T., and M. A. Fischler, 1982. Computational Stereo, *ACM Computing Surveys*, Vol. 14, No. 4, pp. 553–572.

Brady, M., 1982. Computational Approaches to Image Understanding, *ACM Computing Surveys*, Vol. 14, No. 1, pp. 3–71.

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

Fischler, M. A., and R. C. Bolles, 1980. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, *Proceedings: Image Understanding Workshop*, College Park, Maryland, April 1980, pp. 71–88.

Gennery, D. 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.

Grimson, W. E. L., 1981. *From Images to Surfaces: A Computational Study of the Human Early Visual System*, M.I.T. Press, Cambridge, Mass.

Gülch, E., 1988. Results of Test on Image Matching of ISPRS WG III/4, *International Archives of Photogrammetry and Remote Sensing*, Vol. 27-III, pp 254–271, and accompanying poster presentation, XVI ISPRS Congress, Kyoto, Japan, July 1988.

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

———, 1980. Bootstrap Stereo, *Proceedings: Image Understanding Workshop*, College Park, Maryland, April 1980, pp. 201–208.

———, 1988. Digital Stereo Image Matching Techniques, *International Archives of Photogrammetry and Remote Sensing*, Vol. 27-III, pp 280–293, XVI ISPRS Congress, Kyoto, Japan, July 1988.

Hildreth, E. C., 1980. *Implementation of a Theory of Edge Detection*, M.Sc. Thesis, Massachusetts Institute of Technology, Artificial Intelligence Laboratory Technical Report AI-579.

Hueckel, M., 1971. An Operator Which Locates Edges in Digital Pictures, *Journal of the Association for Computing Machinery*, Vol. 18, pp. 113-125.

Marr, D., and T. Poggio, 1976. Cooperative Computation of Stereo Disparity, *Science*, Vol. 194, pp. 283-287.

Moravec, H. 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. K., and T. Poggio, 1983. Stereo Vision for Robotics, *Proceedings of the International Symposium of Robotics Research*, Bretton Woods, New Hampshire, September 1983.

Ohta, Y., and T. Kanade, 1985. Stereo by Intra- and Inter-Scanline Search Using Dynamic Programming, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-7, No. 2, pp. 139–154.

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

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

———, 1984. Hierarchical Warp Stereo, *Proceedings: Image Understanding Workshop*, New Orleans, Louisiana, October 1984, pp. 149–156.

Smith, G. B., 1984. *A Fast Surface Interpolation Technique*, SRI International Artificial Intelligence Center Technical Note 333, August 1984.

(Received 6 March 1989; revised and accepted 20 March 1989)