

# SRI's Baseline Stereo System

Marsha Jo Hannah

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

## Abstract

We have implemented a baseline system for automated, area-based stereo compilation. This system, STEREO SYS, 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 STEREO SYS and give examples of the results it produces. We find that these results agree quite well with the best available benchmark—results produced on the interactive DIMP system at the U.S. Army Engineer Topographic Laboratories.

## 1 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 and 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 forms the core of an ever-improving stereo system. We have also tested the baseline system [Hannah, 1985-a] against a "challenge data base" [Hannah, 1985-b] of image areas where conventional stereo techniques encounter difficulty.

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

- Construct hierarchies for stereo images
- 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 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 and match evaluation.

## 2 The Stereo System

SRI has integrated existing pieces of stereo code into a baseline system for automated area-based stereo compilation, then improved the system to its present form. 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 STEREO SYS (STEREO SYSTEM). It allows the user to invoke 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 STEREO SYS 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.

### 2.1 Preliminary Processing

Before the actual stereo matching can begin, some preliminary image processing is necessary. This includes the creation of the image hierarchy and the selection of the interesting points to be matched.

#### 2.1.1 Creating the Image Hierarchy

The basis for the image matching techniques is a hierarchy of images, as shown in Figure 1. The module of STEREO SYS that forms this hierarchy from the original images is called REDUCE. In the example used for the figures, the original images are a pair of image "chips" digitized from standard 9" × 9" mapping pho-

tos taken over Phoenix South Mountain Park, near Guadalupe (a suburb of Phoenix), Arizona. These images are  $2048 \times 2048$  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  $1024 \times 1024$ ,  $512 \times 512$ ,  $256 \times 256$ ,  $128 \times 128$ ,  $64 \times 64$ ,  $32 \times 32$ , and  $16 \times 16$  pixels in size. (Figure 1 shows only the  $256 \times 256$  through  $16 \times 16$  image pairs.)

REDUCE ordinarily produces pixels in each reduced image by convolving the image with a Gaussian, then sub-sampling [Burt, 1981]. Older code also exists to reduce images by simple averaging of the pixels in an  $N \times N$  square from the next-largest image (in most cases,  $N=2$ ). It is known that this technique can produce artifacts in the data, and the more sophisticated Gaussian technique is preferred.

### 2.1.2 Selecting Interesting Points

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 strength 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 various correlation techniques. (A possible enhancement to STEREO SYS 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.)

The module INTEREST permits the user to specify the interest operator to be used [Hannah, 1980], the window size over which it is calculated, and the minimum spacing for interesting points. 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) in 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  $6 \times 6$  grid of cells.

## 2.2 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 a stereo pair taken with handheld cameras), an unstructured hierarchical matching algorithm is used on the most interesting points. The results of these matches are used in seek-

ing a solution for a simplistic relative camera model (5 angles describing the relative positions and orientations of 2 ideal pin-hole cameras [Hannah, 1974]), which can then be used for the epipolar constraint in further matching. 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.

### 2.2.1 Unconstrained Hierarchical Matching

Unconstrained hierarchical matching is done by the module 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  $2048 \times 2048$  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  $16 \times 16$  image for the  $11 \times 11$  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 plausibility of the final match is then checked by reversing the roles of the right and left images and repeating the unconstrained hierarchical search, starting with the just-found matching point. In order for the match to be believed, this reverse search must produce a match at (or immediately adjacent to) the original interesting point. The addition of this constraint, which effectively requires co-operative results between the two images, has greatly reduced the number of bad matches in photographs which violate one or more of the assumptions around which STEREO SYS was designed. 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 image chips covering the matched area in the  $2048 \times 2048$ ,  $1024 \times 1024$ , and  $512 \times 512$  images; these are shown in the upper right corner of each hierarchy. The matching began in the right image in the  $2048 \times 2048$  chip, traced this point through the right-hand hierarchy (approximately clockwise in the figure) to the  $16 \times 16$  right image, matched that to the  $16 \times 16$  left image, then refined the match back through the left image hierarchy until reaching the left  $2048 \times 2048$  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  $256 \times 256$  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  $2048 \times 2048$  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  $1024 \times 1024$  images, the highest resolution image in which the noise was considered tractable. Once processing is complete, STEREOSYS can be used to refine the final matches from this level down to the original  $2048 \times 2048$  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 STEREOSYS.

### 2.2.2 Relative Camera Model Calibration

If no camera calibration information is available, the module C2MODEL can be used to calculate 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 pinhole 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 pinhole cameras, thus totally ignoring the internal geometry of real cameras. However, in many cases, it is suitable for approximating the epipolar constraint to simplify further matching.

### 2.2.3 Epipolar Constrained Hierarchical Matching

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 two dimensions (all over the image) to one dimension (back and forth along the epipolar line), as implemented in the module LMATCH.

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 in each grid cell and on any already-matched 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.

## 2.3 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 neighbors. Thus, the disparity at 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. Note that a point is considered to be well-matched if it has a correlation above a user-settable absolute threshold, usually 0.5, and above a variable threshold, based on the autocorrelation function around the point in the first image (see Table 1 for examples); in addition, a well-matched point cannot deviate more than a specified distance from the epipolar line.

### 2.3.1 Matching the Rest of the Interesting Points

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, searching along a portion of the epipolar line, to find these matches. Figure 7 shows the results of this module. Only points found to be well-matched are recorded.

### 2.3.2 Matching a Grid of Points

STEREOSYS permits the user to produce matched points on a closely spaced grid, if desired. The module GMATCH also uses the anchored match technique, searching along the epipolar line, to calculate matches on a user-specified grid. Figure 8 shows the results of this module on a  $20 \times 20$  grid. The smaller marks indicate matches in which STEREOSYS has little confidence; these are currently not recorded in the data structure, leaving holes in the grid. This points up a problem with grid matching—not all areas of an image have information suitable for matching, and forcing a match at such areas can lead to poor results.

## 2.4 Post Processing

Although not strictly stereo processing, there are follow-up processes which are necessary to turn stereo disparities into more meaningful 3-dimensional quantities. These processes include interpolation and terrain modelling.

### 2.4.1 Interpolation

Often, it is not feasible to apply correlation matching at points on a pre-determined grid. Even when grid matching is feasible, there will be areas of the images that cannot be matched, due to noise in the data, insufficient information, or changes such as moving vehicles; this will result in “holes” in the grid of terrain

data, which must be filled in somehow. And, frequently, a terrain model is desired that has its points more closely spaced than that provided by the stereo matching process. In all of these cases, interpolation of the matched data points is necessary to provide information at other points. STEREOSYS incorporates an efficient interpolation scheme [Smith, 1984], permitting the user to construct elevation data grids from either randomly spaced points or a widely spaced grid of points.

### 2.4.2 Terrain Modeling

Given the dense grid of matched points and the camera calibration, it is possible to derive a digital terrain model. If absolute external camera information and internal camera calibration is available, the module STERDTM 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.

## 3 Evaluation

Evaluation of the accuracy of STEREOSYS is difficult, as there do not seem to exist stereo data sets with known ground truth against which to compare 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  $5 \times 5$  grid in the  $2048 \times 2048$  images of the Phoenix data set, and which 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 randomly spaced interesting points in the  $1024 \times 1024$  images without warping. Comparing them is a little like comparing apples and oranges, but we did so in the following manner.

Comparisons were made only for those points for which STEREOSYS recorded an answer. Points were said to have the same answer if the STEREOSYS result and the result at the closest DIMP grid point (scaled into the  $1024 \times 1024$  image in which STEREOSYS produced its results) were within one pixel of having the same disparity. Points about which there was disagreement were examined manually. An analyst looked at both results, overlaid on the images at a variety of resolutions, both monocularly and using a stereoscopic viewer, then decided which algorithm appeared to be in error and, based on experience with correlation algorithms, attempted to determine why the mistake had been made.

On the Phoenix data set, STEREOSYS found 5545 "interesting points," of which it thought it could reliably match 4676. Of these, only 43 disagreed significantly with the DIMP results for nearby points. Closer examination showed 15 of these to be uncorrected DIMP errors, 15 were STEREOSYS errors, 5 were points on which both systems appear to have made errors, and 8 were points for which the analyst could not determine which system was in error. In most of the cases, the DIMP errors seemed to

result from its algorithm having drifted gradually off track (usually starting in an area with little information), and its operator not catching it soon enough. The STEREOSYS approach of first providing a context in which to work, so that the code interpolates disparities, instead of extrapolating them, should remedy this problem. Most of the STEREOSYS errors (and almost all of the points for which the analyst could not determine which algorithm was at fault) appeared to have resulted from an inappropriate threshold on the interest value: STEREOSYS was trying to match areas in which there was not enough information to make reliable matches. Some of the STEREOSYS errors were due to not using warped correlation windows to account for the slopes. In these cases, most of the information in a window would be in a corner of the window, so the disparity that was calculated was that of the corner, not the center of the window; using warped correlation or exponentially weighted interest operators and correlation windows [Quam, 1984] would solve this problem. A fair number of the mistakes (particularly the ones in which both systems arrived at different wrong answers) were because of artifacts in the data—film grain, scratches, lint, hairs, fiducial marks, and the like; we are a long way from being able to understand, let alone automate, the human ability to identify offending objects and then ignore them in processing stereo data.

STEREOSYS has also been used on several other data sets in our "challenge data base", described in Hannah [1985-b]. For data sets with no DIMP results, a much smaller number of points were matched. These were then compared with the human viewer's perception of what were the correct matches. Only the more blatant mistakes were detected and further analyzed; the results of which are presented in Hannah [1985-a].

## 4 Discussion

Our objective in constructing STEREOSYS was to implement a state-of-the-art, area-based system for stereo compilation operating on aerial photography. Along the way, we hoped to remedy some of the obvious problems we had seen with existing systems, such as DIMP's tendency to extrapolate itself off track. In this we have succeeded.

Because STEREOSYS uses fairly independent judgment on each match, it tends to avoid the problems we have seen in the DIMP results; indeed, on the Phoenix data set, STEREOSYS was able to duplicate DIMP's correct results (for the points tried) and rectify a number of DIMP's mistakes. Although it happens rarely, it is still possible for STEREOSYS to make mistakes in the early stages of its processing, then propagate these mistakes into later matches. To avoid this, more work needs to be done on algorithms for detecting improperly matched points, so they can be removed before further processing.

The major criticism we have heard of STEREOSYS is that it produces matches at randomly spaced points (only where adequate information is present), when what is usually wanted is a closely spaced regular grid of elevation points, regardless of image content. So far, attempts at blindly interpolating the disparity data (ignoring the image data) as reported in Smith [1984] have proven less than satisfying. Marriage of the STEREOSYS techniques with something like DIMP, or with hierarchical warp correlation [Quam, 1984], or with image intensity-based interpolation [Smith, 1985 or Baker, 1982] might be profitable.

We have performed one experiment as a preliminary study in how to integrate the strengths of STEREOSYS with those of an edge-based matcher. The results of STEREOSYS were

used as seeds for an edge-based matching system [Baker, 1982], which propagated these matches along the nearby zero-crossing contours, then did one iteration of edge matching. Because determining disparity constraints is a large part of the edge-based matcher's processing, introducing this information from STEREOSYS's results produced a significant reduction in computation time used by the edge-based matcher. The number of matched points also increased by about an order of magnitude over the results of STEREOSYS alone. Although we have not yet finished a quantitative evaluation of these match accuracies, a qualitative analysis indicates that the results from the combined technique are significantly more accurate than the results of the edge-based system alone.

Overall, we have found that STEREOSYS performs credibly on the low-resolution aerial imagery for which it was designed. It has difficulties when processing areas that violate its premises about the continuity of the world, but linking it with an edge-based matcher (which would excel in these types of areas) seems to be a promising approach.

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

### References

- Baker, H. Harlyn, 1982.** "Depth from Edge and Intensity Based Stereo," Ph.D. Thesis, Stanford University Computer Science Department Report STAN-CS-82-930, September 1982.
- Burt, Peter J., 1981.** "Fast Filter Transforms for Image Processing," *Computer Graphics and Image Processing*, Vol. 16, pp. 20-51, 1981.
- 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.
- Hannah, Marsha Jo, 1985-a.** "Evaluation of STEREOSYS vs. Other Stereo Systems", SRI International Artificial Intelligence Center Technical Note 365, October, 1985.
- Hannah, Marsha Jo, 1985.** "The Stereo Challenge Data Base", SRI International Artificial Intelligence Center Technical Note 366, October, 1985.

**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.

**Quam, Lynn H., 1984.** "Hierarchical Warp Stereo," *Proceedings: Image Understanding Workshop*, New Orleans, LA, October, 1984, pp. 149-156.

**Smith, Grahame B., 1984.** "A Fast Surface Interpolation Technique," *Proceedings: Image Understanding Workshop*, New Orleans, LA, October, 1984, pp. 211-215.

**Smith, Grahame B., 1985.** "Stereo Reconstruction of Scene Depth," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, June 9-13, 1985.



Figure 1—Reduction Image Hierarchy.

