Performance Evaluation of Scene Registration and Stereo Matching for Cartographic Feature Extraction

Yuan C. Hsieh, David M. McKeown, Jr., Member, IEEE, and Frederic P. Perlant

Abstract—One goal of automated cartography is to generate an accurate 3-D model of man-made structures and natural terrain. Some of the most challenging problems in cartographic feature extraction occur in dense urban areas where the level of detail and scene clutter greatly complicate traditional map compilation techniques. In this paper, we describe experiments toward a comprehensive stereo analysis system to recover the 3-D description of an urban area using high-resolution aerial imagery. Given an area of interest in terms of geographic coverage, our system can automatically find the appropriate stereo pair using a spatial database, select control points to register the two images so that epipolar geometry is satisfied, and recover disparity information using two complementary matching techniques.

In our research, we do not assume that the initial input images satisfy the epipolar geometry constraint because this is rarely the case in rectified aerial imagery. Therefore, we argue that stereo mapping research must explicitly address error and uncertainty in both scene registration and stereo matching and that we need techniques to evaluate such errors in a rigorous manner. We also argue that in order to achieve robust behavior, multiple methods for scene feature extraction should be utilized, and if possible, their results should be integrated into a consistent framework.

We describe techniques for scene registration using five different features that can be automatically extracted to provide control points for fine image registration. In the stereo matching process, two techniques are utilized: an area-based and a feature-based stereo matcher to generate a disparity map for a scene. We also present some preliminary results on a technique to merge the results of the stereo matching algorithms to provide improved information regarding height estimates.

Finally, we describe techniques to generate rigorous performance analysis metrics to compare stereo matching algorithms based on a manually derived 3-D ground truth segmentation. The analysis includes the error estimation metrics for both height and delineation accuracy based on the measurements of deviations from manual estimates. These estimates are computed globally over the entire scene and locally on a structure-by-structure basis.

Relative accuracy of the area-based, feature-based, and merged disparity estimates are provided for several different test scenes.

Index Terms—Analysis of aerial imagery, cartography, computer vision, digital mapping, image understanding, performance evaluation, scene registration, stereo matching.

I. INTRODUCTION

O NE GOAL OF automated cartography is to generate an accurate 3-D model of man-made structures and natural terrain. Some of the most challenging problems in cartographic feature extraction occur in dense urban areas where the level of detail and scene clutter greatly complicate traditional map compilation techniques [1]. The traditional method for obtaining a 3-D model of the terrain and man-made structures involves matching of stereo-pair imagery. Algorithms for stereo correspondence can be grouped into two major categories: area-based and feature-based matching [2].

Both classes of techniques have advantages and disadvantages that depend on the task domain and the 3-D accuracy required. Area-based approaches tend to be more robust in scenes containing a mix of buildings and open terrain. However, for complex urban scenes, feature-based techniques appear to provide more accurate information in terms of locating depth discontinuities and in estimating height. No single technique performs well in both circumstances. It is precisely for this reason that we are investigating both methods of stereo matching with a goal of utilizing multiple results to achieve more accurate and robust 3-D interpretations.

In both the area-based and feature-based techniques, the epipolar constraint is used to simplify stereo matching by reducing it to a 1-D problem. This is usually achieved by registering the two stereo pairs. The assumption that the scene registration is ideal and that the epipolar constraint is totally satisfied is rarely warranted in imagery digitized from aerial photography. Careful local registration is often required after the scenes have been coarsely aligned. Local registration needs a set of control points that are abundant, are well distributed throughout the scene, and can be matched in the stereo pair. Typically, features such as road intersections have been proposed for urban areas. However, our experience indicates that no single man-made or natural feature can satisfy all of these criteria across a variety of complex urban scenes.

In Section II, we present some recent results on automatic control point detection and matching. In Section III, we briefly describe our two stereo matching techniques: S1 and S2. We also describe our initial results in combining disparity maps
derived from both stereo matchers, and we show some results on a variety of scenes. Three of the scenes contain complex man-made structures, whereas a fourth is a traditional open terrain scene. In Section IV, we introduce a quantitative method for evaluating stereo matching results with respect to a manually derived 3-D ground truth segmentation. Estimates of disparity accuracy and delineation accuracy are presented for our test scenes for both terrain and man-made buildings.

A. Previous Work

Stereo reconstruction is a critical task in automated cartography. It has received a great deal of attention in the computer vision community as well as in the traditional photogrammetric and remote sensing communities. A classical survey of computational stereo can be found in Barnard and Fischler [2].

PHOTOGRAMMETRY [3], [4]

More recent updates, with an increased focus on correspondence and motion, is available in Dhond and Aggarwal [3]. Classical photogrammetric approaches to computational stereo can be found in the Manual of Photogrammetry [4] and in a variety of publications of the Society for Photogrammetry and Remote Sensing.¹

Within the remote sensing community, there is a long history in terrain mapping using aerial mapping photography. More recently, there has been increased activity using satellite scanner data such as SPOT [5], [6] and airborne SAR data. Proposals for dedicated satellite systems for mapping have recently been put forth [7], based on the recognition that from a global perspective, gaps still exist in detailed topographic models for large areas of the earth. The accuracy and detail of such terrain maps is ultimately limited by the spatial resolution of these sensors. In the case of the SPOT system, stereo mapping under ideal acquisition and ground control conditions is limited to approximately a 10-m spatial resolution. Our focus, as well as that of much of the computer vision community, has been to work at a higher spatial resolution in open terrain and, more recently, in areas containing complex man-made structures. This imagery is generated by digitizing photography acquired using aerial mapping cameras. Depending on the scale of the photography and the digitization aperture selected, we can utilize ground accuracies between 1 ft to several meters. For example, the imagery used in this paper was digitized at an aperture of 100 μm from imagery at a scale of 1:12,000, yielding a ground sample size of approximately 1.3 m/pixel.

Despite considerable research effort using aerial imagery, there is no complete system that can reliably perform autonomous stereo mapping in a wide variety of scene domains. This is particularly true in complex urban areas containing a range of buildings with different shapes, heights, and roof textures as well as in complicated underlying terrain. However, some progress has been made. At SRI [8], systems like STEREOSYS [9] use a combination of approaches including correlation and feature-based matching that demonstrate high reliability and robustness against a variety of industrial and close-range photogrammetric test images. Another system (CYP-CLORS) [10], has been demonstrated on open-terrain modeling problems using a stochastic-optimization approach to stereo matching. At USC, several different techniques have been developed to solve the stereo correspondence problem, ranging from the use of area-based and feature-based processing [11] to working with segments [12] and complex primitives resulting from a perceptual grouping stage [13]. Post processing to improve the disparity map [14] and the use of active contours to obtain accurate boundaries of roof tops in aerial views of urban areas [15] have been also explored.

For the most part, this previous work has been performed on imagery containing an isolated building on flat terrain under imaging conditions where roof texture is nearly homogeneous and where the contrast between the building and its background supports good edge/line detection/delineation. In the examples shown in this paper, we have gone beyond these cases toward a more complex scenario with 50 to 60 individual buildings in each scene, with complex underlying terrain, and with buildings exhibiting a variety of building shape, roof shape, roof textures, and heights. We also provide results on a scene containing complex terrain without any man-made structures. We show comparable performance results across both the simple and more complex cases.

A second contribution of this paper is the introduction of a set of performance evaluation metrics based on the comparison of automated disparity maps with those generated by manual compilation using 3-D visualization. We examine issues in both global and local statistics, particularly with respect to accuracy of height estimation and localization of disparity for man-made structures in the scene. Finally, we argue that stereo analysis for cartographic feature extraction must include consideration of the full end-to-end compilation process. That includes automated scene registration, stereo matching, fusion and refinement, and performance evaluation. In this processing chain, we show that the integration of several techniques at all phases of analysis may be necessary to provide sufficient robustness.

II. SCENE REGISTRATION

The primary goal of stereo photogrammetry is to determine the 3-D position of any object point that is located in the overlap area of two images taken from two different camera positions. The determination of the orientation of each camera at the moment of exposure and the relationship between the cameras is a necessary step in the photogrammetric process. The camera orientation determines the relationship between the image points and ground points in the scene. The classical epipolar geometry for stereo imagery establishes a very simple spatial relationship between corresponding points in the left and right images. The solution to the general camera orientation problem has four components: the interior orientation, the exterior orientation, the relative orientation, and the absolute orientation.

The interior orientation refers to the perspective geometry of the camera. The parameters of the camera are generally known a priori and can be determined by precise calibration.
This includes the focal length, the position of the principal point in the image plane of the camera, and the geometric distortion characteristics of the lens system. These parameters are intrinsic to the camera and are generally detailed on a standard camera certificate.

The exterior orientation characterizes the orientation of the camera during the image event. It is defined by the geographic position of the optical center in a 3-D rectangular coordinate system and the direction of the optical axis. Therefore, the exterior orientation determines the projective relationship that exists between the image coordinates of the image points and the ground coordinates of the corresponding object points in the scene. In the context of stereo photogrammetry, the exterior orientation can be decomposed into the relative orientation and the absolute orientation.

The relative orientation determines the relative 3-D position of the two images in the stereo pair with respect to each other. After the relative orientation is accomplished, the stereo model must be scaled, translated, and leveled with respect to a ground reference system. The process of orienting a stereo model into an absolute reference system is called the absolute orientation. It relates the absolute coordinates of an object point in the ground reference system to its coordinates in the model coordinate system of the camera. Each control point, for which the ground coordinates are known, gives rise to three projective transformation equations when its model coordinates are measured. Three control points are necessary to define all the parameters of the absolute orientation.

In our work, we assume that the interior orientation or calibration has already been performed since we assume an ideal pinhole camera. Therefore, the calculation of the relative orientation is the registration problem [16]. Knowing the relative position and attitude of the two images in the stereoscopic pair with respect to each other defines the relationship between the two images of the scene. All of the results presented in this paper will be relative measurements. However, these relative measurements could be used to calculate absolute metrics, such as height, length, and area, by using 3-D ground control points to establish the absolute orientation and thereby convert disparity estimates into height estimates.

After we have established the relative orientation of the two images, it is possible to reformulate this relationship into the epipolar geometry. The epipolar geometry defines a constraint based on the geometric relationship that the plane containing the two optical centers of the cameras, and a ground point \( P \), intersect the two image planes on two lines. These lines are called the conjugate epipolar lines. Thus, the corresponding points in the left and right image correspond to a single 3-D scene point \( P \) and are on the same conjugate epipolar lines. After this relationship has been established, it is common to register the two images so that the conjugated epipolar lines become corresponding scanlines in the left and right image. Therefore, the corresponding points are on the same scanline in each image, and the displacement between the points, or the disparity, corresponds to the relative height of the 3-D scene point.

The epipolar geometry constraint causes conjugate points to lie on the same scanline in the left and right image. This means that stereo matching can be restricted to a 1-D search along the common scanline. Knowledge of the maximum disparity in the scene and ordering of matches can be used to further restrict the search along the epipolar line. This constraint is used as a common framework for most stereo matching algorithms [17]–[20]. These stereo matching techniques, however, assume that the registration is ideal and that the epipolar constraint is completely satisfied. However, this is rarely the case, and this problem becomes acute as we increase the spatial resolution of the imagery. The assumption of an accurate scene registration may also be unrealistic since for many applications in aerial image analysis, one is often simply given overlapping images or partial image areas where the epipolar geometry must be derived. In the case of stereo mapping using satellite imagery such as SPOT, a more complicated sensor model that accounts for the imaging scanner coupled with the orbital characteristics of the platform is required [21], [22].

In the following section, we present two methods for scene registration given overlapping stereo imagery. The first method performs a coarse registration using landmarks from a spatial database. The second method uses pairs of corresponding points in the two images to perform a relative orientation. As we will see, many of the techniques used in computer vision to establish scene registration are approximations to the photogrammetric ideal. These approximations cause the scene registration to be inaccurate and must be taken into account by the matching process.

A. Coarse Registration Using a Spatial Database

The most common method to establish the relative orientation between two images is to select pairs of corresponding points in the two images. One alternative method is to independently tie each image to a common frame of reference. A cartographic coordinate system such as <latitude,longitude,elevation> is one possible frame of reference. Thus, the two images are related to a ground coordinate system, or map. The use of landmarks with known <latitude,longitude,elevation> is a common method to orient each image. The overall accuracy of the registration is dependent on the accuracy of the 3-D position of the landmark and the accuracy with which we can recover the image position of the landmark. We use the landmark database component of CONCEPTMAP, which is a spatial database system that integrates imagery, terrain, and map data to provide landmark descriptions [23], [24]. Typically, CONCEPTMAP provides a registration accuracy of between 10 and 30 m for imagery digitized to a 1.3-m ground sample distance.

Figs. 2-1 and 2-2 show a stereo image pair of an industrial area taken from the CONCEPTMAP database. These images were digitized from standard 9-in format mapping photography taken at an altitude of 2000 m using a camera with a 153-mm lens. One pixel in the image corresponds to approximately 1.3 m on the ground. The left image is a 512 × 512 subarea selected from a 2300 × 2300 image. The right image subarea was generated by calculating the <latitude,longitude> for the corner points of the left image and projecting those points onto the complete right image. This projection is then used to extract the image subarea to form the complete right image.
We have superimposed a set of gridlines on both images in order to make it easier to see the actual misregistration. This misregistration includes a large translation and a smaller scale difference.

1) Fine Registration Using Image Control Points: As we have seen, the computation of the relative orientation can be accomplished by selecting pairs of corresponding points in the two images. After the relative orientation is calculated, the two images can be transformed so that they satisfy the epipolar constraint. We begin the fine registration with the coarse registration described in the previous section. We assume the transformation between the left and right image is isometric (i.e., only translation and rotation). After the transformation, the epipolar lines correspond to the scanlines. However, problems with the accuracy of point selection led us to develop a polynomial transformation adjusted by least squares to fit the selected corresponding points.

a) Automatic selection using different features: Clearly, one requirement for automated registration is the automatic selection of corresponding points in the stereo pair images. There are two problems that must be solved. First, we must automatically detect potential landmarks in each image, and then we must determine those landmarks that have been found in both images. General landmark matching is an unresolved problem, and most automatic registration techniques rely on the matching of "interesting" points that often have no physical significance or relationship with the landmarks. In the case of aerial mapping photography, most stereo pairs are acquired along flightlines taken within a few minutes of each other. This makes the variations in sun angle and illumination between two overlapping images very small.

There are some important criteria for automated control point selection. First, since the elevation of the control points is not known and we are using a simple geometrical model, it is important that the set of selected control points lie approximately in the same elevation plane. Second, the selection of control points should not rely on a single type of scene domain feature, such as road intersections, since not all control point features are abundant in all scenes. For example, in urban scenes, there are often many buildings and shadow regions available as candidate control points, and they are usually well distributed throughout the imagery. However, in airport scenes, elongated line pairs and uniform intensity regions appear to be a better choice. In any case, we use an iterative selection algorithm [26] that converges to a consistent set of control points that are usually a small subset of all of the possible matches in the stereo pair.

Another advantage of using multiple features for control point estimation is that the results of feature matching can be used to estimate the disparity range of the scene. Once the scene is registered, all matched features can be remapped to the new coordinate frame. It is then possible to calculate the disparity of each feature. Since all features are not at the same height, we automatically have a rough estimate of the disparity range for this scene. This disparity range estimate is directly used by the stereo matching algorithms to control the search for corresponding points and can greatly reduce initial matching errors.

For this experiment, we assume that a coarse registration of the two images, such as described in Section II-A, has already been performed. Using this coarse correspondence, we are able to limit the search to find corresponding features in the images. Most of the remaining error is translational rather than rotational, which simplifies the determination of corresponding points. Candidates for automatic control point generation include shadow corners, shadow regions, BABE [27] monocular building hypotheses, uniform intensity regions, and elongated line structure pairs.

Shadow Corners: Shadow corners are good candidates for automatic detection and correspondence as well as for manual
selection. We use corners produced by the BABE system. After removing corners that are inconsistent with shape and orientation constraints imposed by the sun direction angle and estimated shadow intensity, we select sets of shadow corners in both the left and right images. Fig. 2-3 shows the corners found in the left image in white. The right image corners are shown in black and are projected onto the left image using the coarse registration. Those pairs of shadow corners that are matched are shown to be connected by a white line whose endpoint circles indicate the conjugate points provided to the registration process.

**Building Hypotheses:** Control points can also be defined geometrically with respect to features or structures extracted from the imagery. Building hypotheses generated by a monocular analysis system (BABE) can be used as match features. The center of mass of these structures is defined as the corresponding control points. Compared with shadow corners, control points defined by hypothesized buildings are not always accurate, but it is easier to disambiguate one building from another. Properties such as shape, size, and perimeter are good criteria that are not available for point features such as shadow corners. Fig. 2-4 shows the BABE boxes in the left and right images with the matched features in the same manner as in Fig. 2-3.

**Other Scene Features:** We performed experiments to obtain control points from shadow regions, edges, and segmented regions using simple histogram analysis. In each case, control points are defined as the center of mass of the structures. Shadow regions are extracted using the traditional connected component extraction technique with an estimate of the shadow intensity provided by building detection [28]. Due to variation of shape of the shadows, shadow regions usually gives poor results in complex urban scenes with very high buildings. This variation of shape is caused by occlusion of shadow by tall structures. However, shadow regions can be very reliable in suburban house images where buildings are separated and have simple roof profiles. Under many conditions, shadow analysis can also provide a direct estimate of building heights but not with the same accuracy or reliability as direct stereo measurement [28].

Edges provide another possible matching feature for registration. Only edges with significant length are used as candidates for matching. The criteria for edge matching are edge orientation, length, and the intensity gradient across the edge. Fig. 2-5 shows the significant lines extracted and matched in the industrial scene. Unique bright points in the scene can be used to form bright blob regions. The intensity threshold for blob regions is determined by successively decreasing the intensity scale until enough regions are extracted. These features turned out to be useful for scenes with few or no man-made structures, where shadow corners, hypothesized buildings, and shadow regions failed to generate sufficient matching candidates.

One problem that might be encountered in high-resolution aerial imagery is that moving objects such as cars, buses, and trucks become large enough that they may be detected as control points. In our approach, most of the features selected, such as shadow corners and building hypotheses, require a fairly high-level analysis of the scene and are unlikely to produce potential match points that are actually small moving objects. Techniques, such as interest operators [25], that select points based on local intensity properties may be more sensitive to moving objects. The use of a range of potential match features that correspond to significant structural objects in the scene coupled with a registration process that looks for a consistent subset of points that lie in a plane seems to mediate this potential problem.

Fig. 2-6 shows the superposition of BABE results using the refined registration from Fig. 2-4. The offset between building hypotheses is now primarily in the column direction and can be attributed to the displacement of the building in the left and right image due to height. In many cases, we have been able to automatically reduce the row offset error to subpixel accuracy from an initial displacement of 15 to 20 rows in the coarse CONCEPTMAP registration.

**B. Evaluation of Automatic Registration:** Table II-1 shows the local accuracy of the different scene registrations performed on the industrial scene shown in Figs. 2-1 and 2-2. Coarse registration is the result of CONCEPTMAP image-to-database scene registration as described in Section II-A. ISO means that the image registration is accomplished by an isometric fit, whereas POLY means that actual registration is performed using a polynomial fit. As has been described previously, since we are dealing with aerial mapping photography, we assume that the two identical metric cameras have parallel axes and that the two images were taken at the same altitude. Therefore, the transformation between the left and the right images should preserve the distances between corresponding points. This allows for an isometric transformation involving translation and rotation without scaling. Since these assumptions are not always satisfied, due to to scaling and perspective distortion effects, we also evaluated a polynomial registration model that fits a second-order polynomial using least square minimization and analysis of the residual error for the automatically selected control points. The polynomial model can compensate for small scaling effects and perspective distortion.

Using a set of manually selected control points, we are able to evaluate the accuracy of each registration generated by the five feature extraction methods in terms of row offset compared with the ideal epipolar geometry (corresponding points on the same scanlines). The manual results are shown as ISO manual and POLY manual. The five feature extraction results (corners, structure, edge, shadow, and blob) are given for both the ISO and POLY registrations. Overall, the polynomial approximation performs better than the isometric solution. In all cases, the polynomial solution generates subpixel accuracy in terms of row and column offset with the exception of the solution obtained using BABE structures. The overall accuracy is close to that achieved using the manual registration.

The isometric solution, although it is inferior to the polynomial, greatly improves the scene registration when compared with the coarse database registration. Thus, it may be important to include the possibility of isometric registration in an end-to-end stereo analysis system. First, the isometric approximation needs a minimum of three control points since we only have three parameters to recover: two translation and one planar
rotation. Thus, in the case of sparse match features (if there are three good points available), the isometric approximation may recover a reasonable registration when there are insufficient matching points available for a polynomial fit. Second, the polynomial approximation is more sensitive to noise, whereas with a sufficient number of points, an isometric approximation can be formulated to detect largely inconsistent points. Finally, the solution of both the isometric and polynomial registrations can be used as a simple check on the general consistency of each solution.

For DC38008 and all of the urban scenes described in this paper, there are sufficient points from any single feature detection process to support both an isometric and polynomial registration. The overall accuracy is much better than the coarse database registration and is generally comparable with the registration achieved using a manual selection of control points.

Scene registration is a key initial step in many tasks involving the automated interpretation of aerial images. Stereo analysis requires particular care in scene registration because of the geometric assumptions made by most stereo matching algorithms and their inability to recognize and recover from registration errors. Such registration errors are usually reflected as gross errors in the stereo match. As part of our goal to produce 3-D interpretations of complex urban scenes, we have found it necessary to develop registration techniques that are accurate and robust across a variety of scene domains. We have tested our system on airport scenes, urban scenes, and suburban housing developments with varying degrees of
success. We are currently investigating ways to evaluate distribution of control points and to incorporate this evaluation into the registration system. We are also looking into improving our ability to recover a more accurate position for control detection and tracking with an estimate generated at every point in the image. Feature-Area-based techniques generally provide a dense disparity map where the features are generated; these are often points of intensity discontinuity that may correspond to discontinuities found in urban areas. For complex urban scenes, robust enough to perform well under the diverse set of image conditions found in urban areas. For complex urban scenes, feature-based techniques appear to provide more accurate information in terms of locating depth discontinuities and in estimating height. However, area-based approaches tend to be more robust in scenes containing a mix of isolated buildings and open terrain. They degrade gracefully, can be improved by using heuristics [11], [9], and have seen renewed interest for both aerial and satellite-based mapping.

For this reason, we have developed two stereo matching systems based on each of the general types of matching algorithms. S1 uses an area-based algorithm based on the method of differences matching technique developed by Lucas [30], [31]. S2 is feature based using a scanline matching method that treats each epipolar scanline as an intensity waveform. The technique matches peaks and troughs in the left and right waveform [32]. Both are hierarchical and use a coarse-to-fine matching approach. Each is quite general because the only matching constraint imposed by either technique is the order constraint for the feature-based approach. The order constraint should generally be satisfied in our aerial imagery except in the case of hollowed structures.

### III. Two Stereo Matching Techniques

Algorithms for stereo correspondence can be grouped into two major categories: area-based and feature-based matching. Area-based techniques generally provide a dense disparity map with an estimate generated at every point in the image. Feature-based approaches provide depth information only at points where the features are generated; these are often points of intensity discontinuity that may correspond to discontinuities in depth.

We do not believe that any one technique is likely to be robust enough to perform well under the diverse set of image conditions found in urban areas. For complex urban scenes, feature-based techniques appear to provide more accurate information in terms of locating depth discontinuities and in estimating height. However, area-based approaches tend to be more robust in scenes containing a mix of isolated buildings and open terrain. They degrade gracefully, can be improved by using heuristics [11], [9], and have seen renewed interest for both aerial and satellite-based mapping.

For this reason, we have developed two stereo matching systems based on each of the general types of matching algorithms. S1 uses an area-based algorithm based on the method of differences matching technique developed by Lucas [30], [31]. S2 is feature based using a scanline matching method that treats each epipolar scanline as an intensity waveform. The technique matches peaks and troughs in the left and right waveform [32]. Both are hierarchical and use a coarse-to-fine matching approach. Each is quite general because the only matching constraint imposed by either technique is the order constraint for the feature-based approach. The order constraint should generally be satisfied in our aerial imagery except in the case of hollowed structures.

### Table II-1

<table>
<thead>
<tr>
<th>Type of Registration</th>
<th>Number of Points</th>
<th>Avg. row offset</th>
<th>Std. row offset</th>
<th>Min/Max row offset</th>
<th>Avg. col offset</th>
<th>Std. col offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coarse</td>
<td>-20</td>
<td>1.6</td>
<td>-2.16</td>
<td>0.4</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>ISO manual</td>
<td>11</td>
<td>-0.4</td>
<td>0.6</td>
<td>-1.1</td>
<td>0.6</td>
<td>1.4</td>
</tr>
<tr>
<td>ISO corner</td>
<td>20</td>
<td>1.0</td>
<td>0.5</td>
<td>0.3</td>
<td>2.7</td>
<td>1.3</td>
</tr>
<tr>
<td>ISO structure</td>
<td>14</td>
<td>-1.7</td>
<td>0.9</td>
<td>-3.0</td>
<td>2.9</td>
<td>1.2</td>
</tr>
<tr>
<td>ISO edge</td>
<td>17</td>
<td>1.3</td>
<td>0.9</td>
<td>0.4</td>
<td>0.9</td>
<td>1.2</td>
</tr>
<tr>
<td>ISO shadow</td>
<td>12</td>
<td>-0.2</td>
<td>1.1</td>
<td>-2.2</td>
<td>3.8</td>
<td>1.7</td>
</tr>
<tr>
<td>ISO blob</td>
<td>17</td>
<td>0.6</td>
<td>1.6</td>
<td>-2.5</td>
<td>1.4</td>
<td>1.2</td>
</tr>
<tr>
<td>POLY manual</td>
<td>11</td>
<td>0.1</td>
<td>0.3</td>
<td>-1.1</td>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>POLY corner</td>
<td>20</td>
<td>0.5</td>
<td>0.6</td>
<td>0.2</td>
<td>-0.5</td>
<td>1.2</td>
</tr>
<tr>
<td>POLY structure</td>
<td>14</td>
<td>-0.8</td>
<td>0.8</td>
<td>-2.2</td>
<td>-5.2</td>
<td>2.0</td>
</tr>
<tr>
<td>POLY edge</td>
<td>17</td>
<td>0.8</td>
<td>0.7</td>
<td>0.3</td>
<td>0.0</td>
<td>1.8</td>
</tr>
<tr>
<td>POLY shadow</td>
<td>12</td>
<td>-0.6</td>
<td>0.8</td>
<td>-2.1</td>
<td>-0.4</td>
<td>0.9</td>
</tr>
<tr>
<td>POLY blob</td>
<td>17</td>
<td>0.6</td>
<td>0.6</td>
<td>0.2</td>
<td>-0.6</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Both matching algorithms assume the imagery has been registered into the epipolar geometry, as discussed in the previous section, and each algorithm produces a disparity map that is registered to the left stereo pair image. Both matching algorithms also need an estimate of disparity range found in the image. This estimate need not be precise, but the accuracy of the disparity range estimate directly affects the quality of stereo matching. An accurate disparity estimate limits the search range and therefore reduces possible mismatches. The disparity range can be provided by a user or from the result of the registration process. To generate disparity ranges automatically, we first need to select a type of registration to use and register all features described in Section II-B-1-a using the same type of registration. Once the control points from both left and right images are in the same coordinate frames, the column offset between a pair of matched control points is the disparity. We can find the scene disparity range by finding the disparity range for all the control points [26]. Table III-1 shows the disparity range calculated from the detailed stereo scene reference model, the disparity range selected automatically using the scene registration column offsets, and that calculated by sampling isolated control points. The disparity range used for all stereo matching examples shown in this paper are those selected by manual sampling. However, researchers need to address the problem of accurate automatic disparity range estimation since one must currently scan the entire scene making the process both tedious and prone to error. This is especially true for cartographic applications where the disparity range may fluctuate widely across a scene due to combinations of terrain and man-made structures.

Space concerns prohibit us from giving a full description of the S1 and S2 stereo matching systems [32]. In the following sections, we give a technical sketch of each system. Since the primary focus of this paper is the development of performance evaluation techniques, in some sense, results from any state-of-the-art, area-based, or feature-based method could be applied. However, for the work in merging of stereo estimates, it is important to have methods that may exhibit different or complimentary strengths and weaknesses in complex urban scenes.

#### A. S1: Method of Differences

The stereo matching algorithm in the S1 system was developed by Lucas [30], [31] and uses a hierarchical set of reduced frequency resolution images to perform coarse-to-fine matching on small windows in the two images. At each level, the size of the windows for the matching process depends on the spatial frequency resolution (i.e., smoothness)
of the image. An initial disparity map is generated at the first level, corresponding to a coarse level of image detail, by matching image intensity. Subsequent matching results, which are computed at successively finer levels of detail, are used to refine the disparity estimate at each level. Therefore, the amount of error in the scene registration that can be tolerated by this matching algorithm depends on the size of the matching windows. However, since there is a relationship between the matching window size and the level of accuracy, simply using larger matching windows may not be desirable. Consider a point in the left image of the stereo pair; the difference between the correct disparity value and our initial estimate is the amount by which the stereo process must correct the disparity. Initially, this difference will be relatively large because the initial disparity estimate is not particularly accurate. Because of this, the method of differences requires that we start out with smoothed images to accommodate these large differences. As the disparity estimate improves, we can use less smoothed images because the magnitude of the matching error decreases.

SI has several desirable properties that are shared by other area-based stereo matchers. It does not require a specific feature extraction, and matching is accomplished at every pixel, matching can be done on scenes devoid of features, and the matching does not require perfectly registered stereo pairs. SI also shares the drawbacks of such systems, however. The matcher has difficulties finding the correct pixel correspondence in areas containing nondescript surfaces. Disparity estimates at sharp boundaries are smoothed, and it requires the specification of a disparity range to constrain epipolar search.

B. S2: A Feature-Based Approach

S2 is a feature-based system that treats the problem of stereo matching as a 1-D signal matching, which is similar to Witkin's scale space signal matching [33]. S2 extracts intensity and gradient signals from each epipolar scanline and matches peaks and valleys in the left and right signals. S2 is hierarchical and uses a coarse-to-fine matching approach to guide successive refinements of the waveform approximation. This work represents a similar approach to previous work in stereo matching. For example, Baker [34] used a coarse-to-fine approach with the Viterbi dynamic programming algorithm to match edges. Ohta [20] tried to find an optimal matching surface in a 3-D search space using interscanline search and dynamic programming. However, edge matching is difficult, especially in complex scenes, and often requires thresholds based on edge strength to avoid combinatorial explosion. Edge matching methods also rely heavily on the quality of the edge detector, and the resulting disparity maps are usually very sparse. The S2 approach is to view intensity profiles of scanlines as waveforms. There has also been some previous work using this paradigm in the pattern analysis area to perform waveform correlation by representing the waveforms as trees [35] or by tracking zero crossings of intensity signals through scale space [33].

S2 matches epipolar scanlines in the left and right image using a hierarchical approximation of the scanline intensity waveforms to match peaks and valleys at different levels of resolution. It uses interscanline consistency to enforce a linear ordering of matches without order reversals. It also applies an interscanline consistency check that considers the matches in adjacent scanlines. Application of the interscanline constraint is used to increase the confidence of matches found to be consistent across multiple scanlines and to delete improbable matches. Since disparity discontinuity usually occurs at the intensity discontinuity, the gradient waveform is matched after the intensity matching phase to localize disparity jumps. Efforts are made to detect occlusions and correct them, and the sparse disparity map is converted to a complete disparity map by step interpolation.

C. Stereo Test Results

In this section, we show stereo matching results for two complex urban scenes (DC38008 and DC37405), for a classic stereo test image (PENTAGON), and for a scene containing complex natural terrain (DENVER). S1 and S2 have been tested on approximately 15 stereo scenes including airports containing hangars, runways, and tarmacs, suburban house areas with complex terrain and buildings, and industrial areas with large and complicated buildings. The results presented here are representative of problems and successes achieved in these tests. Most of the previous work using high-resolution aerial imagery has shown results on isolated buildings in flat terrain [12], [14], [36], [37]. Our test data clearly extends previous work to conditions of multiple structures with various heights embedded in complex terrain. In the case of DC37405, many of the building heights are lower than the surrounding terrain. Further, many of the buildings do not have strong contrast boundaries with their background. This may have an adverse impact on techniques based on perceptual grouping because our experience with BAMB indicates that many key structural features such as lines and corners may be missed by the low-level feature extraction component. Finally, there are no shape constraints embedded in the stereo matchers that would restrict their applicability to a limited class of building shapes. One can see, however, that most buildings are rectangular or are composed of collections of rectangular structures.

We present the reference left images, reference disparity maps that have been manually compiled using an interactive 3-D editing system, and the S1 and S2 results. In all of the disparity map results presented in this paper, brighter regions are closer to the observer and have greater height. Darker regions are at or below the relative terrain ground plane established by the scene registration process. Thus, the disparity map encodes relative height. Given several points with known absolute elevation in the scene, we could calculate the absolute height at each point in the disparity map.

Fig. 3-1 is a complex industrial area scene. This scene contains many of the difficulties found in stereo matching, including occlusion, complicated textures, large depth discontinuities, and complicated 3-D objects. Fig. 3-2 shows a manually compiled disparity reference map. Figs. 3-3 and 3-4 show the result of the S1 and S2 matching, respectively. S1 performs well on textured, smoothed, and continuous regions. However, the depth discontinuities are not well captured, and
therefore, the delineation of the high buildings is not crisp. Most of the terrain relief is present, particularly the slope of the land toward the water in the bottom portion of the scene. One advantage of S1 matching is that it is not overly reliant on the initial scene registration and can accommodate small errors without producing artifacts in the disparity map.

Fig. 3-5 is a complex urban scene with hilly terrain and a large number of small buildings with low height. In fact, there are many buildings in the scene whose roofs are lower than nearby terrain features. There is also a great variety of building shapes including large apartment structures, town homes, and low-density commercial buildings. Thus, this is a very good test site for evaluating the relative strengths and weaknesses of area-based and feature-based matchers. Fig. 3-6 is the manually generated reference disparity map, whereas Figs. 3-7 and 3-8 are the results of S1 and S2, respectively. Both S1 and S2 do a good job of recovering the complicated terrain, including the diagonal road and the terrain valley in which it runs. Both methods captured the general ground relief, with S1 providing a better terrain estimate and S2 capturing the building shapes more crisply.

Fig. 3-9 shows a view of the Pentagon in Washington, DC. This image has been used by a variety of researchers in stereo vision and has become an informal standard. It is also a good example of the difficulties one may encounter in the manual generation of an accurate reference disparity map. For instance, there are many small structural details on the roof that have not been completely picked up in our reference...
disparity map. These structures do not account for a large percentage of the pixels in the scene, but they do provide key perceptual cues to human observers. More importantly, due to the occlusions in the scene, it is not a simple matter to select the correct disparity height for the central interior of the Pentagon as well as for the other holes in the structure due to the interior corridors. Even if perceptually we know that the center of the building is at the "ground level," it is difficult to select points such as the bottom of the interior wall to determine the ground height when the bottom of the wall can not be seen in the other image.

Both stereo methods do a good job at recovering the Pentagon building structure, although the S2 disparity map subjectively appears sharper. The step interpolation performed by S2 does not completely recover the smooth terrain variations but fits very well with the complex roof structure. The road and overpass on the right-hand side of the scene is well modeled by S2. The left/right image pairs for the Pentagon scene were already in epipolar geometry; therefore, no manual or automatic registration was performed. However, one artifact of this registration is that there are no points in the scene having zero disparity. This is probably due to an artificial shifting of the images in this stereo pair and is not related to the matching of actual physical features in this scene.

Fig. 3-13 is a scene containing open terrain without any significant man-made structures. This scene has a large disparity range, and the imagery was already provided in epipolar geometry. Fig. 3-14 shows a manually created disparity map.
for this scene. Figs. 3-15 and 3-16 show the result for s1 and s2. Both methods recovered the general relief of the scene. s2 had difficulties at the right-hand edge of the image because that part of the scene is not visible in the right image of the stereo pair. However, as we will discuss in Section IV, it is a little surprising that s2 performs better than s1. The s2 result has an average error of 9% as compared with s1's average error of 22%. This is most likely due to the very large disparity range in this scene where the area-based technique used by s1 has difficulties recovering from initial mismatches in later stages of processing.

Such qualitative statements of performance, while not inaccurate, do not actually allow us to understand the impact of small algorithmic changes to either matching technique, the effect of various registration methods on the overall scene interpretation, or the effect of various analysis methods such as merging in a way that is quantitative. The lack of accurate ground-truth information makes it quite difficult to evaluate stereo matching algorithms performed by various researchers, even on identical imagery. We address a more quantitative approach to performance evaluation in Section IV. In the following section, we describe a technique for merging the
D. Merging S1 and S2 Results

The disparity map produced by area-based methods have characteristics of smooth relief and give good disparity estimates. On the other hand, feature-based methods produce disparity that is noisy and sparse but with good delineation of the disparity discontinuity. There are other systems that tried to take advantage of this, such as the stereo vision system [37] from the University of Southern California (USC). Although they used features such as intensity edges to refine their area-based disparity map, we have taken the approach of disparity hypothesis evaluation from both methods at every point. The idea behind this is that at the disparity edge, the feature-based method should be more accurate than the area-based method; at smoothly changing terrain, the area-based disparity value should generally be considered to be more reliable.

Presently, a very simple technique is used. The system looks at the disparity hypothesis generated from S1 and S2, decides which one is better, and uses the best disparity value.

disparity maps generated by S1 and S2 in order to produce an improved scene disparity map.
To avoid bias, the goodness of the disparity is measured using a technique similar to the area-based and waveform-based methods. An image patch is extracted from both the left and right stereo pairs. The position of the right image patch extracted depends on the disparity value. The goodness of the disparity is simply the result of the difference in the two intensity patches. A similar measure is also used for the waveform estimate.

When the disparity estimates from both methods differ by about 20% of the disparity range, we do not try to generate any hypotheses. Such points are set to black in the disparity map images. These areas usually correspond to occluded regions. These regions should not be matched to anything since we do not have the corresponding section in the other part of the stereo pair. In S1, these regions were simply filled in by matching to areas with similar characteristics outside the occluded region. In S2, since there were no matchable features, the disparity values were interpolated.

Fig. 3-17 shows the result of merging the S1 and S2 results for the DC38008 industrial scene. The black regions within the disparity map show the regions of possible occlusion. Qualitatively, it appears that this result has a better delineation of the disparity jumps when compared with the S1 result in Fig. 3-3 and is less noisy when compared with the S2 result in Fig.
3-4. Similar results are obtained with the suburban house scene as shown in Fig. 3-18 and the Pentagon scene in Fig. 3-19. Fig. 3-20 shows some striking results in areas where S1 and S2 disagree on a scene with no man-made occlusions.

These preliminary results of merging two stereo matching algorithms indicate that they generate disparity estimates that are quite complementary. We believe that it is possible to take advantage of the different failure modalities in order to form a composite disparity map that gives a more accurate 3-D representation of the scene. It may be possible to use areas of disagreement to guide further processing aimed at refining the stereo disparity estimate using other matching methods. The detection of occlusion boundaries/regions can also be used as a guide for building delineation. In the following section, we describe some quantitative evaluation methods based on the manual generation of a detailed disparity map and a detailed scene segmentation.

IV. PERFORMANCE EVALUATION

It is difficult to evaluate the results of any stereo matching algorithm working on real, rather than synthetic, stereo image data. Qualitative methods for the evaluation of stereo results include the use of 3-D perspective wire-frame models or image texture maps to visualize the disparity map derived from stereo matching. Quantitative methods have been primarily applied to synthetic scenes where the disparity levels are directly known. Although synthetic scenes, such as those produced by random dot stereograms, can provide controlled 3-D scene structure, we do not believe that they provide sufficient complexity to predict behavior of stereo matching algorithms used in real-world imagery containing natural terrain and man-made structures. For cartographic applications, synthetic images can be generated using a digital elevation model (DEM). However, in creating such imagery, it is difficult to model many of the imaging conditions such as geometric and radiometric differences as well as natural textures that cause the majority of problems for stereo matching.

We argue that a true evaluation of stereo analysis for cartographic applications requires the use of a reference ground-truth disparity map for comparison. Figs. 4-1 and 4-3 show a wireframe and image texture mapped representation of a 3-D ground reference disparity map. Both types of graphics have been used extensively in the literature to illustrate qualitative stereo results. The graphics are not particularly useful for quantitative performance evaluation and, more importantly, can easily give a false impression of the quality of the stereo result. In the case of the wire frame display, the 3-D points in the reference disparity map are mapped onto a regular grid and can be plotted from an arbitrary viewpoint. We have chosen a low oblique view looking from beyond the bottom of the vertical scene toward its center. Fig. 4-3 is generated from
the same viewpoint with the original vertical image used as a texture map under perspective projection.

One can observe that the texture mapping provides powerful perceptual cues to understand the structure of the scene that are not available in the wireframe model. This appears to be true across a wide range of viewing angles, especially when the texture map is derived from a near vertical view of the scene. Figs. 4-2 and 4-4 show the S2 results for this scene as both wireframe and texture mapping graphics. One can visually compare these results and notice that the texture mapped visualization appears to provide a better sense of the scene although it represents exactly the same underlying disparity map. We have observed that it is quite difficult to understand the differences between the disparity maps produced by S1, S2, or MERGE simply by looking at perspective views or at the intensity scaled disparity maps. This observation has led us to a first attempt at performance evaluation that would support quantitative comparison of different stereo matching systems, or the same system using different parameters, on difficult aerial imagery. Specifically, our goal was to provide metrics for the global accuracy of a stereo matching result over the entire scene and for man-made structures such as buildings, bridges, and elevated roads that are not well modeled by the terrain.

A. Construction of a Reference Disparity Map

It is actually very difficult to get a good reference disparity map for an arbitrary test scene. One could imagine resorting to the use of existing digital elevation models or paper maps with terrain contours. Unfortunately, unless one is fortunate enough to find an area with high-resolution ground truth, the accuracy of standard digital products or maps is insufficient, especially with imagery having a ground sample distance of around 1 m/pixel. We have developed a display tool to manually generate disparity maps, allowing a user to select points on the registered images and generate accurate disparity values. The user views the scene using a Tektronics 920 stereo display monitor with the imagery registered using a manual ground point selection. Once a sufficient number of points has been selected (usually around 200 but dependent on the complexity of the underlying terrain), we can generate a dense reference disparity map of the terrain by interpolation. Similarly, we add disparity regions that correspond to man-made structures to the terrain disparity map. In some sense, these manual disparity maps are detailed cartographic descriptions of the scene and can be much more accurate than most traditional paper-based maps. Figs. 3-2, 3-6, 3-10, and 3-14 show the manually produced disparity maps for the industrial, suburban house, Pentagon, and Denver terrain scenes, respectively.

At least three different performance measures can be calculated to evaluate a stereo disparity result. We can evaluate the general performance on a scene, the performance for all the buildings, or the performance on a building-by-building basis. The global average disparity error is computed by finding the error for each point between an estimated disparity value and the reference disparity map. This single statistic provides a quick quantitative measure of the quality of the disparity map. One can further categorize points in the reference disparity map as high gradient points, low gradient points, points with high disparity, or points with low disparity. Based on this classification, it could be interesting to evaluate the performance of various stereo matching algorithms for specific problems such as smoothing over depth discontinuities or sensitivity to disparity range.

<table>
<thead>
<tr>
<th>TABLE IV-1</th>
<th>STATISTICS FOR DIFFERENT STEREO MATCHING METHODS ON DC38008</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stereo Method</td>
<td>Min/Max Disparity</td>
</tr>
<tr>
<td>S1</td>
<td>-12/15</td>
</tr>
<tr>
<td>S2</td>
<td>-15/15</td>
</tr>
<tr>
<td>S1+S2</td>
<td>-15/15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE IV-2</th>
<th>STATISTICS FOR DIFFERENT STEREO MATCHING METHODS ON DC37405</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stereo Method</td>
<td>Min/Max Disparity</td>
</tr>
<tr>
<td>S1</td>
<td>-12/15</td>
</tr>
<tr>
<td>S2</td>
<td>-15/15</td>
</tr>
<tr>
<td>S1+S2</td>
<td>-15/15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE IV-3</th>
<th>STATISTICS FOR DIFFERENT STEREO MATCHING METHODS ON PENTAGON SCENE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stereo Method</td>
<td>Min/Max Disparity</td>
</tr>
<tr>
<td>S1</td>
<td>-16/9</td>
</tr>
<tr>
<td>S2</td>
<td>-13/5</td>
</tr>
<tr>
<td>S1+S2</td>
<td>-15/13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE IV-4</th>
<th>STATISTICS FOR DIFFERENT STEREO MATCHING METRIC ON DENVER SCENE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stereo Method</td>
<td>Min/Max Disparity</td>
</tr>
<tr>
<td>S1</td>
<td>-22/9</td>
</tr>
<tr>
<td>S2</td>
<td>-20/3</td>
</tr>
<tr>
<td>S1+S2</td>
<td>-20/3</td>
</tr>
</tbody>
</table>
Fig. 4-5. Average error in pixel disparity at each disparity level in DC38008.

Fig. 4-6. Percent of points within ±1 pixel of ideal disparity in DC38008.

Fig. 4-7. Average error in pixel disparity at each disparity level in DC37405.

Fig. 4-8. Percent of points within ±1 pixel of ideal disparity in DC37405.

Tables IV-1, IV-2, IV-3, and IV-4 give the global error estimates for each of the four test scenes. These global statistics show that S1 (the area-based method), S2 (the feature-based method), and S1@S2 (the combination of S1 and S2) give very similar results across each of the four scenes. Interestingly, these measures do not seem to statistically reveal the apparent perceptual improvement achieved by merging the results of S1 and S2. We believe that this argues for a more structural analysis in addition to global scene measures.

The global errors show an interesting consistency in the percentage error, ranging from 4 to 7%. These correspond to an error in disparity of about 1 pixel. The percentage of points within our ±1 pixel disparity band around the ground-truth estimate is consistently between 60 and 70%. However, results on the Pentagon scene are somewhat better, ranging from 75 to 83%. This can be attributed to the fact that there is little terrain relief and that a single structure composed of large horizontal surfaces dominates the scene. This might argue for increased research and evaluation emphasis on scenes with structure similar to DC38008 and DC37405. A 3-D view of the disparity map compared with the actual 3-D scene using a stereo display gives the distinct impression that the S2 results are consistently better than the S1 measurements. This seems to be borne out in the global statistics.

1) A Refinement of Global Disparity Accuracy: One way to
address some of the issues that are hidden by the global statistics discussed in the previous section is to measure the influence of the actual disparity value on matching accuracy for each of the methods, that is, as we attempt to recover larger and larger disparity values, does our matching error increase? Alternatively, is the matching error independent of the disparity range? The graphics in Figs. 4-5-4-12 plot error rates sorted by reference disparity. Figs. 4-5, 4-7, 4-9, and 4-11 show the average error in pixel disparity at each disparity level for each of the test scenes. Each contains three graphs showing the results for $s_1$, $s_2$, and the merged result of $s_1$ and $s_2$. Figs. 4-6, 4-8, 4-10, and 4-12 show the percentage of points within $\pm 1$ pixel of the ideal pixel disparity over each disparity range.

In general, these graphs indicate that the greater the actual disparity, the more likely the various matching algorithms will make a mistake. This is true for both larger positive and negative disparity. These errors are reflected in both a higher average error and a lower percentage of points within $\pm 1$ pixel of the actual disparity. These global metrics also show that in areas of low disparity, $s_1$, $s_2$, and their merger give similar results. For higher disparities, $s_1$ has much more of a problem in correctly estimating the disparity than does $s_2$. Further, in most cases, the result of $s_1$ and $s_2$ merging produces an improved estimate, causing errors to decrease. This is probably
due to two factors. The first is the improvement made by the merging algorithm by selecting the better estimate. Second, the merge estimate will reject points found to be unreliable and remove them from the statistical population by not giving them a height estimate. Since these points most often correspond to areas of occlusion, this improves the global statistics.

C. Performance Evaluation with Man-Made Structures

In areas with man-made structures, global accuracy statistics such as those described in the previous section do not adequately convey the quality of the stereo matching system with respect to the buildings in the scene. In most cases, buildings may cover only a small portion of the scene, and the background terrain will statistically dominate the scene-wide estimate of disparity quality. Thus, we require a method that allows buildings to be evaluated independently or as a class of objects in the scene. Additionally, there are several metrics that can be used to evaluate both the disparity estimate and the quality of the depth jumps. We discuss these metrics in the following sections. Figs. 4-13 and 4-14 are hand segmentations of the left image where we have associated a reference building ID’s. Figs. 4-15 and 4-16 are graphs showing the actual building heights referenced to the building ID’s. We have also computed, for each building in the ground truth, the height of the building over its surrounding terrain. We have assigned
building ID’s based on the ground truth disparity map so that taller buildings have larger numeric ID’s. Using the numeric ID, the reader can index into each of the performance graphs to determine the results for a particular building of interest.

D. Quality of Building Disparity Estimate

In order to evaluate the performance of S1, S2, and the merged result on buildings in the scene, we can gather statistics on the disparity estimate for each pixel considered to be on the roof of the building. As before, the average disparity error in pixel disparity and the percentage of points within ±1 pixel of the ground-truth estimation are good measures for performance. Fig. 4-17 shows the quality of the disparity estimate for each of the buildings in the DC38008 industrial scene. The x axis represents the ID number for each building, and the y axis shows the errors in estimated disparity for a particular building across S1, S2, and the merged result.

This graphic, although a bit cluttered, shows no clear trend of performance advantage; both S1 and S2 produce a comparable result, although S2 appears to perform better, especially on buildings with greater disparity. For most buildings, the error is bounded between ±2 pixels. The result of merging generally appears to improve the average error. As we have assigned building ID’s sorted by disparity, we can observe a trend toward increased error as we move along the x axis.
We can also represent results using the disparity jump instead of the building ID to index the results. These graphics represent the integration of the average disparity error over all buildings with the same disparity jump. Figs. 4-18 and 4-21 show the effect of disparity jump on the disparity estimate and allow us to determine whether the actual height of a building over its neighborhood (disparity jump) affects the disparity estimate produced by stereo matching. It appears that S1 is comparable with S2 for smaller buildings. This is because low buildings can satisfy the continuity constraint of the area-based method. S2 performs better on scenes with buildings having significant height because low buildings can be easily masked by random mismatches in the feature-based analysis. The merge of S1 and S2 produces results that combine the best properties of both methods.

Figs. 4-19, 4-20, and 4-22 provide similar statistics for the suburban house scene DC37405. As in DC38008, the average error for each building appears to be bounded by ±2 pixels, S2 appears to have slightly better performance than S1, and the result of the merger almost always improves the average error. Whereas S2 always appears to perform much better than S1 with respect to the percentage points (to within ±1 pixel of the correct disparity in DC38008 (see Fig. 4-21)), this is not the case for DC37405, as shown in Fig. 4-22.

These statistics allow us to pinpoint problems at a much finer grain of detail than can be accomplished with global analysis. Thus, we can identify specific buildings in the scene and try to understand, at the algorithmic level, whether there are specific situations where matching could be improved. Once identified, these improvements should have an overall positive effect on the rest of the scene. The result, of course, can be subjected to the same rigorous performance analysis. Once we commit to working on complex scenes, as opposed to synthetic controlled images, the visual inspection of disparity results to discover small variations in performance becomes very unsatisfactory, except possibly at the earliest stages of experimentation. Such manual inspection greatly limits our ability to detect subtle conceptual bugs or recognize possibilities for algorithmic improvement. We can perform systematic analysis across multiple scenes. For example, in applying statistics that take into account the disparity jump for individual buildings, we can aggregate performance information for all buildings across all scenes to achieve a larger statistical sample.

**E. Quality of the Delineation Estimate**

In the previous section, we described techniques to measure the accuracy with which we can recover the height of buildings in the scene. For cartographic applications, it is equally important that we generate an accurate delineation of the buildings with respect to their surroundings. In this section, we discuss another metric, which is the quality of the stereo delineation of each building in the scene. We compute edge location,
which measures the distance of the estimated disparity jump from that in the ground-truth disparity. We also measure edge sharpness, which corresponds to the shape of the disparity jump in the estimated disparity map. Ideally, we would expect the stereo matcher to generate a step disparity jump at the point where the actual disparity jump occurs in the reference disparity map. As before, we assume that the ground-truth disparity map accurately captures the location and the height of the building edges. In order to allow for measurement error, we tolerate some uncertainty in both the location of the edge (± 1 pixel) and the height estimate on both sides of the edge (edge sharpness). The uncertainty in edge sharpness is somewhat difficult to quantify since it depends on both the height estimate on each side of the building roof edge and on the height estimate of the neighboring ground. These estimates may be biased since, in some cases, we are interpolating the ground elevation from a sparse network of points. We can alleviate this error by making sure that we select representative ground points as close to the buildings as possible.

Fig. 4-23 shows how we compute the edge location and sharpness for each building in the scene. The two waveforms represent the gradient of the reference disparity map and the disparity result being evaluated. The peaks in the reference disparity map gradient represent the true edge of the building in the scene. The evaluation process finds the best matching peaks in the s1, s2, or merged disparity map.
gradient within a neighborhood of the reference edge. The distance $P$ corresponds to the position error of the edge in the resulting disparity map. The ratio $H_d/H_r$ corresponds to the sharpness evaluation of the edge. A ratio of one is perfect. The value $H_d$ and $H_r$ correspond to the amplitude of the gradient related to the reference zero gradient. Both the position error and the edge sharpness metric require that an edge point in the reference disparity map be matched with an edge point produced by the stereo matcher under evaluation. In many cases, no such match is possible, that is, there is no suitable match for the reference disparity edge. In the following examples, between 35% (DC37405) and 50% (DC38008) of the reference points are not matched; hence, the matchable edges represent between 50–65% of the reference points in the scene. Figs. 4-24 and 4-26 represent the average position error for the matchable edges across all buildings in DC38008 and DC37405, respectively. Figs. 4-25 and 4-27 show the percentage of edges produced by the stereo matchers that are within ±1 pixel of a reference disparity map edge. These graphs are the subset of points lying in the band ± one position error from Figs. 4-24 and 4-26, respectively, plotted with respect to all edges in the reference disparity map. In both cases, the position error metric shows that the ability to accurately delineate the disparity depth jump appears to be much weaker than visual examination of the disparity maps might indicate.
For the evaluation of disparity sharpness, we calculate the average edge ratio and the sharpness of edge points whose edge position is within ±1 pixel of the reference edge. Fig. 4-28 represents the average edge sharpness ratio for all matchable edges across all buildings in DC38008. A ratio of one indicates a perfect step edge. Fig. 4-29 shows the sharpness of edge points that are within ±1 pixel of the reference position for all buildings in DC38008. Figs. 4-30 and 4-31 show the same results applied to the buildings in DC37405.

We can make several observations based on this performance data. First, it is clear from this analysis that S1 does not perform as well as S2 in terms of disparity delineation. Its ability to estimate the sharpness of the disparity jump (edge ratio) is likewise poorer than that of S2. However, there are some comparative advantages. S1 gives comparable results in the case of buildings with low disparity. On the DC37405 scene, the S1 and S2 results are similar because the buildings in this scene do not have large disparity jumps.

It is interesting to note that errors in delineation, position, and sharpness increase as the height of the buildings increases. This is an artifact of occlusion, where higher buildings will occlude a larger area, making it more difficult to detect the exact position of the disparity jump. Edge errors seem to be comparable for both S1 and S2 for buildings with low disparity. As expected, S1 does not delineate tall buildings well, and the merged result combining S1 and S2 sometimes produces a result that is an improvement over each individual method, but more often, it simply decreases the maximal error.

F. Limitations of Performance Evaluation

The common theme in this section on performance evaluation is to describe a variety of quantitative measures that allow us to objectively judge how well a particular set of registration/matching techniques perform with respect to a manually compiled 3-D ground-truth model and, by comparison, how well they perform with respect to one another. The reference disparity map is generated using monocular and stereoscopic visualization and is a representation of the scene within a certain accuracy. In most cases, the ground-truth segmentation can be constructed with enough care to provide for accurate detection of gross errors and as a common basis for general comparison between matching methods. However, the actual accuracy of the reference disparity map has to be considered if we attempt to use it for the analysis of scene microstructure, such as roofs with shallow pitch that are modeled as flat surfaces, small super structures such as building air conditioner units, stair well towers, and other small roof structures. These superstructures can add an error bias into the overall statistics. This bias is likely to be small; consider the fraction of error introduced in the case of a nine-story building where we have not correctly modeled an air conditioner unit that rises another story over 15% of the total roof surface.

Nevertheless, we are sampling only a small subset of the actual 3-D points in the scene. If we count all of the building edge pixels and terrain web points manually selected for scenes such as DC38008 and DC37405, less than 3% of the scene points are used to produce the dense reference disparity map. These points are represented in a triangulated irregular network (TIN) for the terrain, on which is superimposed the building roof structures. We linearly interpolate the TIN network in order to calculate the dense disparity map. Interestingly, S2 gives us matches for approximately 12% of the scene points, which is typical for feature-based matching algorithms. As such, our performance analysis is subject to possible errors in the evaluation of the S2 matching algorithm introduced due to interpolation from the sparse disparity map.

Given the lack of performance evaluation techniques in computer vision for 3-D scene modeling, we are probably content simply to know the height of the buildings and the general shape of the underlying terrain. However, we should understand that if we attempt to push performance analysis to detail the small effects of subtle algorithmic changes, we may run up against fundamental limits in our ability to recover these microstructures. Thus, as we have discussed previously, we have added an uncertainty of ±1 pixel of disparity to the ideal ground-truth value and feel that this covers a large fraction of the inherent inaccuracies. In summary, our disparity performance evaluation has to be considered as a method to easily detect large mismatches by the stereo analysis system; it may have some limitations in the fine evaluation of disparity values. Nevertheless, we see such techniques as the only method for effective comparison of disparity results. We also argue that it is far superior to traditional visualization techniques such as perspective views of wireframes or texture mapped imagery.

V. CONCLUSIONS

Fully automated stereo analysis in complex urban scenes is a difficult research problem. In this paper, we have discussed three major areas in the development of competent 3-D scene interpretation systems. First, we discussed the importance of accurate automatic scene registration and the difficulty in automated extraction and matching of scene reference points. We showed several results in fully automated scene registration including the estimation of the scene disparity range as a necessary parameter for stereo matching algorithms.

Second, we briefly described two stereo matching algorithms: S1, which is an area-based matcher previously used in the SPAM system [31], and S2, which is a new feature-based matching algorithm based on hierarchical waveform matching. We also introduced a technique to merge the results of the two matching algorithms, which appears to give an improved disparity map and indicates areas where occlusion and gross mismatches may have occurred.

Finally, we introduced several performance evaluation metrics that allowed us to measure the quality of the overall scene recovery, the building disparity estimate, and the quality and sharpness of the building delineations. We argue that such manually generated scan reference models are critical for understanding strengths and weaknesses of various matching algorithms as well as in the incremental development of improvements to existing algorithms.

We performed these experiments on difficult examples of aerial imagery containing complex urban scenes with varia-
tions in terrain, building shape, size, and height, as well as in an example of open terrain. Our future work is directed toward improvements in the basic s1 and s2 matching algorithms, the refinement of our ground-truth disparity maps to allow for a finer detail of analysis, and in techniques that will allow us to merge and refine our 3-D scene interpretations using information available from monocular analysis of the scene.

ACKNOWLEDGMENT

The DENVER scene stereo pair and associated ground-truth disparity map were provided by R. Norvelle and G. Lukes of the Autonomous Technologies Division, Research Institute, U.S. Army Topographic Engineering Center, Fort Belvoir, VA. The other imagery presented in this paper was digitized from aerial mapping photography provided by the Washington Gas Light Company, Springfield, VA.

We thank our colleagues in the Digital Mapping Laboratory for an interesting and congenial working environment. We particularly acknowledge W. Harvey and J. Shufelt, who helped in the preparation of this paper, and E. Burke, who generated several of the scene reference disparity maps and organized our racquetball schedules.

The authors also thank the anonymous reviewers and the editors of this Special Issue for their valuable suggestions, which improved the content and organization of this paper.

REFERENCES

David M. McKeown, Jr. (M'72) received the B. S. degree in physics and the M. S. degree in computer science from Union College, Schenectady, NY. Prior to joining the faculty, he was a researcher at Carnegie-Mellon University, Pittsburgh, PA, from 1975 to the present, a research associate at George Washington University and a member of the technical staff at NASA Goddard Space Flight Center, Greenbelt, MD, from 1974–1975, and an instructor in computer science and electrical engineering at Union College from 1972–1974. He is currently a Senior Research Computer Scientist in the School of Computer Science at Carnegie-Mellon University. His research interests in computer science are in the areas of image understanding for remote sensing and cartography, digital mapping and image/map database systems, computer graphics, and artificial intelligence. He is the author of over 35 papers and technical reports and is an active consultant for government and industry in these areas. He has been the principal investigator on research programs sponsored by the U. S. Army Engineer Topographic Laboratories, the Defense Mapping Agency, the Air Force Office of Scientific Research, and the Defense Advance Research Projects Agency.

Mr. McKeown is a member of ACM, AAAI, the American Society for Photogrammetry and Remote Sensing, and Sigma Xi.

Frederic P. Periant received the french engineering diploma from Ecole Nationale Superieure de l'Aeronautique et de l'espace in 1986. Before working as a research staff member at ARECOM, Paris, on SPOT stereo reconstruction and 3-D simulation (from 1987–1988), he completed the french masters degree in image processing at Ecole Nationale Superieure des Telecommunications, Paris, in 1987. From 1988 to 1990, he was a VSNA visiting scientist at the Digital Mapping Laboratory, Carnegie Mellon University, Pittsburgh, PA. He is currently a member of the technical staff of ISTAR, Sophia Antipolis, France, working on the analysis of radar imagery, particularly ERS-1. His research interests in computer science are in the areas of image understanding, cartography, and remote sensing.