Scene Registration in Aerial Image Analysis

Frederic P. Perlant and *David M. McKeown*
Digital Mapping Laboratory, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

ABSTRACT: In this paper we discuss the importance of scene registration for several tasks in the automated interpretation of aerial imagery. These tasks are structure matching, stereo matching, and stereo visualization. While the processes of registration and matching have traditionally been treated as separate problems, particularly in the case of stereo matching, we describe techniques that may unify these processes. We also demonstrate the automatic generation and matching of control points in complex aerial imagery and show that the resulting registration is comparable to that achieved using manual control point selection. Finally, methods for the generation and visualization of stereo disparity images and stereo ground-truth scene segmentations are described.

INTRODUCTION

SCENE REGISTRATION is a fundamental requirement for a number of image analysis tasks such as stereo matching, multi-image matching for temporal changes, and image sequence or motion analysis. As a result, there exists a rich variety of techniques to perform scene registration. For example, scene registration can be accomplished by identification of image points to a common frame-of-reference using control points whose three-dimensional location is accurately known. Registration can also be accomplished in a relative manner by identifying corresponding points between one or more images, i.e., the establishment of image-to-image control points. The positions of these points need not be known in the three-dimensional world. For some applications registration can be accomplished with respect to a cartographic map, a photomosaic, or an orthophoto that has been warped in order to remove position distortions due to terrain relief.

Depending on the type of registration, there arise many issues in accuracy. Accuracy is generally evaluated within the context of a particular task requirement. Often, techniques with some inherent inaccuracy can suffice in many task situations. If we look at traditional photogrammetric techniques for recovering the position of an image, we require detailed information regarding the camera (i.e., its focal length and lens distortion characteristics) and the platform coordinates in terms of its three-dimensional position and attitude (pitch and yaw). Inaccuracies arise when we are unable to know any of these parameters precisely. In the case of digital imagery, the image formation process can introduce additional errors due to the digitization of the photographic film. Even once we have accurate image, sensor, and platform information, our ability to locate ground control points accurately in digital imagery is independent of the inherent accuracy of those control points.

This paper raises issues in how scene registration can be achieved in digital imagery and illustrates the importance of accurate registration for three analysis tasks. We discuss some general issues in registration in computational vision, including the use of a spatial database to provide coarse ground control information, the selection of manual control points, and the automatic determination of control points. Then we show the importance of registration to three particular tasks in the interpretation of aerial imagery. These tasks are the correlation (or fusion) of monocular analysis from partially overlapping views, computational stereo matching techniques, and the visualization of stereo matching results. Finally we present the results of a fully automatic scene registration from initial coarsely registered stereo pair to a final three-dimensional interpretation.

SCENE REGISTRATION

The primary goal of stereo photogrammetry is to determine the three-dimensional 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 relationship between the image points and ground points in the scene is determined through the camera orientation. The relative orientation determines the relative three-dimensional position of the two images in the stereo pair with respect to each other. According to simple assumptions, the calculation of the relative orientation is in itself the registration problem (Horn, 1988). 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 three-dimensional ground control points to establish the absolute orientation. The relative orientation is classically reformulated into the epipolar geometry for stereo imagery. When two images are registered in the epipolar geometry the spatial relationship between corresponding points in the left and right images is greatly simplified. The corresponding points are on the same scanlines in the left and right image and the displacement between the points, or disparity, corresponds to the relative height of the three-dimensional scene point.

Epipolar geometry is a common framework for most stereo matching algorithms (Arnold, 1978; Barnard and Fischler, 1982; Barnard, 1988; Nasrabadi, et al., 1988; Ohta and Kanade, 1985). These stereo matching techniques assume that the registration is ideal and that the epipolar constraint is completely satisfied. Some researchers have attempted to explicitly account for the inaccuracy of image registration and have attempted to improve it by preprocessing the imagery before beginning the matching process (Hannah, 1985; Brooks *et al.*, 1988; Chen and Boult, 1988; Weinshall, 1988). Modeling inaccuracy in image registration has most often been studied within the context of robotic applications where it is common to have a good deal of control over the cameras (Faugeras and Toscani, 1986) and where a detailed preregistration and calibration step is possible. However, 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 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 performs a fine registration using pairs of corresponding points to get a precise 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. The effect and implications of these inaccuracies will be explored within the context of two matching tasks.

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

PHOTOGRAMMETRIC ENGINEERING AND REMOTE SENSING,
Vol. 56, No. 4, April 1990, pp. 481–493.

0099-1112/90/5604–481$02.25/0

in the two images. One alternative method is to independently tie each image to a common frame of reference. A cartographic (geographic) coordinate system such as latitude, longitude, and 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, and elevation is a common method to orient each image. The overall accuracy of the registration is dependent on the accuracy of the three-dimensional 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, a spatial database system that integrates imagery, terrain, and map data to provide landmark descriptions (McKeown 1984, 1987). Each landmark description in the database has a reference image fragment; a ground position definition which contains the latitude, longitude, and elevation information and its position in the reference image fragment; and a brief textual description of the landmark for the user. Each image in the CONCEPTMAP database is put into correspondence using a polynomial model derived by manual selection of landmarks.

Figure 1 shows a stereo image pair of an industrial area DC38007 taken from the CONCEPTMAP database. These images were digitized from standard nine-inch format mapping photography taken at an altitude of 2000 metres by a camera with a focal length of 153 millimetres. One pixel corresponds to 1.3 metres on the ground. The left image is a 512 by 512 sub-area selected from 2300 by 2300 image. The right image sub-area was generated by calculating the latitude and longitude for the corner points of the left image using a polynomial approximation and projecting those points onto the complete right image. This projection is then used to extract the image sub-area from 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. Typically CONCEPTMAP provides a registration accuracy of between ten to thirty meters for imagery digitized to a 1.3-metre ground sample distance.

Fine Registration Using Image Control Points

We begin the process of fine registration with the coarse registration described in the previous section. We make several assumptions that simplify the relative orientation model. We assume that the camera is metric, and that the optical axes are parallel and are, in fact, vertical. Because we are using aerial imagery taken by the same camera along the same flightline, these assumptions are not unreasonable. The largest source of error is whether the camera platforms were at precisely the same altitude and orientation at each imaging event. Given these assumptions, the transformation between the left and right image is only a translation and a rotation (ISOMETRY), because the image planes are the same. In such a transformation the absolute distances in the two images are preserved and the epipolar lines are already parallel. After the transformation the epipolar lines correspond to the scanlines.

To get the parameters of the isometric transformation, we need to select points in the same plane that, when transformed, preserve their relative distances with the images. Problems with the accuracy of point selection lead us to use more points than necessary to determine the transformation between the two images. However, on some images this model was not flexible enough to account for the variations to our ideal sensor model. As a result, we developed a polynomial transformation up to the third order, adjusted by least squares to fit the selected corresponding points.

Manual selection of common points. The classical method to select corresponding points is by the manual identification of landmark points in stereo imagery. Typically, man-made features such as road intersections, boundary corners of fields or parking lots, or markings such as road centerlines are used because of the ease with which they can be found in the imagery. We chose to manually select shadow corners since these points were the focus of our experimentation in automatic landmark detection. Given that we are working in an urban environment, shadow corners have the advantage that they are generally on the ground and therefore in the same plane, assuming only small changes in terrain elevation. Although the shadow position changes as the sun moves, if we have imagery taken at nearly the same time, as is common in aerial mapping photography, the shadow corners will fall on the same point in the three-dimensional scene. Such corners also tend to be uniformly distributed in scenes containing large numbers of buildings. The manually selected shadow corners give us a baseline against which we could measure the accuracy of the automatic landmark selection

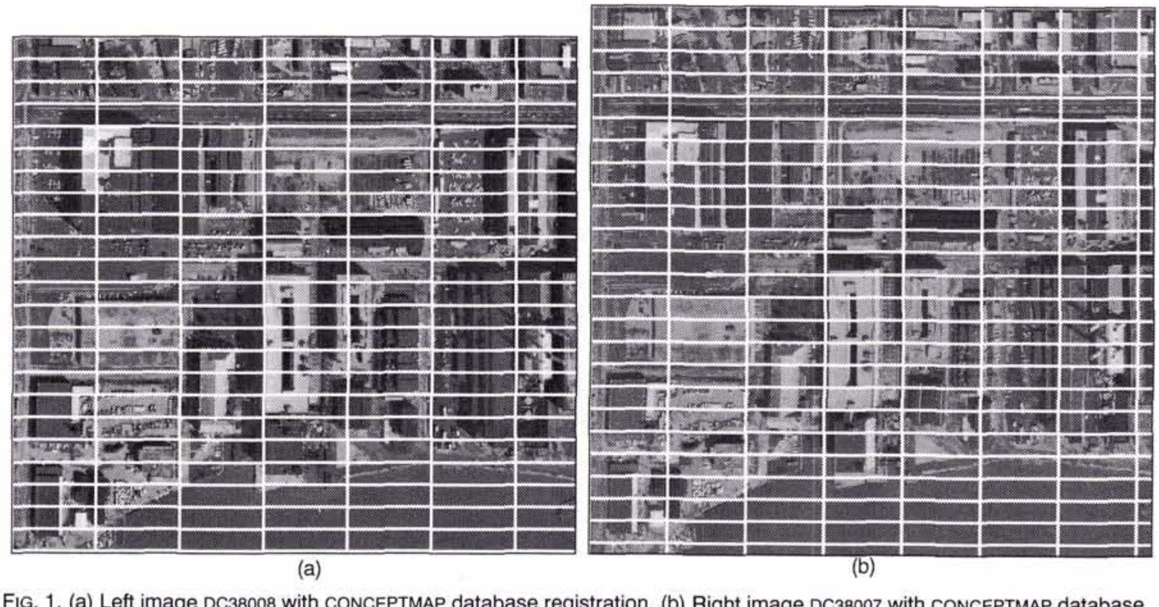

FIG. 1. (a) Left image DC38008 with CONCEPTMAP database registration. (b) Right image DC38007 with CONCEPTMAP database registration.

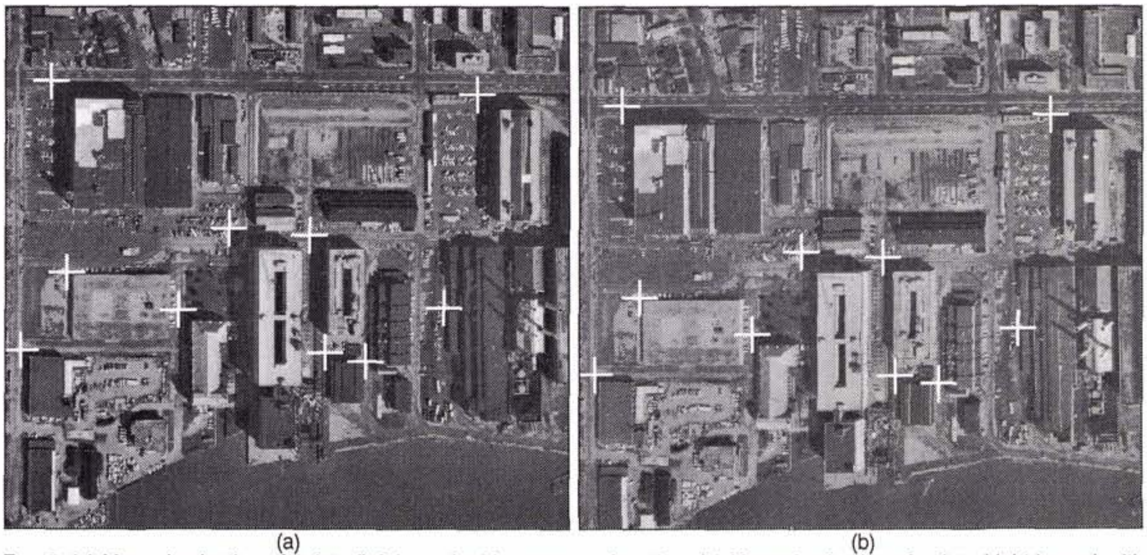

FIG. 2. (a) Manual selection of points (left image) with coarse registration. (b) Manual selection of points (right image) with coarse registration.

process. Figure 2 shows the manually selected shadow corners in the left and right images, respectively.

Automatic selection using shadow corners. Clearly, one requirement for automated registration is the automatic selection of corresponding points in the stereo pair images. There are actually 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 unsolved problem, and most automatic registration techniques rely on the matching of characteristic points (Moravec, 1980) that often have no physical significance or reference with respect to landmarks.

For this experiment, we assume that a coarse registration of the two images, such as previously described, 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.

Shadow corners are good candidates for automatic detection and correspondence as well as for manual selection. We use a monocular detection of shadow regions and determine the boundary line between the shadow and the building (Irvin and McKeown, 1989). This boundary is used to determine the position of the shadow corner in the left and the right images (Aviad, 1988). After removing corners that were inconsistent with shape and orientation constraints imposed by the sun angle, we selected sets of shadow corners that were detected in both images. Figure 3 shows these corresponding shadow corners on the two images. Note that the corners selected differ from those selected manually.

Figure 4 shows the results of the fine registration using shadow points selected automatically. This registration is obviously better than the coarse registration using the CONCEPTMAP database shown previously in Figure 1. In the following section we quantify the registration quality.

Quality of registration. Tables 1 and 2 show the local accuracy of the different scene registrations performed on DC38008 and LAX stereo image pairs. The first three rows of each table characterize the quality of the coarse registration using CONCEPTMAP database. Because the polynomial model stored in the CONCEPTMAP database is derived over the entire scene (2300 by 2300 for DC38008 and 2000 by 4000 for LAX), it is interesting to evaluate the quality of the local fit for the 512 by 512 image sub-areas using each set of independently derived control points. We used three sets of control points: (1) the points selected manually, (2) the points generated by automatic detection of shadow corners, and (3) the points derived from monocular structure matching discussed in the next section. In the case of DC38008, 11 corresponding point pairs were manually selected, 26 shadow corner pairs were automatically extracted, and 16 point pairs were automatically derived using structure matching. In the case of the LAX stereo pairs, these numbers are 14, 13, and 16, respectively. For each of the three sets of test control points, the CONCEPTMAP polynomial registration produced a vertical offset between the left and right images to within approximately 12 pixels (16 metres) for both stereo pairs.

We evaluated the quality of registration for the isometrical and polynomial models with respect to the manually derived control points. That is, the solutions for manual, corner, and structure matching control points were validated using the manual points. In all cases both registrations achieved significantly better results than the CONCEPTMAP coarse registration. In several cases the registrations achieved by matching shadow corners and structures is quite comparable to the manual registration. However, manual registration is in all cases as good as any of the automatic control point experiments. In all cases, the manual selection of corresponding points produced a registration of less than one metre, or subpixel accuracy. In some cases, similar subpixel results were achieved using the automatic point selection. Finally, the polynomial approach led to better results although the simpler isometrical model gave comparable results.

One additional issue is how well our local solution performs as a global registration in other areas of the complete stereo pair. In Table 3 we show the results of using our local fine registration for both the isometrical and polynomial methods in four quadrants of the complete stereo pair. In each of the four quadrants we manually selected 12 control points and used the manual solutions for DC38008 to calculate residual errors. Because of the large variation in the row and column offsets, it is clear that the local model can not be treated as a global model even though the row residuals stay within reasonable bounds. However, it is the case that the fine solutions should give a better global solution than the current CONCEPTMAP polynomial registration.

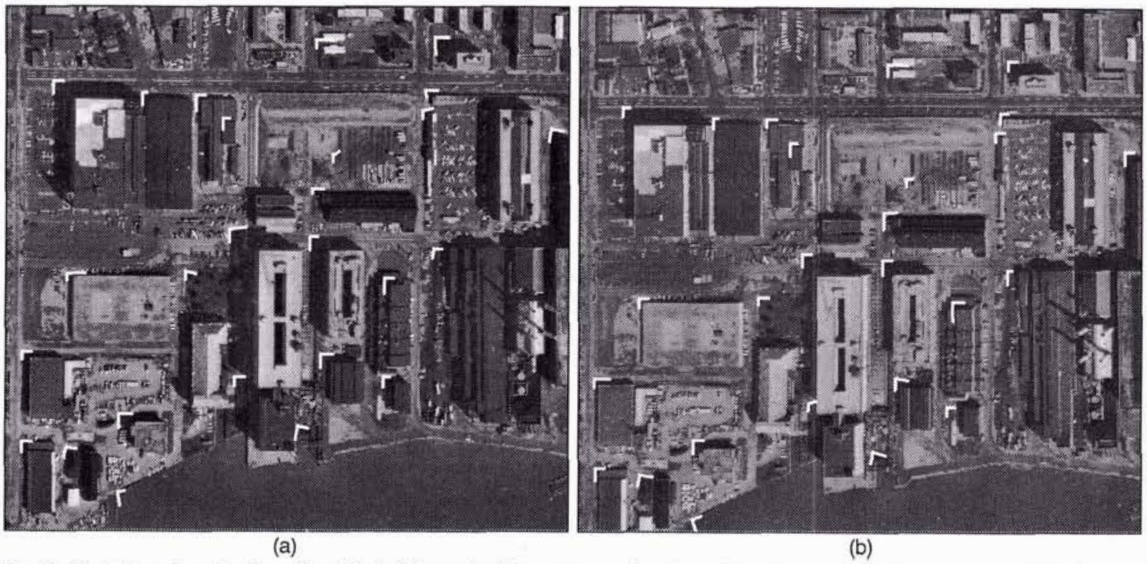

FIG. 3. (a) Automatic selection of points (left image) with coarse registration. (b) Automatic selection of points (right image) with coarse registration.

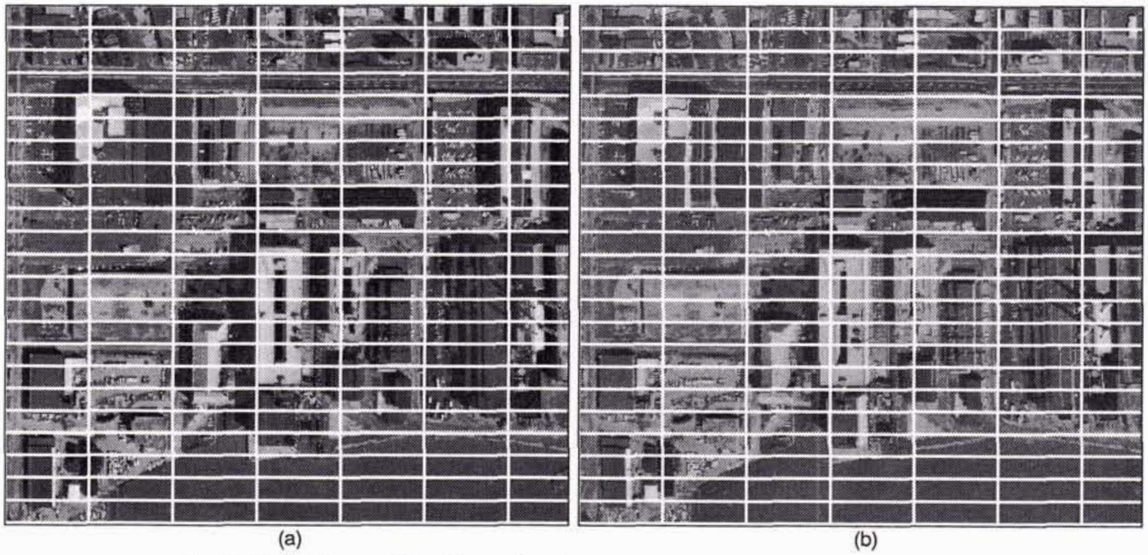

FIG. 4. (a) Left image of the fine registration. (b) Right image of the fine registration.

TABLE 1. STATISTICS FOR DIFFERENT REGISTRATIONS ON DC38008 STEREO PAIR

Statistics on the quality of the different registrations for DC38008						
Type of registration	Number of points	Avg. row offset (pixels)	Std. dev. row offset (pixels)	Min/Max row offset (pixels)	Avg. col offset (pixels)	Std. dev. col offset (pixels)
Coarse manual	11	−12.4	1.6	−15/− 8	905.5	1.2
Coarse corner	26	−13.2	1.6	−18/−10	905.7	1.7
Coarse structure	16	−12.1	1.6	−15/− 8	909.3	3.4
ISO manual	11	0.1	0.7	−1/2	1.3	1.3
ISO corner	11	1.7	1.7	−1/7	5.1	1.2
ISO structure	11	0.5	0.6	0/2	−3.4	1.4
POLY manual	11	0.1	0.3	−1/1	0.1	0.5
POLY corner	11	−0.2	1.8	−3/4	0.0	1.6
POLY structure	11	0.1	0.5	−1/1	−3.3	1.5

TABLE 2. STATISTICS FOR DIFFERENT REGISTRATIONS ON LAX STEREO PAIR

Statistics on the quality of the differnt registrations for LAX						
Type of registration	Number of points	Avg. row offset (pixels)	Std. dev. row offset (pixels)	Min/Max row offset (pixels)	Avg. col offset (pixels)	Std. dev. col offset (pixels)
Coarse manual	14	10.6	0.9	9/13	1866.6	0.7
Coarse corner	13	10.9	0.7	9/12	1866.8	1.4
Coarse structure	16	10.9	0.4	10/12	1869.3	1.7
ISO manual	14	−0.4	0.9	−2/2	0.6	0.7
ISO corner	14	−0.4	0.9	−2/2	1.6	0.7
ISO structure	14	−0.4	0.9	−2/2	−2.4	0.7
POLY manual	14	0.0	0.1	−1/1	0.1	0.7
POLY corner	14	1.3	1.0	−1/3	1.5	0.9
POLY structure	14	−0.3	0.7	−1/2	−2.9	0.9

TABLE 3. STATISTICS FOR DIFFERENT REGISTRATIONS ON DC38008 STEREO PAIR

Quality of the registrations for the complete image DC38008						
Type of registration	Region in image	Avg. row offset (pixels)	Std. dev. row offset (pixels)	Min/Max row offset (pixels)	Avg. col offset (pixels)	Std. dev. col offset (pixels)
ISO manual	North	1.7	1.5	0/4	4.5	0.8
	West	−1.3	0.4	−2/−1	1.4	1.2
	East	1.3	0.6	0/2	−2.4	1.5
	South	−2.5	0.8	−5/−1	3.1	1.1
POLY manual	North	−1.6	1.2	−4/0	−70.5	16.6
	West	−1.3	0.7	−2/0	−2.4	3.0
	East	−0.1	0.3	−1/1	2.7	3.7
	South	0.8	0.7	−2/2	−51.5	14.3

Although traditional error analysis can give us an idea of relative accuracy for each of these approaches, this does not necessarily translate into the effectiveness of the registration. That is, for many tasks in scene analysis a coarse-grain registration to within 10 to 30 metres is quite adequate, especially considering that the imagery covers several square kilometres. For instance, tasks that require assembling a collection of image subareas taken over time for change detection and analysis can be supported using this level of accuracy. However, for other tasks, such as matching and stereo analysis, the effect of mis-registration may become more critical. In the following section we see how such tasks are affected by coarse and fine registration.

TASKS REQUIRING ACCURATE SCENE REGISTRATION

We describe three scene analysis tasks that require or support scene registration. These tasks are matching structures derived by monocular analysis of overlapping imagery, traditional stereo matching using area-based and feature-based matching techniques, and the construction of a three-dimensional image to present matching results to a human using a stereo display monitor.

In the case of matching monocular structures, we can acquire additional information about the actual structure of the objects, including their height, as a result of the matching process, and also generate new automatic control points to refine the registration. The goal is to match high-level structures in two overlapping images, where the actual detection and delineation of the structures is likely to contain significant errors, and matching is complicated due to a large number of false alarms produced by the structure generation process.

In stereo analysis the goal is to automatically match points in the left and right images of the stereopair in order to establish a *disparity* (parallax) between these points. This disparity can be used, along with the camera model, to calculate the actual height of the matched points in the three-dimensional scene.

Finally, it is becoming increasingly important for researchers to be able to visualize the three-dimensional models that their analysis programs are generating. Such a visualization tool allows us to directly compare these results to three-dimensional ground-truth models for performance evaluation.

CORRELATION OF MONOCULAR ANALYSIS

There are many situations where overlapping coverage imagery is available, but may not be suitable for stereo matching due to sensor acquisition parameters, temporal or seasonal changes, or image scale. The issue then becomes one of how to relate the results of independent monocular analysis. One of the first examples in the literature was symbolic change detection (Price, 1976; Price and Reddy, 1979) and the matching of coarse regions such as lakes, fields, and forests based upon relationships that were largely invariant over small rotations in the image plane (<45 degrees) and relatively large scale changes (factor of ten resolution). These techniques have been generalized to the matching of semantic network descriptions generated by separate monocular analysis or from a baseline cartographic description (Price, 1985).

Our interest in matching of monocular interpretations arises from our desire to relate structure descriptions generated from a building hypothesis system. The BABE (Built-up Area Building Extraction) system (Aviad *et al.*, 1989) performs monocular analysis on an image by extracting lines and corners and generating structure hypotheses. This work is similar to Huertas and Nevatia (1988), but differs in that a large number of hypotheses are purposely generated such that buildings are rarely missed. These structures are then evaluated by a number of techniques such as shadow verification, shadow prediction, and shadow grouping (Irvin and McKeown, 1989). The processes of verification, prediction, and grouping are used to rank order or prune the large number of BABE structure hypotheses.

Monocular matching can also be viewed as a form of structure verification. That is, sets of independently derived hypotheses from different images are matched using the scene registration model to relate absolute ground position in the two images. This provides information that is not available in a single image: an estimate of structure height and the reliability of each hypothesis. For example, because matching allows multiple hypotheses in one image to correspond with a single hypothesis in the second image, we can use this fact to guide a re-examination of the structure delineation in the first image. The fragmentation of structures is a common source of error in computer vision, and understanding fragmentation requires some external process to predict its occurrence or to identify situations where it has occurred. Even in cases where there is a good one-to-one match between structures, different viewing angles, accidental alignments of objects in the scene, or differences in imaging conditions will produce differences in the segmentation that can be used as cues to guide further interpretation.

A Matching Experiment. We performed an experiment in structure matching using a portion of a stereo pair of the Los Angeles International Airport (LAX) used by Huertas and Nevatia (1988) in their building extraction research. The goal of this experiment is two-fold. First, we want to explore the effects of registration errors in structure matching. Second, given the results of automatic structure matching, could we achieve a local registration comparable to that generated using manual selection of control points?

Figure 5a shows the superimposition of BABE results using the CONCEPTMAP coarse registration while Figure 5b shows the results using the polynomial model generated by manual selection of control points. The building hypotheses outlined in white were generated for the left image while those outlined in black are hypotheses independently generated for the right image. This superimposition allows us to see the structure correspondence as well as the differences between the two monocular analysis results. The horizontal displacement between the hypotheses for the fine registration can be related to the relative height of the structures because of the epipolar constraint. Coincident structure hypotheses give very strong support for hypotheses of buildings. This is due to the fact that the feature extraction process rarely fails in exactly the same way in each of the images.

To automate matching between the hypotheses in the left and right images, we utilize geometric constraints. We take each BABE hypothesis from the left image and find the best corresponding hypothesis on the right image. To evaluate the effects of registration on our matching algorithm, we performed structure matching on using both the coarse and fine registration results. Figure 6a shows the matching results of the coarse registration while Figure 6b shows the matching results of the fine scene registration. In both cases we have chosen a small area from the left-center of the LAX scene to illustrate the details of matching. The light (dotted) structures are hypotheses from the left image while the dark (solid) structures are from the right image.

The matching process is a global search between two sets of boxes according to local limitations in the search area. The epipolar geometry of the fine registered images can be used to constrain the search area to a range of scanlines in the images. Then we use the following simple criteria to select potential matches; the position of the hypothesis center-of-mass projected into a rectangular search space and amount of overlap between the pairwise structures.

This simple matching process allows us to consider arbitrarily complex polygonal structures because we are not performing discrete vertex or structure matching to establish a stereo correspondence such as in Mohan and Nevatia (1988). This is required given the relatively complex imagery and imprecise segmentation delineation provided by BABE. In many cases detailed high-level structure matching will be defeated by errors in monocular feature detection due to occlusion, texture, and accidental alignments of objects and background. These are precisely the errors that cause area-based and feature-based matching to fail, although propagated to a high-level matching process.

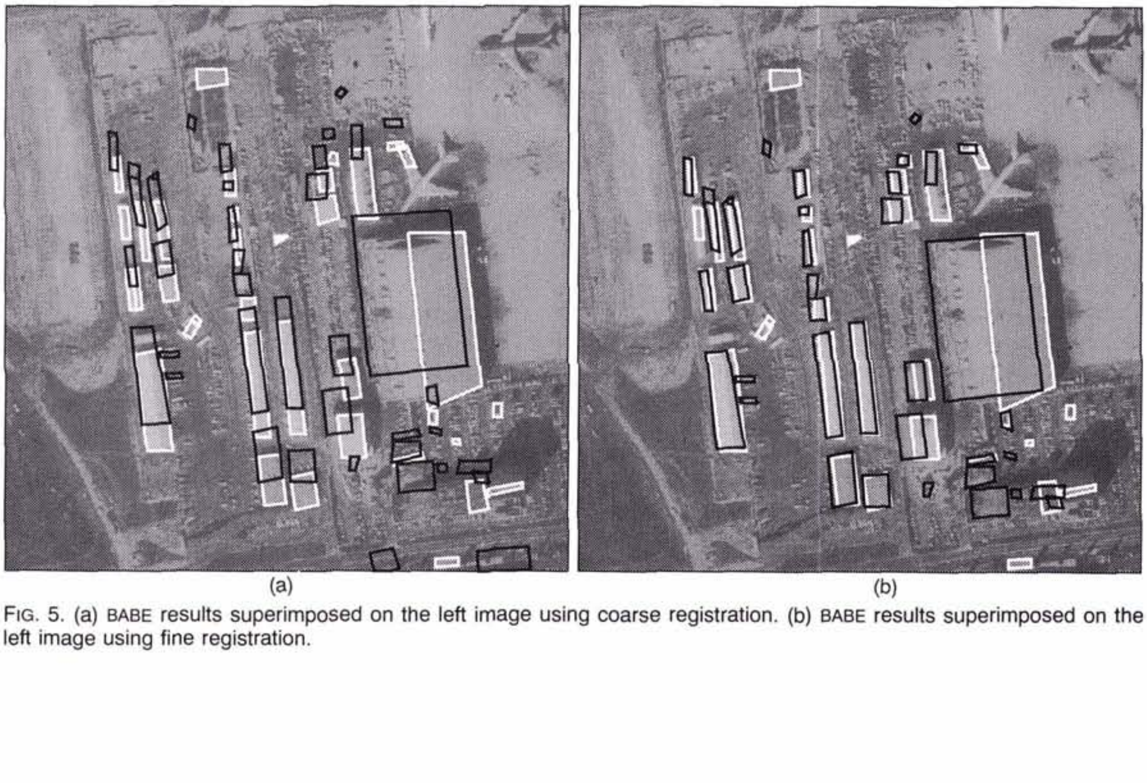

FIG. 5. (a) BABE results superimposed on the left image using coarse registration. (b) BABE results superimposed on the left image using fine registration.

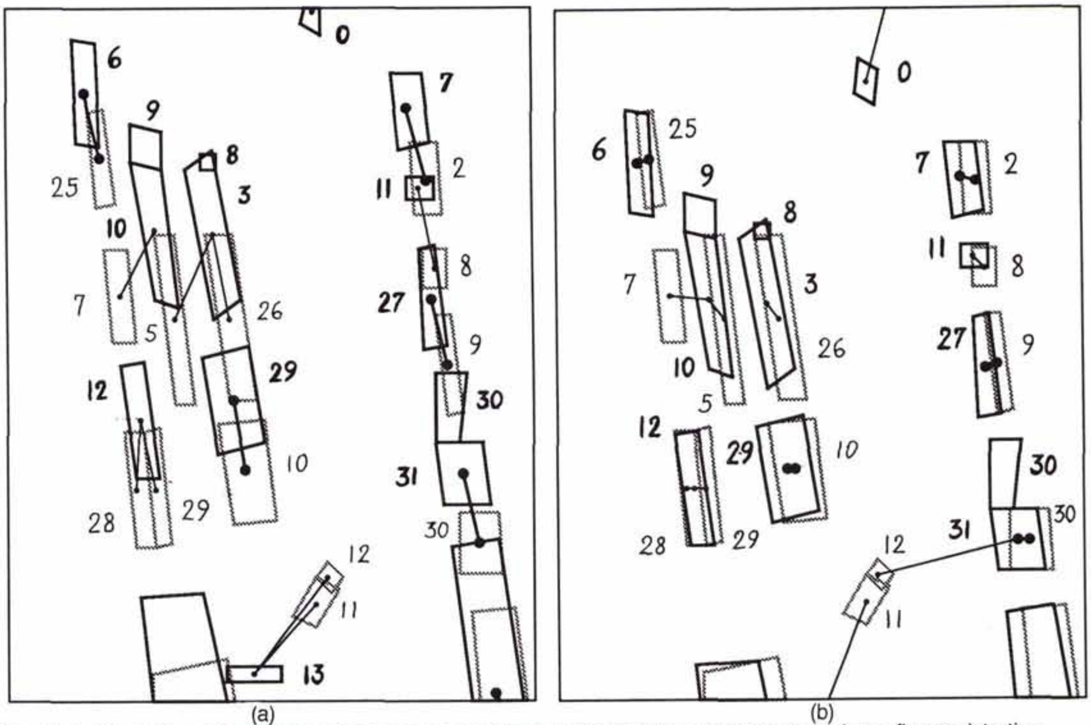

FIG. 6. (a) Matching of buildings using a coarse registration. (b) Matching of buildings using a fine registration.

Because our matching criteria are not very selective, we must disambiguate among many plausible matches. However, even if we devised a more specific set of match criteria, it is unclear whether we could account for situations in built-up urban areas where the buildings are very close, and have very similar shapes, orientations, and heights simply by using a set of local optimal matches. There are several examples of the alignment of similar buildings even within the LAX imagery. Thus, there is a virtue in the application of weak constraints because they do not require detailed high-level knowledge abut the mis-registration. Instead of trying to disambiguate the matches locally, we use global considerations based on the plausibility of the matched sets of structures. We define four different situations that occur depending on whether several structure hypotheses share the same correspondence with a hypothesis structure in the other image.

Tables 4 and 5 show the results of the monocular matching using the coarse and fine scene registration. In this experiment we search for the best match for each of the building hypotheses produced by BABE in the left image. A similar analysis could be performed on the structures generated from the right image. Four different situations can occur during matching that correspond to the application of local and global properties.

TABLE 4. MATCHING RESULTS FOR LAX BUILDING HYPOTHESES WITH COARSE REGISTRATION

Results of the matching of boxes for the coarse registration					
Type of corres.	Left box	Right box ———	Corres. score	rel. height estimate (pixels)	rel. line offset (pixels)
1	2	7	0.92	4.0	18.2
	8	11	0.73	3.5	21.0
	9	27	0.85	4.4	16.4
	10	29	0.94	3.7	18.3
	25	6	0.83	3.8	17.4
	30	31	0.93	4.2	18.2
2	7	10	0.73	−9.3	16.9
	28	12	0.89	−0.1	17.5
	29	12	0.80	4.4	17.1
3	5	3	0.75	−10.0	22.0
	26	3	0.87	4.5	21.5
4	11	13	−44.5	19.6	−24.9
	12	13	−34.5	16.0	−18.5

- *Type 1*. This case corresponds to the "ideal" situation where we have a unique correspondence between a single hypothesis in the left and right image. The score of the correspondence gives us an estimate of the quality of the match between the two structures. A match score greater than 0.9 indicates that the two structures are quite similar. Pair (10,29) is an example of such a good match. A lower match score, as in the case of Pair (8,11), indicates that, while there is a correspondence between structures, their BABE delineation is not completely consistent between the left and right image.
- *Type 2*. This case occurs when a structure in the left image shares a right structure correspondence with other structures in the left image. This correspondence is therefore ambiguous. In this case the match score is not sufficiently different to disambiguate between the multiple choices. However, knowledge of a reasonable height range for these structures could be used to select the correct correspondence. For example, Pair (28,12) and Pair (29,12) have a significant difference of 4.5 metres in their height estimate, although neither height is sufficiently unusual to prefer one interpretation over another without some external information. However, in the case Pair (7,10) for the coarse registration, this match could be discounted due to a height interpretation that is below the local terrain.
- *Type 3*. This case occurs when several matches were possible with different structures in the right image, for example, Pair (26,3). The correspondence selected has the highest confidence match but other correspondences are possible, i.e., Pair (26,10). Once again, knowledge about the reasonable height range for structures could be used to select the appropriate correspondence.
- *Type 4*. This case occurs when structures in the left image do not have any reasonable correspondence in the right image, such as in Pair (11,13). Nevertheless, the best correspondence is given.

Table 5. Matching Results for LAX Building Hypotheses with Fine Registration

Results of the matching of boxes for the fine registration					
Type of corres.	Left box	Right box ———	Corres. score	rel. height estimate (pixels)	rel. line offset (pixels)
1	2	7	0.92	3.1	0.2
	8	11	0.73	2.5	3.0
	9	27	0.86	3.4	−1.6
	10	29	0.94	2.7	0.3
	25	6	0.83	2.8	−0.6
	30	31	0.93	3.2	0.2
2	5	10	0.75	3.7	4.9
	28	12	0.89	−1.1	−0.5
	29	12	0.81	3.4	−0.9
3	26	3	0.7	3.5	3.5
4	7	10	−11.4	−10.3	−1.1
	11	31	−44.6	−35.5	−1.6
	12	13	−51.5	15.0	−36.5

There appears to be no major difference between the matching results derived starting with either the coarse or fine scene registration. In the case of good matches, i.e., Type 1, having a high score, the results are identical for the coarse and the fine registration. Even if, according to the registration, different types of matching are possible for the ambiguous matches, this technique still seems to be quite robust. However, different scene clues can be derived by the analysis of the absolute match score, the match type, the estimated height, and the relative row offset for the different registrations. For example, in Type 2 matching a choice between two competing interpretations must be made. A verification process could be invoked to locate a better matching and delineation of the structures. In the following section we show that the results of structure matching can be used to automatically generate a refined scene registration.

Structure Matching for Registration. One requirement for automated registration is the automatic selection of corresponding points in the stereo pair images. Classically, these points are physical features of the images such as shadow corners, road intersections, or specific unique landmarks. However, we can also derive "virtual geometric points" that are solely defined by their geometrical relationship to real features in the images. Structure matching provides an estimate of the local offsets in rows and columns for the center-of-mass of the structures generated by BABE in the left and right images. We can consider these corresponding points as control points selected automatically and then perform a registration of the stereopair exactly as with the shadow corners.

Figure 7a shows the control points automatically selected using structure matching based upon the coarse registration as previously shown in Figure 5a. As before, structures generated by BABE in the left image are shown outlined in white, and the right image structures are shown outlined in black. The automatic control points selected correspond to good structure matches. They are shown as linked black circles. Because BABE does a fairly good job in structure delineation and generates consistent hypotheses in both images, the set of control points considered is quite reliable. In fact, the residuals of the registration (ISO, structure and POLY, structure) shown in Table 2 are comparable to the registration derived using manual control points (ISO, manual and POLY, manual). Finally, Figure 7b shows the fully automated registration achieved using the control points in Figure 7a. The results are nearly identical to those derived using manual registration and show a good superimposition of the building structures.

Registration for Stereo Matching

The central issue in computational stereo is the solution of the correspondence between features visible in two overlapping images. The correspondence of a point feature visible in the left and right image of a stereo pair can be used to generate the three-dimensional description of that point in the scene. Points need not be the only feature matched. As we have seen, the result of matching structures generated by monocular analysis can be considered as stereo matching and yields a relative height estimate.

The epipolar constraint is used to simplify stereo matching by reducing it to a one-dimensional problem because the epipolar lines are registered to be corresponding scanlines in the left and right image. 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. In the following section we discuss the effect of coarse and fine scene registration on two stereo matching algorithms.

Two stereo correspondence algorithms. Algorithms for stereo correspondence can be grouped into two major categories: area-based and feature-based matching (Barnard and Fischler, 1982). Both classes of techniques, area-based and feature-based, have advantages and drawbacks that primarily depend on the task domain and the three-dimensional accuracy required. 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 buildings and open terrain. For this reason we have developed two stereo matching algorithms. S1 is an area-based algorithm and uses the method of differences matching technique developed by Lucas (Lucas, 1984; McKeown *et al.*, 1986). 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. Both are hierarchical and use a coarse-to-fine matching approach. Each is quite general as the only constraint imposed is the order constraint for the feature-based approach. The order constraint should generally be satisfied in our aerial imagery except in cases of hollowed structures.

Both matching algorithms assume the epipolar geometry but have different sensitivity to its accuracy. The S1 area-based approach uses a hierarchical set of reduced resolution images to perform a coarse-to-fine matching of small windows in the two images. At each level the size of the windows for the matching process depends on the resolution of the reduced image. An initial disparity map is generated at the first level. Subsequent matching results 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, because there is a relationship between the matching window size and the level of accuracy, simply using larger matching windows may not be desirable.

The S2 feature-based approach matches epipolar lines in the left and right image. It uses a hierarchical approximation of the intensity waveforms to match peaks and valleys at different levels of resolution. To avoid mismatches, it uses inter-scanline consistency that enforces a linear ordering of matches without order reversals. It also applies an intra-scanline consistency that considers the matches in adjacent scanlines. Application of intra-scanline constraint is used to increase the confidence of matches found to be consistent across multiple scanlines and to delete improbable matches.

Figure 8a is the complex industrial area scene that was the focus of our previous discussion on coarse and fine scene registration. This scene contains many of the difficulties found

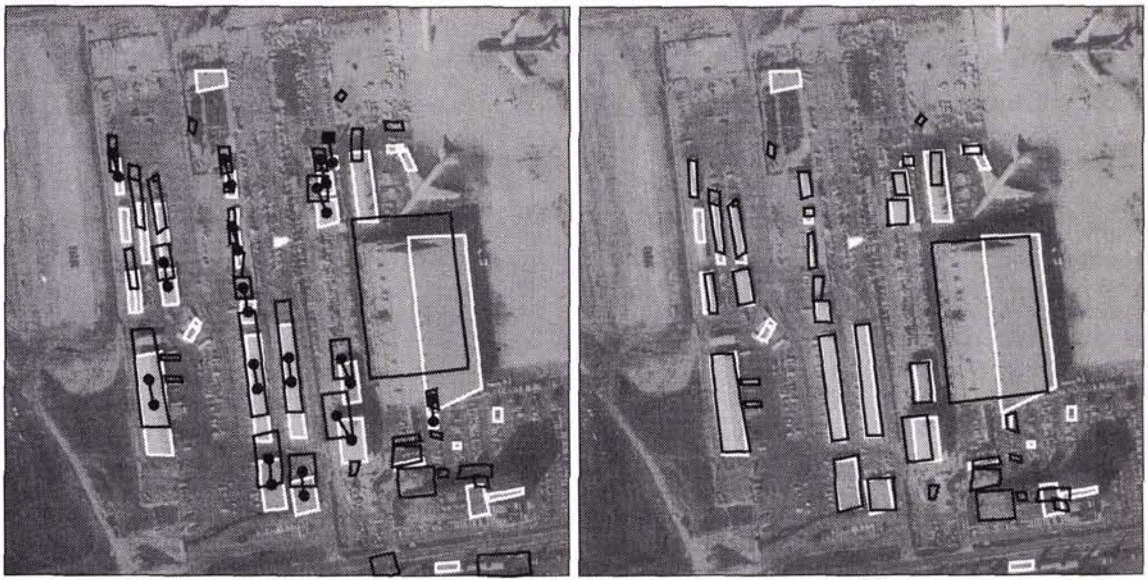

FIG. 7. (a) Structure matching using coarse scene registration. (b) Super-position of structures using structure matching registration.

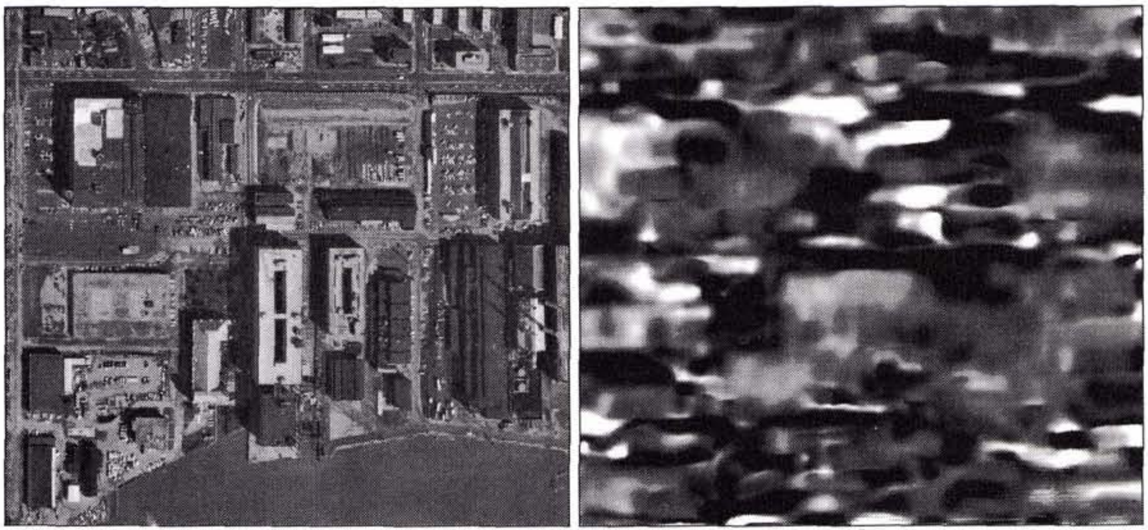

FIG. 8. (a) Intensity image of the scene. (b) S1 disparity map using coarse registration.

in stereo matching, including occlusion, complicated textures, large depth discontinuities, and complicated three-dimensional objects. Figure 8b shows the results of the matching for the CONCEPTMAP coarse registration using the area-based algorithm S1. In all of the disparity match results presented in this paper, brighter regions are closer to the camera and have greater height. Darker regions are at or below the relative terrain ground plane. The results using the coarse registration contain huge errors. We can barely discern the general shape of the taller buildings in the middle and the upper left areas of the scene. The S2 algorithm is completely unable to use a coarse registration because the scanline matching asssumes that the epipolar constraint is satisfied.

Figure 9a shows the results of the matching using the S1 algorithm with the fine registration produced by the manual selection of shadow corner points. The matching results are significantly better, with the bright areas again representing the highest regions and corresponding to most of the buildings in the scene. Although the delineation is not crisp, there are few major mismatches, and we have an adequate impression of the range of heights in the scene. The S1 algorithm has many of the advantages and drawbacks of any area-based technique. As we can see in Figure 9a, most of the errors are due to abrupt changes in height due to man-made structures.

Figure 9b shows the results of the matching using the S2 algorithm with the same fine scene registration in Figure 9a. This technique performs very well on the disparity discontinuities caused by man-made structures, and therefore we have a much better delineation of the buildings than in Figure 9a. Nevertheless, despite post processing of the disparity results, the resulting disparity image is noisy. As expected, the S2 algorithm may not provide robust matches in areas of uniform intensity or in highly textured areas.

The results of the two stereo matching algorithms are quite

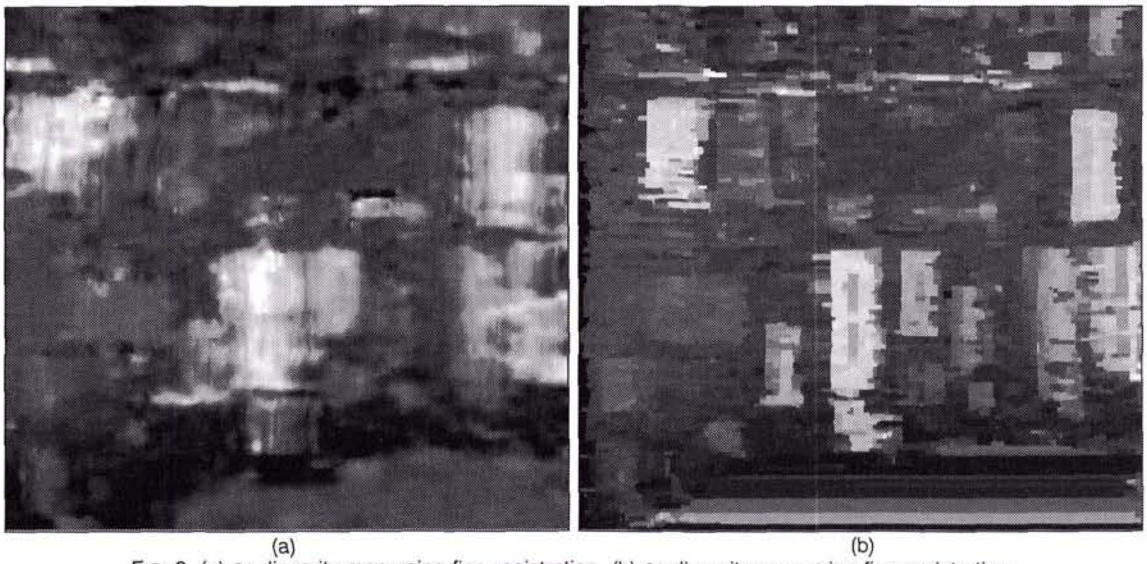

FIG. 9. (a) S1 disparity map using fine registration. (b) S2 disparity map using fine registration.

complimentary, and 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 three-dimensional representation of the scene. It is also clear that stereo matching relies on a more accurate scene registration than is provided by the CONCEPTMAP coarse registration. Even when the matching window size for an area-based stereo algorithm is larger than the inherent mis-registration, it may be difficult for the matching algorithm to recover from mis-matches due to poor scene registration. This is in contrast to the results in monocular matching that appear to be much less sensitive to a coarse scene registration.

REGISTRATION FOR VISUALIZATION

A practical means for three-dimensional visualization is needed in order to understand the quality of the stereo results. This can be achieved by the construction of a left and right stereo image from the original unregistered overlapping imagery. There are two common techniques to present stereo images to a user separating the left and right image: anaglyph and polarized stereo. For our experiments, we have used the anaglyph method; however, the construction of a synthetic stereo image is identical regardless of viewing method.

Construction of the right image from the left image. Given a left and right image registered into the epipolar geometry, as are the intensity images after fine registration, we can display them using either anaglyph or polarized stereo techniques. A more interesting approach is to generate a stereo pair (left and right image) from the three-dimensional information we have computed by stereo matching. In this way we can directly visualize the matching results as a three-dimensional scene. We call the generation of a stereo pair from three-dimensional information derived from stereo analysis *synthetic stereo reconstruction*. Synthetic reconstruction can be used to visualize and compare the results of stereo matching by direct visualization. A relative height computation, or disparity, is the result of most stereo matching algorithms with the disparity encoded relative to the geometry of the left image. In such a disparity map, the values of each point in the map correspond to the relative height of that point in the left image. In order to generate a synthetic reconstruction containing the information extracted by the matching process, we simply generate a new right image where each point in the right image is the corresponding point in the left image displaced by the relative height estimate in the disparity map.

The computed right image is, by definition, perfectly registered since there are only local horizontal shifts between the left and the right image. Thus, we satisfy the epipolar constraint. Figures 10a and 10b show reconstructed synthetic stereo images for the stereo matching results produced by S2 in Figure 9b.

3D segmentation for ground truth determination. The visualization of scenes and stereo matching results is a powerful tool for the qualitative comparison of different scene interpretation techniques. One technique is the side-by-side comparison of the original stereo scene and the automated reconstruction. Such a comparison allows us to quickly see those buildings that are missing or have errors in height or ground position.

However, a quantitative approach is also possible and is potentially more useful. Using stereo display techniques, we can generate a three-dimensional segmentation that allows us to represent the structure of each building in the scene. The form of the data is simply a segmentation description file containing collections of left image points and their relative height. From that representation we can infer a partial three-dimensional representation of the buildings, guessing the shape of the invisible parts, much as is done with simple wireframe models. We can also use this representation as a baseline reference representation for buildings in order to compare and contrast the various processing results. Figures 11a and 11b show how this technique can be used to construct a simple three-dimensional ground truth segmentation that can be visualized as a stereo scene.

FULLY AUTOMATIC SCENE REGISTRATION

As we have seen, structure matching, stereo matching, and visualization all rely on the quality of the stereo registration. The registration and the matching process are therefore interdependent. Structure matching appears to be a task that we can reliably perform even with a coarse scene registration. Further, the results of structure matching provide a method to automatically refine the initial coarse scene registration. In this section, we demonstrate a complete end-to-end scenario of automatic structure matching, fine registration, and stereo analysis. Thus, we can automatically generate a three-dimensional representation of the scene starting from the coarse scene registration provided by the CONCEPTMAP database.

We began with the DC38008 test area corresponding to Figures 1a and 1b. We then utilize the BABE structure results and per-

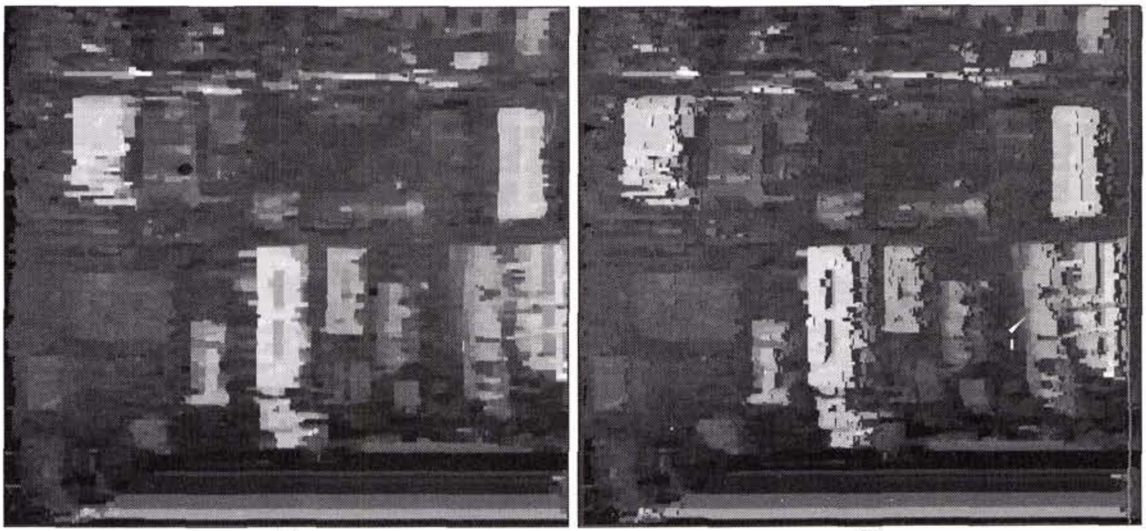

FIG. 10. (a) S2 stereo matching result (left). (b) S2 stereo matching result (right).

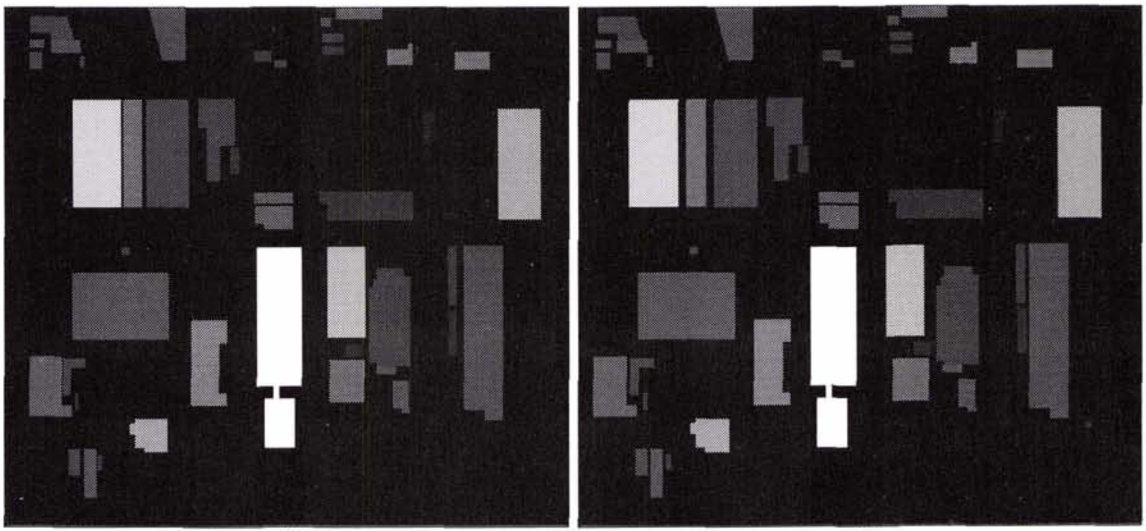

FIG. 11. (a) Ground truth left image for DC38008 scene. (b) Ground truth right image for DC38008 scene.

form structure matching to select reliable control points. The structures generated by BABE are fragmented and are not as consistent as those generated for the LAX stereo pair. Nevertheless, we are able to find a number of good matches, well distributed across the image, as shown in Figure 12a. Subjectively, the registration quality is good, as seen in Figure 12b, where many of the building fragments are now aligned. The overall registration quality is detailed in Table 1 (ISO structure). While it is not as accurate as the registration derived by manual ground control selection, it is clearly comparable.

Finally, using this automatically registered stereo pair, we performed stereo matching to get a dense disparity map of the scene. Figures 13a and 13b show the results for the S1 and the S2 matchers. The results are comparable to those in Figures 9a and 9b achieved using manual selection of control points for the fine registration. Thus, we have shown the feasibility of end-to-end processing to establish precise local registration using automatic ground control point estimation.

CONCLUSIONS

The importance of scene registration in the automated interpretation of aerial imagery can not be overstated. Scene registration is required for monocular matching, stereo analysis, scene visualization, accurate mensuration, and many other photo-interpretation tasks. Most work in computational stereo has ignored the problem of scene registration, assuming that the left/right image pairs were already in epipolar geometry. As we have seen, this may limit the utility of many feature-based and some area-based matching techniques, especially in cases where there are significant residual errors in the registration process.

Traditionally, we have separated the stereo analysis of digital images into two problems, registration and matching, and have attempted to solve each independently. However, the results of matching, whether structural or using a stereo model, are actually the ultimate form of scene registration because the matching solves the correspondence between different objects in the images. And while registration is generally necessary to

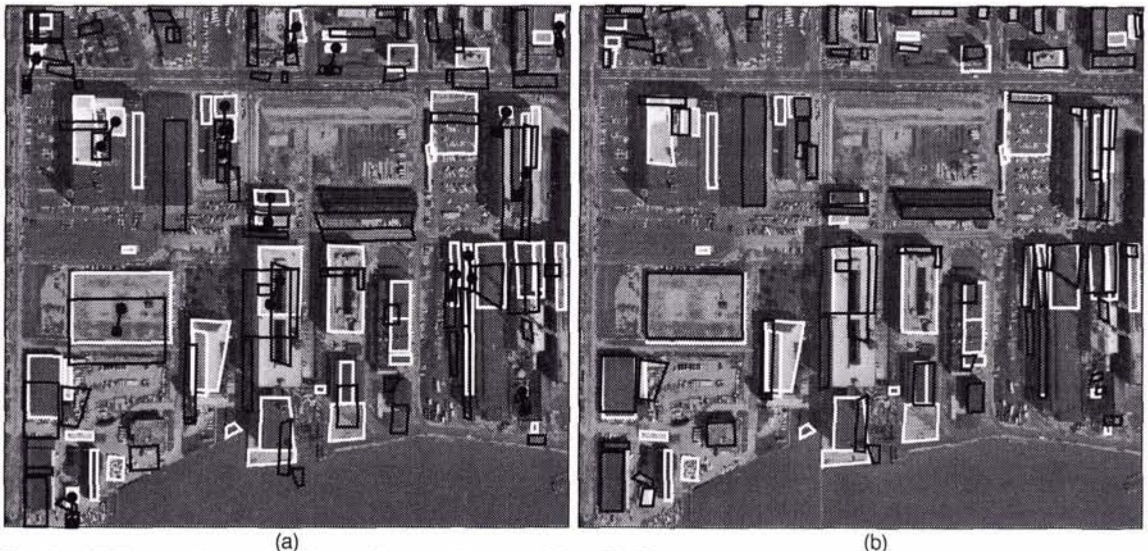

FIG. 12. (a) Automatic control points using structure matching. (b) Super-position of structures using structure matching registration.

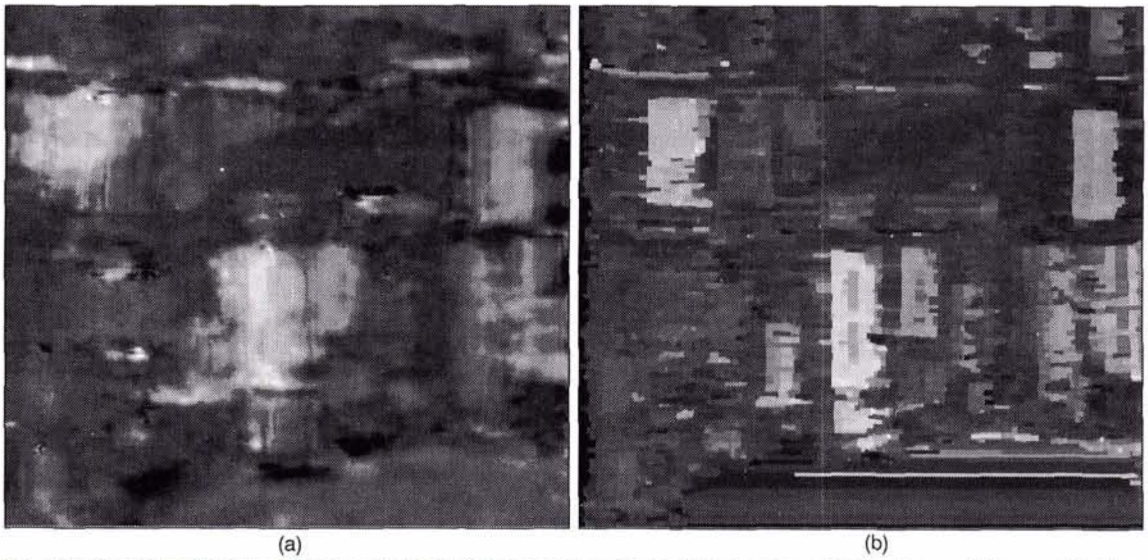

FIG. 13. (a) S1 disparity map using structure matching registration. (b) S2 disparity map using structure matching registration.

constrain search during matching, at least a sparse matching is necessary to perform the registration.

There are several areas for future work focused on improving techniques for scene registration. More research needs to be performed in the utilization of complex landmarks such as road networks for automated image-to-map scene registration (McGlone, 1989). For direct image-to-image correspondence, we have seen some limitations in automatic extraction of shadow corner points in complex urban imagery for registration. Additional sources of reliable registration points should be available using monocular extraction of man-made structures such as the road networks. From our previous work in road detection (Aviad and Carnine, 1988) and tracking (McKeown and Denlinger, 1988), it seems quite reasonable to use these structures as potential landmarks for scene matching.

ACKNOWLEDGMENTS

We thank our colleagues in the Digital Mapping Laboratory for an interesting and congenial working environment. We particularly acknowledge the contribution of Yuan Hsieh and Aviad Zlotnick to the development of several of the examples used in this paper. This research was primarily sponsored by the U.S. Army Engineer Topographic Laboratories under Contract DACA72-87-C-0001 and partially supported by the Defense Advanced Research Projects Agency, DoD, through DARPA order 4976, and monitored by the Air Force Avionics Laboratory Under Contract F33615-87-C-1499. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Engineer Topographic Laboratories, or the Defense Advanced Research Projects Agency, or of the United States Government.

REFERENCES

Arnold, R.D., 1978. Local Context in Matching Edges for Stereo vision. *Proceedings: DARPA Image Understanding Workshop*, May, pp. 777–791.

Aviad, Z., 1988. *Locating Corners in Noisy Curves by Delineating Imperfect*

Sequences. Technical Report CMU-CS-88-199, Carnegie-Mellon University, December.

Aviad, Z., and P. D. Carnine, 1988. Road Finding for Road Network Extraction. *Proceedings: Computer Vision and Pattern Recognition* , Ann Arbor, Michigan, June, pp. 814–819.

Aviad, Z., D. M. McKeown, Y. Hsieh. 1989. *The Generation of Building Hypotheses From Monocular Views*. Technical Report, Carnegie-Mellon University, January (to appear).

Barnard, S. T., and M. A. Fischler, 1982. Computational stereo. *Computing Surveys* 14(4):553–572.

Barnard, S. T., 1988. Stochastic stereo matching over scale. *Proceedings: DARPA Image Understanding Workshop*, April, pp. 769–778.

Brooks, R. A., A. M. Flynn, and T. Marill, 1988. Self calibration of motion and stereo vision for mobile robot navigation. *Proceedings: DARPA Image Understanding Workshop*, April, pp. 398–410.

Chen, L-H, and T. E. Boult, 1988. An integrated approach to stereo matching, surface reconstruction and depth segmentation using consistent smoothness assumptions. *Proceedings: DARPA Image Understanding Workshop*, April, pp. 166–176.

Faugeras, O. D., and G. Toscani, 1986. The Calibration Problem for Stereo. *Proceedings of Computer Vision and Pattern Recognition*, June, pp. 15–20.

M. J., Hannah, 1985. SRI's Baseline Stereo System. *Proceedings: DARPA Image Understanding Workshop*, December.

B. K. P., Horn, 1988. Relative Orientation. *Proceedings: DARPA Image Understanding Workshop*, April, pp. 826–837.

Huertas, A., and R. Nevatia. 1988. Detecting Buildings in Aerial Images. *Computer Vision, Graphics, and Image Processing* 41:131–152.

Irvin, R. B., and D. M. McKeown, 1989. Methods for Exploiting the Relationship between Buildings and their Shadow in Aerial Imagery. *SPIE Proceedings Image Understanding and the Man-Machine Interface II*. January. Also available as CMU Computer Science Technical Report CMU-CS-88-200.

Lucas, B. D. 1984. *Generalized Image Matching By The Method of Differences*. PhD thesis, Carnegie Mellon University, July.

McGlone, Chris. 1989. Automated image-map registration using active contour models and photogrammetric techniques. *SPIE Proceedings on Reconnaissance, Astronomy, Remote Sensing, and Photogrammetry*, January.

McKeown, D. M. 1984. Digital Cartography and Photo Interpretation from a Database Viewpoint. *New Applications of Databases*, G. Gargarin, and E. Golembe, (editor). Academic Press, New York, N.Y., pp. 19–42.

———, 1987. The Role of Artificial Intelligence in the Integration of Remotely Sensed Data with Geographic Information Systems. *IEEE Transactions on Geoscience and Remote Sensing* GE-25(3):330–348.

McKeown, D. M., C. A. McVay, and B. D. Lucas, 1986. Stereo Verification In Aerial Image Analysis. *Optical Engineering* 25(3):333–346.

McKeown, D. M., and J. L. Denlinger, 1988. Cooperative Methods for Road Tracking in Aerial Imagery. *Proceedings IEEE Computer Vision and Pattern Recognition Conference*, June, pp. 662–672.

Moravec, H. P., 1980. *Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover*. PhD thesis, Stanford University, September.

Nasrabadi, N. M., Y. Liu, and J-L. Chiang, 1988. Stereo vision correspondence using a multi-channel graph matching technique. *IEEE international Conference on Robotics and Automation*, April.

Mohan, R., and R. Nevatia, 1988. Perceptual grouping for the detection and description of structures in aerial images. *Proceedings: DARPA Image Understanding Workshop*, April, pp 512–526.

Ohta, Y., and T. Kanade, 1985. Stereo by Intra- and Inter-scanline Search using Dynamic Programming. *IEEE Transactions* PAMI-7(2):139–154.

Price, K. E., 1976. *Change Detection and Analysis in Multi-Spectral Images*. PhD thesis, Carnegie-Mellon University, December.

———, 1985. Relaxation Matching Techniques – A Comparison. *IEEE Transactions* PAMI-7(5):617–623.

Price, K., and D. R. Reddy, 1979. Matching Segments of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1(1):110–116.

Weinshall, D., 1988. Qualitative vs. quantitative depth and shape from stereo. *Proceedings: DARPA Image Understanding Workshop*, April, pp. 779–785.