Fusion of Monocular Cues to Detect Man-Made Structures in Aerial Imagery

Jefferey A. Shufelt
David M. McKeown, Jr.

September 27, 1990
CMU-CS-90-194 2

This report is a revised and extended version of a paper presented at the *IAPR Workshop on Multisource Data Integration in Remote Sensing*, College Park, MD, June 14-15, 1990.

Digital Mapping Laboratory
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

This research was primarily sponsored by the U.S. Army Engineering 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 Engineering Topographic Laboratories, or the Defense Advanced Research Projects Agency, or of the United States Government.

Keywords: Cartographic feature extraction, computer vision, aerial image interpretation, information fusion, building detection, monocular image analysis

Table of Contents

List of Figures

List of Tables

Abstract

The detection and delineation of man-made structures from aerial imagery is a complex computer vision problem. It requires locating regions in imagery that possess properties distinguishing them as man-made objects in the scene, as opposed to naturally occurring terrain features. The building extraction process requires techniques that exploit knowledge about the structure of man-made objects. Techniques do exist that take advantage of this knowledge; various methods use edge-line analysis, shadow analysis, and stereo imagery analysis to produce building hypotheses. It is reasonable, however, to assume that no single detection method will correctly delineate or verify buildings in every scene. As an example, a feature extraction system that relies on the analysis of cast shadows to predict building locations is likely to fail in cases where the sun is directly above the scene.

In this paper we introduce a cooperative-methods paradigm for information fusion that is shown to be highly effective in improving the system performance over that achieved by individual building extraction methods. Using this paradigm, each extraction technique provides information that can be added or assimilated into an overall interpretation of the scene. Thus, our research focus is to explore the development of a computer vision system that integrates the results of various scene analysis techniques into an accurate and robust interpretation of the underlying three-dimensional scene.

We briefly survey four monocular building extraction, verification, and clustering systems that form the basis for the research described here. A method for fusing the symbolic data generated by these systems is described, and it is applied to both monocular image and stereo image data sets. A set of performance evaluation metrics are developed, described, and applied to the fusion results. Several detailed analyses are presented, as well as a summary of results on 23 monocular and 5 stereo scenes. These experiments show that a significant improvement in building detection is achieved using these techniques.

1. Introduction

The detection and delineation of man-made structures from aerial imagery is a complex computer vision problem [10]. It requires locating regions in imagery that possess properties distinguishing them as man-made objects in the scene, as opposed to naturally occurring terrain features. The building extraction process requires techniques that exploit knowledge about the structure of man-made objects. Techniques do exist that take advantage of this knowledge; various methods use edge-line analysis, shadow analysis, and stereo imagery analysis to produce building hypotheses. It is reasonable, however, to assume that no single detection method will correctly delineate or verify buildings in every scene. As an example, a feature extraction system that relies on the analysis of cast shadows to predict building locations is likely to fail in cases where the sun is directly above the scene.

In this paper we introduce a cooperative-methods paradigm for information fusion that is shown to be highly effective in improving the system performance over that achieved by individual building extraction methods. Using this paradigm, each extraction technique provides information that can be added or assimilated into an overall interpretation of the scene. Thus, our research focus is to explore the development of a computer vision system that integrates the results of various scene analysis techniques into an accurate and robust interpretation of the underlying three-dimensional scene.

In the cooperative-methods paradigm we assume that no single scene analysis method can provide a complete set of building hypotheses for a scene. Each method, however, may provide a subset of the information necessary to produce a more meaningful interpretation of the scene. For instance, a shadow-based method might provide unique information in situations where ground and roof intensity are similar. An intensity-based method can provide boundary information in instances where shadows were weak or nonexistent, or in situations where structure height was sufficiently low that stereo disparity analysis would not provide reliable information. The implicit assumption behind this paradigm is that the symbolic interpretations produced by each of these techniques can be integrated into a more meaningful collection of building hypotheses.

It is reasonable to expect that there will be complications in fusing real monocular data. In the best case, the building hypotheses will not only be accurate, but complementary. It is just as likely, however, that some building hypotheses may be unique. Further, it is rare that building hypotheses are always accurate, or even mutually supportive of one another. For a cooperative-methods data fusion system to be successful, it must address the problems of redundant and conflicting data.

1.1. Previous work

There are many interesting building detection and extraction techniques in the contemporary literature. We briefly mention some recently developed methods, to illustrate the variety of techniques that produce building hypothesis information. Each of these techniques is one possible source of building segmentation. None of this previous work, to the best of our knowledge, addresses the problem of hypothesis fusion across multiple feature extraction systems.

Fua and Hanson [3] described a system that used generic geometric models and noise-tolerant geometry parsing rules to allow semantic information to interact with low-level geometric information, producing segmentations of objects in the aerial image. The system used region-based segmentations as input, and applied the geometry rules to connect simple image tokens such as edges into more complex rectilinear structures.

Nicolin and Gabler [12] described a system for analysis of aerial images. The system had four components: a method-base of domain-independent processing techniques, a long-term memory containing *a priori* knowledge about the problem domain, a short-term memory containing intermediate results from the image analysis process, and a control module responsible for invocation of the various processing techniques. Gray-level analysis was applied to a resolution pyramid of imagery to suggest segmentation techniques, and structural analysis was performed after segmentation to provide geometric interpretations of the image. These interpretations were then given confidence values based on their similarity to known image features such as roads and houses.

Mohan and Nevatia [11] present a method by which simple image tokens such as lines or edges could be clustered into more complex geometric features consisting of parallelopipeds. They used constraint-satisfaction networks to decide which features were mutually supportive and which features subsumed or eliminated other features. They also applied set operations to the segments of features to merge pairs of features.

Huertas and Nevatia [7] discuss a technique for detecting buildings in aerial images. Their method detected lines and corners in an image and labeled these corners based on detected shadows. Then, object boundaries were traced by grouping corners that shared line segments. The position and orientation of these chains of segments were then examined, and the appropriately aligned chains were connected to form boxes representing the structures in the image. Shadow analysis was used to verify the remaining chains by adding lines as necessary.

The key contribution of our work is a demonstration of the effectiveness of simple information fusion techniques as applied to the problem of building detection in complex aerial imagery. These techniques significantly improve performance, compared to any of the component feature extraction systems. We demonstrate this by using several building analysis systems, each of which uses a different image domain cue to generate and evaluate building hypotheses.

1.2. Building extraction techniques

For the experiments described in this paper, a set of four monocular building detection and evaluation systems were used. Three of these were shadow-based systems; the fourth was line-corner based. The shadow based systems are described more fully by Irvin and McKeown [8], and the line-corner system is described by Aviad, McKeown, and Hsieh [2]. A brief description of each of the four detection and evaluation systems follows.

BABE (Builtup Area Building Extraction) is a building detection system based on a line-corner analysis method. BABE starts with intensity edges for an image, and examines the proximity and angles between edges to produce corners. To recover the structures represented by the corners, BABE constructs chains of corners such that the direction of rotation along a chain is either clockwise or counterclockwise, but not both. Since these chains may not necessarily form closed segmentations, BABE generates building hypotheses by forming boxes out of the individual lines that comprise a chain. These boxes are then evaluated in terms of size and line intensity constraints, and the best boxes for each chain are kept, subject to shadow intensity constraints similar to those proposed by Nicolin [12] and Huertas [7],

SHADE (SHAdow DEtection) is a building detection system based on a shadow analysis method. SHADE uses the shadow intensity computed by BABE as a threshold for an image. Connected region extraction techniques are applied to produce segmentations of those regions with intensities below the threshold, i.e., the shadow regions. SHADE then examines the edges comprising shadow regions, and keeps those edges that are adjacent to the buildings casting the shadows. These edges are then broken into nearly straight line segments by the use of an imperfect sequence finder [1]. Those line segments that form nearly right-angled corners are joined, and the corners that are concave with respect to the sun are extended into parallelograms, SHADE's final building hypotheses.

SHAVE (SHAdow VErification) is a system for verification of building hypotheses by shadow analysis. SHAVE takes as input a set of building hypotheses, an associated image, and a shadow threshold produced by BABE. SHAVE begins by determining which sides of the hypothesized building boxes could possibly cast shadows, given the sun illumination angle, and then performs a walk away from the sun illumination angle for every pixel along a building/shadow edge to delineate the shadow. The edge is then scored based on a measure of the variance of the length of the shadow walks for that edge. These scores can then be examined to estimate the likelihood that a building hypothesis corresponds to a building, based on the extent to which it casts shadows.

GROUPER is a system designed to cluster, or group, fragmented building hypotheses, by examining their relationships to possible building/shadow edges. GROUPER starts with a set of hypotheses and the building/shadow edges produced by BABE. GROUPER back-projects the endpoints of a building/shadow edge towards the sun along the sun illumination angle, and then connects these projected endpoints to form a region of interest in which buildings might occur. GROUPER intersects each building hypothesis with these regions of interest. If the degree of overlap is sufficiently high (the criteria is currently 75% overlap), then the hypothesis is assumed to be a part of the structure which is casting the building/shadow edge. All hypotheses that intersect a single region of interest are grouped together to form a single building cluster.

These four building extraction systems, each with particular strengths and weaknesses, provide an interesting set of feature extraction primitives. Their individual performance is, we believe, typical of the current state-of-the-art in automated building extraction. They are mature systems whose performance is not likely to improve significantly and therefore provide a 'best effort' comparison against which fusion results can be compared.

2. Building hypothesis fusion using monocular imagery

Building hypotheses generated from monocular imagery typically take the form of two-dimensional polygonal boundary descriptions. One can imagine "stacking" sets of these polygonal boundary descriptions on the image: in the process, those regions of the image that represent man-made structure in the scene should accumulate more building hypotheses than those regions of the image that represent natural features in the scene. The merging technique developed here exploits this idea.

Figure 2-1: DC37405 image with ground-truth segmentation superimposed

The basic fusion method takes as input an arbitrary collection of polygons. An image is created that is sufficiently large to contain all of the polygons, and each pixel in this image is initialized to zero. Each polygon is scan-converted into the image, and each pixel touched during the scan is incremented. The resulting image then has the property that the value of each pixel in the image is the number of input polygons that cover it. Segmentations can then be generated from this "accumulator" image by applying connected region extraction techniques. If the image is thresholded at a value of 1 (i.e, all non-zero pixels are kept), the regions produced by a connected region extraction algorithm will simply be the geometric unions of the input polygons. It is the case, however, that the image could be thresholded at higher values. We motivate thresholding experiments in Section 4.

There are several variations to the basic hypothesis fusion technique:

1. Fusion of hypotheses generated by a single feature extraction method on a monocular image.
2. Fusion of hypotheses generated by multiple feature extraction methods on a monocular image.
3. Fusion of hypotheses generated by multiple feature extraction methods across a stereo image pair.
4. Fusion of hypotheses generated by multiple feature extraction methods on multiple views taken over time.

The careful reader may notice that two variations are missing from this list; namely, the fusion of hypotheses generated by a single feature extraction method across a stereo image pair, and on multiple views taken over time. These are simply special cases of fusion on multiple feature extraction methods, and do not merit separate treatment. We describe the application of the first three fusion variations as applied to the results of four building detection and evaluation systems (BABE, SHADE, SHAVE, and GROUPER). The first two variations, primarily monocular, are described in the following section. Experiments on the third variation, stereo fusion, are described in Section 3, along with a brief discussion of the fourth variation, multi-temporal fusion.

2.1. Fusion of hypotheses from a single view

Figure 2-1 shows a section of a suburban house scene in the Washington, D.C. area. This scene is quite complex; it contains a wide variety of buildings ranging from small individual houses and townhouses to large apartment buildings. There are a variety of roof shapes including pitched and flat roofs, and the roof colors vary due to surface materials with different reflectance properties. Simple intensity-based or shape-based techniques have significant difficulty with such scenes. We use this scene throughout our discussion of monocular hypothesis fusion.

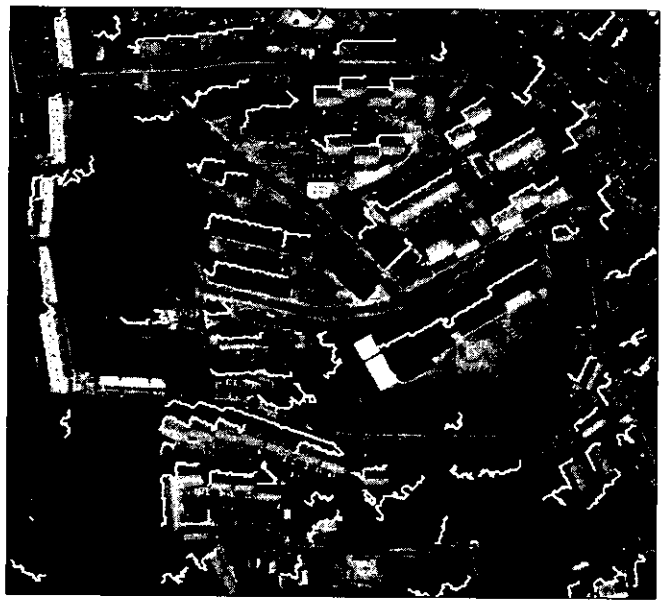

Figure 2-2: DC37405 Shadow/Building Edges

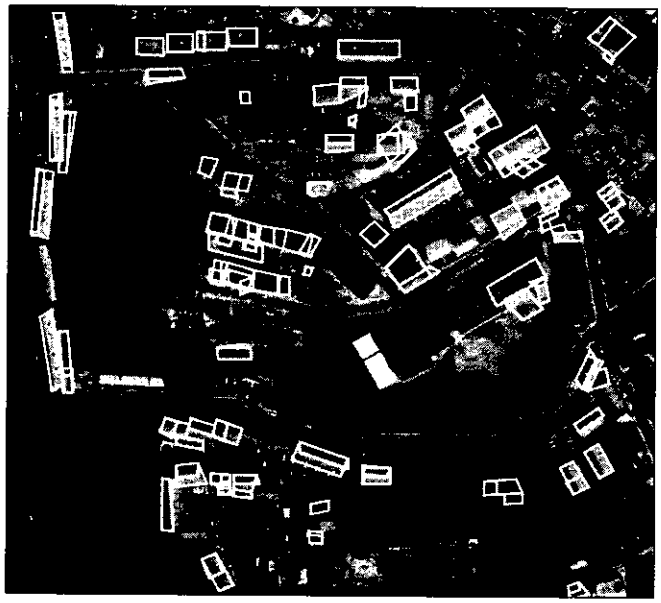

Figure 2-3: DC37405 Grouper Regions

There are two variations on hypothesis fusion using a single monocular image. The first involves the creation of a single hypothesis out of a collection of fragmented hypotheses believed to correspond to a single man-made structure. This problem was addressed by applying the scan-conversion technique to the fragmented clusters produced by GROUPER. Figure 2-2 shows the shadow/building edges generated by SHADE, which are used by GROUPER to select a subset of the building hypotheses produced by BABE that are consistent with buildings casting shadows along each edge. The result of this process is shown

in Figure 2-3, where each shadow/building edge has been used to select and cluster sets of building hypotheses that exhibited a strong relationship with each edge. The scan-conversion technique was applied to each cluster individually, and the resulting accumulator image was thresholded at 1. Connected region extraction techniques were then applied to provide the geometric union of each cluster. These clusters were then used as the building hypotheses produced by GROUPER as shown in Figure 2-6.

The second variation involves the fusion of each of the monocular hypothesis sets created by BABE, SHADE, SHAVE, and that created by fusion of the GROUPER hypotheses, into a single set of hypotheses for the scene. Again, the scan-conversion technique was applied. The four hypothesis sets were scan-converted into a single accumulator image, which was thresholded at a value of 1. Connected region extraction techniques were applied to produce the final segmentation for the image.

Figure 2-4 shows the SHADE results for DC37405, the suburban house scene. Figure 2-5 shows the SHAVE results, Figure 2-6 shows the GROUPER results, and Figure 2-7 shows the BABE results. Figure 2-8 shows the fusion of these four monocular hypothesis sets. Close inspection of each of the four figures indicates that each method produces building hypotheses that are (in most cases) complementary and tend to be mutually supportive, but there exist situations in which only one method arrives at a correct or partially correct building hypothesis. In the following section we discuss techniques for evaluating the performance of the hypothesis merging technique, and, as a side effect, the performance of each of the building hypothesis methods.

2.2. An evaluation of hypothesis fusion

To judge the correctness of an interpretation of a scene, it is desirable to have some mechanism for quantitatively evaluating that interpretation. Unfortunately, there is very little current work described in the computer vision literature that addresses this topic. Our approach is to compare a given set of building hypotheses against a set that is known to be correct, and analyze the differences between the given set of hypotheses and the correct ones. In performing evaluations of the fusion results, we use *ground-truth segmentations* as the correct detection results for a scene. Ground-truth segmentations are manually produced segmentations of the buildings in an image. Figure 2-1 shows the superposition of the manual ground-truth segmentation on the suburban house scene.

There exist two simple criteria for measuring the degree of similarity between a building hypothesis and a ground-truth building segmentation: the mutual area of overlap and the difference in orientation. A correct building hypothesis and the corresponding ground-truth segmentation region should cover roughly the same area, and should have roughly the same alignment with respect to the image. A scoring function can be developed that incorporates these criteria. A region matching scheme such as this, however, suffers from the fact that multiple buildings in the scene are segmented by a single region in the hypothesis set. In these cases, the building hypothesis will have low matching scores with each of the buildings it contains, due to the differences in overlap area.

A simpler coverage-based global evaluation method was developed. This evaluation method works in the following manner. H, a set of building hypotheses for an image, and G, a ground-truth segmentation of that image, are given. The image is then scanned, pixel by pixel. For any pixel P in the image, there are four possibilities:

1. Neither a region in H nor a region in G covers P. This is interpreted to mean that the system producing H correctly denoted P as being part of the background, or natural structure, of the scene.

2. No region in H covers P, but a region in G covers P. This is interpreted to mean that the system producing H did not recognize P as being part of a man-made structure in the scene. In this case, the pixel is referred to as a "false negative".

3. A region (or regions) in H cover P, but no region in G covers P. This is interpreted to mean that the system producing H incorrectly denoted P as belonging to some man-made structure, when it is in fact part of the scene's background. In this case, the pixel is referred to as a "false positive".

4. A region (or regions) in H and a region in G both cover P. This is interpreted to mean that the system producing H correctly denoted P as belonging to a man-made structure in the scene.

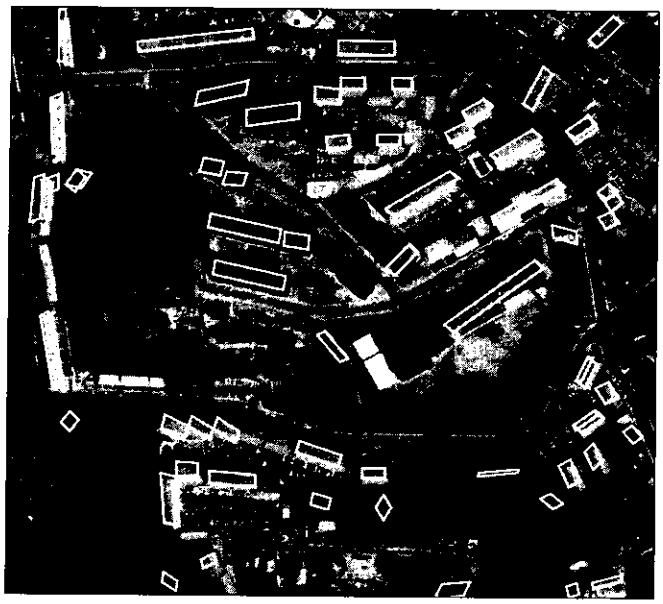

Figure 2-4: DC37405 SHADE results

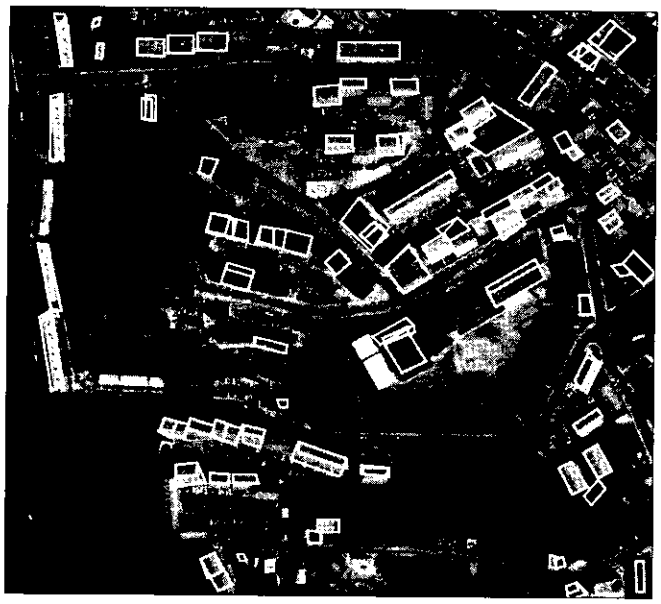

Figure 2-5: DC37405 SHAVE results

Figure 2-6: DC37405 GROUPER results

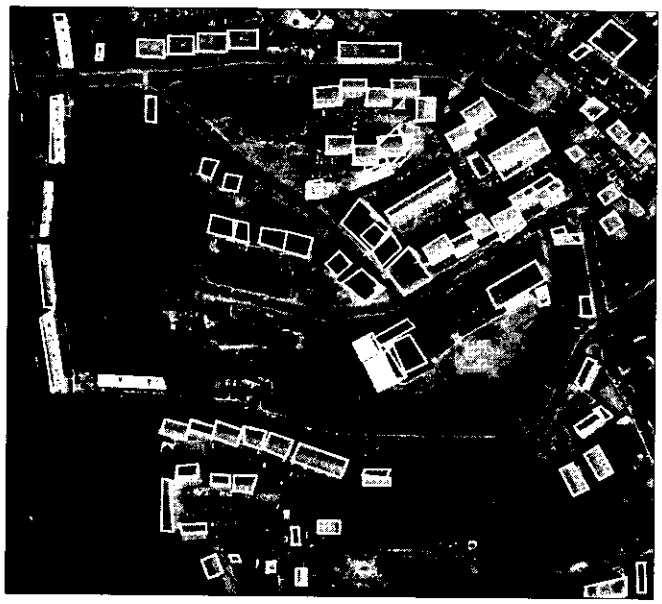

Figure 2-7: DC37405 BABE results

By counting the number of pixels that fall into each of these four categories, we may obtain measurements of the percentage of building hypotheses that were successful (and unsuccessful) in denoting pixels as belonging to man-made structure, and the percentage of the background of the scene that was correctly (and incorrectly) labeled as such. Further, we may use these measurements to define a *building pixel branching factor*, which will represent the degree to which a building detection system overclassifies background pixels as building pixels in the process of generating building hypotheses. The building pixel branching factor is defined as the number of false positive pixels divided by the number of correctly detected building pixels.

2.3. Results and analysis

Table 2-1 gives the performance statistics for monocular building fusion as applied to the suburban house scene in DC37405 shown in Figure 2-8. The first column represents one of the building extraction systems. The next two columns give the percentage of building and background terrain correctly identified as such. The fourth and fifth columns show incorrect identification percentages for buildings and terrain. The next two columns give the breakdown (in percentages) of incorrect pixels in terms of false positives and false negatives. The last column gives the building pixel branching factor.

Examining the results for each extraction method individually, we note that BABE exhibits the best performance. This is not surprising, since the image domain cues that BABE utilizes (lines and corners) are relatively easy to detect in the DC37405 image. BABE also performs its own internal verification step to prune away building hypotheses that do not satisfy its own requirements for shadow support. Thus, BABE presents only those hypotheses in which it has high confidence as its final result. Of the four systems, SHADE is the least effective in terms of building detection; however, it also generates the fewest number of false positive pixels, which is a desirable property.

GROUPER and SHAVE both operate on *all* of the hypotheses produced by BABE, not just those hypotheses that have passed BABE's conservative shadow evaluation; and each produce quantitatively similar results. It is worth noting that the building pixel branching factor for these systems is higher than in BABE or SHADE; this is due to the fact that both GROUPER and SHAVE are required to verify a larger number of hypotheses that are, in fact, incorrect. This has a more dramatic effect on the number of false positive pixels than erroneous line placement errors typically encountered in BABE or SHAVE.

In this case, by performing monocular fusion, we are able to improve the building detection percentage from the best extraction result of 58% (due to BABE) to 77% for the fused results. This implies that the extraction systems as a whole provide more information about building structure than any individual system. We also note, however, that erroneous information accumulates as well. The building pixel branching factor indicates that for every pixel correctly hypothesized to belong building structure, over 0.6 pixels are incorrectly hypothesized as such. Just as each individual system can provide unique information about the presence of man-made objects in a scene, each individual system may also fail in a unique way under the absence of relevant image domain cues.

We believe that the quantitative results generated by the new evaluation method accurately reflect the subjective visual quality of the set of building hypotheses, when taken as a relative measure. Further, the building pixel branching factor provides a rough estimate of the amount of noise generated in the fusion process. Judging by these measures, we note that the final results of the hypothesis fusion process significantly improve the detection of buildings in a scene.

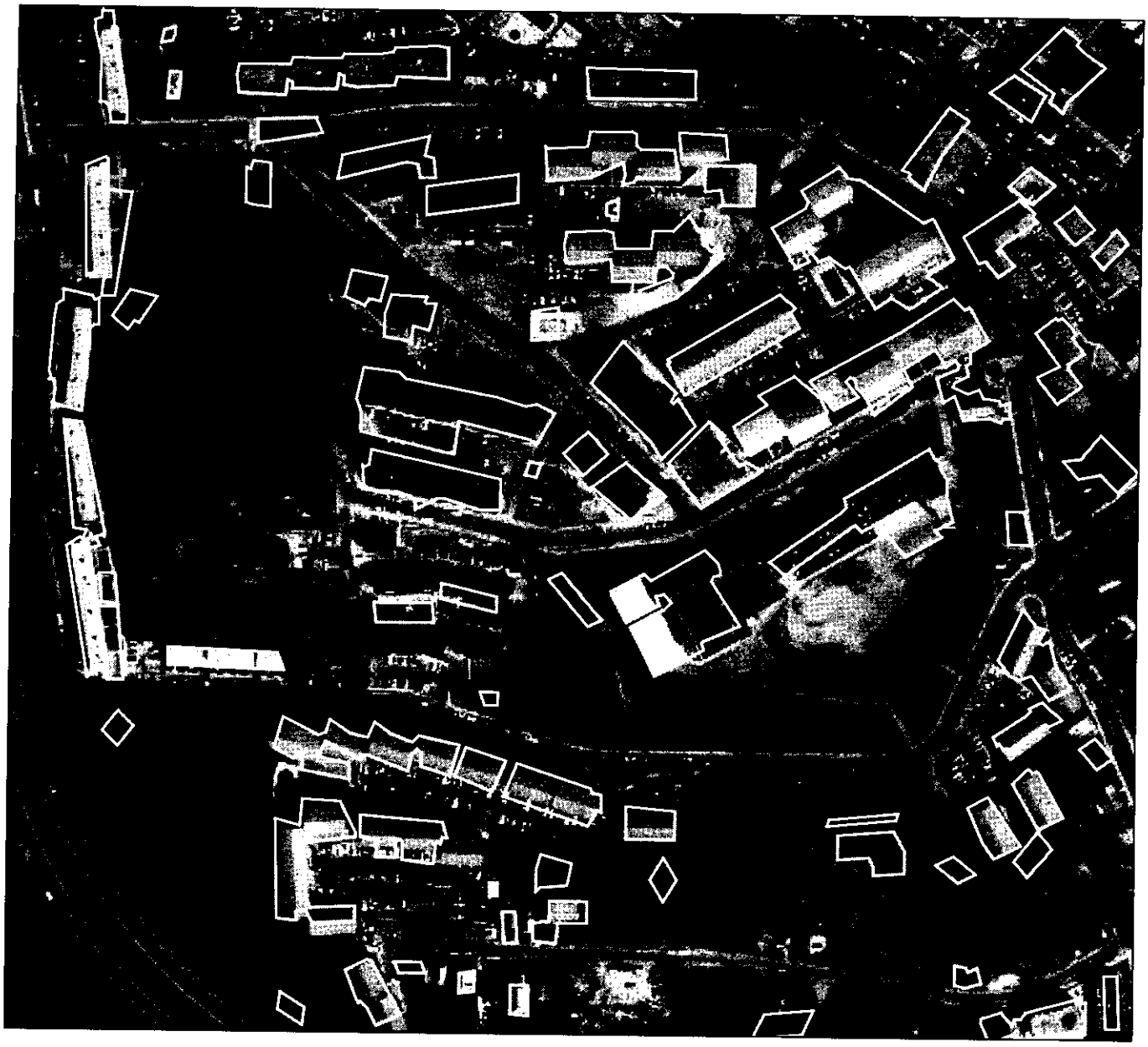

Figure 2-8: Monocular hypothesis fusion for DC37405

Evaluation results for the fusion process on DC37							
System	% Bld Detected	% Bkgd Detected	% Bld Missed	% Bkgd Missed	% False Pos.	% False Neg.	Br Factor
SHADE	37.5	98.2	62.5	1.8	15.0	85.0	0.294
SHAVE	47.2	96.8	52.8	3.2	26.8	73.2	0.408
GROUPER	48.7	95.8	51.3	4.2	32.6	67.4	0.508
BABE	58.9	97.2	41.1	2.8	28.5	71.5	0.278
FUSION	77.7	92.0	22.3	8.0	68.0	32.0	0.611
99 regions in ground truth							

Table 2-1: Evaluation statistics for DC37 hypothesis fusion

3. Building hypothesis fusion using stereo imagery

In many cases, automated feature extraction systems may have multiple views of a scene available for analysis. As discussed in Section 2, there are two variations of information fusion on multiple views; the use of stereo coverage in an image pair, and the use of images acquired over time of a particular geographic area. In the case of multi-temporal acquisition, the viewing geometry may not generate a stereo pair; monocular feature extraction, however, can be employed on each image in the multi-temporal dataset. In both cases, an image-to-image correspondence must be established, preferably by the use of a camera model.

In this section we describe experiments utilizing stereo imagery to perform hypothesis fusion. We suggest that multi-temporal fusion could be performed in a similar way, except that the adjustments due to disparity (discussed in Section 3.1) could not be accomplished. Thus, the multi-temporal case is exactly the same as the stereo case *with* image-to-image registration, but *without* hypothesis position adjustment by the use of stereo disparity estimates. In this section we describe the fusion technique for the case where stereo imagery is available.

Given a stereo pair of a scene, each of the building detection systems can be run on both the left and right images, to produce a set of hypotheses for each image. Since the images will be representations of the scene from different perspectives, and thus will have slightly different geometric features and intensities, the systems should produce slightly different results. Combining the left and right results for a particular system should provide a slightly more complete hypothesis set for a scene, due to these differences.

Since the left results and right results might lie in different coordinate frames, the first step was that of placing both sets of hypotheses in the same coordinate system. Control points were manually selected for the left and right images, and a polynomial-based registration method was then applied to bring points in the right coordinate frame to the left coordinate frame [13]. Then, the scan-conversion technique was applied to the hypothesis pair (now in the same coordinate frame), and the resulting accumulator image was thresholded at 1 and segmented to produce the fused hypothesis set for a single building system.

3.1. Disparity effects on stereo mergers

As part of the overall building hypothesis fusion process, stereo pairs of building hypotheses are fused to provide a single set of hypotheses for a monocular view of the scene. As described earlier, a polynomial-based registration method was applied to bring regions from the right image's coordinate frame to the left image's coordinate frame. This procedure, however, does not take into account the disparity between the left image and the right image, which can cause the translated regions to suffer from displacement errors along the scanline. Since the translated regions may not be accurately located, the fused hypotheses are likely to cover extraneous pixels in the image, and the overall detection rate will decrease. As mentioned earlier, this is the case of fusion for multi-temporal imagery.

To account for the disparity shift, a simple method was used to improve the location of regions translated from one coordinate frame to another. Given a stereo pair of images, a sparse disparity map was produced by S2, a feature-based hierarchical scanline matching system [5, 6]. Step interpolation was used to produce dense disparity maps from the sparse maps, and a vertical median filter algorithm was applied to smooth the dense maps.

Once a smoothed dense disparity map is obtained, it is then possible to compute the disparity shift for a particular building hypothesis, by calculating the average disparity inside the hypothesized region. This average disparity value is then used to shift the region along the scanline. Assuming that the disparity map is relatively good, this procedure will shift the region to match it with the corresponding building in the image. Figure 3-1 shows the smooth dense disparity map for the DC37405 image, and Figure 3-2 shows the BABE right image results registered into the left image coordinate frame, before

Figure 3-1: S2 smoothed dense disparity map for the DC37405 stereo pair

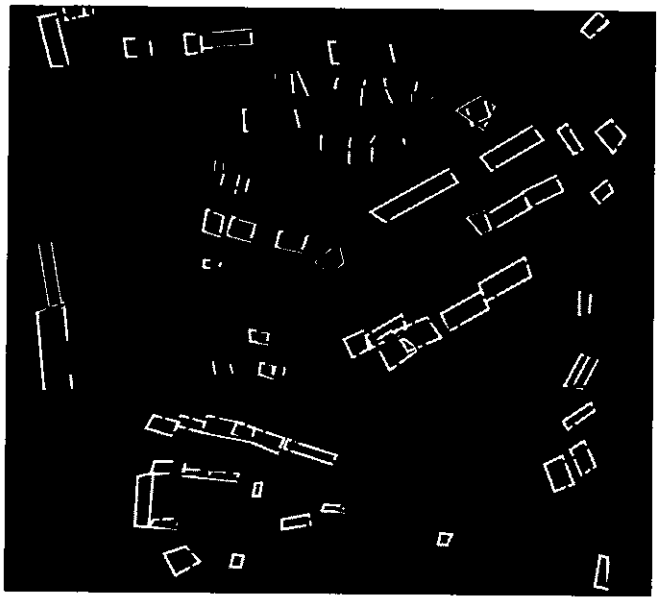

Figure 3-2: DC37405 BABE registered results, before and after disparity shift

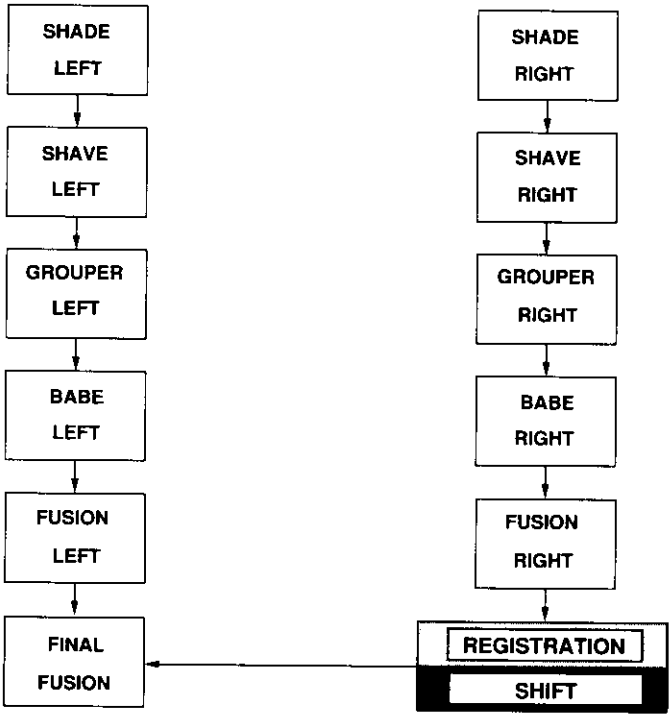

Figure 3-3: Left-right Fusion

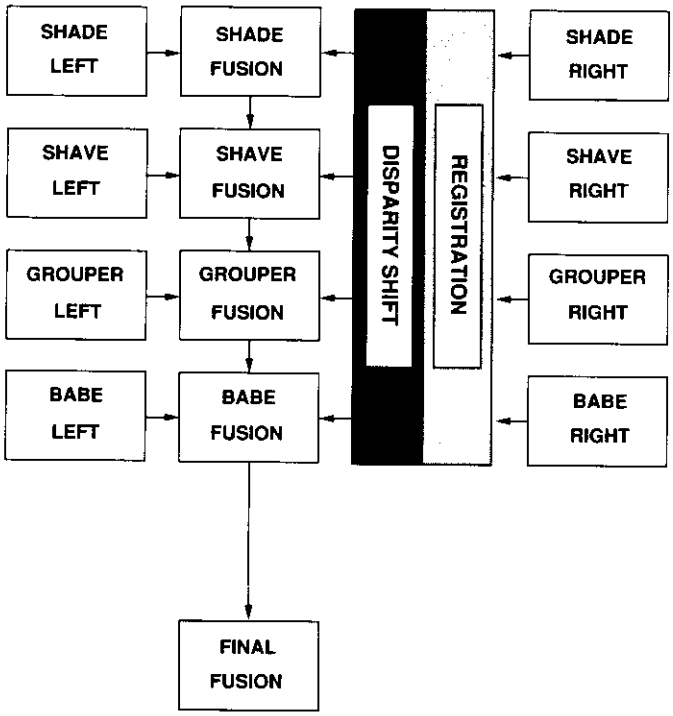

Figure 3-4: Extraction-based Fusion

and after the disparity shifting process. The registered results appear in white, and the shifted results appear in black.

3.2. Stereo fusion experiments

Given the stereo fusion technique described in the previous sections, we can construct two basic processing models for merging building hypotheses. In the first model, which we call *left-right fusion*, all hypotheses for the left image are fused, and all hypotheses for the right image are fused. Then, the stereo fusion technique described in the previous section is applied to fuse the right monocular merger

with its counterpart in the left image. Figure 3-3 gives a pictorial representation of this processing model.

An alternative model, which we call *extraction-based fusion*, applies stereo fusion to the results of each building extraction system, and then performs monocular fusion on these stereo mergers to produce a final result. Figure 3-4 gives a pictorial representation of this processing model.

Evaluation results for DC37 system left/right fusion							
System	% Bld Detected	% Bkgd Detected	% Bld Missed	% Bkgd Missed	% False Pos.	% False Neg.	Br Factor
SHADE	37.5	98.2	62.5	1.8	15.0	85.0	0.294
REG	16.8	97.9	83.2	2.1	13.1	86.9	0.749
SHIFT	20.4	98.4	79.6	1.6	10.6	89.4	0.464
MERGER	39.5	97.0	60.5	3.0	22.9	77.1	0.456
SHAVE	47.2	96.8	52.8	3.2	26.8	73.2	0.408
REG	40.4	94.8	59.6	5.2	34.1	65.9	0.762
SHIFT	44.8	95.5	55.2	4.5	32.4	67.6	0.591
MERGER	65.0	93.3	35.0	6.7	53.1	46.9	0.610
GROUP	48.7	95.8	51.3	4.2	32.6	67.4	0.508
REG	29.0	94.9	71.0	5.1	29.8	70.2	1.037
SHIFT	31.6	95.4	68.4	4.6	28.7	71.3	0.871
MERGER	56.0	92.3	44.0	7.7	51.0	49.0	0.819
BABE	58.9	97.2	41.1	2.8	28.5	71.5	0.278
REG	42.5	95.9	57.5	4.1	29.8	70.2	0.575
SHIFT	51.4	97.4	48.6	2.6	24.4	75.6	0.305
MERGER	74.0	95.1	26.0	4.9	52.8	47.2	0.393
FINAL	86.6	84.6	13.4	15.4	87.2	12.8	1.053
99 regions in ground truth							

Table 3-1: Evaluation statistics for DC37 system fusions

At first glance, one might expect the final results of these processing models to be exactly the same. This would certainly be the case if the building extraction systems produced error-free hypothesis sets, and if the stereo matching algorithms produced perfect disparity maps. In practice, this is not the case, and there will be slight differences between the results. To understand the source of the divergence, recall the stereo fusion algorithm described in the previous section.

In the stereo fusion algorithm, building disparity is taken into account by computing the average disparity inside each polygonal boundary description, and shifting each boundary description along the scanline accordingly. The regions obtained by extraction-based stereo fusion will delineate different areas, and thus have different disparity values, than the areas obtained by left-right stereo fusion; hence, the building hypotheses will be shifted by differing values along the scanline. In practice, these differences are small, since the two types of stereo hypotheses tend to delineate approximately the same regions, and thus have similar disparity values.

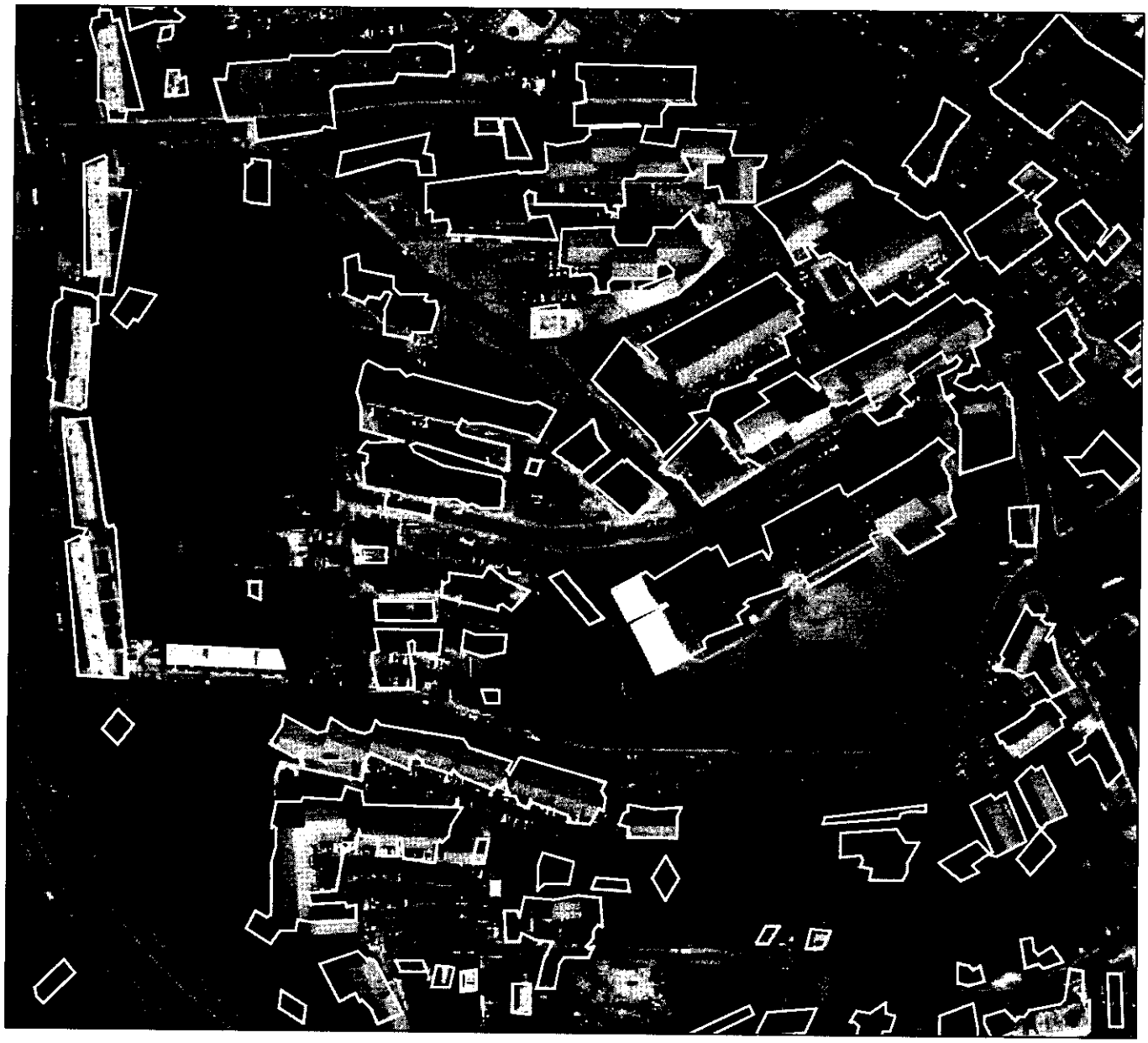

Figure 3-5: Stereo hypothesis fusion for DC37405

Evaluation results for stereo fusion on DC37							
System	% Bld Detected	% Bkgd Detected	% Bld Missed	% Bkgd Missed	% False Pos.	% False Neg.	Br Factor
LEFT	77.7	92.0	22.3	8.0	68.0	32.0	0.611
RIGHT	68.0	90.1	32.0	9.9	67.2	32.8	0.962
REG	58.8	88.6	41.2	11.4	62.2	37.8	1.150
SHIFT	65.5	89.8	34.5	10.2	63.7	36.3	0.927
FUSION	86.4	84.7	13.6	15.3	86.9	13.1	1.049

Table 3-2: Evaluation statistics for DC37 stereo fusion

Table 3-1 gives statistics for extraction-based fusion on the DC37405 scene. Each column gives the same statistics as in previous tables, but the first column bears additional explanation. The rows beginning with boldface names represent the raw results from each of the four building extraction and verification techniques on the left image of the stereo pair. Rows prefaced by REG represent the right image results after registration into the left image coordinate frame. Rows prefaced by SHIFT represent the registered results after shifting due to disparity. Rows prefaced by MERGER represent the results of fusing the left image results with the registered and shifted right image results. The final row of the table gives the results of the final fusion of all four system merger results.

Analyzing these results, we first note that the disparity shifting process provides improved results in all cases, in terms of building detection rate and building pixel branching factor. We also note, however, that the results for the registered and shifted right results are uniformly worse than the corresponding results for the left image. In this case, the decline in performance can be attributed to the fact that the right image of the stereo pair had fewer image domain cues (such as shadow corners and intensity edges) than the left image. In other stereo fusion experiments, the left and right results were comparable in quality.

We further note that the stereo fusion for each system provides a better result in terms of building detection rate than either of its component results, and we also observe that the final fusion provides a better result (again in terms of building detection rate) than any of the component system fusions. It should be noted, however, that the building pixel branching factor has increased as well, indicating that errors in each of the individual hypothesis sets have accumulated in the final result.

Figure 3-5 shows the results of left-right image fusion on the DC37405 scene. Table 3-2 gives the statistics for the left-right image fusion. Again, the row headings bear explanation. The first two rows give the results for monocular fusion of the extraction results on the LEFT and RIGHT images of the stereo pair, respectively. The next two rows give the results of registering (REG) and shifting (SHIFT) the right monocular fusion in the left image coordinate frame. The final row gives the statistics for the merger of the left monocular fusion results with the registered and shifted right monocular fusion results (FUSION).

Comparing the left-right fusion statistics with those of the extraction-based stereo fusion, we observe similar behavior in terms of increased building detection rate (after stereo and monocular fusion), as well as increased error as reflected in the building pixel branching factor. We further note that although the final results do in fact have different statistics, the differences are very minor, and our results for other stereo pairs exhibit only minor differences between left-right fusion and extraction-based fusion. In general, the fusion of stereo information provides improved performance over monocular fusion, just as monocular fusion provides improved performance over any individual building extraction technique.

4. Thresholding the accumulator image

As part of the scan-conversion fusion process, an accumulator image is produced that represents the "building density" of the scene. More precisely, each pixel in the image has a value, which is the number of hypotheses that overlapped the pixel. Pixels with higher values represent areas of the image that have higher probability of being contained in a man-made structure. Theoretically, thresholding this image at higher values and then applying connected region extraction techniques would produce sets of hypotheses containing fewer false positives, and these hypotheses would only represent those areas that had a high probability of corresponding to structure in the scene.

To test this idea, the accumulator images generated by extraction-based fusion for several scenes were thresholded at values of 2, 3, and 4, since four systems were used to produce the final hypothesis fusion. Connected region extraction techniques were then applied to these thresholded images to produce new hypothesis segmentations. The new evaluation method was then applied to these new hypotheses. Brief summaries of the results are shown in Tables 4-1 through 4-5.

Summary of thresholding results for DC36A				
Threshold Value	% Bldgs Detected	% Bkgd Detected	% False Positives	Bld Pixel Br Factor
1	0.904	0.857	0.910	1.071
2	0.768	0.948	0.601	0.455
3	0.626	0.973	0.324	0.286
4	0.472	0.989	0.116	0.147

Table 4-1: Thresholding statistics for DC36A fusion results

Summary of thresholding results for DC36B				
Threshold Value	% Bldgs Detected	% Bkgd Detected	% False Positives	Bld Pixel Br Factor
1	0.654	0.858	0.800	2.109
2	0.300	0.964	0.332	1.159
3	0.122	0.992	0.077	0.598
4	0.020	0.999	0.006	0.292

Table 4-2: Thresholding statistics for DC36B fusion results

Summary of thresholding results for DC37				
Threshold Value	% Bldgs Detected	% Bkgd Detected	% False Positives	Bld Pixel Br Factor
1	0.863	0.847	0.869	1.055
2	0.682	0.956	0.450	0.380
3	0.481	0.981	0.175	0.228
4	0.281	0.994	0.040	0.108

Table 4-3: Thresholding statistics for DC37 fusion results

Summary of thresholding results for DC38				
Threshold Value	% Bldgs Detected	% Bkgd Detected	% False Positives	Bld Pixel Br Factor
1	0.893	0.835	0.816	0.527
2	0.723	0.926	0.433	0.292
3	0.521	0.962	0.183	0.206
4	0.332	0.983	0.067	0.145

Table 4-4: Thresholding statistics for DC38 fusion results

In each of the scenes, increasing the threshold from its default value of 1 to a value of 2 causes a reduction of roughly 20 percent in the number of correctly detected building pixels. This suggests that a fair number of hypothesized building pixels are unique; i.e., several pixels can only be correctly identified as building pixels by one of the detection methods. Another interesting observation is that the

Summary of thresholding results for LAX				
Threshold Value	% Bldgs Detected	% Bkgd Detected	% False Positives	Bld Pixel Br Factor
1	0.931	0.891	0.917	0.817
2	0.759	0.977	0.397	0.208
3	0.506	0.991	0.108	0.119
4	0.354	0.998	0.020	0.038

Table 4-5: Thresholding statistics for LAX fusion results

building pixel branching factor roughly doubles every time the threshold is decremented. These observations suggest that thresholding alone may eliminate unique information produced by the individual detection systems, and that more work will need to be done to limit the number of false positives (and erroneous delineations) produced by each system, and by the final fusion as a whole.

5. Additional results in building hypothesis fusion

The fusion process has been run on 51 monocular scenes (of which 23 have detailed hand segmentations) in addition to the DC37405 scene. We have also run the stereo fusion process on four stereo pairs in addition to the stereo pair for DC37405. In Section 5.1, we show some additional examples that are representative of our results on this larger number of examples, and we give the evaluation statistics for these scenes. We also give a brief analysis of the results on the monocular scenes, as well as graphs charting the performance improvement gained by the fusion process and the cumulative fusion error as represented by the building pixel branching factor. In Section 5.2, we give the evaluation statistics for the four additional stereo pairs, a brief analysis of the results, and performance and error graphs similar to those given in Section 5.1.

5.1. Monocular fusion results

The fusion process was applied to several monocular scenes. Here we show the results for scenes DC36A, DC36B, and DC38, three scenes from the Washington, D.C. area; and LAX, a scene from the Los Angeles International Airport [7]. Figures 5-1 through 5-4 are the ground-truth building segmentations used for performance analysis. The final fusion results for each of these scenes are shown in Figures 5-5 through 5-8. The coverage-based evaluation program was then applied to each of these results to generate Tables 5-1 through 5-4. As in the previous discussion, each of the results tables gives the statistics for a single scene. The first column represents a building extraction system. The next two columns give the percentage of building and background terrain correctly identified as such. The fourth and fifth columns show incorrect identification percentages for buildings and terrain. The next two columns give the breakdown (in percentages) of incorrect pixels in terms of false positives and false negatives. The last column gives the building pixel branching factor.

In all of the scenes, the detection percentage for the final fusion is greater than the same percentage for any of the individual extraction system hypotheses, although the building pixel branching factor also increases due to the accumulation of delineation errors from the various input hypotheses.

It is worth noting that the results for the DC36B scene (Table 5-2) are substantially worse than those of the other scenes. This is in large part due to the fact that the image intensity of the DC36B scene has a small dynamic range. Since the component systems used for these fusion experiments are inherently intensity-based, it is more difficult to detect shadow/building boundaries and building/background contours. As a result the building pixel branching factors reflect the poor performance of the component systems; in GROUPER's case, over 3 pixels are incorrectly hypothesized as building pixels for every

correct building pixel. The fusion process, however, improved the building detection percentage noticeably over the percentages of the component systems.

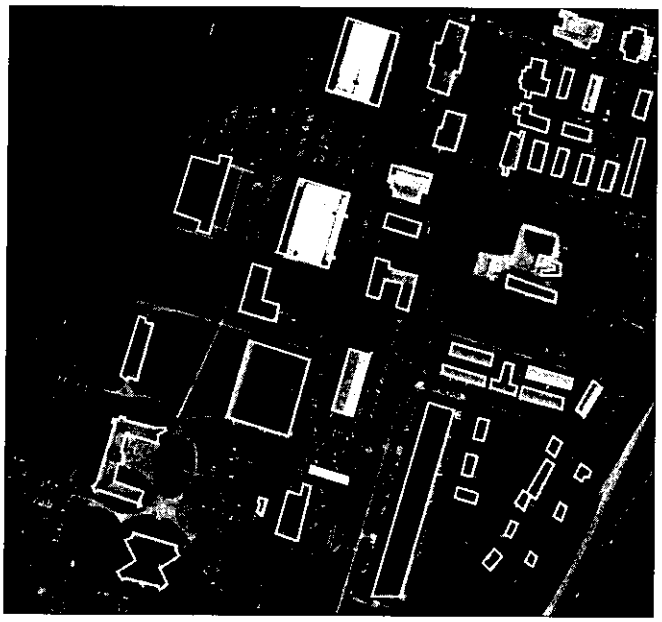

Figure 5-1: DC36A image with ground-truth segmentation

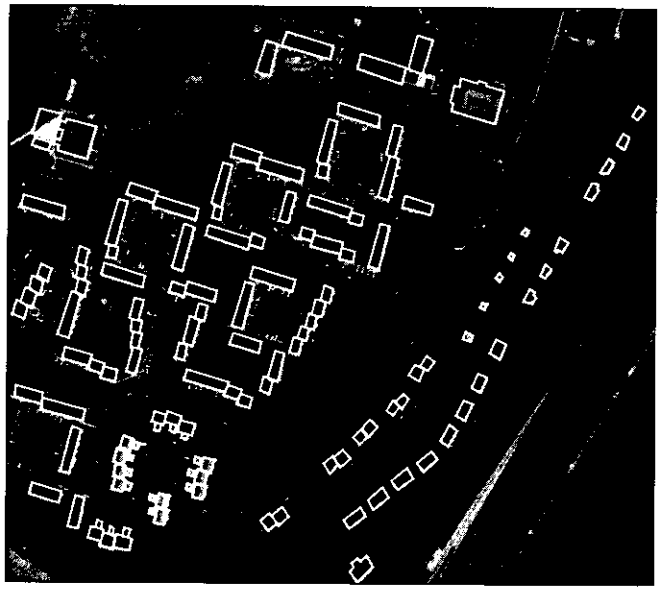

Figure 5-2: DC36B image with ground-truth segmentation

Figure 5-3: DC38 image with ground-truth segmentation

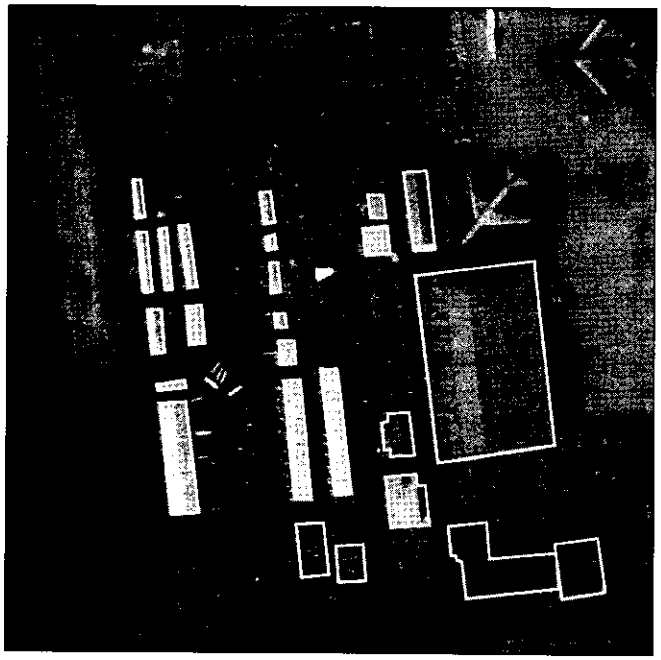

Figure 5-4: LAX image with ground-truth segmentation

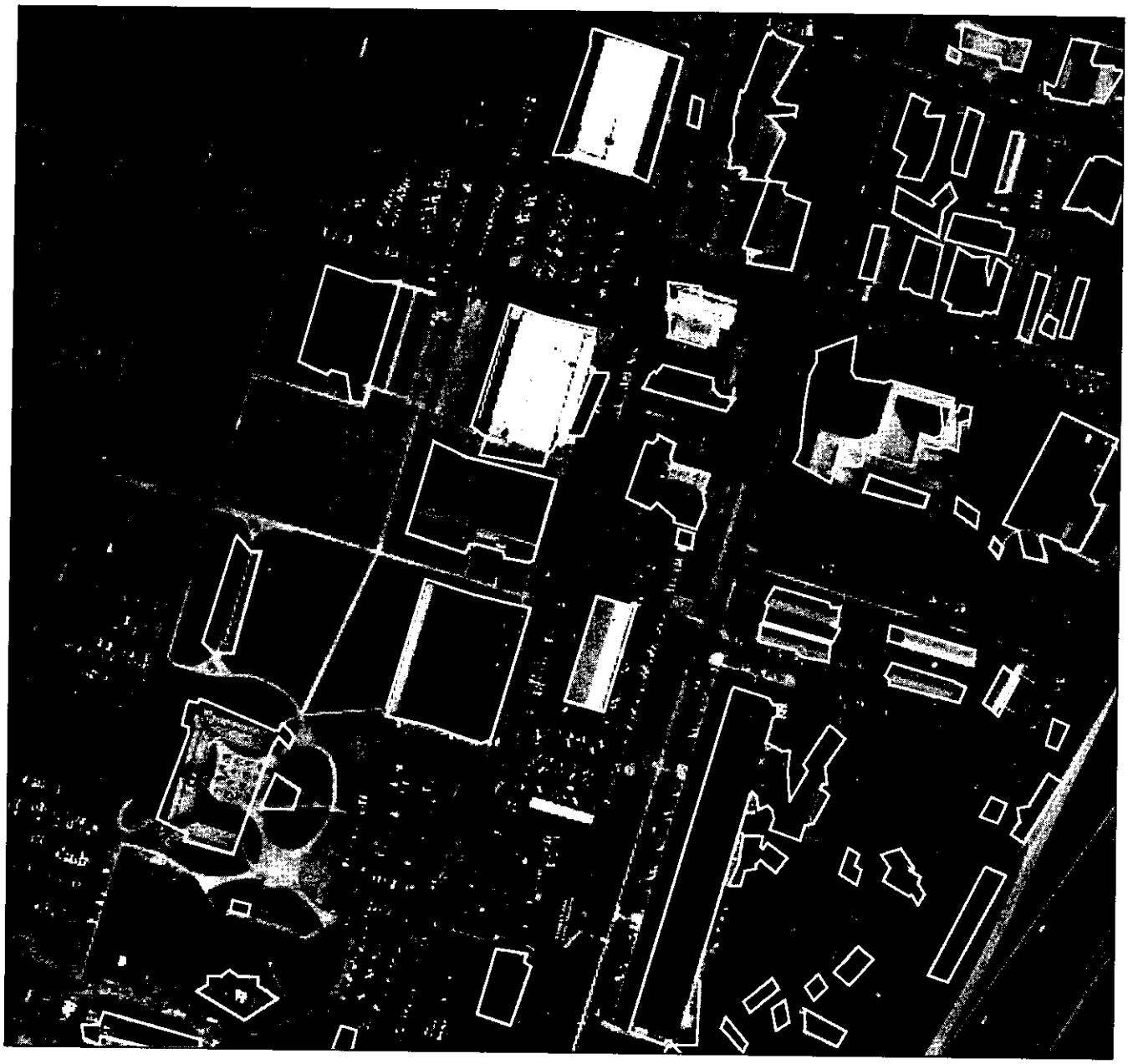

Figure 5-5: Monocular hypothesis fusion for DC36A

Evaluation results for the fusion process on DC36A							
System	% Bld Detected	% Bkgd Detected	% Bld Missed	% Bkgd Missed	% False Pos.	% False Neg.	Br Factor
SHADE	53.8	97.0	46.2	3.0	30.7	69.3	0.381
SHAVE	63.6	96.2	36.4	3.8	41.8	58.2	0.411
GROUPER	58.0	95.8	42.0	4.2	40.6	59.4	0.495
BABE	51.0	97.9	49.0	2.1	22.1	77.9	0.273
FUSION	80.9	91.9	19.1	8.1	74.3	25.7	0.682
51 regions in ground truth							

Table 5-1: Evaluation statistics for DC36A hypothesis fusion

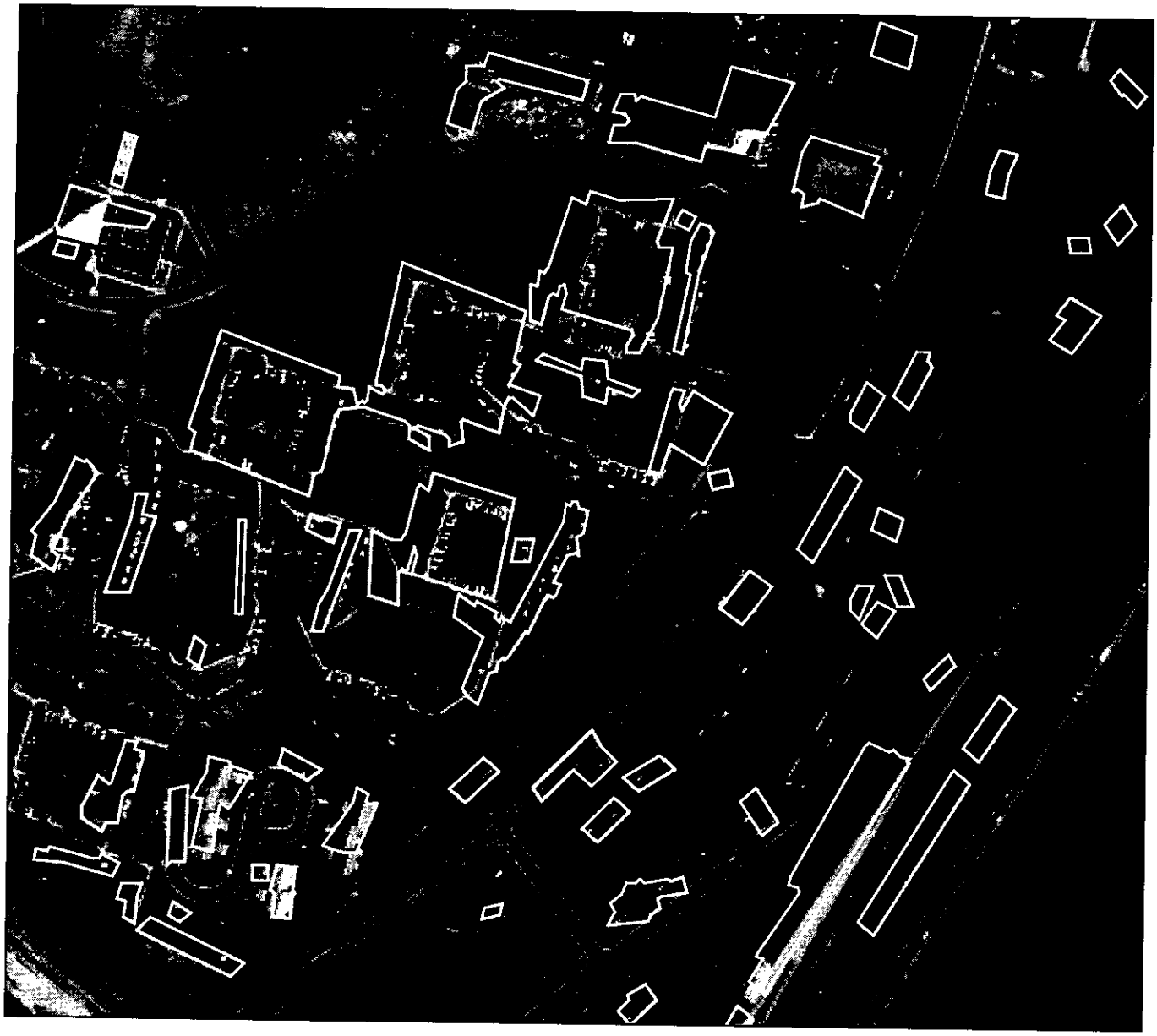

Figure 5-6: Monocular hypothesis fusion for DC36B

Evaluation results for the fusion process on DC36B							
System	% Bld Detected	% Bkgd Detected	% Bld Missed	% Bkgd Missed	% False Pos.	% False Neg.	Br Factor
SHADE	29.8	93.8	70.2	6.2	46.3	53.7	2.034
SHAVE	28.4	96.7	71.6	3.3	31.3	69.7	1.146
GROUPER	10.3	96.8	89.7	3.2	25.9	74.1	3.027
BABE	9.9	98.8	90.1	1.2	11.3	88.7	1.159
FUSION	49.8	89.2	50.2	10.8	67.8	32.2	2.126
133 regions in ground truth							

Table 5-2: Evaluation statistics for DC36B hypothesis fusion

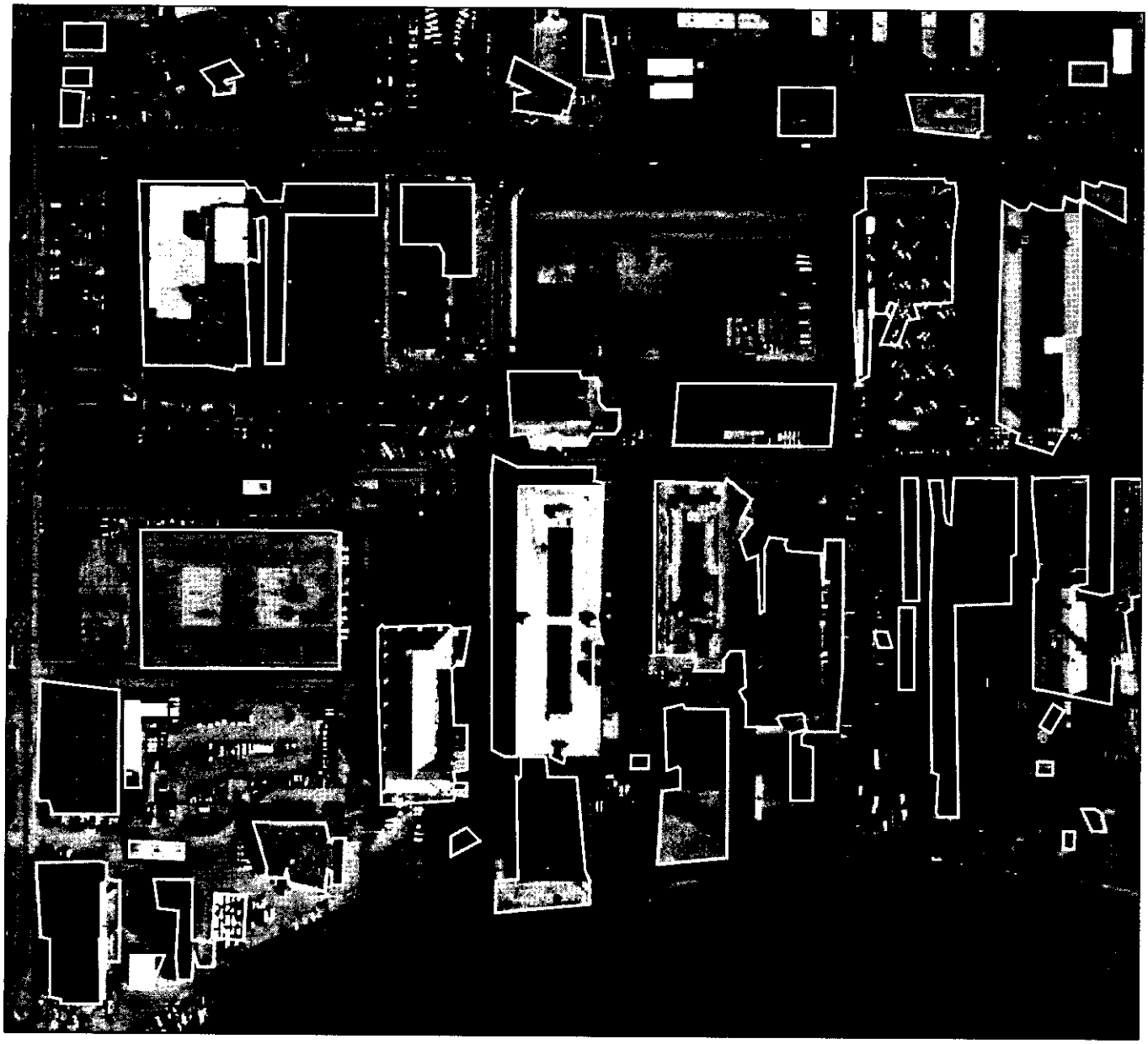

Figure 5-7: Monocular hypothesis fusion for DC38

Evaluation results for the fusion process on DC38							
System	% Bld Detected	% Bkgd Detected	% Bld Missed	% Bkgd Missed	% False Pos.	% False Neg.	Br Factor
SHADE	51.3	97.4	48.7	2.6	13.2	86.8	0.144
SHAVE	43.1	95.3	56.9	4.7	19.1	80.9	0.311
GROUPER	54.6	95.8	45.4	4.2	21.0	79.0	0.221
BABE	44.7	96.0	55.3	4.0	17.3	82.7	0.260
FUSION	74.7	90.6	25.3	9.4	51.5	48.5	0.360
53 regions in ground truth							

Table 5-3: Evaluation statistics for DC38 hypothesis fusion

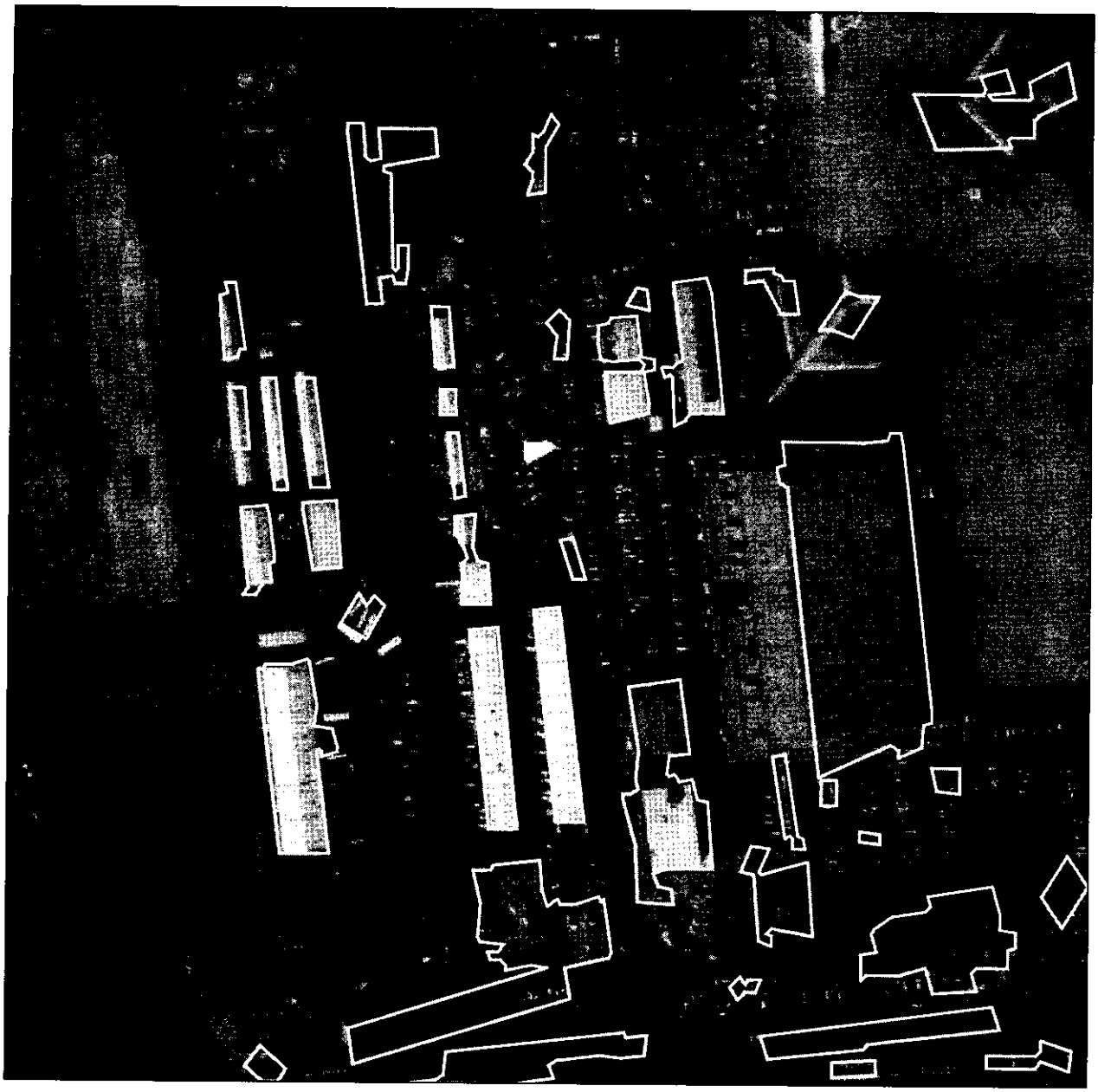

Figure 5-8: Monocular hypothesis fusion for LAX

Evaluation results for the fusion process on LAX							
System	% Bld Detected	% Bkgd Detected	% Bld Missed	% Bkgd Missed	% False Pos.	% False Neg.	Br Factor
SHADE	34.4	99.0	65.6	1.0	10.1	89.9	0.213
SHAVE	54.1	94.9	45.9	5.1	43.6	56.4	0.655
GROUPER	46.0	98.5	54.0	1.5	16.5	83.5	0.232
BABE	63.3	98.8	36.7	1.2	18.3	81.7	0.130
FUSION	73.0	92.9	27.0	7.1	65.0	35.0	0.687
26 regions in ground truth							

Table 5-4: Evaluation statistics for LAX hypothesis fusion

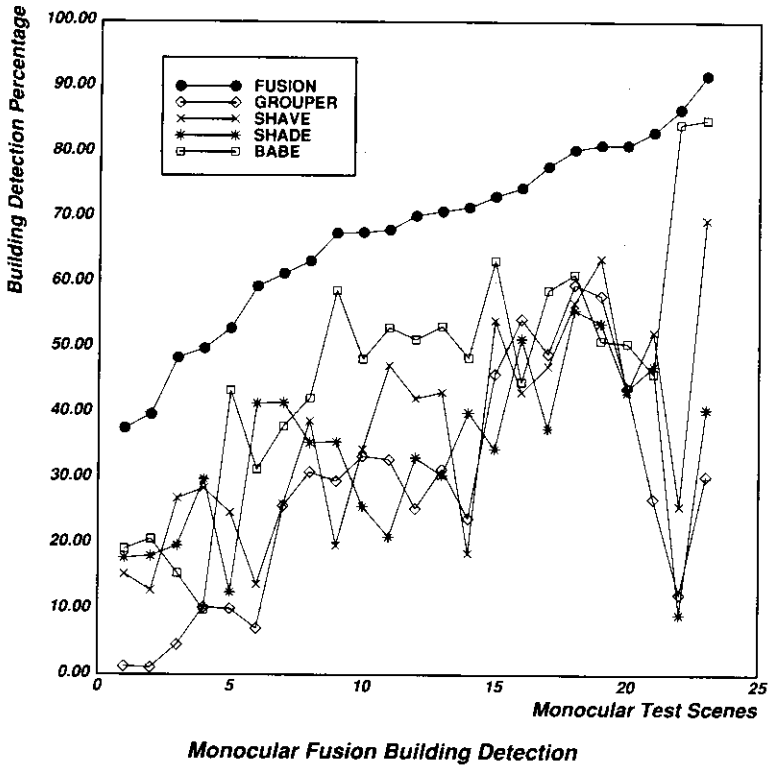

Figure 5-9: Monocular building detection percentages

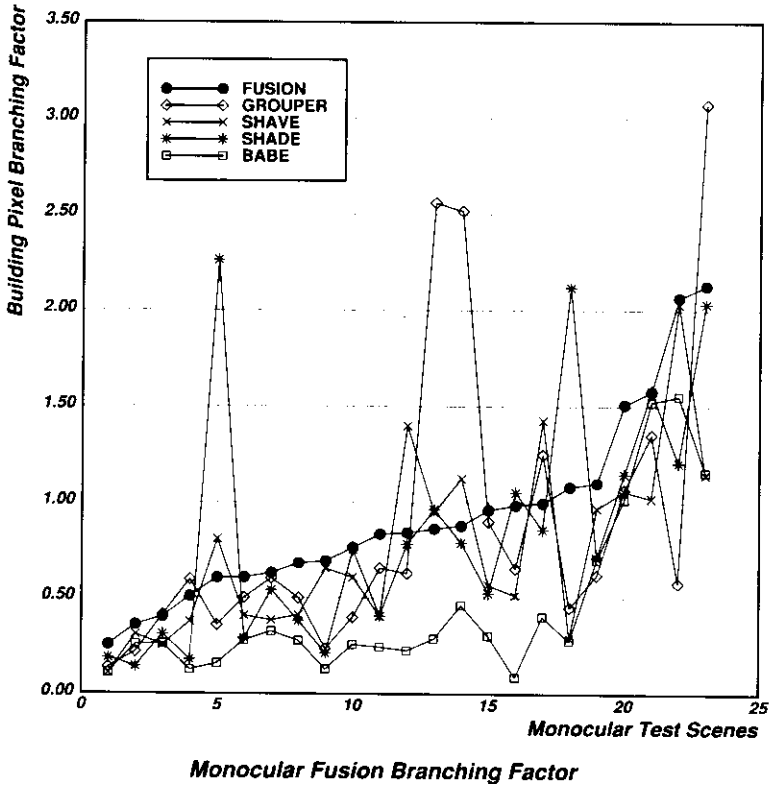

Figure 5-10: Monocular building pixel branching factors

We also note that several difficulties are attributable to performance deficiencies in the systems producing the original building hypotheses:

1. The shadow-based detection and evaluation systems, SHADE and SHAVE, both use a threshold to generate "shadow regions" in an image. This threshold is generated automatically by BABE, a line-corner based detection system. In some cases, the threshold is too low, and the resulting shadow regions are incomplete, which results in fewer hypothesized buildings.

2. GROUPER, the shadow-based hypothesis clustering system, clusters fragmented hypotheses by forming a region (based on shadow-building edges) in which building structure is expected to occur. This region is typically larger than the true building creating the shadow-building edge, and incorrect fragments sometimes fall within this region and are grouped with correct fragments. The resulting groups tend to be larger than the true buildings, and thus produce a fair number of false positive pixels.

3. SHAVE scores a set of hypotheses based on the extent to which they cast shadows, and then selects the top fifteen percent of these as "good" building hypotheses. In some cases, buildings whose scores fell in the top fifteen percent actually had relatively low absolute scores. This resulted in the inclusion of incorrect hypotheses in the final merger.

4. SHADE uses an imperfect sequence finder [1] to locate corners in the noisy shadow-building edges produced by thresholding. The sequence finder uses a threshold value to determine the amount of noise that will be ignored when searching for corners. In some situations, the true building corners are sufficiently small that the sequence finder regards them as noise, and as a result, the final building hypotheses can either be erroneous or incomplete.

Despite these problems, the fusion process outlined here performs well in obtaining improved building detection percentages for many scenes. Figure 5-9 gives the building detection percentages for the 23 monocular scenes with detailed hand segmentations. The percentages for each of the four component systems and the final fusion are given, and the results are sorted by the detection percentage of the final fusion. As the figure shows, building detection is improved for every monocular scene. For some scenes, the fusion process produces smaller improvements, due to the fact that the best performing component system produces very good results. For example, the next to last point on the graph shows a small performance improvement. In this scene, the building edges were consistently strong, so BABE performed very well; and the sun was at zenith when the scene was imaged, so shadow analysis provided little complementary information.

Figure 5-10 gives the building pixel branching factor for each of the 23 monocular scenes. Again, the scenes are sorted by the value produced by evaluating the fusion result. Not surprisingly, the building pixel branching factor for the fusion result is usually greater than the branching factor for each of the component results. In a few cases, this is not true; this can be explained by the fact that a component system performed very poorly, producing a small number of very bad building hypotheses, which results in a very high branching factor. The fusion results have a lower branching factor because other component systems produce better results, alleviating the number of false positive pixels.

5.2. Stereo fusion results

The stereo fusion processes (both left-right and extraction-based) were run on four stereo pairs in addition to the DC37405 scene. In all cases, the final results were quite similar; for brevity, we have omitted the statistics for the extraction-based fusion and present only the statistics for left-right stereo fusion. Tables 5-5 through 5-8 give the statistics for the four scenes.

Evaluation results for stereo fusion on DC36A							
System	% Bld Detected	% Bkgd Detected	% Bld Missed	% Bkgd Missed	% False Pos.	% False Neg.	Br Factor
LEFT	80.9	91.9	19.1	8.1	74.3	25.7	0.682
RIGHT	80.4	90.4	19.6	9.6	77.5	22.5	0.835
REG	78.6	89.7	21.4	10.3	76.5	23.5	0.888
SHIFT	79.3	89.9	20.7	10.1	76.7	23.3	0.861
FUSION	90.5	85.9	9.5	14.1	91.0	9.0	1.060

Table 5-5: Evaluation statistics for DC36A stereo fusion

Evaluation results for stereo fusion on DC36B							
System	% Bld Detected	% Bkgd Detected	% Bld Missed	% Bkgd Missed	% False Pos.	% False Neg.	Br Factor
LEFT	49.8	89.2	50.2	10.8	67.8	32.2	2.126
RIGHT	48.4	92.6	51.6	7.4	59.7	40.3	1.578
REG	51.5	93.4	48.5	6.6	57.0	43.0	1.249
SHIFT	49.7	93.2	50.3	6.8	56.8	43.2	1.333
FUSION	65.1	85.9	34.9	14.1	79.7	20.3	2.114

Table 5-6: Evaluation statistics for DC36B stereo fusion

Evaluation results for stereo fusion on DC38							
System	% Bld Detected	% Bkgd Detected	% Bld Missed	% Bkgd Missed	% False Pos.	% False Neg.	Br Factor
LEFT	74.7	90.6	25.3	9.4	51.5	48.5	0.360
RIGHT	81.1	89.7	18.9	10.3	63.2	36.8	0.402
REG	73.8	88.9	26.2	11.1	54.9	45.1	0.432
SHIFT	76.1	89.9	23.9	10.1	54.6	45.4	0.378
FUSION	88.6	83.4	11.4	16.6	80.5	19.5	0.535

Table 5-7: Evaluation statistics for DC38 stereo fusion

On the LAX scene, we note that the right monocular results are quantitatively better than the left in terms of building detection rate. As with the DC37405 stereo pair, we have a situation where one of the images has more prominent cues than the other image; in this case, the right image of the stereo pair has more prominent building shadows, and the shadow-based analysis systems exhibit improved performance, which is then reflected in the monocular fusion results. As noted earlier, stereo fusion increases the overall building detection rate in all of our test scenes, although the branching factor increases as well due to the accumulation of individual delineation errors and erroneous hypotheses. These trends can be observed in Figures 5-11 and 5-12.

Evaluation results for stereo fusion on LAX							
System	% Bld Detected	% Bkgd Detected	% Bld Missed	% Bkgd Missed	% False Pos.	% False Neg.	Br Factor
LEFT	73.0	92.9	27.0	7.1	65.0	35.0	0.687
RIGHT	91.8	93.5	8.2	6.5	85.1	14.9	0.508
REG	90.3	94.4	9.7	5.6	80.3	19.7	0.438
SHIFT	90.3	94.4	9.7	5.6	80.3	19.7	0.439
FUSION	93.6	89.1	6.4	10.9	92.3	7.7	0.821

Table 5-8: Evaluation statistics for LAX stereo fusion

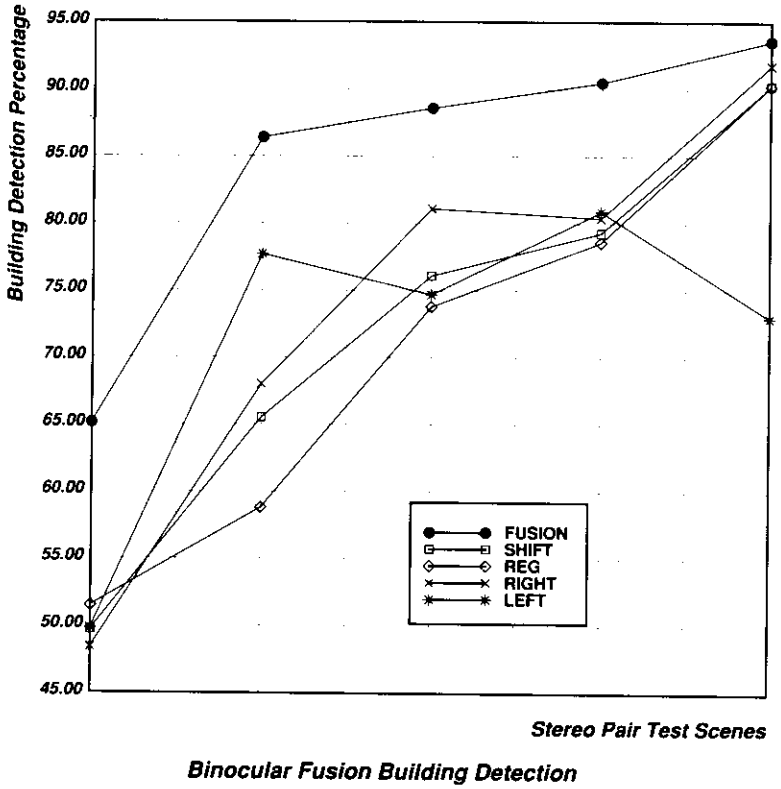

Figure 5-11: Stereo building detection percentages

It is worth noting that stereo fusion provides improved detection results over monocular fusion. In each of the five stereo pairs, the building detection percentage for stereo fusion is greater than the building detection percentage for the corresponding monocular fusion. (Compare Table 2-1 with Table 3-2, and Tables 5-1 through 5-4 with Tables 5-5 through 5-8.) As noted in our initial discussion of stereo fusion, images taken from different vantage points provide different (and in many cases complementary) information. Shadows appear different in stereo imagery, and edges and corners may become more (or less) visible from different perspectives. Fusion of stereo data provides a means for taking advantage of the different results produced for each image.

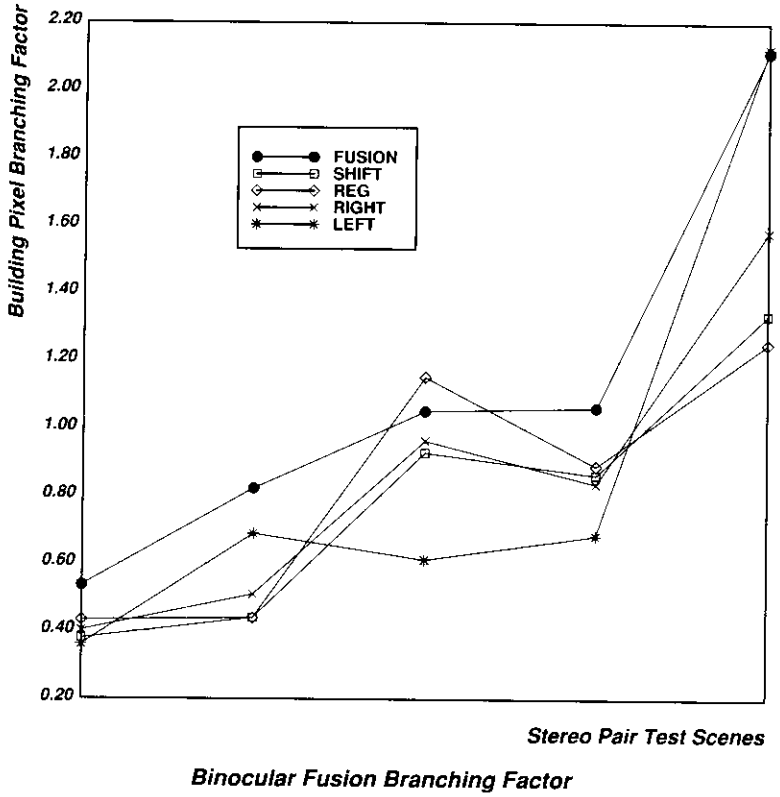

Figure 5-12: Stereo building pixel branching factors

6. Generating three-dimensional representations

The goal of three-dimensional scene analysis is to generate an interpretation of the imagery that is as close as possible to the actual scene under consideration. It is our belief that no individual computer vision technique can reliably provide a complete scene reconstruction. To achieve this goal, we will need to utilize multiple sources of information (which may be incomplete or inconsistent) and integrate them into a consistent interpretation of the scene. The method described in this paper integrates one type of monocular information: building delineations.

There are other types of information that can be integrated with these fused building delineations to allow the formation of three-dimensional representations. Since we have qualitative building boundary information, we can generate three-dimensional views with the integration of height information. This height information can be obtained from several visual cues as well; among these are shadow information and disparity information from the analysis of stereo imagery.

Figure 6-1 shows a perspective view for the DC37405 scene, generated by the use of ground-truth terrain elevation values and building height segmentations. It is an accurate three-dimensional view of the scene structure using manual feature extraction techniques. Figure 6-2 shows a similar perspective view generated without manual height estimates for the terrain. Figure 6-3 shows a perspective view with structural height estimates automatically derived from a disparity map. The disparity map was generated by the fusion of disparity estimates produced by two stereo matchers, one area-based and one feature-based [5]. It is worth noting that height estimates of this nature do not constitute three-dimensional representations of the scene; a true representation would include building delineations, a transportation network of roads, and a digital elevation model. The information fusion approach provides a means for integrating image cues to produce the components of a true three-dimensional representation of the scene.

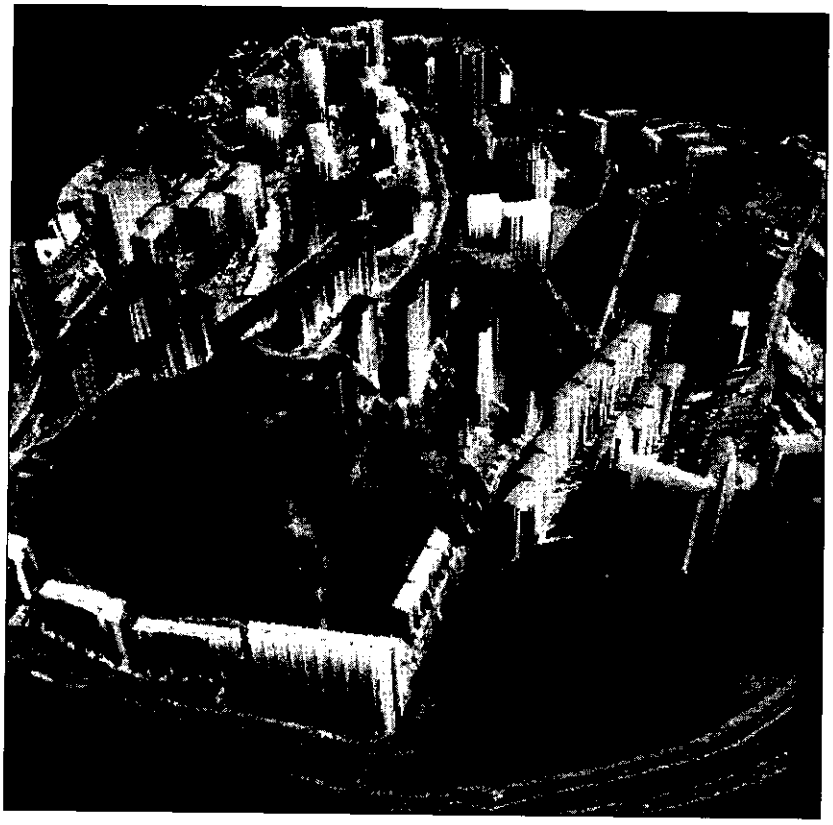

Figure 6-1: Perspective view for DC37405 using ground-truth building and height data

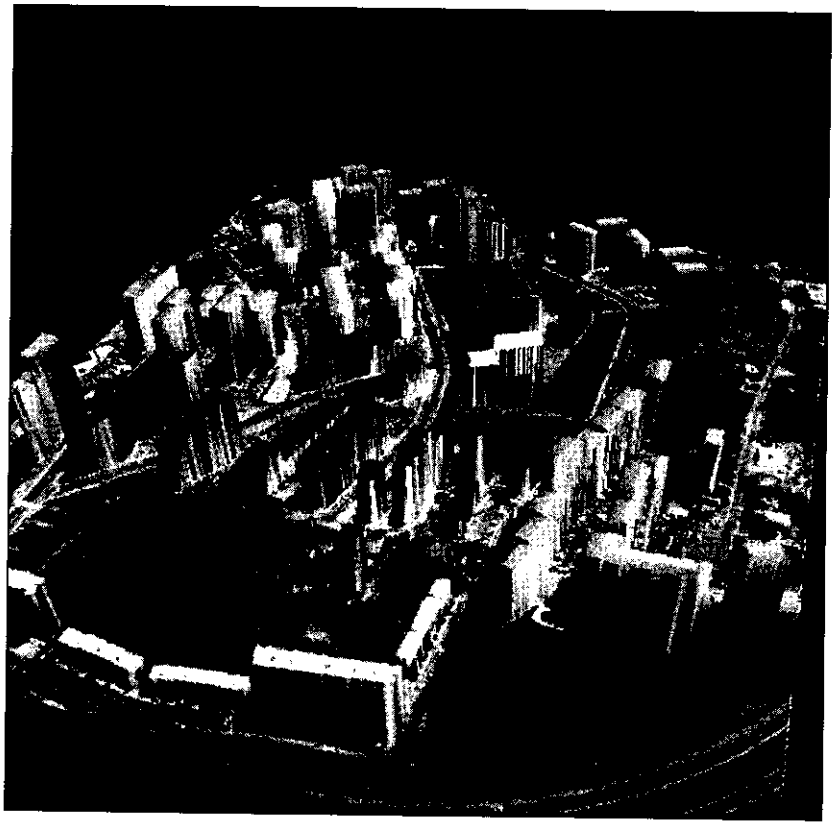

Figure 6-2: Perspective view for DC37405 using ground-truth building data only

In this particular case, the fusion of building boundaries (which are themselves fusions of building hypotheses) with disparity maps provides one component of the three-dimensional representation: qualitatively accurate building delineations and heights. In that sense, Figure 6-3 should be compared with the perspective view in Figure 6-2, since we do not utilize a terrain model in the fusion techniques described here.

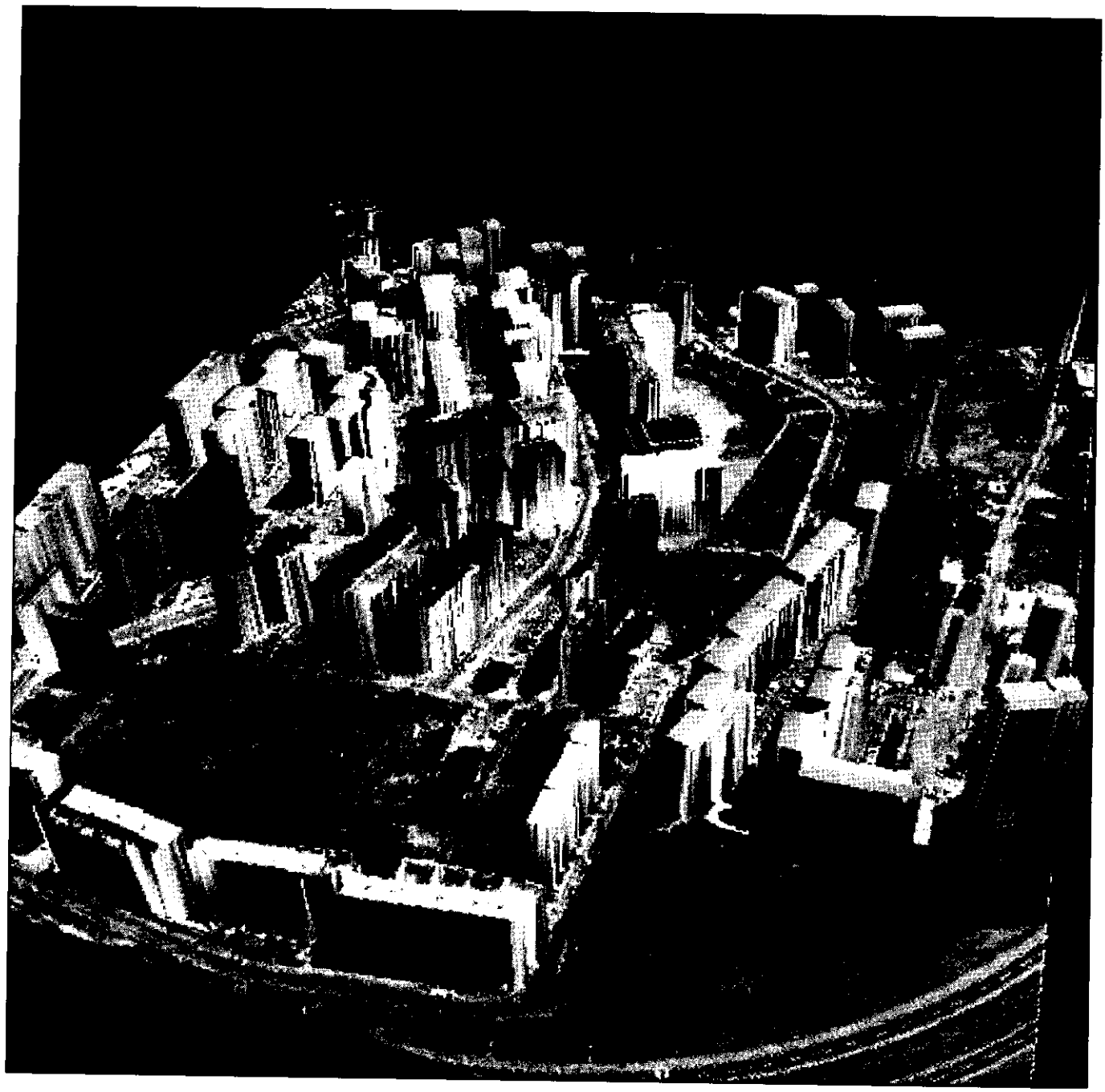

Figure 6-3: Perspective view for DC37405 using stereo disparity information

Figure 6-4 shows another perspective view for the DC37405 scene, with structural estimates derived from SHAVE by analysis of the lengths of the cast shadows of buildings [8]. SHAVE detects and delineates the shadows cast by each of the fusion building regions by walking from the shadow/building edge along the sun direction vector. At each pixel along the shadow/building edge an estimate of the shadow length is computed. The median length of the set of shadow vectors is computed for each building; this becomes the building shadow length estimate. Using the trigonometric relationship

between building height, sun inclination angle, and length of the cast shadow we can estimate the building height with good accuracy. In fact, this procedure is used regularly in manual photo interpretation. It is interesting to note that this view was generated solely from monocular analysis; no stereo information was utilized. Although stereo information is necessary in many situations for accurate height estimation, monocular analysis is capable of providing reasonable qualitative building delineations and heights.

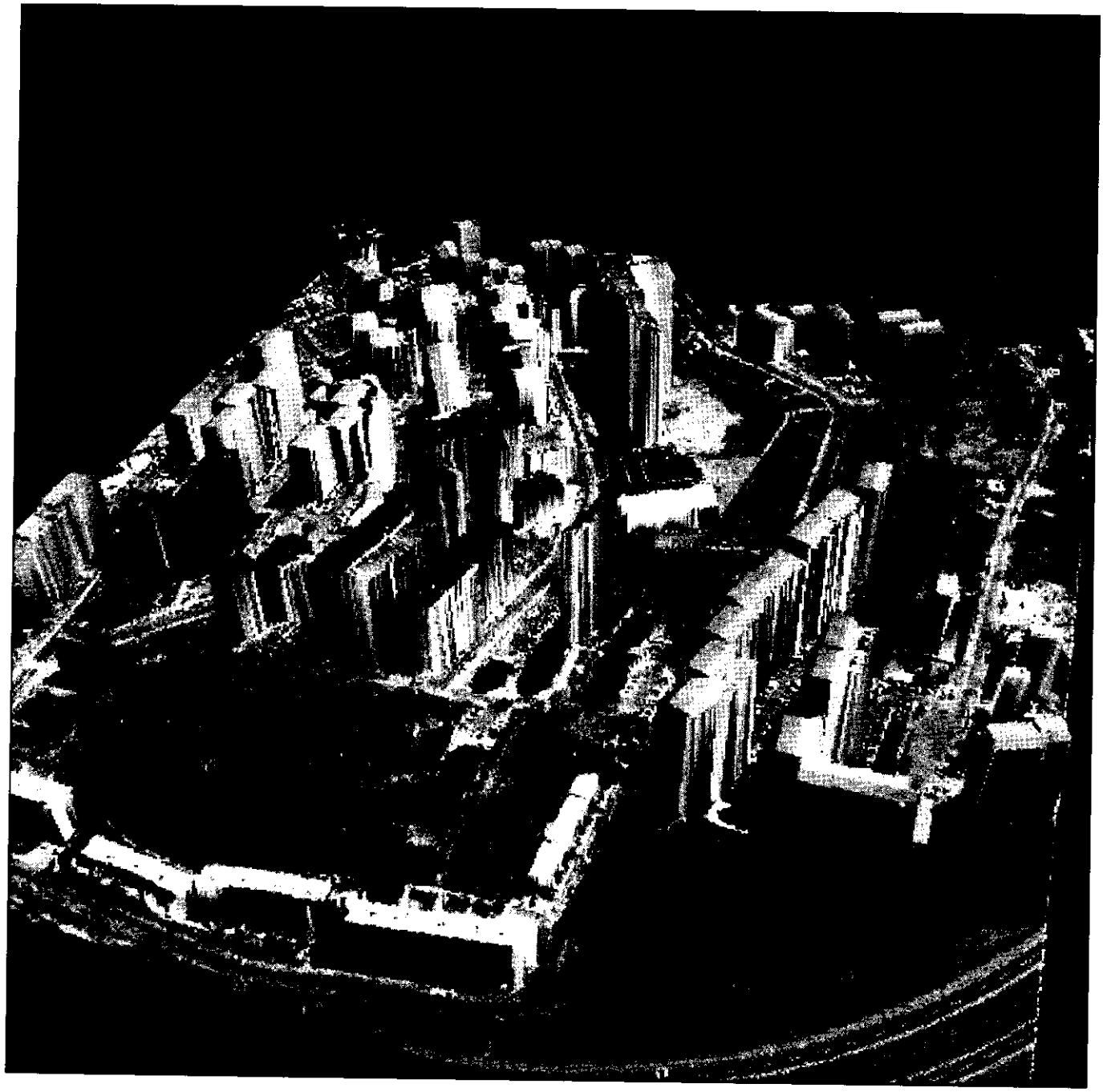

Figure 6-4: Perspective view for DC37405 using monocular shadow analysis

7. Conclusions

This paper has described a simple, yet effective, method for fusing sets of monocular building hypotheses for aerial imagery. Scan-conversion and connected region extraction techniques were applied to produce mergers of sets of building hypotheses, and the results were analyzed by the use of an evaluation technique based on pixel coverage. We also show the ability to merge hypotheses generated from the left and right images of a stereo pair to obtain an improved interpretation of a scene.

The simple hypothesis fusion approach developed here appears promising; the detection rate can be improved significantly by applying it to the results of several building detection systems. Much work remains to be done, however. Analysis of the fusion results has revealed shortcomings in each of the building detection systems, and there are also a number of directions to pursue in terms of improving the intermediate and final fusions generated during the overall fusion process.

1. GROUPER is effective in clustering the fragmented hypotheses that are typically produced by BABE, but several of the grouped fragments do not correspond to building structure in the scene. Experimentation with disparity maps to refine these clusters is currently underway.

2. SHAVE's scoring system is simplistic and sometimes allows hypotheses with low shadow scores to pass as good hypotheses. Alternative scoring schemes might be explored.

3. SHADE's shadow segmentation and corner finding system can be improved. Work is currently underway on a method for iteratively approximating the location of corners in noisy lines by using an imperfect sequence finder to break lines at potential corners, and applying a gradient-based line evaluation function to score the breaks.

4. The fusion steps in the overall fusion process tend to increase the number of false positive pixels, and thresholding alone may not improve this without decreasing the number of correctly hypothesized pixels as well. The use of a refined disparity map, as well as the use of the original intensity image, may aid in eliminating false positive pixels from hypothesized regions in the final fusion. Alternatively, active contour models [9, 4] might be used to refine segmentations, using the fusion segmentations (possibly thresholded) as the initial seed to the process. This may prove difficult, however: fairly accurate estimates of the building boundaries will be necessary, and there may be difficulties in recovering from local energy minima in complex high-resolution scenes.

A more general question concerns the effectiveness of simple fusion approaches such as the one described here. Certainly, one can envision other approaches for combining building hypotheses that would make use of *a priori* information about the systems producing the hypotheses to produce meaningful fusions of the individual hypotheses. It is unclear, however, whether such approaches would ultimately benefit from the additional complexity required to take advantage of such knowledge. Although the results at this stage are rough, the fusion method developed here appears to be a simple and effective means for increasing the building detection rate for a scene, and may eventually provide a means for incorporating several sources of photometric information into a single interpretation of the scene.

8. Acknowledgments

We would like to thank the members of the Digital Mapping Laboratory for providing an interesting and congenial working environment. Particular thanks go to Yuan Hsieh and Frederic Perlant for interesting discussions about information fusion and building extraction methods, as well as assistance with stereo matching and registration techniques. Bertha's Mussels and Annabelle's Big Hunks helped to brighten our lives.

References

[1] Aviad, Z.
Locating Corners in Noisy Curves by Delineating Imperfect Sequences.
Technical Report CMU-CS-88-199, Carnegie-Mellon University, December, 1988.

[2] Aviad, Z., McKeown, D. M., Hsieh, Y.
The Generation of Building Hypotheses From Monocular Views.
Technical Report, Carnegie-Mellon University, 1991.
to appear.

[3] Fua, P., Hanson, A. J.
Resegmentation Using Generic Shape: Locating General Cultural Objects.
Technical Report, Artificial Intelligence Center, SRI International, May, 1986.

[4] Fua, P., Hanson, A. J.
Objective Functions for Feature Discrimination: Theory.
In *Proceedings: DARPA Image Understanding Workshop*, pages 443-460. May, 1989.

[5] Hsieh, Y., Perlant, F., and McKeown, D. M.
Recovering 3D Information from Complex Aerial Imagery.
In *Proceedings: 10th International Conference on Pattern Recognition, Atlantic City, New Jersey*, pages 136-146. June, 1990.

[6] Hsieh, Y., Perlant, F., and McKeown, D. M.
Recovering 3D Information from Complex Aerial Imagery.
In *Proceedings: DARPA IUS Workshop*, pages 670-691. September, 1990.

[7] Huertas, A. and Nevatia, R.
Detecting Buildings in Aerial Images.
Computer Vision, Graphics, and Image Processing 41:131-152, April, 1988.

[8] R. B. Irvin and D. M. McKeown.
Methods for exploiting the relationship between buildings and their shadows in aerial imagery.
IEEE Transactions on Systems, Man and Cybernetics 19(6):1564-1575, November, 1989.

[9] Kass, M., Witkin, A., and Terzopoulos, D.
Snakes: Active Contour Models.
International Journal of Computer Vision 1(4):321-331, 1987.

[10] McKeown, D.M.,
Toward Automatic Cartographic Feature Extraction.
In Pau, L. F. (editor), *NATO ASI Series.* Volume F 65: *Mapping and Spatial Modelling for Navigation*, pages 149-180. Springer-Verlag, Berlin Heidelberg, 1990.

[11] Mohan, R., Nevatia, R.
Using Perceptual Organization to Extract 3-D Structures.
IEEE Transactions of Pattern Analysis and Machine Intelligence 11(11):1121-1139, November, 1989.

[12] Nicolin, B., and Gabler, R.
A Knowledge-Based System for the Analysis of Aerial Images.
IEEE Transactions on Geoscience and Remote Sensing GE-25(3):317-329, May, 1987.

[13] Perlant, F. P., McKeown, D. M.
Scene Registration in Aerial Image Analysis.
Photogrammetric Engineering and Remote Sensing 56(4):481-493, April, 1990.