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

Jefferey Shufelt and David M. McKeown

Digital Mapping Laboratory
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213[1]

Abstract

The extraction of buildings from aerial imagery is a complex problem for automated computer vision. It requires locating regions in a scene that possess properties distinguishing them as man-made objects 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 analysis of cast shadows to predict building locations is likely to fail in cases where the sun is directly above the scene.

It seems clear that a cooperative-methods paradigm is useful in approaching the building extraction problem. Using this paradigm, each extraction technique provides information which can then 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.

This paper describes preliminary research on the problem of building hypothesis fusion in aerial imagery. Building extraction techniques are briefly surveyed, including four building extraction, verification, and clustering systems that form the basis for the work described here. A method for fusing the symbolic data generated by these systems is described, and applied to monocular image and stereo image data sets. Evaluation methods for the fusion results are described, and the fusion results are analyzed using these methods.

[1] 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 Engineering Topographic Laboratories, or the Defense Advanced Research Projects Agency, or of the United States Government.

1. Introduction

In the cooperative-methods paradigm it is assumed that no single method can provide a complete set of building hypotheses for a scene. However, each method 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.

2. Building extraction techniques

At the Digital Mapping Laboratory, we have developed several techniques for the extraction of man-made objects from aerial imagery. The goal of many of these techniques is to organize the image into manageable parts for further processing, by using external knowledge to organize these parts into regions.

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 [5], 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 [4], [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.

There are many other interesting building detection and extraction techniques. We briefly mention some recently developed methods, to illustrate the variety of techniques that produce building hypothesis information. Although this by no means constitutes a comprehensive survey of building detection techniques, it provides some examples of the methods used to generate hypotheses for a scene, as well as examples of the types of data that may eventually be integrated into a cooperative-methods building analysis scheme.

Mohan and Nevatia [6] described a method by which simple image tokens such as lines or edges could be clustered into more complex geometric features consisting of parallelopipeds. Huertas and Nevatia [4] described a method for detecting buildings in aerial images. Their method detected lines and corners in an image and constructed chains of these to form building hypotheses which were then subject to shadow verification.

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. Nicolin and Gabler [7] 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.

3. A simple hypothesis merging technique

Building hypotheses typically take the form of geometric descriptions of objects in the context of an image. One can imagine "stacking" sets of these geometric 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.

The 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 4.4.

4. Merging multiple hypothesis sets

This section outlines the experiments performed with the scan-conversion hypothesis fusion technique. The procedure used to apply this technique to the results of four building detection and evaluation systems (BABE, SHADE, SHAVE, and GROUPER) is described. A technique for quantitative evaluation of building hypotheses is described, and applied to the hypothesis fusion results. These results are analyzed to suggest improvements to the fusion technique.

4.1. The merging technique applied to four extraction systems

There were two merging problems under consideration. The first of these was 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. The technique was applied to each cluster individually, and the resulting accumulator image was thresholded at 1, and connected region extraction techniques were applied to provide the geometric union of each cluster. These clusters were then used as the building hypotheses produced by GROUPER.

The second problem was the fusion of each of these monocular hypothesis sets into a single set of hypotheses for the scene. Again, the scan-conversion technique was applied. The four hypothesis sets were scan-converted, and the resulting accumulator image was thresholded at 1. Connected region extraction techniques were applied to produce the final segmentation for the image.

Figure 4-1 shows a section of a suburban area in Washington, D.C. Figure 4-2 shows the SHADE results for this scene, Figure 4-3 shows the SHAVE results, Figure 4-4 shows the GROUPER results, and Figure 4-5 shows the BABE results. Figure 4-6 shows the fusion of these four monocular hypothesis sets.

4.2. Evaluation of the technique

To judge the correctness of an interpretation of a scene, it is desirable to have some mechanism for quantitatively evaluating that interpretation. One approach is to compare a given set of 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.

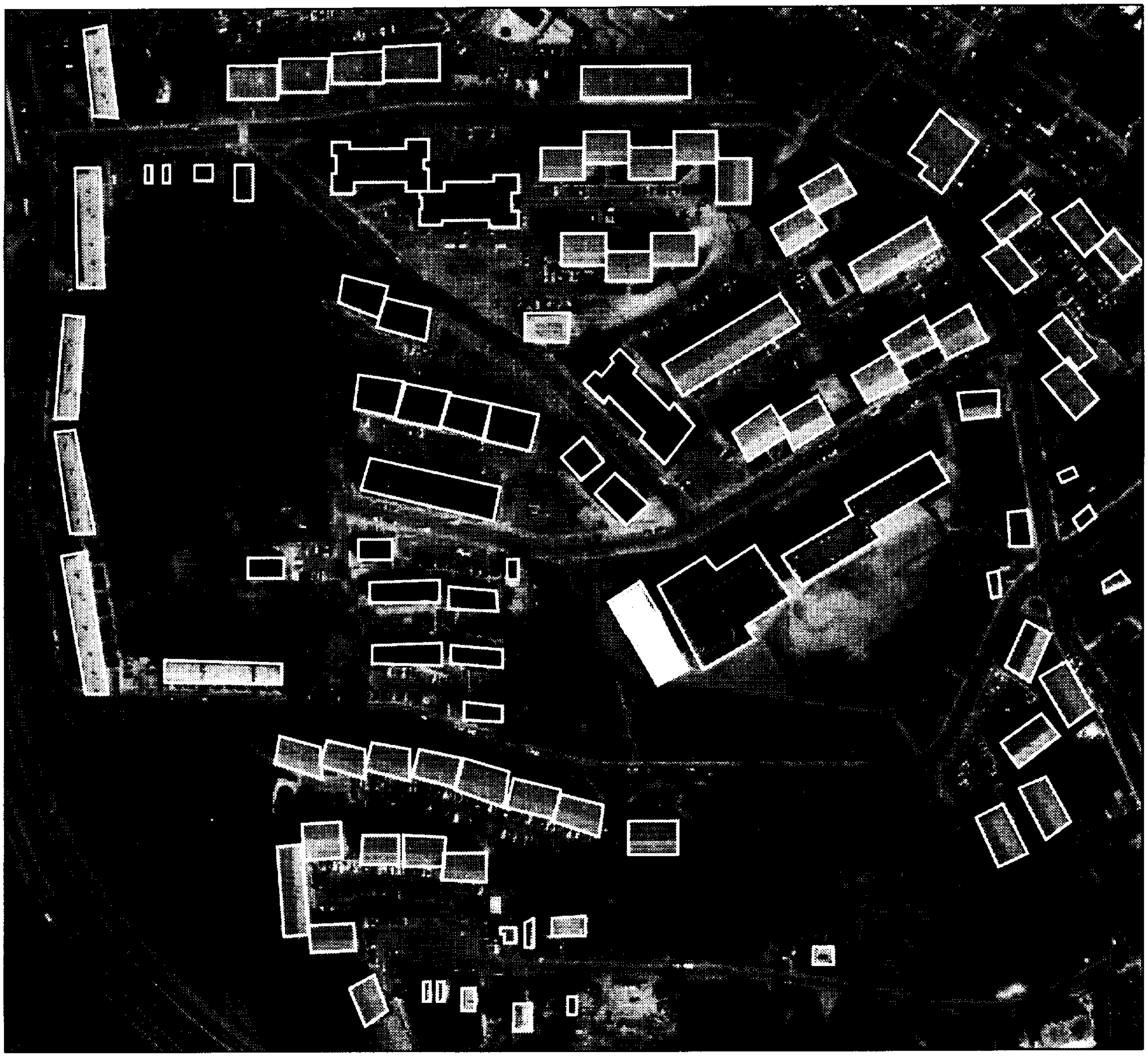

Figure 4-1: DC37 image with ground-truth segmentation

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:

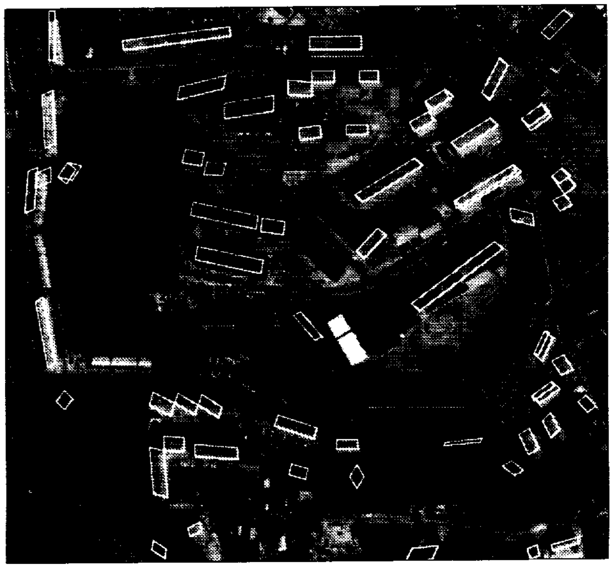

Figure 4-2: DC37 SHADE results

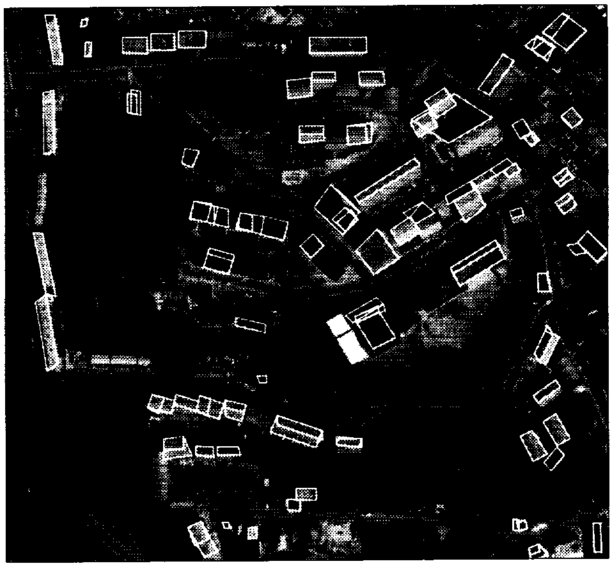

Figure 4-3: DC37 SHAVE results

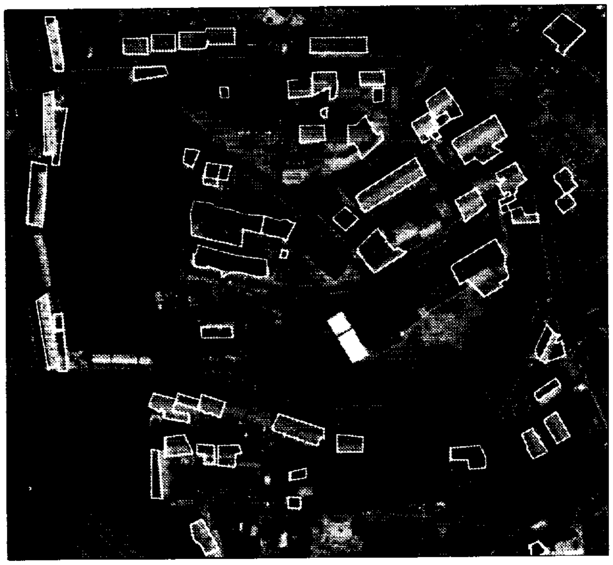

Figure 4-4: DC37 GROUPER results

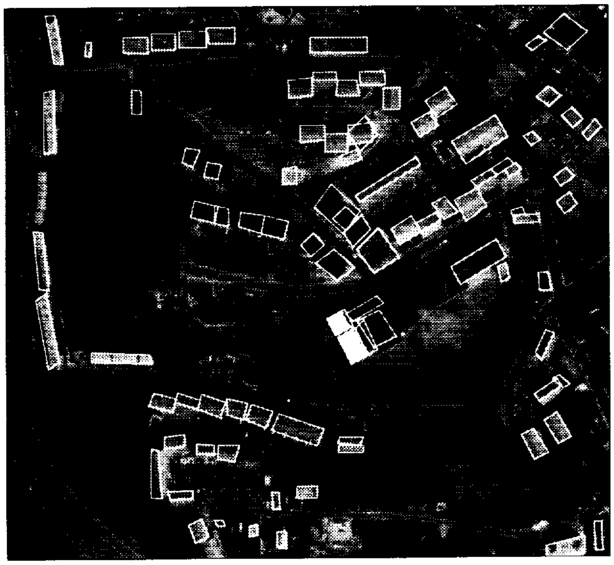

Figure 4-5: DC37 BABE results

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,

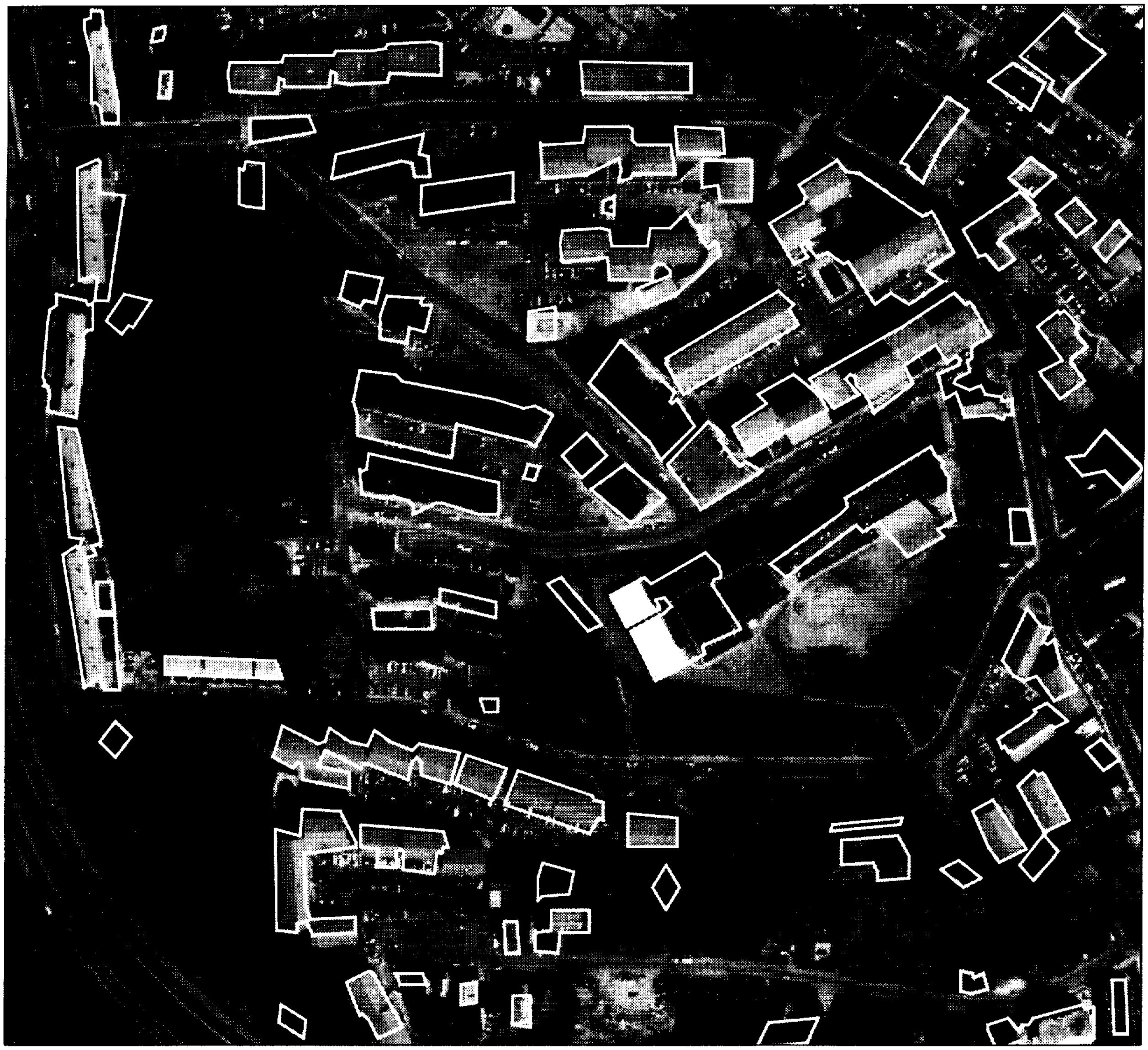

Figure 4-6: Monocular hypothesis fusion for DC37

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.

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.

4.3. Results and analysis

The fusion process was run on other scenes in addition to the DC37 scene: DC36A, DC36B, and DC38, three more scenes from the Washington, D.C. area; and LAX, a scene from the Los Angeles International Airport. The coverage-based evaluation program was then applied to generate Tables 4-1 through 4-5. Each table 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.

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 4-1: Evaluation statistics for DC37 hypothesis fusion

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 4-2: Evaluation statistics for DC36A hypothesis fusion

We note that the quantitative results generated by the new evaluation method accurately reflect the visual quality of the set of building hypotheses. 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. 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.

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 4-3: Evaluation statistics for DC36B hypothesis fusion

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 4-4: Evaluation statistics for DC38 hypothesis fusion

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 4-5: Evaluation statistics for LAX hypothesis fusion

It is worth noting that the results for the DC36B scene (Table 4-3) are substantially worse than those of the other scenes. This is in large part due to the fact that the DC36B scene has a low dynamic range of intensities, and the component systems used for these fusion experiments are inherently intensity-based. 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.

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

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.

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.

SHADE uses an imperfect sequence finder 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.

4.4. Thresholding the accumulator image

As part of the scan-conversion fusion process, an accumulator image is produced which 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 for each of the six 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.

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 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. Conclusions

This paper has described a simple 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.

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. BABE produces two shadow thresholds, only one of which is used by SHAVE and SHADE. It may be the case that the other threshold more accurately reflects the shadow threshold for a given image, or perhaps some combination of the two may prove more effective. Experiments need to be performed in this area.

2. 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.

3. 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.

4. SHADE's 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.

5. 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 might be used to refine segmentations, using the fusion segmentations (possibly thresholded) as the initial seed to the process.

6. Another interesting application of this fusion technique would be on binocular imagery. One could imagine merging hypotheses from the left and right images of a stereo pair to obtain an improved interpretation of a scene, since it is likely that the left and right hypothesis sets would differ due to changes in image perspective. Experiments are underway in this area.

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.

6. 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 techniques.

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, 1990.
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] Huertas, A. and Nevatia, R.
Detecting Buildings in Aerial Images.
Computer Vision, Graphics, and Image Processing 41:131-152, April, 1988.

[5] 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.

[6] Mohan, R., Nevatia, R.
Perceptual Grouping for the Detection and Description of Structures in Aerial Images.
In *Proceedings: DARPA Image Understanding Workshop, April 1988*, pages 512-526. April, 1988.

[7] 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.