From *Proceedings of the 5th IEEE Workshop on Applications of Computer Vision*, pp. 155-162, 2000.

Automatic Description of Complex Buildings with Multiple Images[1]

ZuWhan Kim, Andres Huertas, and Ramakant Nevatia
Institute of Robotics and Intelligent Systems
University of Southern California
3737 USC-Watt way, PHE 204 Los Angeles, CA 90089
{zuwhan,huertas,nevatia}@iris.usc.edu

Abstract

Three-D building detection and description is a practical application of 3-D object description, a key task of computer vision. We present an approach to detecting and describing buildings of polygonal rooftops by using multiple, overlapping images of the scene. First, 3-D features are generated by using multiple images, and rooftop hypotheses are generated by neighborhood searches on those features. For robust generation of 3-D features, we present a probabilistic approach to address the epipolar alignment problem in line matching. Image-derived unedited elevation data is used to assist feature matching, and to generate rough cues of the presence of 3-D structures. These cues help reduce the search space significantly. Experimental results are shown on some complex buildings.

1. Introduction

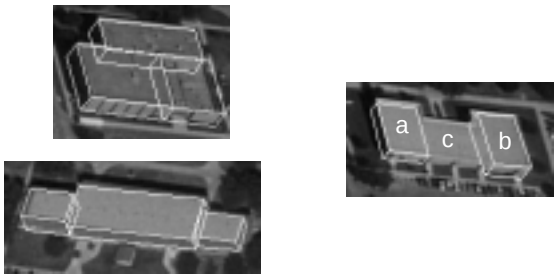
Three-D building detection and description is a practical application of 3-D object description. a key task of computer vision. Description of buildings in 3-D can greatly improve the automation of 2-D and 3-D map generation, that, in turn, can be used in many applications including GPS (global positioning system), virtual and augmented reality, military simulations and training, and radio signal reachability tests for wireless communications. Automatic building modeling has proven to be difficult to automate and has been an active research area for several years[1].

Many of the early building description systems used a single intensity image [2-5]. Shadows and wall vertical and base lines were used as important cues to infer 3-D heights of simple buildings. Most of the recent work has focused on the stereo or multi-view analysis because of the additional information it provides and because these data are widely available.

When two or more images are used, 3-D information is typically obtained by matching intensity values or image features. In [6-9], 3-D descriptions of buildings are generated from matched lines and junctions. In [10], rough building models are generated from laser range data and refined using line features. This approach requires accurate range data that may be difficult to obtain. Baillard and Zisserman [11] use six or more images to find 3-D matched lines, and find the orientations of half-planes for the 3-D lines by using intensity matching. Refined polygonal meshes are obtained by this method but it does not give explicit building models but just a collection of planar surfaces.

In most building description systems, buildings models are constructed by extruding polygonal rooftops. The shapes of rooftops vary from simple rectangles to unrestricted polygons; however, as the complexity of rooftops increases, the computation required for rooftop hypotheses generation grows exponentially. Hence, rectilinear rooftops have been modeled by a collection of rectangular components [7] or by simple blocks with gabled rooftops [12] as these simpler models can be derived with a reasonable amount of computation. While collections of rectangular rooftops can represent many rectilinear buildings (Figure 1a) this representation shows a limitation on the detection of even simple rectilinear buildings. An example is shown in Figure 1b. Components a and b of the building are likely to have low evidence support on the image because major parts of the roof boundaries do not exist. Component c is unlikely to be even detected because it has even lower evidence support; a half of the roof boundary is missing. On the other hand, modeling general rectilinear rooftops imposes large computational demands and/or

1. This research was supported by a MURI subgrant from Purdue University under Army Research Office grant No. DAAH04-96-1-0444.



(a) (b)

Figure 1. (a) Rectilinear buildings that can be described as collections of rectangular buildings and (b) a building where collections of rectangles shows a limitation.

results in poor detection rates [8]. In [10], unrestricted polygonal rooftops are described by using high resolution range data.

We present an approach to describing buildings having multi-layered flat polygonal rooftops. We reduce the computation by using multiple (more than 3) images. In addition, we use image-derived unedited digital elevation models (DEM's). In contrast with [10], these DEM's are not accurate enough to retrieve building model directly from it. Although DEM's can be computed from high resolution (sub meter) images, the underlying correlation methods used have inherent limitations and produce errors at and near building (and other) depth discontinuities. The use of multiple images provides additional advantages for 3-D object description problems over those of stereo images. They provide alternative views for occluded parts of objects and also help compensate for missing evidence and accidental alignments. We can also eliminate accidental wrong line or junction matches by filtering out lines and junctions which have matches in only one or two other views.

Figure 2 shows a flow diagram of our method. The system uses multiple images at about 1 meter resolution and an unedited DEM at a similar resolution. Note that the rough DEM is used primarily to provide cues that help reduce the search spaces and validate feature matches. First, 2-D features, lines, junctions and parallel relationships, are extracted from the images. Next, we derive 3-D features from groups of matched 2-D features over multiple views to generate rooftop hypotheses. All 2-D features may not be present in all views, therefore, to generate 3-D features, pairwise feature matching across all views is performed first, followed by grouping of matched pairs.

The next step is rooftop hypotheses generation. Flat polygonal rooftop hypotheses are generated by neighborhood searches on 3-D features. To reduce search space and get multi-layered rooftops, rough cues for each building components are generated by segmenting the DEM image

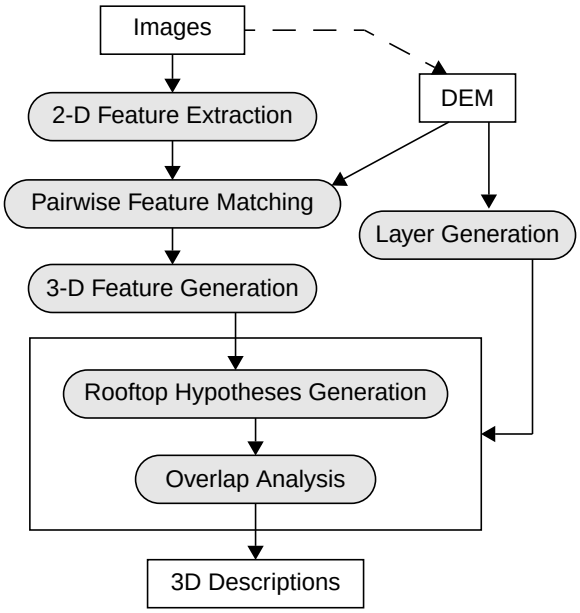


Figure 2. A flow diagram of the suggested approach.

into connected volumetric regions in space. Finally, overlap analysis is performed on the generated hypotheses to give final building descriptions.

Section 2 describes 2-D feature extraction and matching focusing on the use of DEM information. Section 3 deals with the use of multiple images to generate 3-D features. Specifically, we describe the epipolar alignment problem of line matching and suggest a probabilistic approach for it. In Section 4, we describe the rooftop hypothesis generation guided by DEM segmentation. Experimental results are shown in Section 5, and conclusion and future work are given in Section 6.

2. Feature Extraction and Matching

We first extract 2-D features from each image. Example aerial images of a building at 1.0m/pixel resolution are shown in Figure 3.

2-D Feature Extraction. Line segments are extracted from the image and grouped into "linear" features by collapsing collinear lines of small gaps. Building designs are largely geometric and have rectilinear boundaries that meet at restricted angles. The junctions among these lines represent strong point features that are well localized and have reduced ambiguity. T-junctions are also extracted. A T-junction is considered to be two L-junctions for further processing.

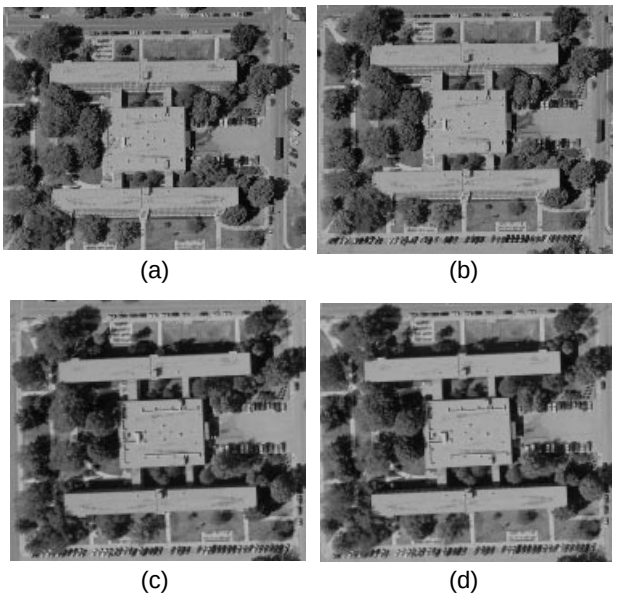


Figure 3. Example aerial images.

Building structures exhibit a great deal of parallelism in their design and construction. We extract parallel relationships from each view.

Epipolar Matching. The next step is to find pairwise matches of junctions and linears. We first find matches by using epipolar geometry. We follow the approach of [7] for the epipolar matching. The 1607 linears and 475 junctions that have epipolar matches in any of the remaining views in Figure 3a are shown in Figure 4.

3-D DEM assisted Matching. We use a commercially available system to compute a digital elevation model (DEM). The correlation-based process uses multiple views to derive the elevation data to generate a raw DEM. The portion of the DEM that corresponds to our example is shown in Figure 5 with lighter pixels denoting higher elevations. Note that the building and the trees are well represented although delineation is not accurate.

We use the DEM to filter out incorrect matches as illustrated in Figure 6. With knowledge of camera geometry a point p on linear l in view V_1 is made to correspond to several points p_1, p_2 and p_3 in another view V_2 by computing the intersections of the epipolar line E_p in V_2 and the candidate match linears for linear l in V_1.

We then calculate, by triangulation, the height for each pair $h_1=H(p,p_1)$, $h_2=H(p,p_2)$ and $h_3=H(p,p_3)$ and use these heights to project the line l onto the DEM a the three computed elevations. Figure 6 illustrates the projection of l onto the DEM at height h_3. The projected line is used to compute statistics of the DEM values in regions (sampling windows) adjacent and on both sides of the projected line.

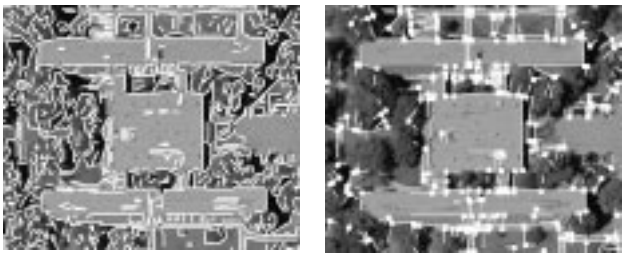


Figure 4. Matched linears and junctions.

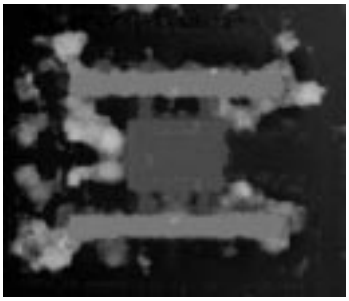


Figure 5. Correlation DEM

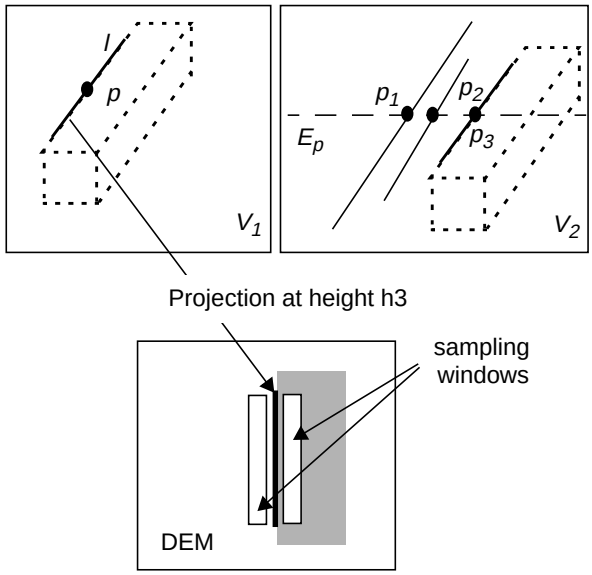


Figure 6. DEM assisted matching

If the calculated height, h_3, is consistent with the DEM values, the match is selected as correct. Note that the other two matches can be filtered out as well.

A similar procedure is applied to matching the junctions and to the parallel relationships to filter out incorrect matches. Figure 7 shows the 200 remaining matching linears and the 329 junction matches for this particular view in our example. Note the order of magnitude reduction in the number of line matches for this view and the signifi-

Figure 7. DEM assisted line matches and junction matches

(a) (b)

Figure 8. Compensation of missing evidence by using multiple views.

cant reduction in matched junctions. The figures are similar for the remaining views.

3. 3-D Feature Generation

Use of multiple (more than 3) images brings more advantages for 3-D object description problems than that of stereo images. Among these, an important advantage is that multiple images help compensate for missing evidence and accidental alignment. Figure 8 (a) shows linears extracted from bottom part of Figure 3 (b). Due to accidental illumination lines are missing for the southern part of the building. Therefore it is hard to describe the building only with the stereo pair of Figure 3 (a) and (b). But when multiple views are used, an alternative view (Figure 3c), provides good support for the line missing in two views (Figure 8 b).

Another serious problem of using stereo pair is that of epipolar alignment. For example, horizontal line segments extracted from Figure 3 (a), which are important features to detect the building, are aligned with epipolar lines with respect to the image of Figure 3 (b). Therefore, a small position error of a horizontal line segment causes very significant height error when it is matched with a segment from Figure 3 (b). But when it is matched with a line of Figure 3 (c) or Figure 3 (d), the height estimation becomes accurate because the epipolar lines are vertical with respect to those images. In fact these "accidental alignments" of important features are not rare for aerial image analysis. In many cases, buildings are aligned with epipolar lines because when aerial photos are taken from a flight, pilots tends to fly along the road. Similar situation can also occur in many other stereo analysis because, usually, epipolar lines are vertical or horizontal to the ground where vertical and horizontal lines are important.

This problem should also be addressed when multiple images are used since the height estimate of a near-epipolar-aligned line pair should be distinguished from the height estimates of other line pairs. There has been little research to address the epipolar alignment problem; in [11], for example, the epipolar lines are diagonal to building sides. In Section 3.1, we present a probabilistic approach to address the epipolar alignment problem by assuming the height estimates as a Gaussian random variable.

Based on this technique, we present the generation of robust and small number of 3-D features in Section 3.2.

3.1. Reasoning on height estimates

To incorporate expected height estimation error caused by epipolar alignment, we regard the height estimation of a line match as a Gaussian random variable. Given a line pair, the height (mean) is estimated from stereo analysis, and a confidence interval (standard deviation) is obtained assuming possible displacement errors of lines in the image space. Based on this formalism, we suggest a method for a height compatibility test and an information fusion from different height estimations.

Height Compatibility Test. Two height measurements, $\hat{X}_1$ and $\hat{X}_2$, are considered to come from the same height distribution, X, when the difference, $\Delta\hat{X} = \hat{X}_1 - \hat{X}_2$, is close to 0. If $\hat{X}_1$ and $\hat{X}_2$ are normal distributions with the means μ_1 and μ_2, and the variance σ_1^2 and σ_2^2 respectively, and independent on each other, then the difference $\Delta\hat{X}$ has a normal distribution with the mean $\mu = \mu_1 - \mu_2$ and the variance $\sigma^2 = \sigma_1^2 + \sigma_2^2$. The height estimates $\hat{X}_1$ and $\hat{X}_2$ are regarded to be compatible when the difference $\Delta\hat{X}$ is close to 0;

$$\mu - \sigma \le 0 \le \mu + \sigma .$$

Information Fusion. Suppose we estimate the height H of an object based on the n measurements $X_1, X_2, \ldots, X_n$. The combined estimate, $\hat{X}$, can be viewed as having a conditional probability density function (p.d.f.) $f_{H|X,\Theta}(h|x_1, \ldots, x_n, \theta_1, \ldots, \theta_n)$ [1] where $\Theta_1, \ldots, \Theta_n$ represents environment parameters (i.e. confidence intervals) for the measurements $X_1, \ldots, X_n$. We get

1. Lower case will be used for assignments of random variables.

$$f_{H|X,\Theta}(h|x_1, \ldots, x_n, \theta_1, \ldots, \theta_n)$$
$$= \frac{f_{X|H,\Theta}(x_1, \ldots, x_n|h, \theta_1, \ldots, \theta_n) f_H(h)}{f_{X|\Theta}(x_1, \ldots, x_n|\theta_1, \ldots, \theta_n)}$$
$$= \frac{f_{X|H,\Theta}(x_1, \ldots, x_n|h, \theta_1, \ldots, \theta_n) f_H(h)}{\int_{h'} f_{X|H,\Theta}(x_1, \ldots, x_n|h', \theta_1, \ldots, \theta_n) f_H(h') dh'}.$$

Here, the prior probability distribution $f_H(h)$ of the object height is unknown. If we assume that the object height has a uniform distribution within the possible object height range; *i.e.*,

$$f_H(h) = \begin{cases} c, & \text{if } h \text{ is within a given height range, and} \\ 0, & \text{otherwise,} \end{cases},$$

we get

$$f_{H|X,\Theta}(h|x_1, \ldots, x_n, \theta_1, \ldots, \theta_n)$$
$$= \frac{f_{X|H,\Theta}(x_1, \ldots, x_n|h, \theta_1, \ldots, \theta_n)}{\int_{h'} f_{X|H,\Theta}(x_1, \ldots, x_n|h', \theta_1, \ldots, \theta_n) dh'}.$$

If we assume conditional independence among the observations given height and conditions, we get

$$f_{H|X,\Theta}(h|x_1, \ldots, x_n, \theta_1, \ldots, \theta_n)$$
$$= \frac{\prod_i f_{X_i|H,\Theta_i}(x_i|h, \theta_i)}{\int_{h'} \prod_i f_{X_i|H,\Theta_i}(x_i|h, \theta_i) dh'} = \alpha \prod_i f_{X_i|H,\Theta_i}(x_i|h, \theta_i),$$

where α is a normalizing constant.

Assume $f_{X_i|H,\Theta_i}(x_i|h, \theta_i)$ is a Gaussian distribution with mean x_i and variance σ_i^2. Then the resulting p.d.f $f_{H|X,\Theta}(h|x_1, \ldots, x_n, \theta_1, \ldots, \theta_n)$ is also a Gaussian distribution (a multiplication of Gaussian functions is also a Gaussian function). Simplifying, we get

$$f_{H|X,\Theta}(h|x_1, \ldots, x_n, \theta_1, \ldots, \theta_n)$$
$$= \alpha \prod_i G(h; x_i, \sigma_i^2) = G(h; \hat{h}, \sigma^2)$$

$$\hat{h} = \frac{\sum_i \frac{x_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}} \qquad \sigma^2 = \frac{1}{\sum_i \frac{1}{\sigma_i^2}}.$$

where $G(h; \mu, \sigma^2)$ is a Gaussian function with mean μ and variance σ^2. Therefore, we use $\hat{h}$ as the estimated height.

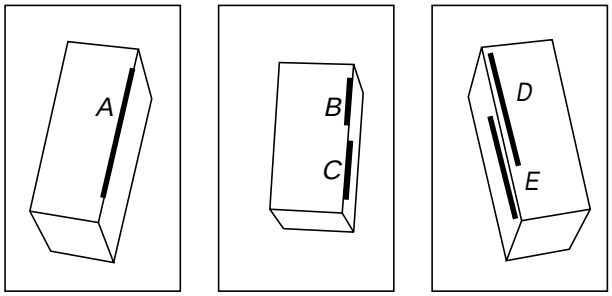


Figure 9. A building from three different views and line support for rooftop.

3.2. 3-D feature generation

We define a 3-D feature to be a group of matched 2-D features. To get a 3-D junction, a matched junction pair is used as a seed. Given a seed, we collect all the junctions from other views which have compatible orientations and heights. We select a single 3-D junction with the most compatible height. A similar method is used to generate 3-D linears, but, we have to deal with possible epipolar alignments. Therefore, only matches with small confidence intervals on the height estimates are used as seeds. Lines also have multiple matches, but the situation is different from that of junctions because of possible presence of collinear lines. For example, in Figure 9, line *A* matches with both *B* and *C* of the second view, as well as *D* and *E* of the third view. Since *B* and *C* are collinears we group them into a single 3-D linear. But *D* and *E* are not; in this case, we choose only one with the most compatible height.

Once 3-D linears and junctions are obtained, *polarities* are assigned to them by looking at the DEM. The polarity of a 3-D junction is positive when the inner side (side of smaller angle) of the junction is higher than the outer side on DEM. The polarity of a 3-D linear is positive when its left-hand side is higher that the right-hand side. When the height difference is not big enough, two 3-D features with positive and negative polarities are generated from a single group of 2-D features. The polarities are used in the hypotheses generation procedure to reduce the complexity of the process.

Figure 10 (a) shows 3-D junctions displayed on one view and Figure 10 (b) is for the 3-D linears. Six images from different view points were used. Out of 1944 2-D junctions of 6 images (324 per an image), 185 3-D junctions are made, and 454 3-D linears are made out of 1080 2-D linears (180 per an image).

(a) (b)

Figure 10. (a) 3-D junctions and (b) linears displayed on an image window.

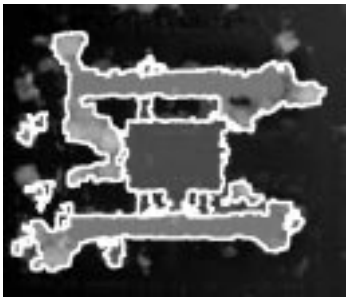


Figure 11. Layers generated by DEM segmentation. 17 layers were found.

4. Hypothesis Generation and Overlap Analysis

Prior to generating rooftop hypotheses, *layers* are generated from the DEM to reduce the search space. We define layers as planar connected surface patches. Layers are generated by segmenting the DEM image. To get layers, adaptive smoothing [13] is applied to an image followed by histogram based segmentation. Figure 11 shows a result of such segmentation.

Once layers are obtained, a search is performed on 3-D features to group them. Given a layer, a search starts from 3-D junctions which are located around the boundary of the layer and have a height compatible with the layer (Section 3.1). A depth-first search is performed on *neighborhood features* which are also consistent with the layer. Neighborhood features consist of forward neighbors and backward neighbors. Forward neighbors of a 3-D linear are either 3-D junctions which have the 3-D linear as their backward branches or forward collinear 3-D linears, and vice versa. The forward neighbor of a 3-D junction is the forward branch. Forwardness and backwardness are determined by the polarities of 3-D features so that hypotheses always generate counter-clockwise rooftops (clockwise for holes). However, grouping only neighbor features does not generate many of desired hypotheses since some of the 3-D features can be missing. Therefore, we also use parallels; two groups which have parallel linears are combined once again. Note that parallels are not used in the neighborhood search because they increase the branching factor enormously.

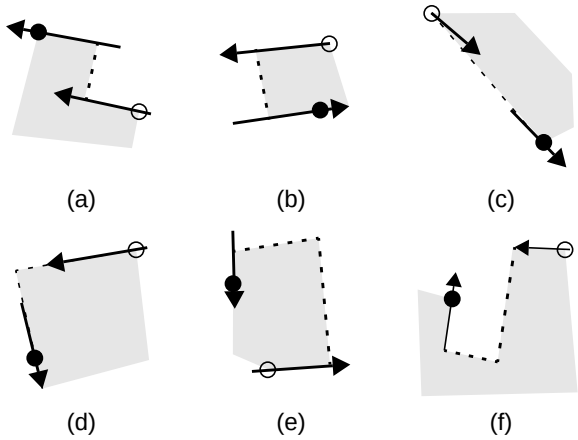


Figure 12. Suggested closures (dashed lines) for (a) parallels with the same polarity; (b) parallels with opposite polarities; (c) collinears; (d) closed L; (e) open L, and; (f) general L. Arrows represent 3-D linears with polarities. Blank circles represent the previous corners and filled circles represent the next corners. Buildings are regarded to lie on the shaded sides.

For every group of 3-D features, we determine proper closures of them to generate rooftop hypotheses. Since junctions always have branches (3-D linears) as neighbors, closures are made between two 3-D linears. Figure 12 shows possible alignments of 3-D linears (arrows) and suggested closures. Since 3-D linears are groups of 2-D linears, the endpoints are not unique. Therefore, we investigate all the closures generated by possible endpoints.

Once a rooftop hypothesis is obtained from a node, supporting evidence is collected for this hypothesis. The evidence consists of basic topological properties, line support, and DEM layer cue coverage. Line support consists of the coverage of the rooftop boundary by actual image lines and the lack of crossing lines [7]. Hypotheses with good evidence are collected as *verified hypotheses*. A linear combination is used to combine the evidence. Since most of the possible combinations of 3-D features are investigated and many of them with good evidence support are verified, a large number of overlapping hypotheses are generated. Overlap analysis is done on the verified hypotheses to get final building hypotheses. Note that we may need to detect more than one building component for a single layer since the DEM is not accurate enough to separate near-by building components of similar heights. Given a layer, first, the rooftop hypothesis of the best evidence support is selected. and overlapping hypotheses are elimi-

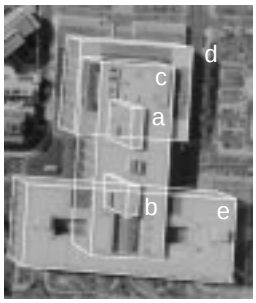


Figure 13. A building at Washington D.C. Building components of each layers interact each other.

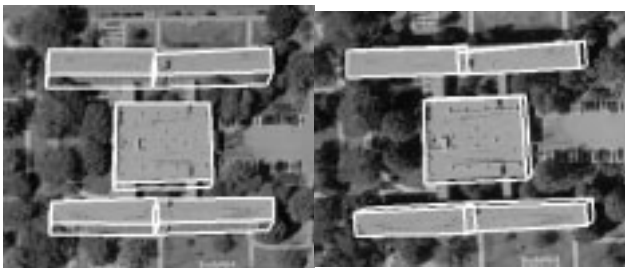
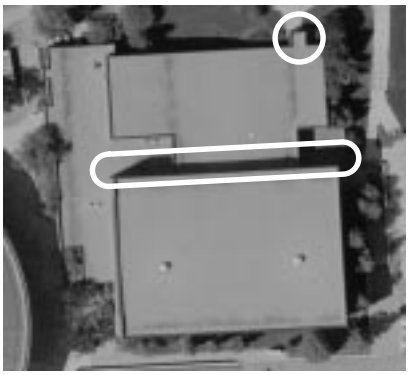


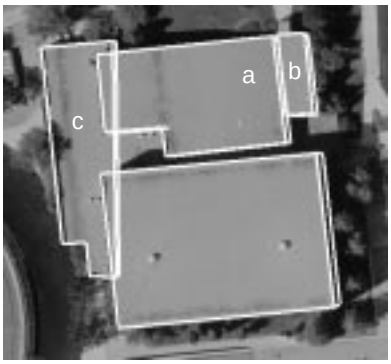
Figure 14. Detected building components. Six views were used.

nated. The same procedure is repeated for the remaining hypotheses until no more hypothesis remains.

The hypothesis generation and overlap analysis described above are performed for each layer. But the processes should not be independent of each other as the results of one layer may affect the results of another layer. The interactions of building components among layers consist of top-down interactions and bottom-up interactions. For example, in Figure 13, a superstructure (c) occludes parts of lower-level buildings (d and e), generates crossing lines, and degenerate line support (top-down). Also, lower-level buildings affects upper-level ones when shadow or wall evidence is used (bottom-up). The shadows of a and b are not cast on the ground but on the roof of c. The wall baseline of a and b also lie on c. Therefore, for a reasoning considering both top-down and bottom-up interactions, both upper-layer and lower-layer building components should be known. One approximation is to apply a relaxation approach. In this paper, we only consider top-down interaction in this paper; shadow and wall evidence is not used although it serves as good features for rooftop hypotheses verification. We perform hypothesis generation and overlap analysis from the highest layer and use the result of upper layer to get proper line support. The line support for possibly occluded part is not used in the verification of lower-layer rooftop hypotheses.



(a)



(b)

Figure 15. (a) A building of the university campus, and (b) a result displayed on another view. Five views were used.

5. Experimental Results

Figure 14 shows the detected building components from the images shown in Figure 3. Two more images (total six views) were used. Although all the rooftops are rectangular, it is hard to detect them correctly without the aid of multiple views because of occlusion by trees and breaking of the roof lines by small superstructures (Figure 4). Also, important lines are missing in some views (Figure 8) and DEM layers are not accurate due to trees of height similar to that of the rooftops (Figure 11). Running time for 3-D feature generation, hypotheses generation and overlap analysis is about 15 minutes on a Sun Ultra 1 workstation. Most of the time was spent on the DEM assisted matching.

A result for another building is shown in Figure 15. Five views were used. The upper building component is broken into two (a, b). This is partly because of the small protrusion of upper side wall (circled). Also, the shadow line in the bottom side prevents 3-D junction and line of a correct height from being made. The right side of the left building component (c) is not correctly described due to

occlusion in all views. Running time for 3-D feature generation, hypotheses generation and overlap analysis is about 18 minutes. More time was spent on the hypothesis generation than that of the previous result (Figure 14) due to the complexity of the building.

6. Conclusion and Future Work

We have shown results on complex buildings with the aid of multiple views. Our system is still under development; however, we feel that the utility of our approach has been demonstrated. In future work, we plan to consider even more complex structures. This will require procedures for analyzing interaction between layers, and the use of shadow and wall evidence for superstructures. Effective methods to describe gabled rooftops also need to be developed.

References

[1] A. Gruen and R. Nevatia (editors), *Computer Vision and Image Understanding: Special Issue on Automatic Building Extraction from Aerial Images*, vol. 72, no. 2, 1998.

[2] A. Huertas and R. Nevatia, "Detecting Buildings in Aerial Images," *Computer Vision, Graphics and Image Processing*, vol. 41, no. 2, pp. 131-152, 1988.

[3] R. Irving and D. McKeown, "Methods for Exploiting the Relationship Between Buildings and Their Shadows in Aerial Imagery," *IEEE Trans. Systems, Man and Cybernetics*, vol. 19, no. 6, pp. 1564-1575, November/December 1989.

[4] J.C. McGlone and J. Shufelt, "Projective and Object Space Geometry for Monocular Building Extraction," *Proc. IEEE Computer Vision and Pattern Recognition*, Seattle, WA, pp. 54-61, 1994.

[5] C. Lin and R. Nevatia, "Building Detection and Description from a Single Intensity Image," *Computer Vision and Image Understanding*, vol. 72, no. 2, 1998.

[6] M. Roux and D. McKeown, "Feature Matching for Building Extraction from Multiple Views," *Proc. IEEE Computer Vision and Pattern Recognition*, Seattle, WA, pp. 46-53, 1994.

[7] S. Noronha and R. Nevatia, "Detection and Description of Buildings from Multiple Aerial Images," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.588-594, 1997.

[8] R. Collins, C. Jaynes, Y.-Q. Cheng, X. Wang, F. Stolle, E. Riseman and A. Hanson, "The Ascender System: Automated Site Modeling from Multiple Aerial Images," *Computer Vision and Image Understanding*, vol. 72, no. 2, pp. 143-162, 1998.

[9] A. Fischer, T. Kolbe, F. Lang, A. Cremers, W. Förstner, L. Plümer and V. Steinhage, "Extracting Buildings from Aerial Images Using Hierarchical Aggregation in 2D and 3D," *Computer Vision and Image Understanding*, vol. 72, no. 2, pp. 185-203, 1998.

[10] M. Cord, M. Jordan, J.-P. Cocquerez and N. Paparoditis, "Automatic Extraction and Modelling of Urban Buildings from High Resolution Aerial Images," *Proc. ISPRS Automatic Extraction of GIS Objects from Digital Imagery*, Munich, Germany, pp. 187-192, September, 1999.

[11] C. Baillard and A. Zisserman, "Automatic Reconstruction of Piecewise Planar Models from Multiple Views," *Proc. IEEE Computer vision and Pattern Recognition*, Fort Collins, Colorado, pp. 559-565, June, 1999.

[12] S. Noronha, *3-D Building Detection and Description from Multiple Intensity Images using Hierarchical Grouping and Matching of Features*, Ph.D. Thesis, Computer Science, University of Southern California, 1998.

[13] P. Saint-Marc and G. Medioni, "Adaptive Smoothing for Feature Extraction," *Proc. DARPA Image Understanding Workshop*, pp. 1100-1113, 1988.