3D Building Detection and Reconstruction from Aerial Images Using Perceptual Organization and Fast Graph Search

Dong-Min Woo† and Quoc-Dat Nguyen*

Abstract – This paper presents a new method for building detection and reconstruction from aerial images. In our approach, we extract useful building location information from the generated disparity map to segment the interested objects and consequently reduce unnecessary line segments extracted in the low level feature extraction step. Hypothesis selection is carried out by using an undirected graph, in which close cycles represent complete rooftops hypotheses. We test the proposed method with the synthetic images generated from Avenches dataset of Ascona aerial images. The experiment result shows that the extracted 3D line segments of the reconstructed buildings have an average error of 1.69m and our method can be efficiently used for the task of building detection and reconstruction from aerial images.

Keywords: Aaerial images, Building detection, Building reconstruction, Perceptual grouping

1. Introduction

Building detection and reconstruction from aerial images is one of the most challenging tasks in computer vision. It has been used widely in various applications including traditional applications such as cartography and photo-interpretation as well as recent applications including mission planning, urban planning, computer graphics, and virtual reality.

There are two main problems that need resolution in any building detection approach. The interested objects must be segmented from the background and the fragmented line segments of the interested object's boundaries should be grouped to human-made structures. These are challenges because the objects of interest could be partly occluded by the presence of vegetation, shadows, roadways, and other objects. Moreover, lines and corners of objects are often fragmented and missed due to the typical failures of low level feature extraction. These tasks have been intensively studied in the field of computer vision.

Early approaches tried to use a singe image only [1], [2]. This direction has some restrictions such as the difficulty in inferring 3D information from one image and the existence of ambiguities in the detected buildings that can be only resolved by feature matching in multiple images. As such, multiple aerial images can only be obtained with extra cost. Most of the recent work in this area has focused on multiple-view analysis [3-5].

Mohan and Nevatia [6] proposed an approach for detecting and describing buildings in aerial images using perceptual grouping. They demonstrated the usefulness of the structural relationships called collated features which can be explored by perceptual organization in complex image analysis. All reasonable feature groupings are first detected and the candidates are then selected by a constraint satisfaction network. But this approach involves all extracted line segments in the image. Consequently, it requires a significant computational effort. It also depends on the accurate extraction of line segments.

Huertas [7] suggested using extracted cues from the IFSAR data, while Kim [8] utilizes commercial DEM (Digital Elevation Map) to solve the problem of segmentation of interested objects. The extracted cues do not give us the shape of the buildings. However, they can give us the idea where the buildings are located in the image. Unfortunately, it is not easy to have IFSAR data or DEM image in all cases.

Some approaches such as Lin [2] and Noronha [4] use hypotheses and verify paradigms based on perceptual grouping to solve the second problems. Hypotheses are generated by a hierarchical perceptual grouping process and verified by the evidence of visible walls and expected shadows. But the system needs to make several decisions in the selection and verification process based on simplicity and intuitive judgments that affects much of the final result. In monocular analysis, Jaynes [9] proposed task driven perceptual organization for extraction of rooftop. Features such as corner and line segments are first extracted and assigned a certainty value. Then features and their grouping are stored in a feature relation graph. Close cycles in the graph represent the grouped polygon

† Corresponding Author: Dept. of Information Engineering, Myongji University, Korea. (dmwoo@mju.ac.kr)
* HSBC, Vietnam. (iavmm@yahoo.com)
Received 15 January, 2008 ; Accepted 21 June, 2008

hypotheses. The independent set of closed groups that have maximum sum of certainty values of its parts is the final grouping choice. This approach is limited on rectangular buildings and tends to have false hypotheses in complexity images.

In this paper, we propose a new method for rectilinear building detection and reconstruction using two overlapping aerial images. We use hypothesis generation and selection based on perceptual organization strategy to solve the building detection task. The key idea is that we use the proposed suspected building regions extracted from the disparity map for obtaining the location of interested objects in the image. This building location information helps to remove the unnecessary line segments in the low level feature extraction result and thus reduces computational complexity and false hypotheses in later steps. Additionally, hypothesis selection is carried out by graph searching for close cycles in an undirected graph. Compared with Jaynes's approach, our method detects corners from filtered low level features before constructing the graph, whereas corners are extracted using pre-defined corner masks and each corner can take part in many different close cycles in his approach. So our system can significantly reduce computational complexity and false hypotheses. Moreover, we expand the condition required for a link between two corners to be formed and thus enable our system to detect the rectilinear buildings that Jaynes's approach does not detect. For building reconstruction, we retrieve 3D information of the buildings using 3D triangulation with the known geometric parameters of image acquisition.

The remainder of this paper is organized as follows: An overview of the system is shown in Section 2. Section 3 describes the generation of epipolar images, disparity map, and suspected building regions. In Section 4, the principle of low level features processing and rooftop hypothesis is introduced. Section 5 presents experimental results on an aerial image data set. At last, the conclusion and future work are given in Section 6.

2. System Overview

Fig. 1 shows the main components in our system. The epipolar images are generated from the aerial images by the epipolar resampling process. We obtain the disparity map between the epipolar pairs by stereo matching using area-based matching with non-parametric technique. From the disparity map, we generate the DEM as a 3D terrain model. The building location information extracted from the disparity map is used to remove the unnecessary line segments extracted in the low level process. After 2D lines are generated, perceptual grouping is applied to the filtered line segments in order to obtain the structural relationship features such as parallel line segment pairs and U-shapes. These can be used to generate rooftop hypotheses. Among the generated hypothesis, the candidate rooftop is selected by searching close cycles in the undirected graph. Finally, we retrieve 3D buildings by using 3D triangulation for each line segment of detected rooftops.

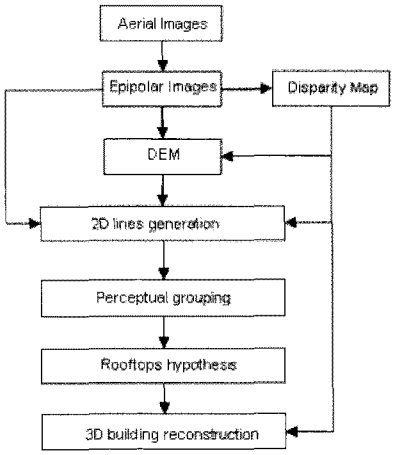


Fig. 1. System overview

3. Building Region Extraction

3.1 Epipolar Images

The goal of this step is to generate the epipolar images from the two original aerial images. For each point in one of the epipolar images, we only need to search for the points in the same row index of the other image for the corresponding point. In this way, matching can be reduced to a one-dimensional search instead of a two-dimensional search. Consequently, the system can significantly reduce computational complexity in solving the correspondence problem in the disparity map step.

To generate epipolar images, we project four corners of the original reference R and wrap W images onto the overlap plane Z. Next, we calculate overlap area, the intersection of the plane Z, with the region in the object space that is visible in both images. Then, epipolar lines that bound the left, right, top, and bottom of the overlap area are found. Each pixel in the row of the resampled image corresponding to the epipolar line segment between the right and left bound are back projected onto the R and W images. The corresponding coordinate of each pixel in the epipolar images can be calculated by resampling the R and W image along a straight line that passes through the back projected start and end point of that line. Finally, we obtain the pixel's intensity value of epipolar images by using interpolation.

3.2 Disparity Map

The goal of the image matching process is finding a match between the pixels in the first (reference) R and second (wrap) W image such that the pixel located at *(i, j)* in the R image and a pixel located at *(i+I(i, j), j+J(i, j))* in the W image view the same point in an object space, i.e., *W(i+I(i, j), j+J(i, j)) -> R(i, j)*, where *I(i, j)* is horizontal disparity map, and *J(i, j)* is vertical disparity map. The index i (column index) is measured along scan lines and the index j (row index) is measured across scan lines. In this paper, we use resampled epipolar images as the input so that $J(i, j) = 0$ for all i and j, and the relation reduces to *W(i+I(i, j), j) -> R(i, j)*.

Considering the correspondence problem, there are two popular approaches. The first one is Normalized Cross Correlation (NCC), which is an area-based matching typical metric approach and the second one is non-parametric technique with census transform [10]. We employ the census transform due to its preservation of the edges and simple computational complexity.

To find the accurate disparity map, we employed a multi-resolution scheme, referred to as hierarchical, or pyramid processing. For each resolution scheme, the correspondence problem is solved by first computing a census transformed image and then using Hamming distance correlation on that image. The census transformation maps the local region surrounding a pixel to a bit string with pixels having lesser intensities. For example, in a window surrounding a pixel, if a particular pixel's value is less than the centre pixel, the corresponding position in the bit string will be set to 1; otherwise it is set to 0. After that, two census transformed images will be compared using a similarity metric based on the Hamming distance, which is the number of bits that differ in the two correlation window bit string. The Hamming distance [11] is summed over the window, as in (1).

$$\sum_{(u,v)\in W} Ham\min g(I_1'(u,v), I_2'(x+u, y+v)) \qquad (1)$$

where I_1' and I_2' represent the census transforms of I_1 and I_2. W is the correlation window.

3.3 Suspected Building Regions

It is usually difficult to separate interested objects from 2D line segment collection obtained in low level features extraction. The boundary of interested objects, the buildings, can be partly occluded by vegetation, shadows, and other objects. In the rooftop hypothesis process, these fragmented boundaries and the presence of roads, vehicles, etc., can make false hypotheses including unwanted rooftops and rooftops of the wrong shape. This causes not only significant computational effort in processing but also erroneous final results. To solve this problem, the system should be able to detect line segments that are within or near buildings in the image. Here, we use suspected building regions that are extracted from the disparity map. The suspected building regions are areas in which pixel values change in comparison with the surrounding area. The difference in pixel values between the suspected building region and surrounding areas indicates the difference of elevation values. It specifies the existence of higher objects such as buildings, trees, etc., in those regions. In other words, these regions can give us the information of where the buildings are located.

These regions could be extracted by using a simple height threshold technique. Their boundaries are extracted by convolving the disparity map with a Laplacian-of-Gaussian filter and then employing connected component analysis to achieve zero-crossing pixels' coordinates in the convolution output. We have LoG as an operator or convolution kernel defined as in (2).

$$\begin{aligned} LoG(x,y) &= \Delta G_\sigma(x,y) \\ &= \frac{\partial^2}{\partial x^2} G_\sigma(x,y) + \frac{\partial^2}{\partial y^2} G_\sigma(x,y) \\ &= -\frac{1}{\pi\sigma^2}\left[1 - \frac{x^2+y^2}{2\sigma^2}\right] e^{-\frac{x^2+y^2}{2\sigma^2}} \end{aligned} \qquad (2)$$

4. 3D Rooftop Model Generation

4.1 Low Level Features Extraction

To detect 2D lines from an epipolar image, edge detection is carried out first and then 2D lines are formed from the edges. To do so, we employed the Canny edge detector, since it is optimal according to the criteria where edge is defined and comes up with thin edges. To obtain a 2D line segment, we used the Boldt algorithm [12] based on token grouping. This method extracts a basic line element, token, in terms of the properties of line A and constructs a 2D line using the grouping process. It is efficient in detecting 2D lines of large structure appearing in urban images.

4.2 Grouping and Filtering Process

Suspected building regions are used to remove line segments that are outside or far from interested object boundaries meanwhile the needed line segments are still kept for later processing steps.

Then, we group the closely parallel linear segments since they usually represent a linear structure of objects in an image, like the border of a roof or the divider between ground terrain and building, by using a "folding space" between two line segments. If both line segments are inside the folding space, two line segments can be replaced by a single line in which orientation is the longer line segment orientation and length is the total length of two segments. After this process, each group of the closely overlapping and parallel line segments is represented by one single line.

Fig. 2 gives the typical case of closely parallel linear segment grouping. These linear segments either are, or nearly, parallel lines. So the first condition is that the angle between them should be from 0^0 to 10^0. If two line segments are fragmented lines from one edge, these line segments must be close and should be inside a folding space created by them.

The U shaped structure in Fig. 3 is used to detect candidates for rooftop hypothesis generation. Any line segment in a set of parallel lines with a U shaped structure is a candidate kept as input for hypothesis generation, otherwise that line segment will be removed.

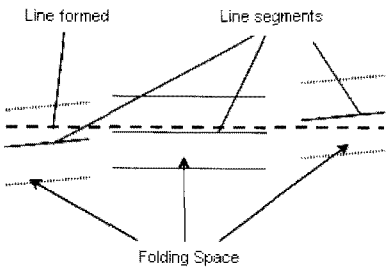


Fig. 2. Folding space

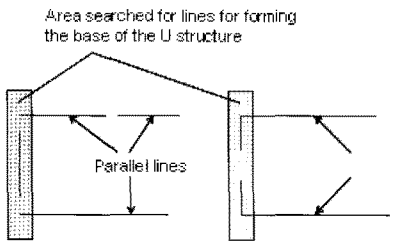


Fig. 3. U-structure

4.3 Corner Detection

Corners are calculated as the intersection of two line segments that have an angle from 80^0 to 100^0 with one of them having the nearest distance to another one. We define four types of corner. They are labeled as I, II, III, and IV, as shown in Fig. 4. Each corner has an attribute to indicate whether it is L-junction or T-junction. This attribute is used to decide whether two different corners have a connection or not. For example, if a corner's label is I and type is L-junction, it connects to any type of corner. However, it prefers connecting to a corner which label is II or IV. If that corner is T-junction, it can only connect to a corner which label is II or IV. This rule is used in hypothesis generation to build collated features.

With the flexible connection between corners, our method is able to detect rectilinear rooftops. Fig. 5 reveals some examples of corner detection. A, B, E, F, and G are L-junctions while C and D are T-junctions.

Fig. 4. Corner labeling

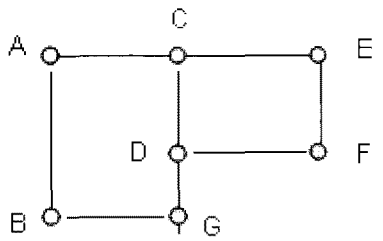


Fig. 5. Corner detection

4.4 Rooftop Hypothesis Generation

A collated feature is a sequence of perceptually grouped corners and line segments. Here, collated features are constructed from filtered line segments and corners obtained from the filtering and grouping process. This reduces computational effort and false hypotheses.

Hypotheses are formed by the alternation of corners and line segments that form collated features. In a collated feature, two corners have connectivity only if they satisfy the corner relation condition and they are the nearest appropriate corner to each other. Beside, every corner connects to only one corner on each of its line segment directions. Hypothesis generation is performed by constructing the feature graph. Construction of the graph can be seen as placing corners as nodes and edges between nodes if there is the relation between the corresponding corners in the collated features. When a node is inserted into the graph, the system looks into the remaining nodes to determine whether any node has a relation with the inserted node. If some nodes satisfy the connectivity relation rules, those nodes are inserted into the graph and the system creates an edge between them. In the example shown in Fig. 5, C is T-junction, and it can connect to D, A, and E. Meanwhile, A can connect to B, C, and E but C is

nearer than E towards A on the line segment AE so that A only connects to B and C. Consequently there are two collated features, ACGB and CEFD, in Fig. 5.

4.5 Rooftop Hypothesis Selection

The graph is the place to store features and their groupings. Feature as corner is node in the graph and relations between corners are represented with an edge between the corresponding nodes. Closed cycles in the graph represent the rooftop candidates. Hypothesis selection can be seen as a simple graph search problem. The close cycles in the graph are rooftops that we need to detect. Fig. 6 shows a graph constructed from the example in Fig. 5. Corner C and corner D are T-junctions so that there are two nodes in the graph for each corner, node C1 and C2 for corner C and node D1 and D2 for corner D. There are two close cycles, C1 and C2, as shown in Fig. 6.

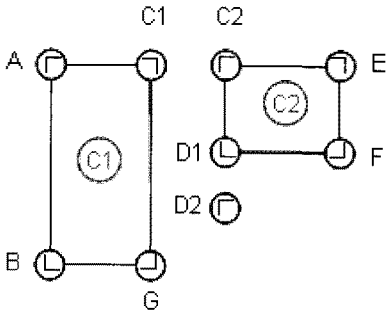


Fig. 6. Feature graph

4.6 3D Building Reconstruction

3D triangulation is used to generate 3D line segments. The relationship between point k located at $X_k = (X_k, Y_k, Z_k)$ in model/object space and the projection of point k located at $x_{Lk} = (x_{Lk}, y_{Lk}, f_L)$ in the image of camera L is depicted in (3).

$$\begin{bmatrix} X_k \\ Y_k \\ Z_k \end{bmatrix} = \begin{bmatrix} X_L^0 \\ Y_L^0 \\ Z_L^0 \end{bmatrix} + \lambda_{Lk} \begin{bmatrix} m_{L11} & m_{L21} & m_{L31} \\ m_{L12} & m_{L22} & m_{L32} \\ m_{L13} & m_{L23} & m_{L33} \end{bmatrix} \begin{bmatrix} x_{Lk} \\ y_{Lk} \\ -f_L \end{bmatrix} \quad (3)$$

Also, it is expressed as a vector form, as in (4).

$$X_k = X_L^0 + \lambda_{Lk} * m_L * x_{Lk} \quad (4)$$

$X_L^0 = (X_L^0, Y_L^0, Z_L^0)$ is the model space coordinates of the focal point of camera L, f_L is the focal length of camera L, λ_{Lk} is the scale factor for point k projected on the focal plane of camera L, and m_L is the rotation matrix between the image space coordinate system and the model space coordinate system.

We have a system of equations for five variables from each pair of points in two images. Solving that system of equations we have the real 3D coordinates of the selected points in two images. As a result, we have 3D line segments from the corresponding 2D line segments.

5. Experimental Results

The experimental environment was set up based on Ascona aerial images of the Avenches area. Since this area's 3D model is supplied as ground truth data, we can evaluate the quantitative accuracy for the 3D rooftop model generated by the proposed method. Two aerial images as revealed in Fig. 7 are used as a set of stereo images for the experiments.

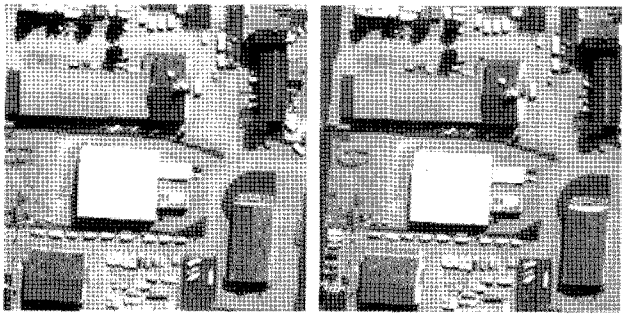


Fig. 7. Aerial images used as a set of stereo images

The result of the epipolar resampling process is shown in Fig. 8. We generate suspected building regions from the map to reduce unnecessary line segments. To find the accurate disparity map, we employed the multi-resolution scheme with four different resolutions, where scaled image sizes are equal to original size divided by 2^n, n = (0,1,2,3). The corresponding correlation window sizes are 3x3, 5x5, 7x7, and 9x9, while the census transform window sizes are 3x3, 5x5, 7x7, and 9x9. The final disparity map is presented in Fig. 9. Fig. 10 gives the generated DEM image by using area-based stereo matching [13], and also shows its ground truth image.

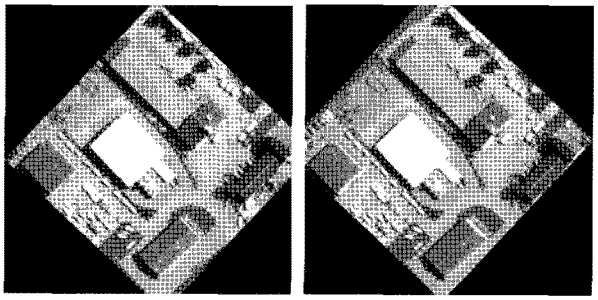


Fig. 8. Example of epipolar images

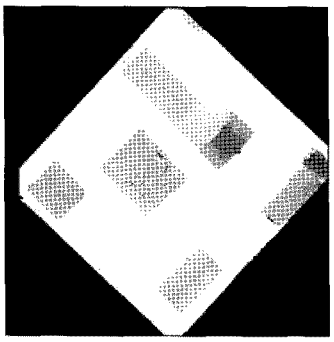


Fig. 9. Example of disparity map

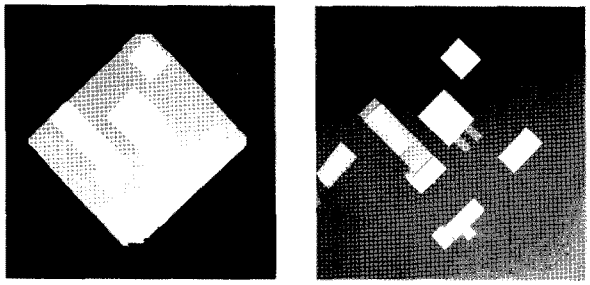


Fig. 10. Generated DEM image and ground truth image

Fig. 11 gives the line segments obtained from the low level feature extraction process. The number of extracted line segments is about 1425. To remove unnecessary line segments, we use the suspected building regions extracted from the disparity map as shown in Fig. 12.

The result of unnecessary line segment removal is presented in Fig. 13. The remaining line segments total about 405. At this point, the perceptual filtering and grouping process is employed to obtain line segments which can be part of any U-structure group. The close parallel line segments that are inside their folding space will be grouped into one representation line. The line segments that are part of a collection of line segments forming the U-structure will be used to generate hypotheses in the next step. Fig. 14 shows the line segments forming U-structures in a collection of line segments. The colors indicate which U-structure group that the line segment belongs to.

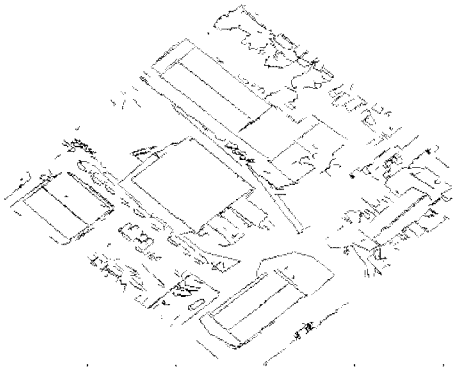


Fig. 11. Example of low level extraction result

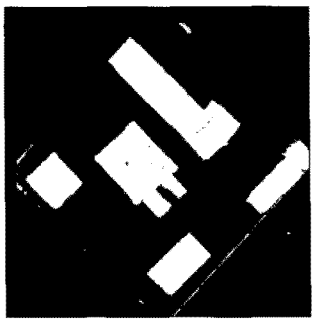


Fig. 12. Example of suspected building regions

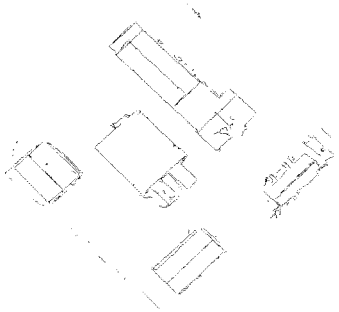


Fig. 13. Example of filtered line segments

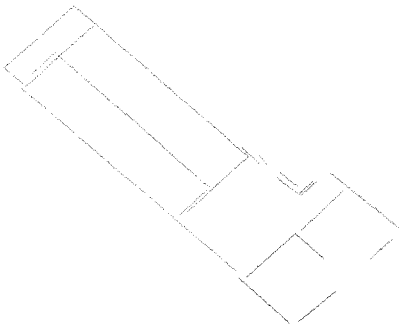


Fig. 14. Example of U-structures

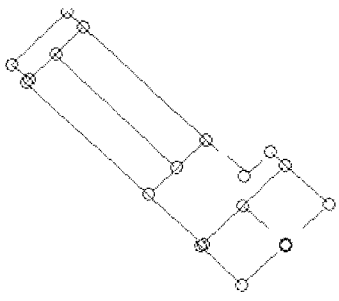


Fig. 15. Example of corner detection

The corners are calculated from the intersection of the line segments that satisfy two conditions: their angle is from 85^0 to 95^0 and one of them has the nearest distance to another one. Fig. 15 provides extracted corners from the line segment collection.

We can use the obtained corners and line segments from the previous steps to build the collated features. In order to have a link between each other, two corners must satisfy

the connecting relation of corner type and the required condition of their distance. Another important rule that helps to define the corner connectivity is that on each line segment of a corner, there is only one corner that has connection with it. Fig. 16 shows the collated features obtained from the line segment collection.

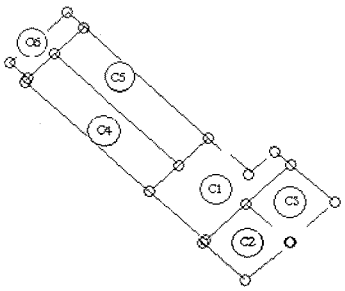


Fig. 16. Example of collated features

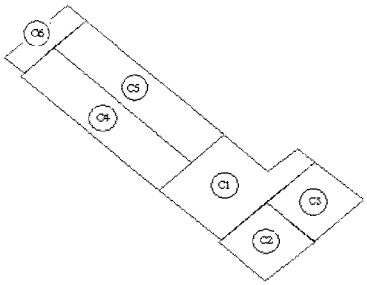


Fig. 17. Example of close cycles

The collated features are used to construct a graph by placing a corner as a node and two corners of a line segment as an edge between two nodes if there is a relation between the corresponding corners in the collated features. Closed cycles in the graph represent the possible rooftops. Hypothesis selection becomes the searching of close cycles in the graph. Fig. 17 shows the close cycles selected from the line segment collection.

Fig. 18 presents the rooftop detection result of the entire area. There is a building located near the border of the epipolar image that the system cannot detect correctly due

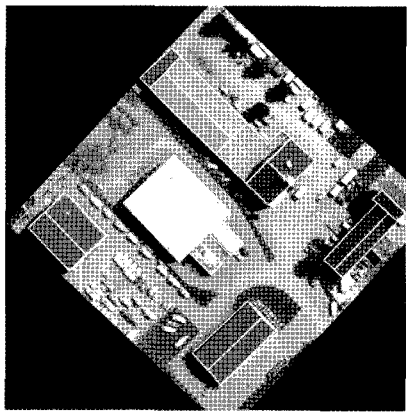


Fig. 18. Example of detected rooftops

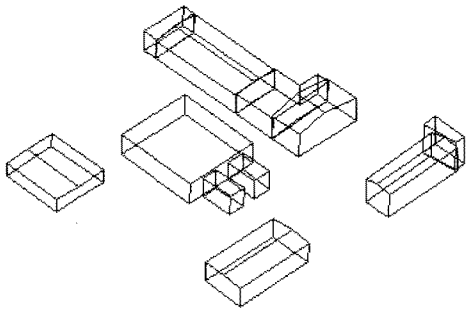


Fig. 19. Example of 3D building reconstruction

to missing line segments in low level extraction step. The result of the remaining building is very good. From the detected rooftop and the known geometric parameters of image acquisition, we reconstructed 3D building using 3D triangulation as shown in Fig. 19.

To represent the quantitative accuracy of 3D building reconstructed by our approach, we obtained the error by calculating the average distance between the extracted 3D line segments and the ground truth line segments as in (5).

$$E = \frac{\sum \frac{e_{1i} + e_{2i}}{2} \times d_i}{\sum d_i} \tag{5}$$

In (5), e_{1i} is the distance from the starting point of line segment i to the ground truth 3D line, while e_{2i} is the distance from the end point of line segment i to the ground truth 3D line and d_i is the length of line segment i. Error calculation indicates that the average error of the reconstructed buildings is 1.65m while the error of the corresponding digital elevation model is 1.93m.

6. Conclusion

A new approach to detect and reconstruct buildings using perceptual organization from two aerial images has been suggested. Low level feature extraction is not applied in the original images but from the epipolar images that help to reduce the search effort in the image matching process. The proposed suspected building regions are used to remove the unnecessary line segments before generating rooftop hypotheses to help in reducing computational complexity and false hypotheses. Using the undirected feature graph, the selection of rooftop hypotheses becomes a simple graph searching for close cycles. The experimental result shows that the proposed method can be very effectively utilized for the rectilinear structures of an urban area.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MOST) (Grant No.: R01-2007-000-20330-0).

References

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

[2] C. A. Lin and R. Nevatia, "Building detection and description from a single intensity image," *Computer Vision and Image Understanding*, vol.72, no.2, pp.101-121, 1998.

[3] A. Fischer, T. Kolbe, F. Lang, A. Cremers, W. Forstner, L. Plumer, 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.

[4] S. Noronha and R. Nevatia, "Detection and modeling of buildings from multiple aerial images," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 23, no. 5, pp. 501-518, 2001.

[5] 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.

[6] R. Mohan and R. Nevatia, "Using perceptual organization to extract 3D Structure," *Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 11, pp. 1121-1139, 1989.

[7] A. Huertas, Z. Kim, and R. Nevatia, "Use of cues from range data for building modeling," *Proc. DARPA Image Understanding Workshop*, 1998, pp. 577-582.

[8] Z. Kim and R. Nevatia, "Automatic description of complex buildings from multiple images," *Computer Vision and Image Understanding*, vol. 96, pp. 60-95, 2004.

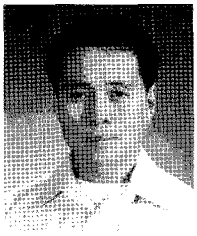
[9] C. Jaynes, F. Stolle, and R. Collin, "Task driven perceptual organization for extraction of rooftop polygons," *IEEE Workshop on Application of Computer Vision*, pp. 152-159, 1994.

[10] R. Zabih and J. Woodfill, "A non-parametric approach to visual correspondence," *Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[11] J. Banks, M. Bennamoun, and P. Corke "Non-parametric technique for fast and robust stereo matching", in *Proceedings of IEEE TENCON – Speech and Image Technologies for Computing and Telecommunication*, 1997, pp. 365-368.

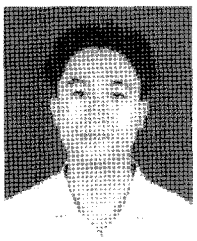
[12] M. Boldt, R. Weiss, and E. Riseman, "Token-based Extraction of Straight Lines," *IEEE Trans. Systems Man Cybernetics*, vol. 19, 1989, pp.1581-1594.

[13] H. Schultz, "Terrain Reconstruction from Widely Separated Images," in *Proceedings of SPIE*, vol. 2486, 1995, pp. 113-122.



Dong-Min Woo

He received his B.S and M.S degrees in Electronic Engineering from Yonsei University, and his Ph.D. in Electrical Engineering from Case Western Reserve University. His research interests are computer vision and satellite image analysis.



Quoc-Dat Nguyen

He received his B.E. degree in Computer Engineering from Ho Chi Minh University of Technology, and his M.S in Information Engineering from Myongji University. His research interests are image processing and analysis.