Task Driven Perceptual Organization for Extraction of Rooftop Polygons*

Christopher O. Jaynes, Frank Stolle and Robert T. Collins
Department of Computer Science
University of Massachusetts
Amherst, MA 01003

Abstract

A new method for extracting planar polygonal rooftops in monocular aerial imagery is proposed. Structural features are extracted and hierarchically related using perceptual grouping techniques. Top-down feature verification is used so that features, and links between the features, are verified with local information in the image and weighed in a graph. Cycles in the graph correspond to possible building rooftop hypotheses. Virtual features are hypothesized for the perceptual completion of partially occluded rooftops.

Extraction of the "best" grouping of features into a building rooftop hypothesis is posed as a graph search problem. The maximally weighted, independent set of cycles in the graph is extracted as the final set of roof boundaries.

1 Introduction

Extraction of polygonal structures from an aerial image is an important step in building detection and model construction. We would like to determine the shape and location of buildings within an aerial image robustly and accurately by extracting the polygons that define rooftop boundaries.

Industrial and urban centers are typically complex and cluttered with structure. Rooftops within the image can be partly occluded by other buildings, shadows, trees, and other features of the surrounding terrain. Rooftop edges can be of low contrast even when they are visible, and may meet several other rooftop boundaries in the projected image. Aerial imagery may contain strong perspective effects and may be captured under a variety of lighting and weather conditions. A single rooftop may be cluttered by air vents, patches of dirt or by other rooftops in the case of multi-layered buildings.

*This work was funded by RADIUS via ARPA contract TEC DACA76-92-C-0041 and NSF grant No. CDA-8922572

Despite these difficulties, a successful system will discover rooftops that can be used for further image understanding tasks. That is, polygon extraction should be detailed (several buildings that lay close together should not be grouped into a single polygon) and accurate (a complex rooftop should not be broken into several polygons).

1.1 Previous work

There has been a large amount of work in interpretation of aerial imagery using both monocular and multiple image strategies. A basic overview of past research appears in [17].

Previous attempts at the building detection problem included contour following techniques augmented with domain specific knowledge about structure [8, 11]. Typically, these methods are restricted to images where edge contours are clear. In many typical scenes, lines are broken and can branch into many different paths, which causes an exponential path following problem.

Segmentation followed by analysis of region boundaries was used to search for human-made structure in aerial images by [5]. These techniques only work well for simple scenes where rooftops appear as complete, uniform regions, or when the expected shape of rooftop regions is known.

The MOSAIC system [6] constructed partial, three dimensional wire frame models from both monocular and stereo imagery of a scene. Trihedral corner junctions were computed and used to hypothesize a complete model through complex geometric reasoning.

Perceptual organization techniques provide the impetus for many building extraction systems. Reynolds and Beveridge [15] defined a focus of attention mechanism for human-made structures based on connected components of spatially proximate parallel and orthogonal line segments. Huertas, Lin and Nevatia [7] extract parallel lines and orthogonal intersections that

are further grouped into rectangular buildings. Building hypotheses are verified through the use of shadow information.

In a similar approach, Mohan and Nevatia [14] extract lines as low level features and compute corners from the line data. Perceptually organized sequences of line and corner features are grouped into *collated features* such as rectangles. All possible sets of collated features are extracted and placed into a constraint satisfaction network. This approach avoids many of the domain assumptions necessary for earlier work and allows for multiple, competing hypotheses. Unfortunately, the work is limited to nadir imagery and on accurate extraction of line segments.

The BABE system [13] performs perceptual grouping of lines and orthogonal corners into chains and rectangles to form building hypotheses that are later verified using shadow information. Although it was originally designed for nadir imagery, recent enhancements to BABE generalize it to oblique views by incorporating rigorous photogrammetric camera models and vanishing point information [12].

2 Task driven organization

Perceptual organization is a powerful method for locating and extracting structure in natural scenes. In our approach, low level features are grouped to form collated features which are then used to hypothesize the final groupings. However, besides this bottom-up approach, each level of the hierarchy may search for features in a top-down manner. For example, suppose we are searching for proximate, parallel groups of three lines as collated features. Then, upon discovery of two parallel lines, a local low level line finder would be invoked as a top-down process to aid in the higher level grouping procedure.

Task driven perceptual organization extends the grouping hierarchies in [7, 14] to a bidirectional approach. The system is not limited to a nadir views; we do, however, assume the camera lens and pose parameters are known or have been precomputed.

2.1 Overview

The system proceeds in three steps; low level feature extraction, collated feature detection, and hypothesis arbitration. Each module generates features that are used at the next phase and interacts with lower level modules through top-down feature extraction. Figure 1 shows the three modules and their basic interactions.

The low level features in this system are perspective image projections of orthogonal corners and straight line segments in the scene.[1] Of course, this restricts us to final groups representing polygonal roofs with orthogonal corners, but a vast majority of urban and industrial building rooftops fall into this category. Mid-

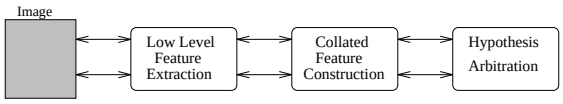


Figure 1: Task Driven Feature Grouping

level collated features are sequences of corners and lines that are grouped together to form *chains*. Collated features are a generalization of the parallels and U-structures discussed in [14]. A single low level feature can take part in many different collated features. High-level polygon hypotheses are formed from closed chains. Chains can be closed during their extraction as collated features, or during the hypothesis generation phase when missing corner or edge features in a chain are searched for in a top-down manner.

Because single collated features can be part of several closed polygons, the final set of closed polygons must be searched for the "best" independent set of closed chains. This is done using certainty measures that are maintained throughout the entire grouping process. Throughout the paper, we use $\kappa(F)$ to denote the certainty of feature F. As each feature is extracted it is assigned a certainty; the final grouping choice is then found as the independent set of closed chains that maximizes the certainty κ.

2.2 The feature relation graph

Features and their groupings are stored in a graph structure called the *feature relation graph*. Low level features are nodes in the graph, and relations between features are represented with an edge between the corresponding nodes. Nodes have a certainty measure that represents the confidence of the low level feature extraction routines. Edges are weighted with the certainty of the grouping that the edge represents.

Chains are represented by paths in the feature relation graph and inherit a certainty measure from the nodes and edges along the path. Cycles in the graph represent grouped polygon hypotheses.

3 Low level feature extraction

[1] That is, while the corners are orthogonal in the world they are not necessarily orthogonal in the image.

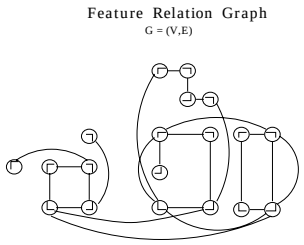
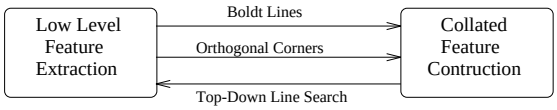


Figure 2: Features are stored in the feature relation graph. Low level features are represented as nodes, collated features as paths, and final polygons as closed cycles.



The importance of particular image features depends on the domain. Because we are interested in the detection of human-made structure, straight lines and orthogonal corners were used as low level features.

3.1 Straight lines

Straight lines are extracted using two different methods. The primary, bottom-up method for extracting low level straight line features is the Boldt algorithm [2]. This algorithm hierarchically groups edgels into progressively longer line segments based on proximity and collinearity constraints. Figure 3 shows the Boldt lines extracted from a typical aerial urban scene.

Boldt lines are used by the higher level features to aid in perceptual grouping tasks and are important clues to human-made structure in the image.

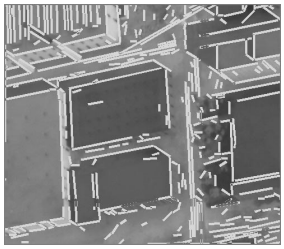


Figure 3: Boldt Lines

Boldt lines are assigned a certainty measure that is calculated during their extraction. The line certainty depends primarily on the contrast of the edge and on the least-squares residual error of the Boldt line fit to the grouped edges. For a detailed description of Boldt line certainty, see [2].

Another line detection scheme is used for top-down grouping verification. These *local lines* are extracted when possible groupings between features are being considered and supporting Boldt lines are absent. This approach provides the power of perceptual grouping without restricting ourselves to a single set of globally extracted features.

For example, while attempting to construct a chain feature, it may be necessary to discover if a local line lies between two corners. Each pixel in the image along a connecting line between the two corners is classified as a supporting edgel or nonedgel according to the image intensity gradient, as computed by an oriented Sobel mask, and the variance in the gradient magnitude. This is performed within a rectangular search window between the two corner features.

The final strength of the local line is determined by dividing the number of edgels, E, by the number of pixels in the search line, N. A threshold can then determine if there is enough edgeness to consider this to be a line. For the results in this paper, local lines were found when $\frac{E}{N} > 0.7$.

This algorithm obviously is only valid for top-down line extraction. It is used to verify that a grouping hypothesis, between two corners for example, is justified by evidence in the image. Figure 4 shows a top-down line search between two corner features. The certainty of local lines is based on the contrast of the edge and the percentage of the area that can be classified as a line. The certainty of a local line, l, is given by the following:

$$\kappa(l) = M \cdot \frac{E}{N} \tag{1}$$

where M is the average contrast of the edge, E is the number of pixels classified as a line, and N is the total number of pixels along the connecting search line.

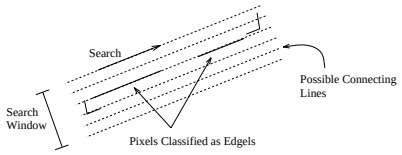


Figure 4: The local line finder is invoked by higher level processes to encourage possible groupings.

3.2 Orthogonal corners

Our domain assumption is that rooftop polygons will be flat with orthogonal corners. This describes a large majority of the building roofs in urban and

industrial centers. Therefore, corner features are orthogonal and parallel to the ground plane in the 3D world. Of course, the apparent shape of an orthogonal corner in the image is not invariant under perspective projection, but varies predictably with respect to image position [10].

To further simplify processing, we assume that a majority of the buildings are aligned according to an approximate city grid. This assumption reduces the set of orthogonal roof corners to be considered to a set of four, which for the purposes of this paper are labeled NorthEast, NorthWest, SouthEast and SouthWest. The relative orientation of the city grid with respect to the camera completely determines how these four cardinal corner types will appear in the image. Currently, we compute this orientation from the given camera pose, however the city-grid orientation can also be computed more generally using vanishing point analysis. [12]

Four different ideal corner masks are generated, one for each cardinal direction. Warping is performed by mapping the lines that define the orthogonal corner through the perspective transformation and into the new expected corner angle. This transformation is performed to sub-pixel accuracy.

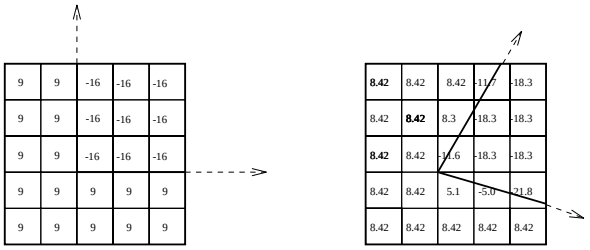


Figure 5: An original 5×5 mask and the corresponding perspectively transformed mask that is convolved with the image.

The final masks, then, are typical $n \times n$ ideal corner detectors that are convolved with the image. In the results shown here 5×5 masks were used. Masks of this small size do well in localizing corners and in detecting non-obvious corners (see section 6), however they are sensitive to noise. We allow a large number of false positives when detecting corners and rely on the higher level grouping processes to discard useless low level features.

Once constructed, each mask is convolved with the image and the correlation value at each pixel in the image is stored. The correlation measure is used as the measure of "cornerness" of each image pixel and is normalized by the maximum change in grey levels in the image over the size of the mask. Normalization is needed because the corners in typical aerial imagery range from high contrast to very dim.

Given a corner mask of size $n \times n$, where $n = 2k+1$ is odd, the correlation at image pixel i,j is given by:

$$C_{i,j} = \sum_{l=-k}^{+k} \sum_{m=-k}^{+k} \frac{I(i+l, j+m)M(l,m)}{\sigma} \qquad (2)$$

where σ is the maximum difference in grey level over the $n \times n$ window, and $I(i,j)$ and $M(i,j)$ are the values of the $(i,j)^{th}$ element of the image and mask respectively. This measure becomes the certainty for the orthogonal corner features that are placed into the feature relation graph. For a corner feature, positioned at i, j in the image, its certainty is given by the correlation measure $C_{i,j}$.

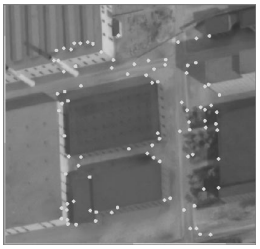


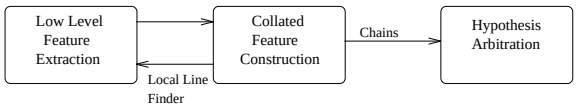
Figure 6: Orthogonal Corners

After convolution of each corner mask with the image, the absolute value of the mask response is thresholded to remove many of the false positives. For the experiments an empirical threshold of 60% on corner certainty was used. Finally, non-maximal suppression over a 5×5 window of pixels is used to eliminate neighboring pixels that respond to the corner mask only partially.

4 Collated features

Collated features are constructed from sets of lines and corners extracted from the image. A collated feature is a sequence of perceptually grouped corners and lines that form a chain. A valid chain group must contain an alternation of corners and lines, and can be of any length. Chains are a generalization of the collated features in earlier work [7, 14] and allow final polygons of arbitrary shape to be constructed from low level features.

Construction of the collated features is analogous to placing edges in the feature relation graph. Low level features are grouped together according to the standard perceptual parameters of smoothness and sym-



metry. When such a group is formed, the corresponding nodes in the feature relation graph are connected with an edge. Paths in the feature relation graph become the chain features.

If a low level feature that is needed to complete a strong perceptual group is missing, a top-down feature detector is invoked and the missing feature is searched for in the image. Currently, the system is able to invoke the local line detector to complete a link of two corners where lines extracted previously were insufficient.

4.1 Feature groups

Perceptual grouping has been shown to be a useful method for extracting high level structure from complex imagery.[15] Standard grouping techniques are applied to the low level features in an attempt to group compatible corners and the line between them. These corner-line-corner groups are the "links" that, when followed as paths in the feature relation graph, form chains. Each link can be thought of as a polygon edge hypothesis, while chains are pieces of a polygon hypothesis. Closed chains are a special case and are treated as completed polygon hypotheses.

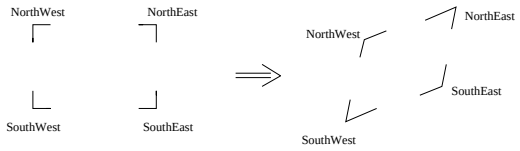


Figure 7: Corner labels, assigned during extraction, are used to help determine proper perceptual groups.

In order for a link to be formed, three conditions must be met (Figure 8). Given two orthogonal corners, they must first be of compatible types, where compatibility is defined according to corner type and axis information. For example, the east-pointing axis of a NorthWest corner cannot be grouped with a corner of type SouthWest. Secondly, corresponding axes of two corners to be linked must be roughly collinear. Finally, a perceptual link can be formed only if there is evidence for a supporting straight line between the corners. Straight line segments extracted from the Boldt line algorithm can be used to contribute to the grouping. However, if the expected Bolt line is missing or is not sufficient to justify the group, the local line

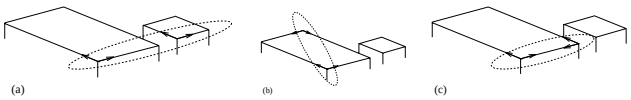


Figure 8: Perceptual grouping of low level features. (a) Incompatible corner types disallow group. (b) Improper alignment of corners. (c) Valid group, supported by line evidence.

finder is invoked and the image is searched directly for the necessary evidence. In this paper, 70% of the pixels between two corners must be classified as either a Boldt or local line in order to create a link. Figure 9 shows an example of perceptually grouped corners.

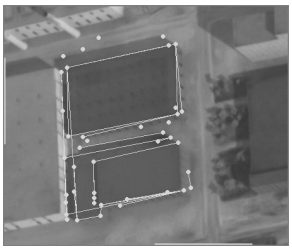


Figure 9: Perceptually Grouped Corner Pairs (Links)

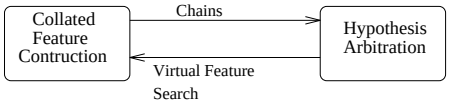
To compute the certainty of a particular chain feature the weight of the corresponding path in the feature relation graph is computed. The certainty of a chain is the sum of the certainties of its parts. Thus, given chain C of length n, the certainty is computed as:

$$\kappa(C) = \sum_{i=0}^{n} \kappa(v_i) + \sum_{i=0}^{n-1} \kappa(e_i) \qquad (3)$$

where v_i is node i in the path corresponding to C and e_i is edge i.

5 Polygon groupings

Extraction of final rooftop polygons proceeds in two steps. First, all possible polygons are computed from the collated features. Then, polygon hypotheses are arbitrated in order to arrive at a final set of non-conflicting, high confidence rooftop polygons. Polygon



hypotheses are simply closed chains, which can be found as cycles in the feature relation graph. All of

the cycles in the feature relation graph are searched for in a depth first manner.

While searching for closed cycles, the collated feature detector may be invoked in order to attempt closure of chains that are missing a particular feature. The system then searches for evidence in the image that such a virtual feature can be hypothesized. Virtual features are hypothesized according to the parameters of perceptual completion. If a cycle is missing a single corner, for example, a virtual corner will be hypothesized at a position that is constrained both by symmetry and smoothness. Currently, the system is able to hypothesize virtual corners and then invoke lower level feature detectors to confirm the hypothesis.

After addition of a virtual corner, the image is searched by the local line finding algorithm for line data that supports this hypothesis. In the event that the evidence is sufficient, the new corner is generated as a low level feature and used to complete the cycle in the feature relation graph. For a corner to be supported by the image, two lines must be found in the appropriate positions within the image that link the virtual feature with other corners in the feature relation graph. In the results presented here, only 60% line evidence for both lines was needed to support a virtual corner as opposed to the 70% threshold used while extracting local lines to link existing corners.

In this way, high level features do not rely on the original set of features that were extracted from the image. Rather, as evidence for a polygon accumulates, tailor-made searches for lower level features can be performed. This type of top-down inquiry increases the robustness of the system.

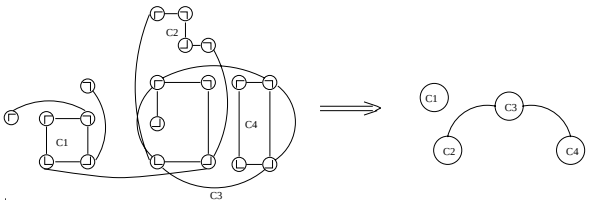


Figure 10: Cycles are extracted from the relation graph and placed, as nodes, into a dependency graph.

Once discovered, all cycles are stored in a dependency graph where nodes represent complete cycles. (See Figure 5) Nodes in the dependency graph contain the certainty of the cycle that the node represents. An edge between two nodes in the dependency graph is created when cycles have lower level features in common. The final set of polygons then, must be the set of nodes that are both independent (have no edges in common) and of maximum certainty.

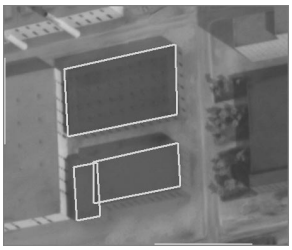


Figure 11: A set of polygon hypotheses extracted from the feature relation graph. When neighboring rooftops partially occlude a polygon, as in the above image, a virtual feature may be generated at the missing corner, which can create final polygons that overlap.

A set of polygon hypotheses extracted from a typical image is shown in 5. Notice that, with the generation of virtual features such as corners, we are able to complete a polygon that is partly occluded by a neighboring building.

6 Results

In addition to the example used above, several more examples are shown in order to demonstrate the system's robustness. Images that were used had a variety of buildings, shadows, and many of the difficulties typically found in aerial imagery.

6.1 Modelboard image

The first set of images used are part of the RADIUS (Research and Development for Image Understanding) model board imagery. As before, the camera model and pose for each image is known. The images were captured at an approximate height of 10,000 feet above the ground plane and the image sizes were 1k $\times$ 1k pixels. In order to speed up processing the image was partitioned into smaller subimages, loosely representing different 'functional areas".

The first subimage, taken from the center of modelboard image J3, contained six distinct buildings of varying sizes. The strong shadows and different rooftop heights make this an interesting area for testing purposes. (See 12) The value of virtual feature extraction is shown by building **A**. A shadow falls across a corner of the building and important low level information is missing.

The second example, a subimage from J6, contains seven buildings of different sizes and complex shapes. Both buildings D and C were not entirely extracted.

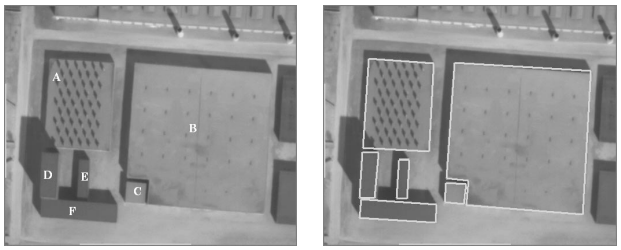


Figure 12: First test image and the results of the building detector.

That is, the final polygons do not match the shape of the underlying structure. Building C is a two level structure and the system failed to discover the lower level rooftop boundary on the right. The corner detector, which relies on dihedral junctions in the intensity image, failed to extract crucial low level features in both buildings.

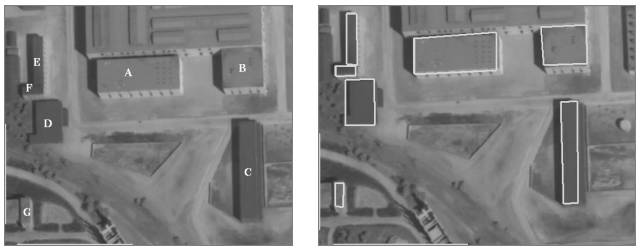


Figure 13: Second test image and the results of the building detector.

The results of the system over an entire image are shown in figure 14. Overall, the system generated 40 rooftop hypotheses. Most major roof boundaries were extracted and the central cluster of buildings (see area **A** in Figure 14 and closeup in Figure 12) were segmented almost perfectly. Several false positives were generated over the large building with many parallel roof vents (marked **B** in Figure 14). False negatives were due mostly to the "all or nothing" nature of the algorithm. That is buildings that are only partially visible (area **C** in Figure 14) cannot be detected and failure in the corner detector may cause a building in full view to be missed, as in area **D**. Finally, some split-level buildings were detected as a single boundry polygon (e.g. **E**) rather than one polygon per roof level. Crucial corner features were undetected due to the sun angle and viewpoint.

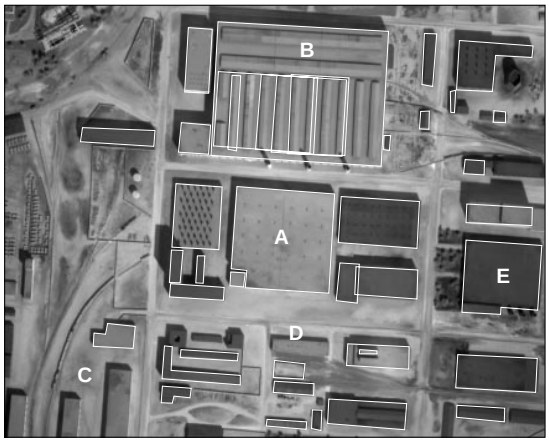


Figure 14: Rooftop polygons extracted over an entire modelboard image. Alphabetic labels are referred to in the text.

6.2 Martin-Marietta site

The system was also tested on a building located at the Martin-Marietta site in Colorado. This is a nadir image, captured at an approximate height of 750 meters. The orientation of the city grid with respect to the world was known and the image was rotated to bring the city grid into alignment with North (see 3.2). The complex, multi-level structure of the building makes it interesting for testing purposes (see Figure 15).

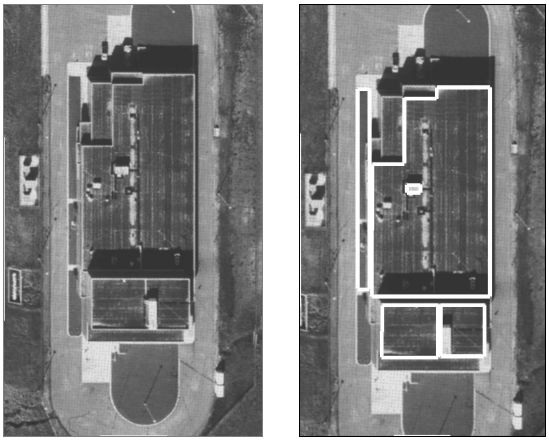


Figure 15: Martin Marietta Building and the Building Detection Results

Five polygons were detected by the system. The leftmost polygon is a false positive that was grouped because corners and lines form a rectangle on the sidewalk adjacent to the building.

The large polygon denotes most of the main roof outline and contains a virtual corner at the lower right. Two roof sections in the upper left corner of the building were missed due to the nature of the final graph search. Features in the final polygon set must be independent and these two polygons shared corner features with the central polygon that had a larger certainty measure. A rectangular roof structure, perhaps an air-conditioning unit, was extracted at the center of the rooftop. Finally, the roof section at the bottom of the building was broken into two polygons because a line separates the sections in the image.

7 Conclusions and future work

The results from the proposed approach are encouraging. In the two modelboard subimages shown here, 87 % of the building rooftops were detected to some degree of accuracy. The polygons extracted are accurate enough to be used in further image understanding tasks.

Currently, polygon detection is a piece of the larger aerial image understanding system being developed at U. Mass, Amherst under the RADIUS project. The hypothesized rooftop polygons are verified by epipolar matching and refined through multi-image triangulation, which computes a height for each polygon in the world. Rooftop polygons that lay too close to the ground plane or are not parallel to the ground plane are discarded by the larger system. The final set of roof polygons are extruded to the ground plane to form final volumetric models of the buildings.

In the future, an improved corner detection mask that incorporates shadow angles from known sun position will be constructed. The task driven approach to perceptual organization will be expanded to cover more general image understanding tasks. Relaxing restrictions such as flat roofs and orthogonal corners will be investigated so that a more general module can be used to group general structures in aerial imagery.

References

[1] R. Bolles and R. Cain, "Recognizing and Locating Partially Visible Objects: The Local-Feature-Focus Method," *International Journal of Robotics Research* Vol 1, No. 3, Fall 1982.

[2] R. Weiss and M. Boldt, "Geometric Grouping Applied to Straight Lines," *IEEE Computer Society on Computer Vision and Pattern Recognition*, 1986.

[3] R. Collins, J. Beveridge, "Matching Perspective Views of Coplanar Structures using Projective Unwarping and Similarity Matching," *Proc. Computer Vision and Pattern Recognition*, June 1993.

[4] M. Fischler and R. Bolles, "Perceptual Organization and Curve Partitioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Jan. 1986.

[5] P. Fua and A. Hanson, "Using generic geometric models for intelligent shape extraction," *Proc. DARPA Image Understanding Workshop*, Los Angeles, CA, Feb. 1987.

[6] M. Herman and T. Kanade, "The 3D MOSAIC Scene Understanding System: Incremental Reconstruction of 3D Scenes from Complex Images," *Proc. DARPA Image Understanding Workshop*, 1984.

[7] A. Huertas, C. Lin, and R. Nevatia, "Detection of Buildings from Monocular Views Using Perceptual Grouping and Shadows," *Proc. DARPA Image Understanding Workshop*, 1993.

[8] A. Huertas and R. Nevatia, "Detecting Buildings in Aerial Images," *Computer Vision, Graphics, Image Processing* vol. 13, 1980.

[9] V. S. Hwang, L. Davis, T. Matsuyama, "Hypothesis Integration in Image Understanding Systems," *Computer Vision, Graphics, and Image Processing* Vol. 36, pg 321-371, 1986.

[10] K. Kanatani, "Constraints on Length and Angle," *Computer Vision, Graphics, and Image Processing*, 1987.

[11] T. Matsuyama and V. Hwang, "SIGMA: A Framework for Image Understanding: Integration of Bottom-Up and Top-Down Processes," *Proceedings of the Ninth IJCAI*, Los Angeles, CA, pp. 908-915, 1985.

[12] J.C. McGlone and J. Shufelt, "Incorperating Vanishing Point Geometry into a Building Extraction System," *Arpa Image Understanding Workshop*, Washington DC, pp.437-448, 1993.

[13] D. McKeown, "Toward Automatic Cartographic Feature Extraction," *Mapping and Spatial Modelling for Navigation,* Nato ASI Series, Vol. F65, pp. 149–180, 1990.

[14] R. Mohan and R. Nevatia, "Using Perceptual Organization to Extract 3D Structures," *Trans. Pattern Analysis and Machine Intelligence*, 1989.

[15] G. Reynolds and J. R. Beveridge, "Searching for Geometric Structure in Images of Natural Scenes," *COINS Technical Report 87-03*, 1987.

[16] A. Singh and M. Shneier, "Grey Level Corner Detection: A Generalization and a Robust Real Time Implementation," *Computer Vision, Graphics, and Image Processing*, 1990.

[17] V. Venkateswar and R. Chellapa, "Intelligent Interpretation of Aerial Images," *University of Southern California, Dept. of Electrical Engineering Technical Report 137* March 1989.