The Ascender System

Automated Site Modeling from Multiple Aerial Images

Robert T. Collins, Christopher O. Jaynes, Yong-Qing Cheng, Xiaoguang Wang, Frank Stolle, Edward M. Riseman and Allen R. Hanson

Department of Computer Science
Lederle Graduate Research Center
Box 34610, University of Massachusetts
Amherst, MA. 01003-4610

The Ascender System[1]

Correspondence To:
Christopher O. Jaynes
Department of Computer Science
Lederle Graduate Research Center
Box 34610, University of Massachusetts
Email: jaynes@cs.umass.edu
Fax: (413)-545-1249

[1] Funded by the RADIUS project ARPA/Army TEC contract number DACA76-92-C-0041, and the National Science Foundation grant number CDA8922572.

Abstract

The Ascender system acquires, extends and refines 3D geometric site models from calibrated aerial imagery. To acquire a new site model, an automated building detector is run on one image to hypothesize potential building rooftops. Supporting evidence is located in other images via epipolar line segment matching in constrained search regions. The precise 3D shape and location of each building is then determined by multi-image triangulation under geometric constraints of 3D orthogonality, parallelness, colinearity and coplanarity of lines and surfaces. Projective mapping of image intensity information onto these polyhedral building models results in a realistic site model that can be rendered using virtual “fly- through" graphics.

As new images of the site become available, model extension and refinement procedures are performed to add previously unseen buildings and to improve the geometric accuracy of the existing 3D building models. In this way, the system gradually accumulates evidence over time to make the site model more complete and more accurate.

An extensive performance evaluation of component algorithms and the full system has been carried out. Two-dimensional building detection accuracy, as well as accuracy of the three-dimensional building reconstruction, are presented for a representative data set.

List of Symbols

$\sum$

θ

r_1

r_i

r_n

1. Introduction

The Research and Development for Image Understanding Systems (RADIUS) project is a national effort to apply image understanding (IU) technology to support model-based aerial image analysis [11]. Automated construction and management of 3D geometric site models is a key component of this effort. Site models enable efficient exploitation of the tremendous volume of information collected daily by national sensors. In all of these applications, the expected benefits are decreased work-load on human analysts, together with an increase in measurement accuracy due to the introduction of digital IU and photogrammetric techniques. When properly annotated, automatically generated site models can also provide the spatial context for specialized IU analysis tasks such as vehicle counting and change detection, and graphical visualization techniques using 3D site models are valuable for training and mission planning. Other applications of the technology presented here include automated cartography, land-use surveying and urban planning.

The long-term goal of our research is an entirely automated system. Given the extreme complexity of some image domains, often rather challenging even for expert human users, this goal may not be fully achievable. However, our focus in this project is to push an automated paradigm as far as possible. We believe that the 3D aerial reconstruction problem can, to a great degree, be automated given a large enough set of images. As related factors become more difficult, such as high building density, complex building and surface shapes (as in European cities), little space between buildings, and/or only a small number of available views, accurate reconstruction becomes much harder. It is generally true that if a sufficiently large number of appropriate image viewpoints are not available, any reconstruction problem can become difficult or impossible. Thus, expectations of a completely automated system must be tempered. However, our goal is to come as close as possible, and as we reach the limits of automation, intelligent interactive tools can be provided for manual specification of constraints or results.

1.1 Ascender AND THE TECHNICAL CHALLENGES

The UMass Automated Site Construction, Extension, Detection and Refinement (ASCENDER) system has been designed to automatically populate a site model with buildings extracted from multiple, overlapping images. There are many technical challenges involved in developing a building extraction system that works reliably on the type of images being considered under the RADIUS program. Images may be taken over significant time spans, and under vastly different weather and lighting conditions. The use of monocular, oblique imagery introduces perspective distortion due both to the obliquity and to the large differences in camera viewpoint. Images taken under different weather conditions and at different times of day introduce large intensity variations between images of the same building surface. There is typically a lot of clutter surrounding buildings (vehicles, pipes, oil drums, vegetation) and on their roofs (roof vents, air conditioner units, ductwork), buildings often occlude each other in oblique views, and shadows falling across building faces break up extracted low-level features such as line segments and regions. Furthermore, nearby buildings can vary greatly in size and shape.

1.2 DESIGN PHILOSOPHY / KEY IDEAS

The Ascender system combines several algorithms into a data flow hierarchy leading from images to a final site model. Image 1 shows the data dependencies between each component and how the overall system is compose. The UMass design philosophy incorporates several key ideas. First, 3D reconstruction is based on geometric features that remain stable under a wide range of viewing and lighting conditions. Second, rigorous photogrammetric camera models are used to describe the relationship between pixels in an image and 3D locations in the scene, so that diverse sensor characteristics and viewpoints can be effectively exploited. Third, information is fused across multiple images for increased accuracy and reliability. Finally, known geometric constraints are applied whenever possible to increase the efficiency and reliability of the reconstruction process. The current Ascender system is designed to perform well at one end of a data-vs-control complexity spectrum, namely a large amount of data and a simple control structure, versus the alternative of using less data but more complicated processing strategies. In particular, while the system can be applied to a single stereo pair, it generally performs better (in terms of number and quality of buildings found) when more images are used.

The design here represents the Ascender I system. New research is underway into more advanced system designs. For example, the system currently extracts polygons from a single image and uses other imagery for verification and height computation. However, a true multi-image scheme would not depend on the accuracy of polygons extracted from this first “reference image”. Suffice it to say, there is not necessarily a single best flow of control for an automated reconstruction system and control may depend on available images, algorithms, and scene context.

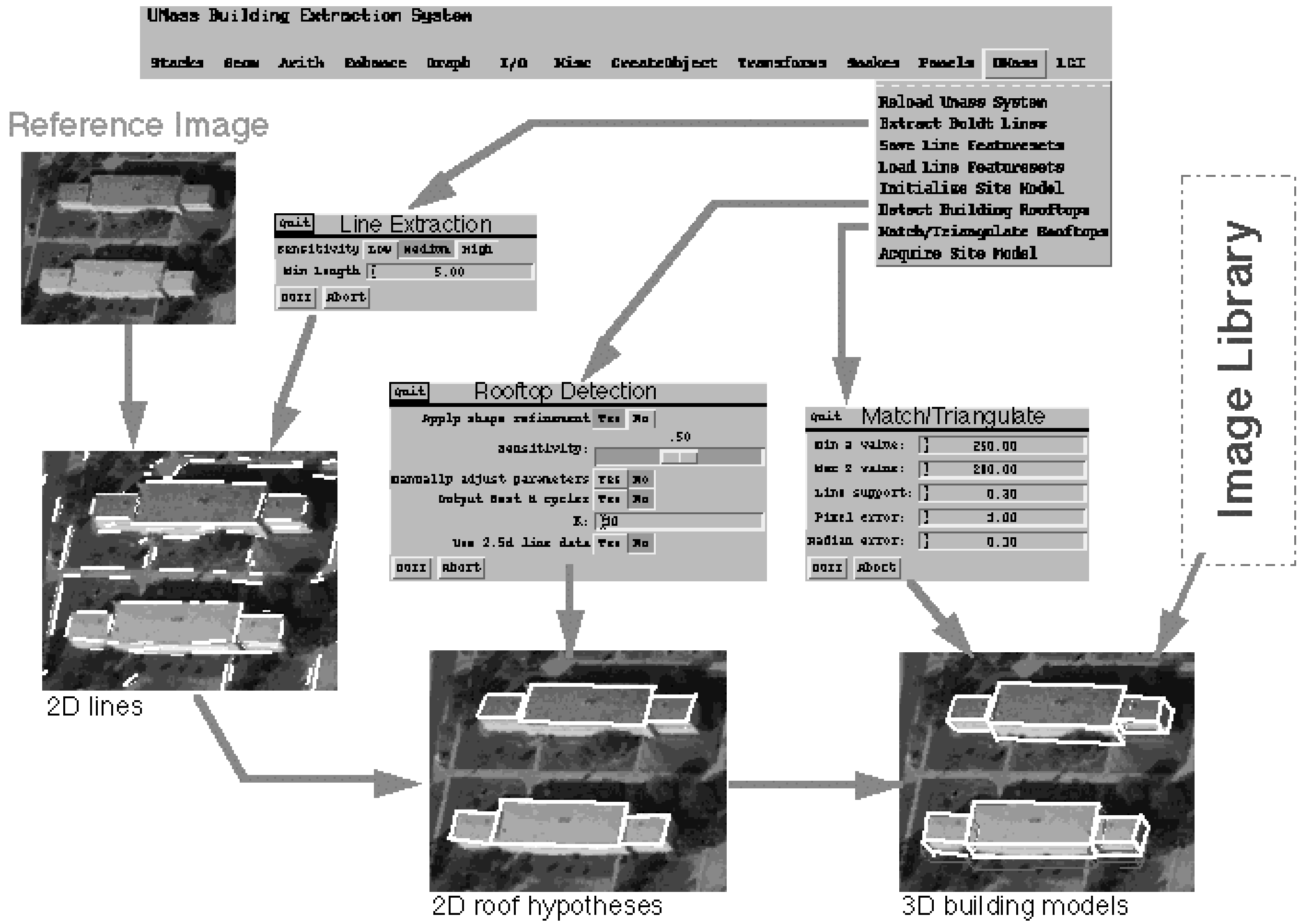


Figure 1: Flow of control in the Ascender 1 system. A reference image is used to detect straight line segements which are grouped into 2D closedrooftop hypotheses. These hypotheses are matched and triangulated with other views in the image library to arrive at a final 3D model.

Ascender I supports three different site modeling tasks, *model acquisition, model extension*, and *model refinement*. Site model acquisition involves processing a set of images to detect both man-made and natural features of interest, and to determine their 3D shape and placement in the scene. Two other important site modeling tasks are model extension -- updating the geometric site model by adding or removing features, and model refinement -- iteratively refining the shape and placement of features as more views become available. Model extension and refinement are ongoing processes that are repeated whenever new images become available, each updated model becoming the current site model for the next iteration. Thus, over time, the site model is steadily improved to become more complete and more accurate.

1.3 OUTLINE OF PAPER

This paper is organized as follows. Section 2 reviews a number of past and present building extraction systems. Section 3 then presents a system-level specification of the Ascender system, followed in Section 4 by a breakdown of the building extraction process into its key algorithmic components. Section 5 presents an in-depth experimental evaluation of system performance using imagery taken over Ft. Hood, Texas. Section 6 discusses the strengths and shortcomings of the current system, proposes future research directions, and concludes the paper with a brief summary.

The Ascender I system is a set of complex algorithms that work together to perform site reconstruction. This paper introduces the system, discusses its components in detail, and presents the results of extensive testing. However, it is important for the reader to realize that many parts of the system have been discussed in previous papers and we suggest that these papers (when referenced) should be read as a useful accompaniment to this paper.

2. Related Work

Over the past decade, automated building detection systems have evolved along many lines, but the trend has always been towards greater generality: from special-case nadir views to general oblique viewpoints, from single image analysis to multi-image techniques, and from purely 2D hypothesis extraction in image-space to rigorous 3D geometric reconstruction in object-space. As a system for extracting precise 3D building models from multiple, oblique views, the Ascender system represents the state-of-the-art in all aspects of this ongoing evolution.

Many early systems were based on the nadir viewpoint assumption, in part because most of the available images at that time were from mapping applications that relied on nadir views. The nadir assumption greatly simplifies building extraction geometry since rectangular building roofs appear as rectangles in the image, and there is very little occlusion of one building by another. The RADIUS project reenforced the need for using oblique images, since even though satellite coverage of the globe is available on a daily basis, only a small fraction of it appears as a nadir view directly underneath the satellite's flight path. The easiest generalization from nadir views to handle obliquity is to

assume weak-perspective or affine views, where rectangular roofs appear as parallelograms [23; 24]. The ultimate generalization is to introduce photogrammetrically rigorous camera projection equations that more accurately model the (typically projective) viewing geometry [26; 27; 14]. Our work takes this latter.
approach.

Early systems were dominated by monocular, image-based approaches, since often only a single view of the area was available. However, buildings are raised 3D structures, and it is difficult to disambiguate roof hypotheses and determine building height without 3D information. One powerful source of inferred 3D information in monocular images is shadows, and indeed, many systems have been designed that exploit the relationship between shadows and roof hypothesis [24; 15; 25; 17]. Shadow analysis is particularly attractive when combined with nadir viewpoints, since building height is directly proportional to the length of the building shadow on the ground. Systems that rely on shadow analysis often assume that the sun position (illumination direction) is known, and always assume that building shadows fall on flat terrain surrounding the building (and not across other buildings or on rocky or hilly terrain). A more general method of deriving 3D height information is to use stereo triangulation across two or more images [7; 16; 31]. The Ascender system uses such multi-image stereo analysis to extract the precise shape and location of each building in the scene. Most notably, the system currently does not use shadow information at all, but derives 3D structure completely from multi-image matching and triangulation.

Several approaches are similar to the technique in which building regions are hypothesized in the Ascender system. These typically organize extracted image features into more complex structures based on geometric constraints [15] and have been used for the grouping of features into 3D models from several views [14]. These approaches to grouping have been improved through the use of a richer set of constraints including the explicit use of knowledge [22] and the fusion of digital surface models (DSMs) with optical images [35]. The use of DSMs has been used for both detection of possible building regions and for constraining a perceptual grouping process [8].

There are at least three current building extraction systems similar to our own, in that they derive 3D building models from multiple, oblique views. Noronha and Nevatia [29] describe a system where hierarchical grouping and matching across multiple images is used to reconstruct 3D building models. Buildings are extracted in hierarchical stages, ranging from line segments, to junction, parallel pairs, U-shapes, and finally, whole parallelograms. At each stage in the extraction hierarchy, the following three steps are performed: 1) two-dimensional perceptual grouping of features at that level in each image, 2) epipolar matching to determine correspondence of features across pairs of views, 3) applying geometric constraints to check the consistency of the 3D structures implied by those feature matches. Final building hypotheses are verified by searching for consistent shadows and evidence of vertical walls. Only rectangular building hypotheses are found -- arbitrary rectilinear structures are formed by merging abutting or overlapping rectangular 3D building hypotheses of similar height. The most notable feature of the system is that information from all views is used in a non-preferential way, as opposed to the Ascender system where one image is used to extract hypotheses, and other views are used to corroborate each hypothesis and compute 3D structure.

In the MULTIVIEW system [31; 27] corner features extracted via vanishing point analysis are matched across a pair of views to get 3D corners. These 3D corners become nodes in a graph, and pairs are linked if image gradient intensity information supports an edge hypothesis. Polygonal surface hypotheses are formed by looking for cycles in the graph that meet certain planarity and perpendicularity constraints. When more views are available, relationships between corners and lines in the graph are updated as each new view is added. Surface detection can be performed after each view, or can be left until after several views have been added (batch mode). The system results are sensitive to the permutation of the views, since the first pair of views is used to initialize the 3D corner-edge graph, and the graph is updated sequentially as each view is added. Both this system and the one of Noronha and Nevatia perform pairwise feature-based stereo to derive 3D features that are then grouped into surfaces.
In contrast, our epipolar matching phase uses all images simultaneously to compute 3D information, even when a building was only detected in a single image, which can result in more accurate localization of 3D features.

In addition to geometric constraints, the use of semantic and domain knowledge can widen the scope of automatic building reconstruction and improve robustness. For example, Fischer, Kolbe, and Lang [9] have emphasized the use of reasoning and knowledge at several levels to constrain the number of possible hypotheses that can be produced from multiple views. As opposed to implicit models used in Ascender (embedded in the 2D grouping constraints), explicit semantic models are used that include simple image features, 3D building terminals (parameterized parts of buildings), and 3D surfaces. A grouping process attempts to construct complete building models from the recognized parts. In a similar effort, researchers within the Amobe project [13] extract trihedral corners and make use of both epipolar imaging constraints and knowledge about building surfaces to group features into complete building models.

The use of a range image registered to an optical image allows the extraction of a rich class of three dimensional features including surfaces, 3D line segments, and 3D trihedral corners. It has been shown that the introduction of three dimensional geometric constraints can allow for a wider range of cultural features to be detected and reconstructed [19].

Kim and Muller [20] combine a monocular building extraction scheme with elevation maps to detect possible building boundaries. Given possible boundaries, interior elevations are used to estimate a height for reconstruction. Foerstner [10] makes use of the range image to both hypothesize buildings and reconstruct the building geometry. Robust techniques select a set of non-contradicting 3D constraints for optimal estimation of the object shape. Haala and Hahn [12] use the elevation map directly to infer the presence of buildings by searching for local maxima, with 3D lines computed in these regions used for parametric model fits to the extracted line segments. The approach estimates the initial parameters for model fitting, but assumes that the buildings at the site can be reconstructed using a single parametric model (e.g. a peaked roof model).

3. The Ascender System

The Ascender system was designed to automatically acquire 3D building models from a set of overlapping aerial images. To maintain tractable research and implementation goals, Ascender I deals only with a single generic class of buildings, namely flat-roofed, rectilinear structures. The simplest example of this class is a rectangular box-shape; however other examples include L-shapes, U-shapes, and indeed any arbitrary building shape such that pairs of adjacent roof edges are perpendicular and lie in a horizontal plane.

The Ascender system was developed using the RADIUS Common Development Environment (RCDE) [28]. RCDE provides a framework for the development of site model acquisition algorithm. The choice of a photogrammetric development environment was constrained by the funding agency.

3.1 IMAGES

Site model acquisition requires a set of images, both nadir and oblique, that view the same area of the site. The system is designed to operate over multiple images, typically five or more, exhibiting a wide variety of viewing angles and sun conditions. The desired number five is chosen arbitrarily to allow ideally one nadir view plus four oblique views from each of four perpendicular directions (e.g. North, South, East and West). This configuration is not a requirement, however. Indeed, some useful portions of the system require only a single image, namely line segment extraction and building rooftop detection. On the other hand, epipolar rooftop matching and wireframe triangulation require, by definition, at least two images, with robustness and accuracy increasing when more views are available.

Although best results require the use of several images with overlapping coverage, the system allows considerable freedom in the choice of images to use. Unlike many previous building extraction systems, this system does not currently use shadow information, and works equally well on images with different sun angles, or with no strong shadows at all. Also, the term "epipolar" as used here does not imply that images need to be in scan-line epipolar alignment, as required by many traditional stereo techniques. The term is used instead in its most general sense as a set of geometric constraints imposed on potentially corresponding image features by the relative orientation of their respective cameras. The relative orientation of any pair of images is computed from the absolute orientation of each individual image (see below).

3.2 SITE COORDINATE SYSTEM

Reconstructed building models are represented in a local site coordinate system that must be defined prior to the reconstruction process. The system assumes this is a "local-vertical" Euclidean Coordinate System, that is, a Cartesian X-Y-Z coordinate system with its origin located within or close-to the site, and the positive Z-axis facing upwards (parallel to gravity). The system can be either right-handed or left-handed. Under a local-vertical coordinate system, the Z values of reconstructed points represent their vertical position or elevation in the scene, and X-Y coordinates represent their horizontal location in the site.

3.3 CAMERA MODELS

For each image given to the system, the absolute orientation of the camera with respect to the local site coordinate system must be known. This is a specification of how 3D locations in the site coordinate system are related to 2D image pixels in each image. One common camera representation is a 3 X 4 projective transformation matrix encoding both the internal orientation (lens/digitizer parameters) and the external orientation (pose parameters) of each perspective camera. Ascender can also handle the fast block interpolation projection (FBIP) camera model used in the RCDE to represent the geometry of non-perspective cameras. Given the absolute orientation for each image, Ascender computes all the necessary relative orientation information needed for determining the epipolar geometry between images (or local approximations to the epipolar geometry in the case of non-perspective cameras).

3.4 DIGITAL TERRAIN MAP

Currently, the Ascender system explicitly reconstructs only the rooftops of building structures, and relies on vertical extrusion to form a volumetric 3D wireframe model of the whole building. The extrusion process relies on knowing the local terrain, namely the ground elevation (*Z* value) at each location in the scene. This can be represented simply as a single plane equation provided the ground is relatively flat, or more generally as an array of elevation values from which terrain values at any horizontal location are interpolated.

3.5 OTHER REQUIRED PARAMETERS

In addition to the general information described above, a few other parameters must be supplied. The most important of these are:

1. **resection-residual-error** -- a number representing the expected residual error (in pixels) present between projected ground truth points and their observed 2D image locations, for the given camera resection. This summarizes, in a single number for each image, the level of geometric error remaining after camera resection has taken place. This parameter is used for generating statistical confidence intervals, for determining the proper relative weights of information from each image, and for generating feature search regions. As new images arrive, a resection-residual error can be over-estimated in order to be sure that evidence gathered from the image is not weighted too greatly. Over-estimation of this parameter will loosen search regions and may create false positives, but will not cause the system to fail to detect features.

2. **max-building-height** -- the maximum possible height of any building that is expected in the site model. This threshold is used to limit the extent of epipolar search regions in each image. The lower this threshold is set, the smaller the search area for rooftop feature matches will be, leading to faster searches with higher likelihood of finding the correct matches.

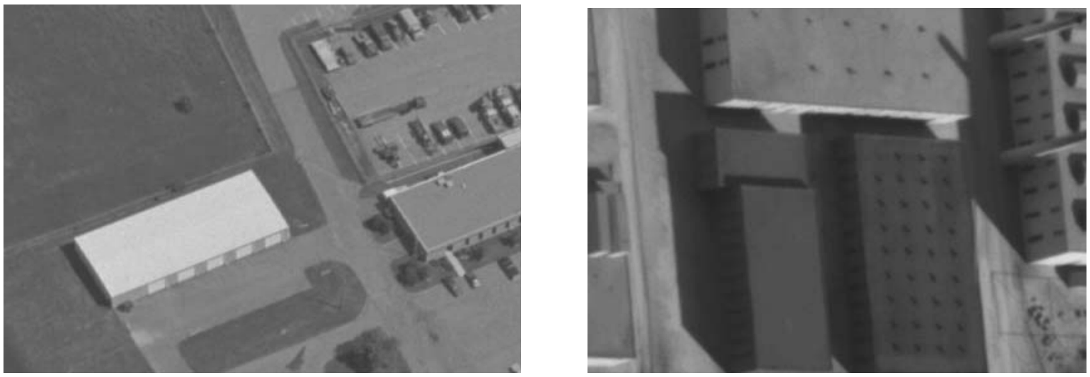
3. **min-building-dimension** -- the minimum extent (width, length or height) of any building that will be included in the site model. This is, loosely speaking, a way of specifying the desired "resolution" of the resulting site model, since any buildings having dimensions less than this threshold will not be found. Setting this value to a relatively long length essentially ensures that only large buildings in the site will be modeled.

4. **feature grouping sensitivity** -- the sensitivity at which image features are progressively grouped into higher level objects. This linear parameter (ranging from "low" to "high") was defined based-on significant experience with the system, and was intended to provide a user interface that is straightforward yet useful. The value of this single grouping sensitivity parameter controls several other component procedures that are part of the system. A low sensitivity will cause the system to group features that strictly comply with the entire set of constraints, while a larger value will loosen the grouping operations to generate more feature aggregations. This parameter influences the grouping behavior of the system but remains independent of the line extraction parameters (see Sections 4.1 and 4.2). For example, the system will only group lines into buildings that are strictly rectilinear at low sensitivity settings, but line extraction filters (on length and contrast) determine the set of features that will be used for grouping.

4. Algorithmic building blocks

The Ascender building extraction system currently follows a simple processing strategy. To acquire a new site model, an automated building detector is run on one image to hypothesize potential building rooftops. Supporting evidence is then located in other images via epipolar line segment matching, and the precise 3D shape and location of each building is determined by multi-image triangulation and extrusion. Image intensity information can be backprojected onto each face of these polyhedral building models, to facilitate realistic rendering from new views.

This section describes the key algorithms that together comprise the model acquisition portion of the system. These algorithms are: line segment extraction, building rooftop detection, epipolar rooftop matching, multi-image wireframe triangulation, and projective intensity mapping. Line segment extraction and building rooftop detection are illustrated with sample results from two sites, the Schenectady County Air National Guard base (Figure 2a), and Radius Model Board 1 (Figure 2b). In the next section, serious system evaluation will be carried out on images of Ft. Hood, Texas.



(a) (b)

Figure 2: Subimages used for reconstruction. (a) Schenectady subimage. (b) Model Board 1 (MB1) subimage.

(a) (b)

Figure 3: (a) Extracted lines for Schenectady subimage. (b) Lines extracted for MB1.

4.1 LINE SEGMENT EXTRACTION

To help bridge the huge representational gap between pixels and site models, a straight line feature extraction routine is applied to produce a set of symbolic line segments representing geometric image features of potential interest such as building roof edges. We use the Boldt algorithm for extracting line segments [4]. At the heart of the Boldt algorithm is a hierarchical grouping system inspired by the Gestalt laws of perceptual organization. Zero-crossings of the Laplacian of the intensity image provide an initial set of local intensity edges. Hierarchical grouping then proceeds iteratively, using measures of colinearity and connectedness. At each iteration, edge pairs are linked and replaced by a single longer edge if their end points are close, their perpendicular offset is small, and their orientation and contrast values are similar (difference in average intensity level across the line). Each iteration results in a set of increasingly longer line segments. The final set of line segment features (Figure 3) can be filtered according to length and contrast values supplied by the user.

Although the Boldt algorithm does not rely on any particular camera model, the utility of extracting straight lines as a relevant representation of image/scene structure is based on the assumption that straight lines in the world (such as building edges) will appear reasonably straight in the image. To the extent that this assumption remains true at the scale of the objects being considered, such as over a region of the image containing a single building, then straight line extraction remains a viable feature detection method. However, very long lines spanning a significant extent of the image, such as the edges of airport runways, may become fragmented depending on the amount of curvature introduced into the image by nonlinearities in the imaging process. Furthermore, image lines may contain contrast changes along their length from illumination differences in the scene, changes in material reflectance, and other properties in the scene. The Boldt algorithm is sensitive to these contrast changes and will produce fragmented lines. The grouping algorithm employed in the 2D polygon detection algorithm addresses this by merging compatible line fragments based on higher level geometric grouping criteria.

4.2 BUILDING ROOFTOP DETECTION/ 2D POLYGON EXTRACTION

The goal of automated building detection is to roughly delineate building boundaries that will later be verified in other images by epipolar feature matching and triangulated to create 3D geometric building models. The building detection algorithm [18] is based on perceptual grouping of line segments into image polygons corresponding to the boundaries of flat, rectilinear rooftops in the scene. Perceptual organization is a powerful method for locating and extracting scene structure. The rooftop extraction algorithm proceeds in three steps; low level feature extraction, collated feature detection, and hypothesis arbitration. Each module generates features that are used during the next phase, and interacts with lower level modules through top-down feature extraction.

Low level features used by the building detection system are straight line segments and corners. Line segments used by the building detection system are produced by the Boldt line algorithm discussed in section 4.1. Edges may be filtered based on length before they are used for detection in a particular image. The shortest expected building edge in the scene is projected into the image to compute a minimum image distance in pixels. Line segments that are shorter are removed.

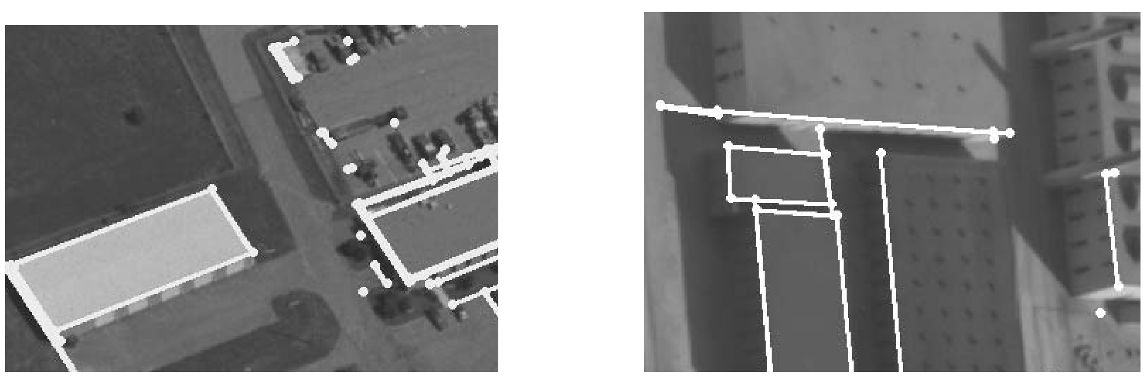
The domain assumption of flat-roofed rectilinear structures implies that rooftop polygons will be produced by flat horizontal surfaces, straight edges, and orthogonal corners.
Orthogonal corners in the world are not necessarily orthogonal in the image; however the known camera geometry can be used to compute a corresponding world angle. To determine a set of relevant corner hypotheses, pairs of line segments with spatially proximate endpoints are grouped together into candidate image corner features. Each potential image corner is then backprojected into a nominal Z-plane in the scene, and the resulting hypothetical *scene corner* is tested for orthogonality. A parameter, tied to the

sensitivity setting of the system, is used to threshold corners based on the angular difference from an orthogonal corner.

Mid-level collated features are sequences of perceptually grouped corners and lines that form a chain (Figure 4). 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 [16] and allow final polygons of arbitrary rectilinear shape to be constructed from low level features.

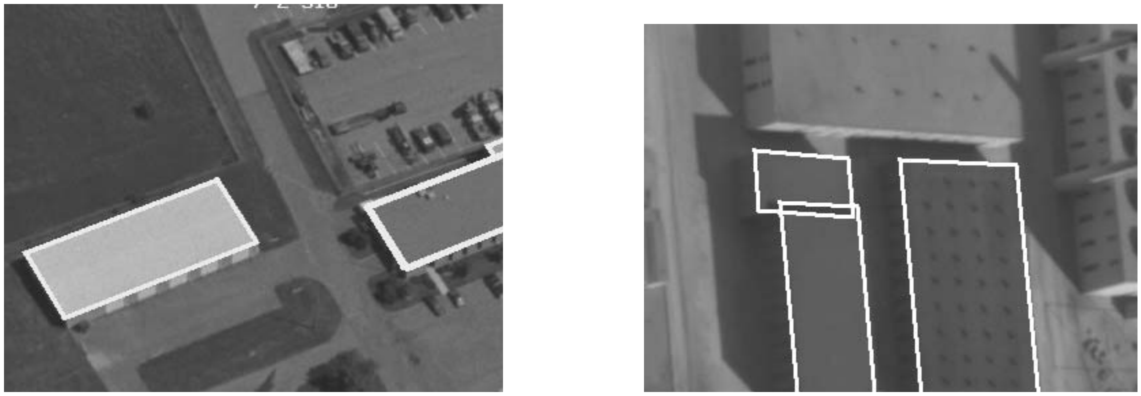
Collated feature chains are represented by paths in a *feature relation graph*. The feature relation graph is an encoding of feature dependencies and perceptual compatibility in the image. Low level features (corners and line segments) are nodes in the graph, and perceptual grouping relations between these features are represented by edges in the graph. 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. For example, an exact alignment of corners in the scene would be represented by an edge in the graph with a large weight, while features that are not exactly aligned but still are grouped together would receive a lower weight edge in the graph. A chain of collated features inherits an accumulated certainty measure from all the nodes and edges along its path.

High level polygon hypothesis extraction 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 (Figure 5).



(a) (b)

Figure 4: (a) Feature chains for Schenectady. (b Feature chains for MB1



(a) (b)

Figure 5: Final rooftop hypotheses. (a) Schenectady subimage. (b) MB1 subimage.

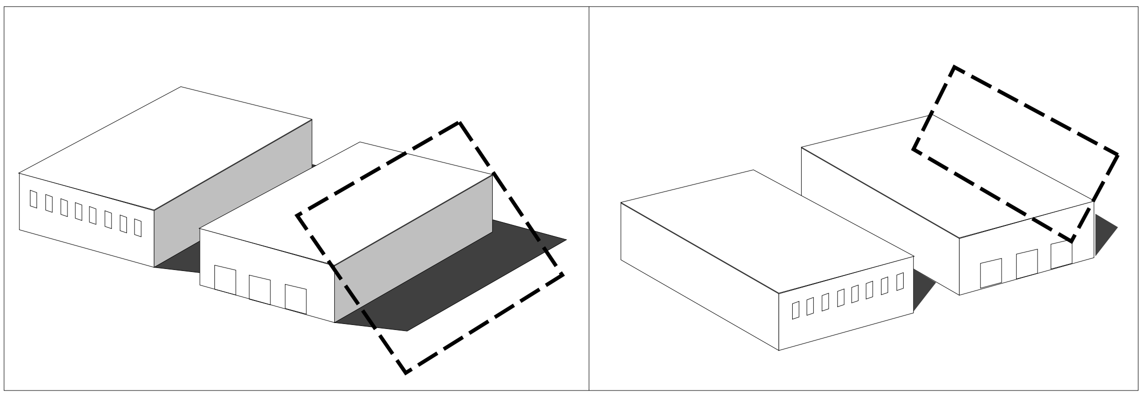
Polygon hypotheses are simply closed chains, which can be found as cycles in the feature relation graph. All of the cycles in the feature relation graph are searched for in a depth-first manner, and stored in a dependency graph where nodes represent complete cycles (rooftop hypotheses). Nodes in the 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 low-level features in common.

The final set of non-overlapping rooftop polygons is the set of nodes in the dependency graph that are both independent (have no edges in common) and are of maximum certainty. Standard graph-theoretic techniques are employed to discover the maximally-weighted set of independent cycles, which is output by the algorithm as the final set of independent high confidence rooftop polygons.

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. An example occurs in Figure 4. The upper-right building corner is missing due to a large gap in the extracted line segments. However, during the graph search, a corner was hypothesized and the extracted line segments provided sufficient support to complete a cycle (figure 5). In this way, the rooftop detection process does not have to 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 system robustness. Currently virtual feature production is only used to fill in a single missing feature, i.e. a missing corner or straight line but not both. Therefore U-shapes will not be hypothesized for completion.

4.3 EPIPOLAR LINE SEGMENT MATCHING

After detecting a potential rooftop in one image, corroborating geometric evidence is sought in other images (often taken from widely different viewpoints) via epipolar feature matching. The primary difficulty to be overcome during epipolar matching is the resolution of ambiguous potential matches, and this ambiguity is highest when only a single pair of images is used. For example, the epipolar search region for a roof edge match will often contain multiple potentially matching line segments of the appropriate length and orientation, one of which comes from the corresponding roof edge, but the others coming from the base of the building, the shadow edge of the building on the ground, or from roof/ base/shadow edges of adjacent buildings (see 6a). This situation is exacerbated when the roof edge being searched for happens to be nearly aligned with an epipolar line in the second image. The resolution of this potential ambiguity is the reason that simultaneous processing of multiple images with a variety of viewpoints and sun angles is preferred.



(a) (b)

Figure 6: Multiple ambiguous matches can often be resolved by consulting a new view.

Rooftop polygons are matched using an algorithm similar to the mutibaseline stereo matching algorithm of Okutumi and Kanade [30], but generalized to handle arbitrary camera poses and line segment image features. For each polygon line segment from one image, an appropriate epipolar search area is formed in each of the other images, based on the known camera geometry and the assumption that the roof is flat. This quadrilateral search area is scanned for possible matching line segments, the disparity of each potential match implying a different roof height in the scene. Results from each line search are combined in a 1-dimensional histogram, each potential match voting for a particular roof height. Each vote is weighted by compatibility of the match in terms of expected line segment orientation and length. This allows for correct handling of fragmented line data, since the combined votes of all subpieces of a fragmented line count the same as the vote of a full-sized, unfragmented line. A single global histogram accumulates height votes from multiple images, and for multiple edges in a rooftop polygon. After all votes have been tallied, the histogram bucket containing the most votes yields an estimate of the roof height in the scene and a set of correspondences between rooftop edges and image line segments from multiple views. Competing ambiguous roof heights will appear as multiple peaks in the histogram; these can be carried forward for disambiguation via future images.

4.4 WIREFRAME TRIANGULATION AND OPTIMIZATION

Multi-image triangulation is performed to determine the precise size, shape, and position of a building in the local 3D site coordinate system. A nonlinear estimation algorithm has been developed for simultaneous multi-image, multi-line triangulation of rectilinear rooftop polygons. Object-level constraints such as perpendicularity and coplanarity are imposed on the solution to assure reliable results. This algorithm is used for triangulating 3D rooftop polygons from the line segment correspondences determined by epipolar feature matching.

The parameters estimated for each rooftop polygon are shown in Figure 7. The horizontal plane containing the polygon is parameterized by a single variable Z. The orientation of the rectilinear structure within that plane is represented by a single parameter θ. Finally, each separate line within the polygon is represented by a single value r_i representing signed distance of the line from a local origin within the roof polygon. The representation is simple and compact, and the necessary coplanarity and rectangularity constraints on the polygon's shape are built in. (A more general approach based on the Plucker coordinate representation of 3D lines has also been implemented for triangulating general wireframe structures [5,6]).

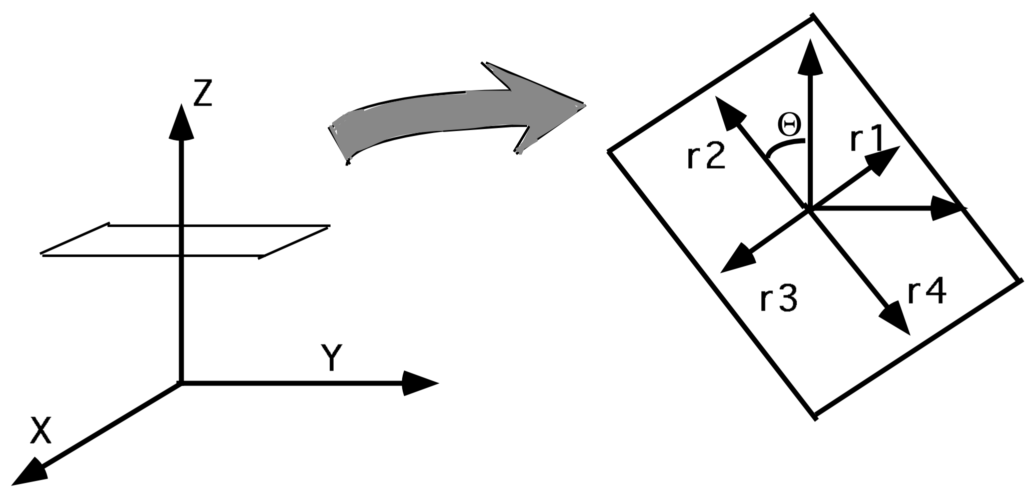


Figure 7: Parameterization of a flat, rectilinear polygon for multi-image triangulation.

The triangulation process minimizes an objective function that measures how well each 3D edge aligns with corresponding 2D line segments in the set of images. Each edge (Z, θ, r_i) of the parameterized 3D roof polygon is projected into an image to form a 2D line, $a x + b y + c = 0$. The endpoints (x1, y1) and (x2, y2) of a corresponding image line segment determined by the prior epipolar matching stage provide a perpendicular distance measure that is squared and added to the function:

$$E_l = \sum_{\substack{\text{endpoints of}\\ \text{corresponding}\\ \text{line segments}}} (ax1 + by1 + c)^2 + (ax2 + by2 + c)^2$$

This is summed over all 3D roof edge lines, and over all images, resulting in an objective function of the form:

$$E_p = \sum_{\text{images}} \sum_{\text{roof lines}} E_l$$

A standard Levenberg-Marquardt algorithm is employed to determine the set of polygon parameters($Z, \theta, r_1 \ldots r_n$) that minimize this objective function. Such nonlinear estimation algorithms typically require an initial estimate that is then iteratively refined. In this system, the original 2D rooftop polygon extracted by the building detector, and the roof height estimate computed by the epipolar matching algorithm, are used to automatically generate the initial flat-roofed polygon estimate. This results in a 2D rectangle with an associated height estimate that best fits all the images simultaneously.

After triangulation, each refined 3D roof polygon is extruded vertically down to the terrain to form a volumetric model. The extrusion process relies on being able to compute a terrain elevation (*Z* value) for each (*X,Y*) vertex location in the scene. This computation is performed by the RCDE, which can handle a number of terrain representations, ranging from a simple plane equation for relatively flat terrain, to a complete digital terrain map (DTM). For representations such as DTMs that represent terrain elevations at a discrete number of sampled locations, the elevation value at any horizontal location between samples is computed via interpolation. We compute the terrain elevation under each of the roof polygon's vertices, and select the minimum elevation as the *Z*-value for the base of the volumetric extrusion.

4.5 VOLUMETRIC HYPOTHESIS ARBITRATION

After building rooftops have been triangulated and extruded to the local DTM, they are represented as a volumetric, 3D model. The final set of buildings are filtered according to spatial overlap in order to generate a complete and consistent site model. Figure 8 shows a reconstruction that resulted in several competing model hypotheses. Arbitration of these overlapping building models is especially important when batch mode processing produces similar or identical models due to 3D reconstructions from multiple overlapping polygons from different images and processing windows. (see Section 3).

The arbitration algorithm is straightforward. Each building model volume, V_M, is intersected with each neighboring, overlapping model volume, V_O, to compute a intersection volume, V_I. If this volume is greater than a certain percentage of both V_M and V_O, then the building model with the lower certainty measure is removed from the site model. That is, if, $\frac{V_I}{V_M} > P$ and $\frac{V_I}{V_O} > P$, then the model with the lowest certainty measure computed from the grouping process (see Section 4.2) is removed. Otherwise, both overlapping models will be retained in the final output.

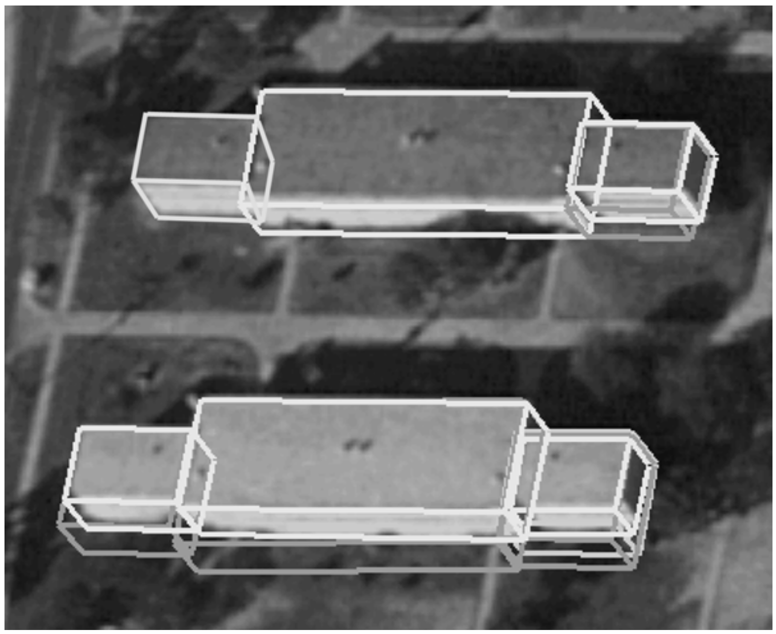


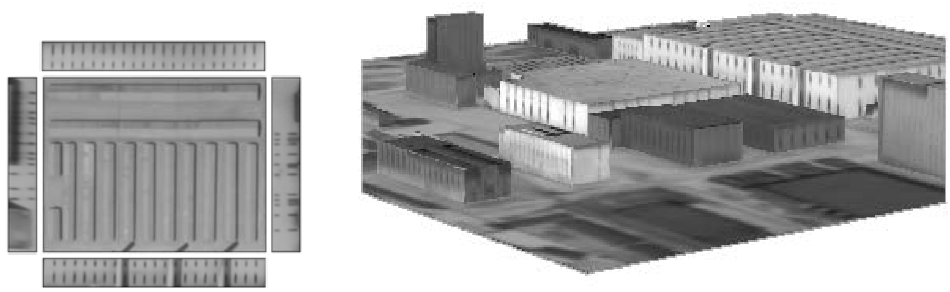
Figure 8: Multiple, overlapping hypotheses generated by the Ascender system. Brightness corresponds to the certainty value of the hypotheses. There are five alternate building models in addition to the six true models. (There are no alternative models for the top center and left models and two alternates for the bottom right model.) Only the best (brightest) of the overlapping models is retained after arbitration.

This criterion is used to filter similar overlapping building models but is not guaranteed to remove all false positives generated in the reconstruction process. For example, a small building completely contained within a larger building model will not be eliminated (since the intersected volume will always remain small with respect to the large model) even though this is not a physically realizable model. Multi-level buildings present another problem since they are often detected as two separate, overlapping polygons. Hypothesis arbitration as currently implemented may filter one of the two polygons and thus correct detail in the site model may be removed. A more sophisticated analysis of the model topology would be required to include complex multi-leveled buildings in the site model, and research into appropriate reconstruction strategies for these cases is underway.

4.6 PROJECTIVE INTENSITY MAPPING

Rapid improvements in the capability of low-end to medium-end graphics hardware makes the use of intensity mapping an attractive option for visualizing geometric site models from any viewpoint, with near real-time interactive virtual reality displays achievable on high-end workstations. These graphics capabilities have resulted in a demand for algorithms that can automatically acquire the necessary surface intensity maps from available digital photographs. We have developed routines for acquiring image intensity maps for the planar facets (walls and roof surfaces) of each building model recovered by Ascender.

Planar projective transformations provide a mathematical description of how surface structure from a planar building facet maps into an image. By inverting this transformation using known building position and camera geometry, intensity information from each image can be backprojected to "paint" the walls and roof of the building model. Since multiple images are used, intensity information from all faces of the building polygon can be recovered, even though they are not all seen in any single image (see Figure 9a). The full intensity-mapped site model can then be rendered to predict how the scene will appear from a new view (Figure 9b).



(a) (b)

Figure 9. (a) Intensity maps are stored with the planar facets of a building model. (b) A complete site site model rendered from a new view.

When processing multiple overlapping images, each building facet will often be seen in more than one image, under a variety of viewing angles and illumination conditions. This has led to the development of a systematic mechanism for managing intensity map data, called the Orthographic Facet Library. The orthographic facet library is an indexed data set storing all of the intensity-mapped images of all the polygonal building facets that have been recovered from the site, tagged with spatial and photometric indices (e.g. viewing angle, resolution, sun angle). The building facets in the library are further automatically partitioned into pieces according to whether they are sunlit, in shadow, or occluded (as determined by the viewpoint, sun angle, and the position and size of the other buildings that are hypothesized to be in the site model). In order to render new views, the multiple intensity-map versions for each building facet are “compiled” into a single, best representative intensity map for that facet. Each pixel in the representative intensity map is backprojected to determine which pieces of the intensity map in the orthographic facet library it is associated with. The set of pieces is then sorted according to a heuristic function [34] that estimates the quality of the pixel data for that piece in terms of resolution, orientation and photometric contrast, and the intensity data from the highest ranked piece is chosen as the representative value for that pixel. Each surface intensity map in the rendered image is thus a composite formed from the best available views of that building face, automatically chosen to avoid as much as possible visual artifacts caused by shadows and occlusions. While pixels are individually ranked, usually larger sets of pixels in connected components are selected from a single image because they are ranked equally in that image.

Although intensity mapping enhances the virtual realism of graphic displays, this illusion of realism is greatly reduced as the observer's viewpoint comes closer to the rendered object surface. For example, straightforward mapping of an image intensity map onto a flat wall surface looks (and is) two dimensional, unlike the surface of an actual wall, windows and doors on a real wall surface are typically inset into the wall surface and are surrounded by framing material that extends out beyond the wall surface. While these effects are barely noticeable from a distance, they are quite pronounced up close. A further problem is that the resolution of the surface texture map is limited by the resolution of the original image. As one moves closer to the surface, more detail should become apparent, but instead, the graphics surface begins to look “pixelated” and features become blurry. In particular, some of the window features on the building models we have produced are near the limits of the available image resolution.

What is needed to go beyond simple intensity mapping is explicit extraction and rendering of detailed surface structures such as windows, doors and roof vents. Our current intensity map extraction technology provides a convenient starting point, since rectangular lattices of windows or roof vents can be searched for in the orthographic facet library without complication from the effects of perspective distortion. Specific surface structure extraction techniques can be applied only where relevant, i.e. window and door extraction can be focused on wall intensity maps, while roof vent computations are performed only on roofs. As one example, a generic algorithm has been developed for extracting windows and doors on wall surfaces, based on a rectangular region growing method applied at local intensity minima in the unwarped intensity map. Extracted window and door hypotheses are used to compose a refined building model that explicitly represents those architectural details. An example is shown in Figure 10. The windows and doors have been rendered as dark and opaque, but since they are now symbolically represented, it would be possible to render the windows with glass-like properties such as transparency and reflectivity that would enhance the dynamic visualization of the scene.

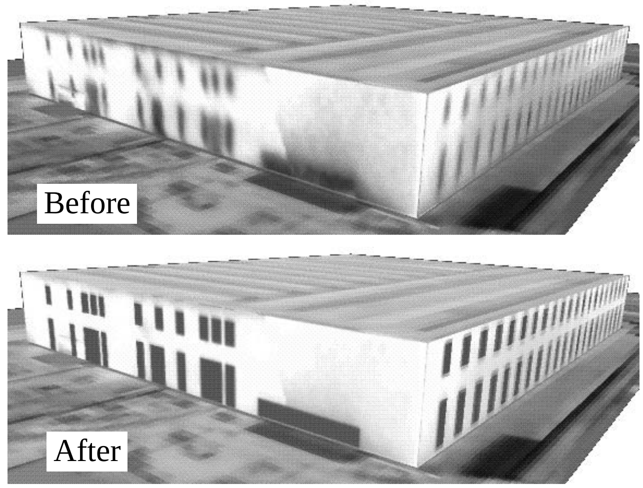


Figure 10: Rendered building model before and after symbolic window extraction.

Future work on extraction of surface structures will concentrate on roof features such as pipes and vents that appear as ``bumps'' on an otherwise planar surface area. Visual cues for this reconstruction include shadows from monocular imagery, as well as disparity information between multiple images. This is a challenging problem given the resolution of available aerial imagery.

4.7 SITE MODEL EXTENSION

The goal of site model extension is to find unmodeled buildings in new images and add them into the site model database. The main difference between model extension and model acquisition is that now the camera pose for each image can be determined via model-to-image registration. Our approach to model-to-image registration involves two components: *model matching* and *pose determination*.

The goal of **model matching** is to find the correspondence between 3D features in a site model and 2D features that have been extracted from an image; in this case determining correspondences between lines in a 3D building wireframe and 2D extracted line segments from the image. The model matching algorithm described in [3] is being used. Based on a *local search* approach to combinatorial optimization, this algorithm searches the discrete space of correspondence mappings between model and image lines for one that minimizes a match error function. The match error depends upon how well the projected model geometrically aligns with the data, as well as how much of the model is accounted for by the data. The result of model matching is a set of correspondences between model edges and image line segments, and an estimate of the transformation that brings the projected model into the best possible geometric alignment with the underlying image data.

Although a set of images with rigorous photogrammetric parameters are required to generate an initial site model, partial site models can be used to compute the pose parameters of new views and extend the capability of the system to handle poorly or partially calibrated imagery. This involves a second aspect of model-to-image registration called **pose determination**. It is important to note that since model-to-image correspondences are being found automatically, the pose determination routine needs to take into account the possibility of mistakes or *outliers* in the set of correspondences found. The robust pose estimation procedure described in [21] is being used. At the heart of this code is an iterative, weighted least-squares algorithm for computing pose from a set of correspondences that are assumed to be free from outliers. The pose parameters are found by minimizing an objective function that measures how closely projected model features overlap with their corresponding image features. Since it is well known that least squares optimization techniques can fail catastrophically when outliers are present in the data, this basic pose algorithm is embedded inside a least median squares (LMS) procedure that repeatedly samples subsets of correspondences to find one devoid of outliers. LMS is robust over data sets containing up to 50% outliers. The final results of pose determination are a set of camera pose parameters and a covariance matrix that estimates the accuracy of the solution.

The model extension process involves registering a current geometric site model with a new image, and then focusing on unmodeled areas to recover previously unmodeled buildings. This process is illustrated using the a partial site model constructed using the Ascender system applied to the Model Board 1 dataset.

Figure 11: An existing model is matched to a new view (thin lines). Areas in the new image are masked if they contain a building and the remaining image is processed for new buildings. New buildings (thick lines) are extracted and merged into a more complete site model.

Results of model-to-image registration of image J8 with the partial site model can be seen in Figure 11, which shows projected building rooftops from the previous site model overlaid on the image. Image areas containing buildings already in the site model were masked off, and the building rooftop detector was run on the unmodeled areas. The multi-image epipolar matching and constrained multi-image triangulation procedures from Sections 4.3 and 4.4 were then applied to verify the hypotheses and construct 3D volumetric building models. These were added to the site model database, to produce the extended model shown in Figure 11 (thick lines). The main reason for failure among building hypotheses that were not verified was that they represented buildings located at the periphery of the site, in an area which is not visible in very many of the eight views. If more images were used with greater site coverage, more of these buildings would have been included in the site model. The utility of this approach is explored in section 5.6 by detecting buildings in multiple views of the Ft. Hood dataset and analyzing the overall building detection rate for the site.

5. System Evaluation

The Ascender system has been delivered to government contractors for testing on classified imagery and for integration into the RADIUS Testbed System [11]. An informal transfer has also been made to the National Exploitation Laboratory (NEL) for familiarization and additional testing. The system has been extensively tested on diverse sets of data. This section presents a of experiments designed to address questions like:

1. How is the rooftop detection rate related to system sensitivity settings?
2. Is the detection rate affected by viewpoint (nadir vs. oblique)?
3. Does 2D detected polygon accuracy vary by viewpoint?
4. Is 2D geometric accuracy related to sensitivity settings
5. How does 3D accuracy vary with the number of images used?
6. Does 3D accuracy vary by the geometry of the images used?
7. How does 3D accuracy vary according to 2D accuracy of the hypothesized polygons?

Experiments were carried out using two different methods. The first set of tests were run on local image patches that were known to contain buildings. This helped to classify system performance and accuracy for a scenario in which a previous focus-of-attention mechanism has detected image regions that may contain buildings. For example, an image analyst may have selected areas in which building reconstruction should take place. Each image patch is selected by creating a bounding volume around each building in the ground truth model (discussed shortly). Each volume is then projected into each of the images using the known camera geometry for those images. This obtains all image patches of every building in the ground truth model for which the entire building appears. The system was then run on each of these projected regions.

The second set of tests deal with the case in which focus of attention regions are not available. In this case, the image is broken into overlapping windows and reconstruction takes place within each image window independently. In this “batch mode” style of processing the final reconstruction undergoes a hypothesis arbitration phase in which redundant buildings, generated from overlapping regions, are filtered (see section 4.5). The size of the window for each of the images was set to be at least as large as the largest ground truth building. The size of the overlapping area between windows was half the width of a window.

Evaluation was carried out on a large data set from Ft.Hood Texas. The imagery was collected by Photo Science Inc. (PSI) in October 1993 and scanned at the Digital Mapping Laboratory at Carnegie Mellon University (CMU) in Jan-Feb, 1995. Camera resections were performed by PSI for the nadir views, and by CMU for the oblique views.

5.1 METHODOLOGY

An evaluation data set was cropped from the Ft.Hood imagery, yielding seven subimages from the views labeled 711, 713, 525, 927, 1025, 1125 and 1325 (images 711 and 713 are nadir views, the rest are obliques). Table 1 summarizes the ground sample distance GSD for each image. The region of overlap within the scene covers an evaluation area of roughly 760x740 meters, containing a good blend of both simple and complex roof structures. Thirty ground truth building models were created by hand using interactive modeling tools provided by the RCDE. Each building is composed of RCDE "cube", "house" and/or "extrusion" objects that were shaped and positioned to project as well as possible (as determined by eye) simultaneously into the set of seven images. This has become a standard procedure for acquiring ground truth data in a domain where ground truth is difficult to obtain. The ground truth data set is shown in Figure 12.

711	**713**	**525**	**927**	**1027**	**1125**	**1325**
0.31	0.31	0.61	0.52	1.10	1.01	1.01

Table 1: Ground sample distances (GSD) in meters for the seven evaluation images. A GSD of 0.3 means that a length of one pixel in the image roughly corresponds to a distance of 0.3 meters on the ground.

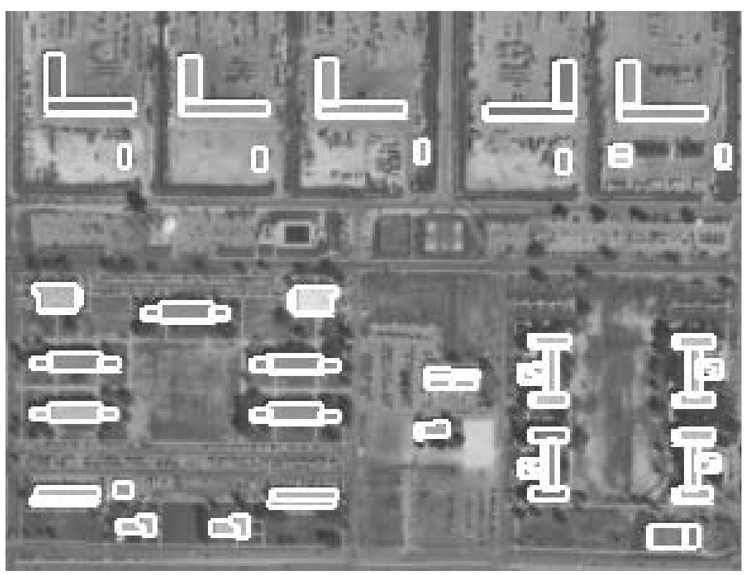


Figure 12: Ft. Hood evaluation area with 30 ground truth building models composed of single and multi-level flat roofs, and two peaked roofs. There are 73 roof facets in all. The size of the image area shown is 2375x1805 pixels.

Since the Ascender system explicitly recovers only rooftop polygons (the rest of the building wireframe is formed by vertical extrusion), the evaluation is based on comparing detected 2D and triangulated 3D roof polygons vs. their ground truth counterparts. In the set of seven images there are 73 ground truth rooftop polygons among the set of 30 buildings. Ground truth 2D polygons for each image are determined by projecting the ground truth 3D polygons into that image using the known camera projection equations.

We have utilized a metric that provides a measure of the average distance between the two polygons boundaries, reported in pixels for 2D polygons, and in meters for 3D polygons. The *Center-Line Distance* measures how well two arbitrary polygons match in terms of size, shape and location[2]. The procedure is to oversample the boundary of one polygon into a set of equally spaced points (several thousand of them). For each point, measure the minimum distance from that point to the other polygon boundary. Repeat the procedure by oversampling the other polygon and measuring the distance of each point to the first polygon boundary. The center-line distance is taken as the average of all these values. We prefer the center-line distance to other comparison measures, such as the one used in [32] since it is very easy to compute and can be applied to two polygons that do not have the same number of vertices.

For polygons that have the same number of vertices, and are fairly close to each other in terms of center-line distance, an additional distance measure is computed between corresponding pairs of vertices between the two polygons. That is, for each polygon vertex, the distance to the closest vertex on the other polygon is measured. For 2D polygons these *Inter-Vertex Distances* are reported in pixels, for 3D polygons the units are meters, and the distances are broken into their planimetric (distance parallel to the X-Y plane) vs. altimetric (distance in Z) components. An Inter-Vertex distance is only computed between vertices for which there is a corresponding ground truth polygon vertex. Therefore statistics involving the inter-vertex distance will not include vertices that are far from ground truth (from a partially detected building, for example).

5.2 EVALUATION OF 2D BUILDING DETECTION

One important module of the Ascender system is the 2D polygonal rooftop detector. If 2D building polygons are not detected in at least one image, then a complete 3D reconstruction is not possible. The detector was tested on images 711, 713, 525 and 927 to see how well it performed at different grouping sensitivity settings, and with different length and contrast settings of the Boldt line extraction algorithm.

The detector was first tested in "bounding-box mode" by projecting each ground truth roof polygon into an image, growing its 2D bounding box out by 20 pixels on each side, then invoking the building detector in that region to hypothesize 2D rooftop polygons. The evaluation goals were to determine both true and false positive detection rates *when the building detector was invoked on an area containing a building*, and to measure the 2D accuracy of the true positives.

---

[2] Robert Haralick, private communication

The detector was also tested in “batch mode” by blindly processing each image in overlapping image windows of size *N* by *N*. Each window overlapped its neighbors by *N/2* pixels. The number *N* was chosen for each image so that the image windows could encompass the largest projected ground truth building. Typically, *N* was much larger than the size of ground truth buildings.

5.3 2D DETECTION RATES

The polygon detector typically produces several roof hypotheses within a given image area, particularly when run at the higher sensitivity settings. Determining true and false positive detection rates thus involves determining whether or not each hypothesized image polygon is a good match with some ground truth projected roof polygon. To automate the process of counting true positives and tabulating their associated error, each hypothesized polygon was ranked by its center-line distance from the known ground truth 2D polygon that was supposed to be detected. Of all hypotheses with distances less than a threshold (i.e. polygons that were reasonably good matches to the ground truth), the one with the smallest distance was counted as a true positive; all other hypotheses were considered to be false positives. The threshold value used was 0.2 times the square root of the area of the ground truth polygon, that is: $Dist(hyp, gt) \leq 0.2 * \sqrt{Area(gt)}$ where *hyp* and *gt* are hypothesized and ground truth polygons, respectively. This empirical threshold allows 2 pixels total error for a square with sides 10 pixels long, and is invarient with respect to the scale of the image.

The total numbers of roof hypotheses generated for each of the images 711, 713, 525 and 927 for bounding-box processing are shown at the top of Figure 13. Total polygons per image were computed for nine different sensitivity settings of the building detector ranging from 0.1 to 0.9 (very low to very high). The line segments used for each image were computed by the Boldt algorithm using length and contrast thresholds of 10. The second graph in 13 plots the number of true positive hypotheses. For the highest sensitivity setting, the percentage of rooftops detected in 711, 713, 525 and 927 using the bounding-box strategy were 51%, 59%, 45% and 47%, respectively. The same test was performed for the system using batch-mode processing and the results are shown in figure 14. For the highest sensitivity, results similar to the bounding-box processing mode were produced. Detection rates of 46%, 55%, 42%, and 39% for each of the 711, 713, 525, 927 images respectively.

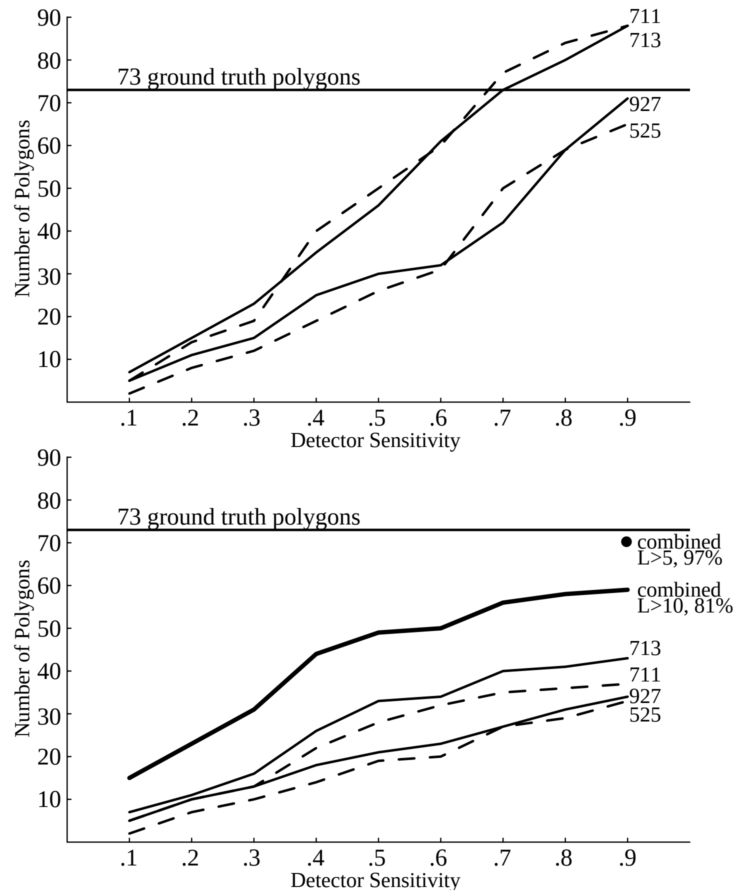


Figure 13: Bounding-Box processing detection rates.Top: Building detector sensitivity vs. total number of generated roof hypotheses per image. Bottom: Sensitivity vs. number of true positives. Horizontal lines show the actual number of ground truth polygons. Combining results from all four view yields a detection rate of 81% with lines of L > 10 , and 97% with lines of L > 5.

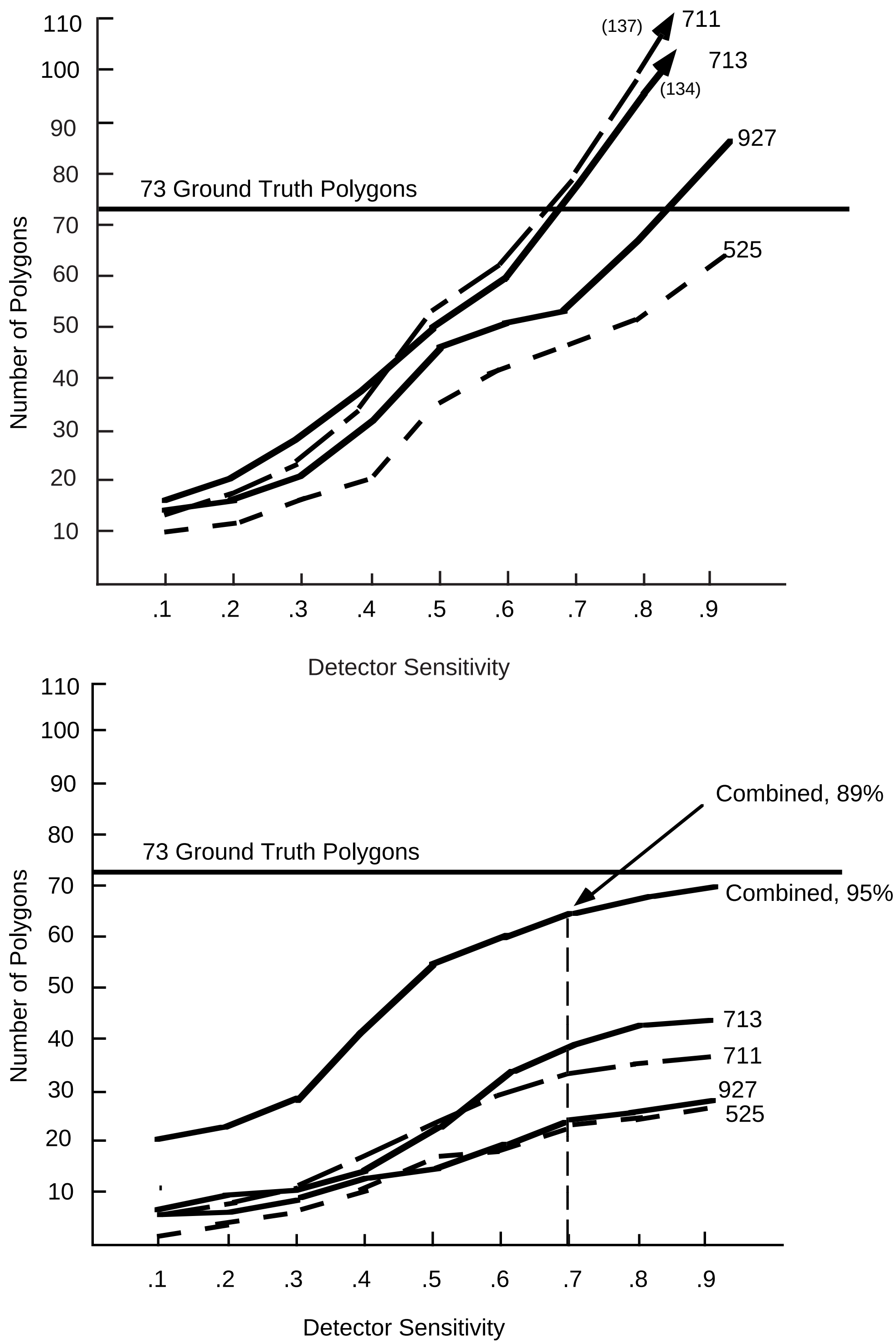


Figure 14: Batch-mode processing detection rates. Top: Building detector sensitivity vs. total number of roof hypotheses. Bottom: Sensitivity vs. number of true positives. Horizontal lines show the actual number of ground truth polygons. Combining results from all four view yields a detection rate of 95%. Combining results from a more feasible setting of 0.7 yields a combined detection result of 89% with a false positive percentage of 46%.

A significant difference between the two modes of processing is in the number of false positives generated by each technique. Because batch-mode processing involves blind application of the building detector to the entire image an increase in the number of polygons detected is expected. At the mid-to-higher range of sensitivities (0.5-0.7) the number of false positives produced is not significant, however, at the highest sensitivities, batch-mode processing produces a large number of false positives. Without a prior focus-of-attention mechanism, the batch-mode extraction is only feasible at middle-range sensitivities, which limits the number of true positives achievable.

The detection rates seem to be sensitive to viewpoint. More total hypotheses and more true positives were detected in the nadir views than in the obliques. This may represent a property of the building detector, but it is more likely that most of the discrepancy is due to the difference in GSD of the images for this area (see Table 1). Each building roof simply occupies a larger set of pixels in the nadir views than in the obliques for this data set, and therefore the nadir view of buildings has a significantly higher resolution.

To measure the best possible performance of the rooftop detector on this data, it was run on all four images at sensitivity level 0.9, using Boldt line data computed with the lowest length and contrast thresholds of 5. These were judged to be the highest sensitivity levels for both line extractor and building detector that were feasible, and the results represent the best job that the current building detector can do with each image. The percentages of rooftops detected in each of the four images under these conditions in bounding-box mode were 86%, 84%, 74%, and 67%, with a combined image detection rate of 97% (71 out of 73). Under these same conditions (ignoring false positives) the batch-mode system reconstruction percentages were 85%, 83%, 72%, and 66%, with a combined image detection rate of 95%. Using the highest possible *feasible* sensitivity for batch-mode processing at 0.7 produces 62%, 51%, 34%, and 32% detection rates for each of the images and a combined rate of 89% while limiting the false positive rate to 46% (see Figure 13). This represents the best possible performance in batch-mode while limiting the number of false positives.

Finally, the rooftop detector was run in batch-mode on all four images at a sensitivity of 0.7, using Boldt line data with length and contrast thresholds of 10. These settings were deemed to be the most feasible for batch-mode processing and were chosen to maximize the detection rate versus false positives. This reflects the proper setting of the system without specific focus-of-attention mechanisms. The set of buildings extracted in the batch-mode experiments at a sensitivity of 0.7 were combined, yielding an overall detection rate of 89%. It is interesting to note that although not all buildings are detected in one image, the use of multiple images improves results significantly. Figure 15 shows a view of the groundtruth with the number of times each of the buildings was detected in the dataset. Nearly all buildings were detected in more than one image.

The reader should understand that if a building polygon was only detected in a single image, all line correspondences in an epipolar constrained search region in the other images will contribute to the 3D building triangulation even though a complete polygon was not detected in other images. The peaked roof building at the right of the image was not detected in any image because it does not conform to the class of buildings currently built into the system (see Section 6). The center roof polygon was missed because tree cover breaks up both line and corner features. Other reasons for failure included too-low contrast between the building and ground in all images, resolution problems (as in the small second story polygon at the right of the image), and accidental alignment of surrounding clutter causing large error in the final polygon (a polygon that included the rooftop at the far left of the image with surrounding walkways was generated in two images and was eliminated because of the introduced error).

Figure 15: The ground truth model projected into image 713. The number of times each roof polygon was detected over the four different views is overlayed to depict an overall 2D detection rate of 89% for batch-mode processing at 0.7 sensitivity setting.

5.4 QUANTITATIVE ACCURACY

To assess the quantitative accuracy of the true positive 2D roof polygons, each was compared with its corresponding 2D projected ground truth polygon in terms of center-line distance. Figure 16 plots the median of the center-line polygon distances between detected and ground truth 2D polygons for different sensitivity settings. Polygons detected at low sensitivity levels seem to be slightly more accurate than those detected at the high sensitivity settings. This is so because the detector only finds clearly delineated rooftop boundaries at the lower settings, and is more forgiving in its grouping criteria at the higher settings (i.e. accepting less accurate line and polygon data) with the obvious benefit of a higher detection rate.

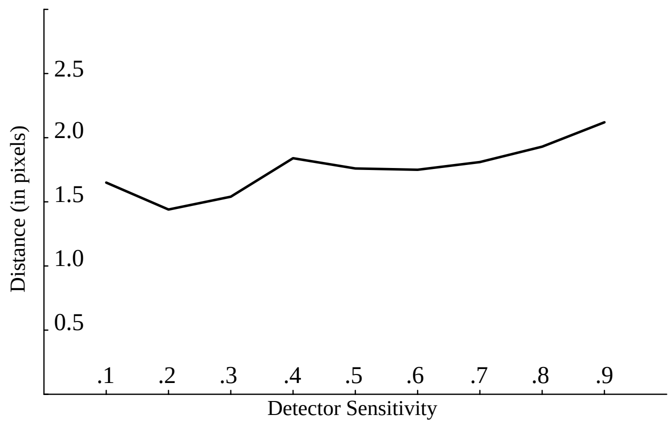


Figure 16: 2D polygon accuracy vs. Building detector sensitivity. Accuracy is represented in pixels for both modes of processing (see text).

For pairs of detected and ground truth polygons having the same number of vertices, their set of inter-vertex distances were also computed, and the medians of those measurements are broken down by image in Table 2. The average distance is around 2.7 pixels. Polygons detected in image 927 appear to be a little more accurate. This difference may or may not be significant; however, image 927 was taken in the afternoon, and all the other images were taken in the morning, so the difference in sun angle may be the cause. An interesting result is that the reconstruction accuracy of the two modes of processing is similar. The differences shown in Table 2 are statistically insignificant.

InterVertex Results for Bounding-Box Mode

	711	**713**	**525**	**927**
IV Distance	2.75	2.82	2.71	2.22

InterVertex Results for Batch Mode

	711	**713**	**525**	**927**
IV Distance	2.78	2.87	2.73	2.24

Table 2: 2D vertex accuracy. Median inter-vertex distances (in pixels) between detected polygon vertices and projected ground truth roof vertices, for four images.

5.5 EVALUATION OF 3D RECONSTRUCTION

The second major subsystem in Ascender takes 2D roof hypotheses detected in one image and reconstructs 3D rooftop polygons via multi-image line segment matching and triangulation. Two different quantitative evaluations were performed on this subsystem. The 3D reconstruction process was first tested in isolation from the 2D detection process by using 2D projected ground truth polygons as input. This initial evaluation was done to establish a baseline measure of reconstruction accuracy, that is, to see how accurate the final 3D building models would be given perfect 2D rooftop extraction. A second evaluation tested end-to-end system performance by performing 3D reconstruction using a set of automatically detected 2D image polygons.

5.5.1 Baseline 3D Reconstruction Accuracy

The baseline measure of reconstruction accuracy was performed using 2D projected ground truth roof polygons. Since these 2D polygons were generated from the same 3D ground truth 3D polygons, presumably they would optimally regenerate the initial 3D polygon model. For each of the 7 images in the evaluation test set, all the ground truth 2D polygons from that image were matched and triangulated using the other 6 images as corroborating views. The accuracy of each reconstructed roof polygon was then determined by comparing it with its 3D ground truth counterpart in terms of center-line distance and inter-vertex distances. Table 3 reports, for each image, the median of the center-line polygon distances between reconstructed and ground truth polygons in pixels for that image. Also reported are the medians of the planimetric (horizontal) and altimetric (vertical) components of the inter-vertex distances between reconstructed and ground truth polygon vertices in meters. Horizontal placement accuracy was about 0.3 meters, which is in accordance with the resolution of the images. This baseline error provides a measure of inherent 2D noise effects and pose errors in the 3D reconstruction process.

	711	713	525	927
CL Distance	0.57	0.46	0.45	0.53
IV planimetric	0.29	0.25	0.33	0.35
IV altimetric	0.49	0.42	0.37	0.43

Table 3: Evaluation of baseline accuracy of the 3D reconstruction process. Median center-line distances (in pixels) as well as inter-vertex planimetric and altimetric errors are shown (in meters) for four images. See text.

Another suite of tests was performed to determine how the number of views affects the baseline accuracy of the resulting 3D polygons. These tests were performed using image 711 as the primary image, and all 63 non-empty subsets of the other 6 views as additional views. For each subset of additional views, all 2D projected ground truth polygons in image 711 were matched and triangulated, and the median center-line and inter-vertex distances between reconstructed and ground truth 3D polygons were recorded. Figure 17 graphs the results, organized by number of images used (including 711), ranging from only two views up to all six views.

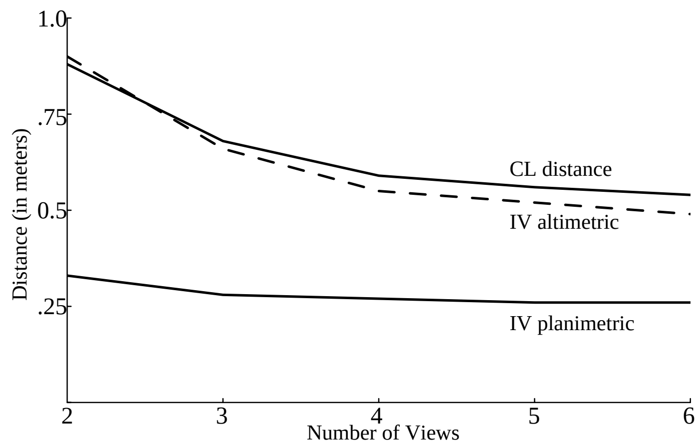


Figure 17: Number of views used vs. 3D reconstruction accuracy in meters. (see text).

The distances reported under label “2” are averaged over the 6 possible image sets containing 711 and one other image, distances reported under “3” are averaged over all 15 possible image sets containing 711 and two other images, and so on. There is a noticeable improvement in accuracy when using three views instead of two, but the curves flatten out after that, and there is only modest improvement in 3D accuracy to be gained by taking image sets larger than four.

5.5.2 ACTUAL 3D RECONSTRUCTION ACCURACY

In actual practice, Ascender reconstruction techniques are applied to the 2D image polygons hypothesized by its automated building detector. Thus, the final reconstruction accuracy depends not only on the number and geometry of the additional views used, but also on the 2D image accuracy of the hypothesized roof polygons. The typical end-to-end performance of the system was separately evaluated by taking the 2D polygons detected through both bounding-box and batch-mode processing and performing matching and triangulation using the other six views. The median center-line distances between reconstructed and ground truth 3D polygons are plotted in Figure 18 for different sensitivity settings of the polygon detector. The accuracy is slightly better when using polygons detected at the lower sensitivity settings, mirroring the better accuracy of the 2D polygons at those levels (compare with Figure 16).

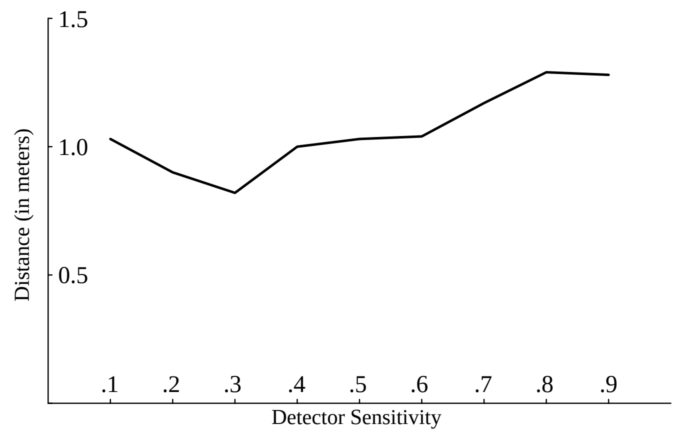


Figure 18: Building detector sensitivity vs. 3D polygon accuracy, computed as the median of center-line distances between reconstructed 3D polygons and ground truth polygons

For pairs of detected and ground truth polygons having the same number of vertices, the set of inter-vertex planimetric and altimetric errors were computed, and the medians of those measurements are shown in Table 4, broken down by the image in which the 2D polygons feeding the reconstruction process were hypothesized. Unlike the baseline error data from Table 3, where the horizontal accuracy of reconstructed polygon vertices was better than their vertical accuracy, here the situation is reversed, strongly suggesting that the planimetric component of reconstructed vertices is more sensitive to inaccuracies in the 2D polygon detection process than the altimetric component. This result is consistent with previous observations that the corners of Ascender's reconstructed building models are more accurate in height than in horizontal position [8].

Bounding-Box Mode 3D Accuracy

	711	**713**	**525**	**927**
IV planimetric	0.68	0.73	1.09	0.89
IV altimetric	0.51	0.55	0.90	0.61

Batch-Mode 3D Accuracy

	711	**713**	**525**	**927**
IV planimetric	0.67	0.75	1.11	0.90
IV altimetric	0.53	0.55	0.91	0.60

Table 4: Evaluation of actual reconstruction accuracy. Median planimetric and altimetric errors (in meters) between reconstructed 3D polygon vertices and ground truth roof vertices for the two different modes of processing.

6. Summary and Future Work

6.1 EVALUATION SUMMARY

The previous section presented results of a comprehensive evaluation of the Ascender system using an unclassified data set of Ft. Hood. While the results of the analysis are inevitably tied to this specific data set, they give some indication of how the system should be expected to perform under different scenarios.

Single-Image Performance: The building detection rate varies roughly linearly with the sensitivity setting of the polygon detector. At the high sensitivity level, roughly 50% of the buildings are detected in each image using Boldt lines extracted at length and contrast > 10, and about 75%-80% when using Boldt lines extracted with length and contrast > 5. Although line segments and corner hypotheses are localized to subpixel accuracy, the median localization error of 2D rooftop polygon vertices is around 2-3 pixels, due in part to grouping errors, but also in part to errors in resected camera pose. Note that even a perfectly segmented polygon boundary will not align with the projected ground truth roof if the camera projection parameters are incorrect.

Multiple-Image Performance: One of our underlying research hypotheses is that the use of multiple images increases the accuracy and reliability of the building extraction process. Rooftops that are missed in one image are often found in another, so combining results from multiple images typically increases the building detection rate. By combining detected polygons from four images, the total building detection rate increased to 81% using medium-sensitivity Boldt lines, and to 97% using high-sensitivity ones. Matching and triangulation to produce 3D roof polygons, and thus the full building wireframe by extrusion, can perform at satisfactory levels of accuracy given only a pair of images, but using three views gives noticeably better results. After four images, only a modest increase in 3D accuracy is gained.

Of course, any of these general statements depends critically on the particular configuration of views used. Further testing is needed to elucidate how different camera positions and orientations affect 3D accuracy. Nadir views appear to produce better detection rates than obliques, but this can be explained by large differences in the ground sample distance for this image set and may not be characteristic of system performance in general -- again, more experimentation is needed. For this data set, 3D building corner positions were recovered to well within a meter of accuracy, with height being estimated more accurately than horizontal position. The accuracy of the final reconstruction depends on the accuracy of the detected 2D polygons, as one might expect; however horizontal accuracy is more sensitive to 2D polygon errors than vertical accuracy. Also, the version of Ascender tested here uses only a simple control strategy for detecting flat-roofed buildings. More complex control strategies under development (see next section) may yield more reliable results.

6.2 FUTURE WORK

The building reconstruction strategies used in the Ascender system provide an elegant solution to extracting flat-roofed rectilinear buildings, but extensions are necessary in order to handle other common building types. Examples are complex multi-level flat roofs, peaked-roof buildings, juxtapositions of flat and peaked roofs, curved-roof buildings such as Quonset huts or hangars, as well as buildings with more complex roof structures containing gables, slanted dormers or spires.

To develop more general and flexible building reconstruction systems, a significant research effort is underway at UMass to explore alternative strategies that combine a wider range of 2D and 3D information. The types of strategies being considered involve generation and grouping of 3D geometric tokens such as lines, corners and surfaces, as well as techniques for fusing geometric token data with high-resolution digital elevation map (DEM) data. By verifying geometric consistencies between 2D and 3D tokens associated with building components, larger and more complex 3D structures are being organized using context-sensitive, knowledge-based strategies.

In addition to work that addressses a wider class of building models, improvements to the Ascender system have been implemented in order to increase the overall detection rates. Changes to the control structure that allow polygons to be detected in any of the available images and improvements to the perceputal grouping routine have increased overall detection rates. For example, in recent tests, three additional buildings have been detected at the Fort Hood site without increasing in the number of false positives.

Our symbolic building extraction procedures is being be combined with Terrest [33], a correlation-based terrain extraction system developed at UMass. The two techniques clearly complement each other: symbolic processing and triangulation of 2D lines produces 3D line features, complementing area correlation techniques that produce DEMs to which planar surfaces can be fit. Another way that they complement each other is that the terrain extraction system can determine a digital elevation map upon which the volumetric building models rest, and the symbolic building extraction procedures can identify building occlusion boundaries in exactly the locations where correlation-based terrain recovery is expected to behave poorly. A tighter coupling of the two systems is also being investigated, to allow correlation-based surface extraction to be applied to building rooftop regions to identify fine surface structure like roof vents and air conditioner units.

Bibliography

[1] J.R. Beveridge and E. Riseman, "Hybrid Weak-Perspective and Full-Perspective Matching," *Proc. Computer Vision and Pattern Recognition*,Champaign, IL, 1992, pp. 432-438.

[2] J.R.Beveridge and E.Riseman, "Optimal Geometric Model Matching under Full 3D Perspective," CVGIP: Image Understanding, Vol.61(3), 1995, pp.351-364.

[3] M.Boldt, R.Weiss and E.Riseman, "Token-Based Extraction of Straight Lines," IEEE Transactions on Systems, Man and Cybernetics, Vol.19(6), 1989, pp.1581-1594.

[4] Y.Cheng, R. Collins, A. Hanson and E. Riseman, "Triangulation Without Correspondences," ARPA Image Understanding Workshop, Monterey, CA, 1994, pp.993-1000.

[5] Y. Cheng. "Acquisition of 3D Models from a Set of 2D Images", PhD Thesis, University of Massachusetts, Computer Science Dept. Sept. 1997.

[6] R. Chung and R. Nevatia, "Recovering Building Structures from Stereo", *Workshop on the Applications of Computer Vision,* Dec. 1992, pp. 64-73.

[7] R. Collins, A. Hanson, E. Riseman. "Site Model Acquisition under the UMass RADIUS Project", ARPA Image Understanding Workshop, Monterey, CA., Nov. 1994, pp.351-358.

[8] T. Dang, O. Jamet, and H. Maitre, "Applying Perceptual Grouping and Surface Models to the Detection and Stereo Reconstruction of Buildings in Aerial Imagery", ISPRS Comm. III Symposium, pp. 165-172, 1994.

[9] A. Fischer, T. Kolbe, and F. Lang, "Integration of 2D and 3D Reasoning for Building Reconstruction", *Workshop on Semantic Modeling for the Acquisition of Topographic Information from Images and Maps*, pp. 159-180, Bonn, Germany, 1997.

[10] W. Foerstner, "Mid-Level Vision Processes for Automatic Building Extraction", *Workshop on Automatic Extraction of Man-Made Objects from Aerial and Space Images*, Monte Vertia, Switzerland, 1995.

[11] D. Gerson and S. Wood, "RADIUS Phase II - The RADIUS Testbed System", *Arpa Image Understanding Workshop*, Washington, DC. April 1993, pp. 253-260.

[12] N. Haala and M. Hahn "Data fusion for the detection and reconstruction of buildings", *Workshop on Automatic Extraction of Man-Made Objects from Aerial and Space Images*, Monte Vertia, Switzerland, 1995.

[13] O. Hennricsson, F. Bignone, W. Willuhn, F. Ade, O. Kubler, E. Baltsavias, S. Mason, and A. Grun. “Project Amobe, Strategies, Current Status, and Future Work”, International Archives of Photogrammetry and Remote Sensing, XXXI, Part B3, pp. 321-330, 1996.

[14] M Herman and T. Kanade, “Incremental Reconstruction of 3D Scenes from Multiple, Complex Images,” AI, 30(3), 1986, pp.289-341.

[15] A. Huertas and R. Nevatia, “Detecting Buildings in Aerial Images”, CVGIP, 42(3),. 2988, pp.131-152, 1988.

[16] A. Huertas, C. Lin and R. Nevatia, ``Detection of Buildings from Monocular Views of Aerial Scenes using Perceptual Grouping and Shadows," *Proc. Arpa Image Understanding Workshop*, Washington, DC, April 1993, pp.~253--260.

[17] R. Irving and D. McKeown, “Methods for Exploiting the Relationship between Buildings and their Shadows”, IEEE SMC , 19(6), 1 989, pp.1564-1575.

[18] C.Jaynes, F.Stolle and R.Collins, “Task Driven Perceptual Organization for Extraction of Rooftop Polygons,” *IEEE Workshop on Applications of Computer Vision, Sarasota*, FL, Dec.1994, pp.152-159.

[19] C. Jaynes, F. Stolle, H. Schultz, R. Collins, A. Hanson and E. Riseman, “Three-Dimensional Grouping and Information Fusion for Site Modeling from Aerial Images” *Proc. Arpa Image Understanding Workshop*, Palm Springs, CA, Feb. 1996, pp.479-490.

[20] T. Kim, J. Muller, “Building Extraction and Verification from Spaceborne and Aerial Imagery using Image Understanding Fusion Techniques”, *Workshop on Automatic Extraction of Man-Made Objects from Aerial and Space Images*, Monte Vertia, Switzerland, 1995.

[21] R.Kumar and A.Hanson, “Robust Methods for Estimating Pose and Sensitivity Analysis,” CVGIP: Image Understanding, Vol.60(3), Nov.1994, pp.313-342.

[22] F. Lang and W. Forstner, “Surface Reconstruction of Man-Made Objects using Polymorphic Mid-Level Features and Scene Knowledge”, International Archives of Photogrammetry and Remote Sensing, XXXI, Part B3, pp. 415-420, 1996.

[23] C.Lin, A.Huertas and R.Nevatia, "Detection of Buildings using Perceptual Grouping and Shadows," IEEE CVPR, Seattle, 1994, pp.62-69.

[24] C.Lin and R.Nevatia, "Building Detection and Description from Monocular Aerial Images," DARPA IUW, Palm Springs, CA, 1996, pp. 461-468.

[25] Y. Liow and T. Pavlidis, “Use of Shadows for Extracting Buildings in Aerial Images”, CVGIP, v.49, 1990, pp. 242-277.

[26] C. McGlone and J. Shufelt, "Projective and Object Space Geometry for Monocular Building Extraction", CVPR, 1994, pp. 54-61.

[27] D.McKeown et.al. "Research in the Automated Analysis of Remotely Sensed Imagery: 1994-1995," DARPA IUW, Palm Springs, 1996, pp. 215-245.

[28] J.Mundy, R.Welty, L.Quam, T.Strat, W.Bremner, M.Horwedel, D.Hackett and A.Hoogs, "The RADIUS Common Development Environment," DARPA Image Understanding Workshop, San Diego, CA, Jan.1992, pp.215-226.

[29] S. Noronha and R. Nevatia, "Detection and Description of Buildings from Multiple Aerial Images," DARPA IUW, Palm Springs, CA, 1996, pp. 469-478.

[30] M. Okutumi and T. Kanade, "A Multiple Baseline Stereo", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 4, pp. 353-363, 1993.

[31] M. Roux and D. McKeown, "Feature Matching for Building Extraction from Multiple Views", CVPR '94, 1994, pp. 46-53.

[32] M. Roux, Y. Hsieh, and D. McKeown, "Performance Analysis of Object Space Matching for Building Extraction using Several Images", *Integrating Photogrammetric Techniques with Scene Analysis and Machine Vision II*, SPIE Vol. 2486, 1995.

[33] H. Schultz, "Retrieving Shape Information from Multiple Images of a Specular Surface", *IEEE Trans. on Parttern Analysis and Machine Intelligence,* Vol. 16, No. 2, 1994.

[34] X.Wang, W.J.Lim, R.Collins and A.Hanson, "Automated Texture Extraction from Multiple Images to Support Site Model Refinement and Visualization," Proc. Computer Graphics and Visualization, Plzen, Czech Republic, 1996.

[35] U. Weidner, "An Approach to Building Extraction from Digital Surface Models", International Archives of Photogrammetry and Remote Sensing, XXXI, Part B3, pp. 924-929, 1996.