EXPLOITING PHOTOGRAMMETRIC METHODS FOR BUILDING EXTRACTION IN AERIAL IMAGES

Jefferey A. Shufelt
Digital Mapping Laboratory
School of Computer Science, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213-3891 USA
Email: *js@maps.cs.cmu.edu*

Commission III, Working Group 2

KEY WORDS: Vision, photogrammetry, recognition, extraction, modeling, image understanding, geometric constraints, automated building detection

ABSTRACT

Traditional computer vision techniques for automated building extraction have neglected the use of photogrammetric camera modeling as a source of geometric information. By incorporating knowledge about the image acquisition geometry at every phase of a building detection process, robust performance can be achieved on a wide variety of scenes. This paper describes the role of rigorous photogrammetric camera modeling in PIVOT, a fully automated building extraction system that uses only a single view to generate three-dimensional structure hypotheses. We present both qualitative and quantitative results on a varied set of complex aerial imagery.

KURZFASSUNG

Traditionelle Techniken aus dem Computer-Vision Bereich zur automatischen Gebäudeextraktion haben die Verwendung photogrammetrischer Kameramodelle als geometrische Information vernachlaessigt. Durch die Einbeziehung von Wissen über die Geometrie der Bildaufnahme auf jeder Stufe der Gebäudeerkennung können robuste Ergebnisse für eine Reihe von Szenen gewonnen werden. Dieser Beitrag beschreibt die Rolle der Kameramodellierung in PIVOT, einem vollautomatischen Gebäudeerkennungssystem, das Einzelbilder zur Ableitung dreidimensionaler Strukturhypothesen verwendet. Wir präsentieren sowohl qualitative als auch quantitative Ergebnisse für eine Reihe verschiedener, komplexer Luftbilder.

1 INTRODUCTION

Building extraction from aerial images has been a topic of great interest in the computer vision community for several years. The compilation of detailed digital cartographic databases over suburban and urban areas requires accurate modeling of manmade structures, a task currently accomplished by tedious and error-prone manual techniques. Systems capable of partially or fully automating the building extraction process would permit more efficient generation of accurate building models. From a research standpoint, building extraction also presents a challenging test for computer vision techniques. A system which achieves robust performance on aerial imagery must be able to address a wide variety of viewing angles and object shapes, correctly interpret object and shadow occlusions, and distinguish natural and manmade features.

Traditionally, computer vision techniques for building extraction have neglected the use of photogrammetric camera modeling, instead treating the image as the sole source of information. This restrictive view of the problem mandates the use of constraints on the image and the scene, to make existing vision algorithms tractable. Both region-based and feature-based techniques make strict assumptions about image geometry and scene content, and consequently exhibit poor performance on imagery where buildings are not easily segmented by intensity criteria alone, or where complex shapes are prevalent and oblique viewing angles violate assumptions about image acquisition geometry.

The central idea behind the research described in this paper is that rigorous photogrammetric camera modeling not only allows generation of building hypotheses in object space, a necessity for realistic cartographic applications [McKeown and McGlone, 1993], but also serves as a valuable source of geometric constraints for building extraction. A particularly attractive feature of these constraints is that they do not limit the scope of a building extraction system, since the constraints are intrinsic to the imaging acquisition process. Recent preliminary work illustrated the effectiveness of the combination of photogrammetric modeling with computer vision techniques [McGlone and Shufelt, 1994].

In this paper, the effects of photogrammetric modeling are discussed in the context of PIVOT (**P**erspective **I**nterpretation of **V**anishing points for **O**bjects in **T**hree dimensions), a fully automated monocular building extraction system under development at the Digital Mapping Laboratory. PIVOT employs a canonical data-driven approach to building detection, constructing intermediate features from raw edge data, and generating building hypotheses from those intermediate features. A major distinction between PIVOT and the systems preceding it is the thorough integration of photogrammetric modeling in all phases of the building extraction process.

2 VANISHING POINTS AND BUILDING PRIMITIVES

Under a central projection camera model, a set of parallel lines in a scene projects to a set of lines in the image which converge on a single point, known as a *vanishing point*. Because each vanishing point corresponds to a unique orientation in 3–space, detecting these points leads to a powerful approach for inferring 3D structure from 2D images. The classical technique for detecting vanishing points [Barnard, 1983] utilizes a *Gaussian sphere*, a unit sphere with origin at the perspective center. The endpoints of each line segment in the image form planes with the perspective center, known as *interpretation planes*. Using the sphere as an accumula-

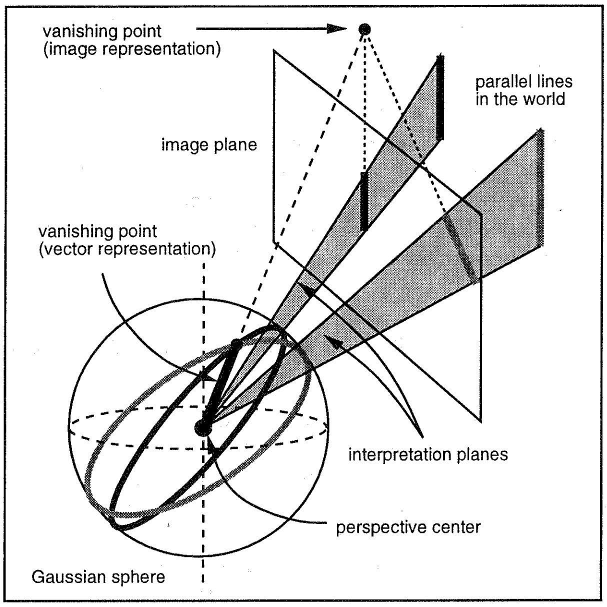

Figure 1: The geometry of vanishing points, interpretation planes, and the Gaussian sphere

tor array, the intersections of interpretation planes with the sphere (great circles) are histogrammed; maxima in the histogram then correspond to orientations shared by several line segments, and can be hypothesized as vanishing points. The geometry of vanishing points, interpretation planes, and the Gaussian sphere is depicted in Figure 1.

A difficulty with the classical approach is its sensitivity to noise. Texture edges caused by natural features in the scene can lead to spurious maxima on the Gaussian sphere, resulting in incorrect vanishing point solutions. However, these short edges exhibit greater uncertainty in image position and orientation, which can be modeled and incorporated into the sphere histogramming process. In recent work, we have proposed two edge error models which use rigorous camera modeling to treat great circles as swaths of variable width on the sphere, where the width corresponds to the uncertainty of the edge. These models locate vanishing points reliably in the presence of large amounts of noise, and are described in detail elsewhere [Shufelt, 1996].

Another difficulty with the classical technique is that it makes no provision for using knowledge about the shapes of objects of interest to guide the search for maxima on the Gaussian sphere. Rather than searching for maxima one at a time with no knowledge of scene structure, we seek a method which utilizes the expected shape of objects to find vanishing points. To develop such a method, it is first necessary to make a choice of representation for buildings.

There exists a wide spectrum of 3D representations, ranging from CAD-based models, in which shape and size are fixed, to highly parametric representations such as superquadrics and physically-based models, in which shape and size are variable. However, the immense variety of manmade constructions renders a model-based representation impractical, and highly parametric representations have proven difficult to reliably extract from complex imagery. Instead, we choose a set of 3D wireframes as "building blocks" for manmade structures; these units, which have fixed shape and topology but variable size, are known as *primitives* [Biederman, 1985; Braun *et al.*, 1995]. Geometric constraints provide sufficient leverage for extracting primitives from aerial imagery, while the combination of primitives provides representational flexibility for modeling complex manmade structures.

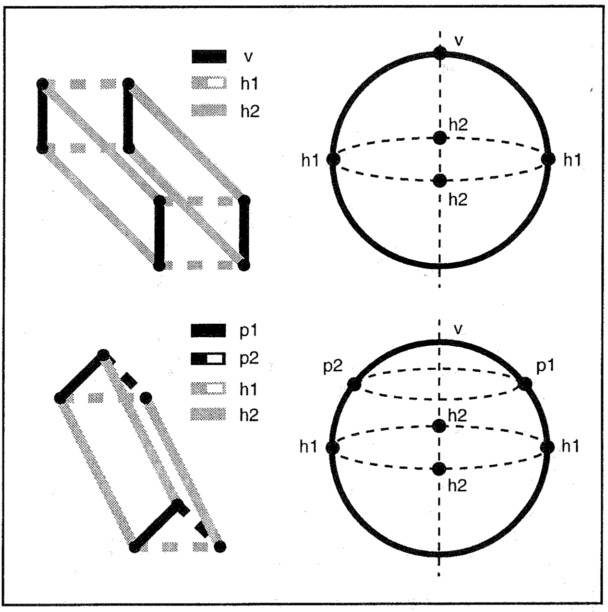

Figure 2: Rectangular and triangular primitives and their vanishing point patterns on the Gaussian sphere

PIVOT currently uses two primitives to model buildings, rectangular volumes and triangular prisms, shown in Figure 2. Rectangular volumes are composed of 3D lines with vertical and orthogonal horizontal orientations in object space (**v**, **h1**, and **h2** respectively); triangular prisms are composed of lines with two symmetric slanted "peak" lines in a vertical plane and two orthogonal horizontals (**p1**, **p2**, **h1**, and **h2** respectively). Figure 2 also depicts the orientation patterns created by these primitives on the Gaussian sphere.

Exploiting this knowledge about object shape is now simple. Rather than scanning the sphere for individual vanishing point maxima, PIVOT scans for pairs of orthogonal horizontals, with respect to the vertical vanishing point which can be computed directly from the camera parameters. Once horizontal vanishing points have been located, PIVOT scans for slanted "peak" lines which lie in vertical planes of the horizontals. This approach leads to robust vanishing point solutions for all primitive edge orientations [Shufelt, 1996].

3 GEOMETRIC CONSTRAINTS FOR CORNERS

Given a set of vanishing points, each line segment in the image can be tested to see if it lies along a line with a vanishing point. If so, the line segment can be labeled with the 3D orientation in object space corresponding to the vanishing point. At the conclusion of this process, each line segment has a set of vanishing point labels which can be used as a means of ensuring that PIVOT's intermediate corner representations are geometrically consistent. This section gives a brief discussion of the use of vanishing point labelings for detecting legal corners.

Corners are generated in PIVOT by performing a range search on all line segments, and linking together pairs of segments

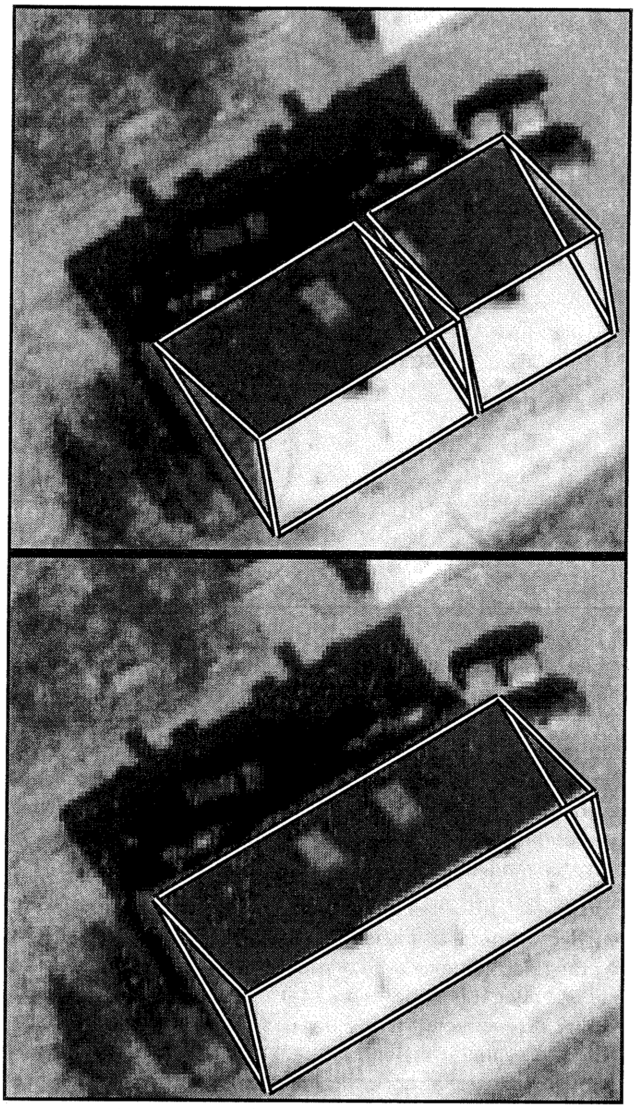

Figure 3: Automatic attachment of two adjacent triangular prisms in the FLAT_L image, before and after

whose endpoints lie in close proximity. Each corner can then be tested geometrically by comparing its labels with the primitives in Figure 2. For example, if both arms of a corner are labeled **v**, that interpretation can be discarded since a **v–v** corner does not occur in either primitive. If a corner has no legal labeling (recall that each line segment can have multiple labelings) then it can be discarded. This method efficiently prunes geometrically inconsistent features, an important aspect of mid-level feature generation [Förstner, 1995].

The basic idea is not new [McGlone and Shufelt, 1994]. However, PIVOT extends the idea to a new intermediate representation: the *2–corner*, which is formed by two corners which share a common line segment, and corresponds to a portion of the boundary of a primitive facet. For example, consider a 2–corner with the labeling **v–h2–v**. Such a 2–corner is legal, since it occurs as part of the wall boundary of a rectangular primitive (although at this stage, it has yet to be determined whether the **h2** line segment is the roof or ground segment).

2–corners are useful intermediate features because each building face can be partially represented by a 2–corner. Another intermediate feature which could have been employed in PIVOT is the *trihedral vertex*, in which three line segments meet at a point. The difficulty with trihedral vertices is that they are not visible from certain viewing angles; for example, trihedrals are often not present in conventional nadir mapping photography of rectangular structures. 2–corners can be found in both nadir and oblique imagery, allowing PIVOT to operate over a wide range of viewing angles. The combination of the 2–corner representation and vanishing point information derived from photogrammetric modeling gives PIVOT a useful intermediate representation for hypothesis construction.

4 CONSTRUCTING 3D BUILDING HYPOTHESES

Since each 2–corner corresponds to a portion of the boundary of a primitive facet, PIVOT can use the 2–corner as a starting point for locating the remainder of the primitive edges. First, PIVOT resolves ambiguities in the 2–corner interpretation. Recall that a 2–corner with the labeling **v–h2–v** is ambiguous; the **h2** segment could be on the roof or ground. This ambiguity is resolved by determining which ends of the vertical segments of the 2–corner are closer to the vertical vanishing point in image space; slanted peak roof lines can be resolved by a similar method. Once ambiguities are resolved, PIVOT then executes another search to find line segments with the correct vanishing point labelings at each of the points in the 2–corner. At the conclusion of this process, several of the edge and point slots in a primitive have been filled in with edge and point measurements from the image.

For a rectangular primitive, only one vertical and two orthogonal horizontal line segments need to be present for the positions of all eight points of the primitive to be computed in image space by intersecting vanishing lines; for a triangular prism, only the long horizontal and one of the triangular facet edges need to be present. PIVOT tries all possible combinations of the edges in the primitive slots to generate complete 2D primitives, discards any completions which do not obey the vanishing line geometry, and selects the best one with respect to the underlying edge data for the image, using a chamfer distance metric.

After this process, PIVOT has a set of fully-specified primitives, measured in image space. PIVOT then uses the camera model and a DEM (digital elevation model) to compute the object space positions of the floor points; the lengths of verticals and horizontals can then be measured in object space to obtain the 3D positions of the remaining points in each primitive. This process results in a set of 3D object space primitives, derived automatically from a DEM, the use of a central projection camera model, and monocular cues.

However, edge fragmentation can cause a single building in a scene to be modeled by several primitives. Further, depending on the viewpoint, primitives may not be found for components of the building. These problems require the ability to join primitives to form composite building structures, and the ability to extrapolate from existing primitives, respectively. PIVOT solves the first problem by joining primitives which have similarly shaped faces in close proximity; the second problem is solved for peaked roof buildings by using vertical edges and shadow analysis to estimate the height displacement of triangular prisms from the ground.

Figure 3 illustrates an example of primitive attachment on the FLAT_L scene, an image distributed as part of a test on image understanding techniques [Fritsch *et al.*, 1994]. PIVOT initially generates two triangular prisms for a single build-

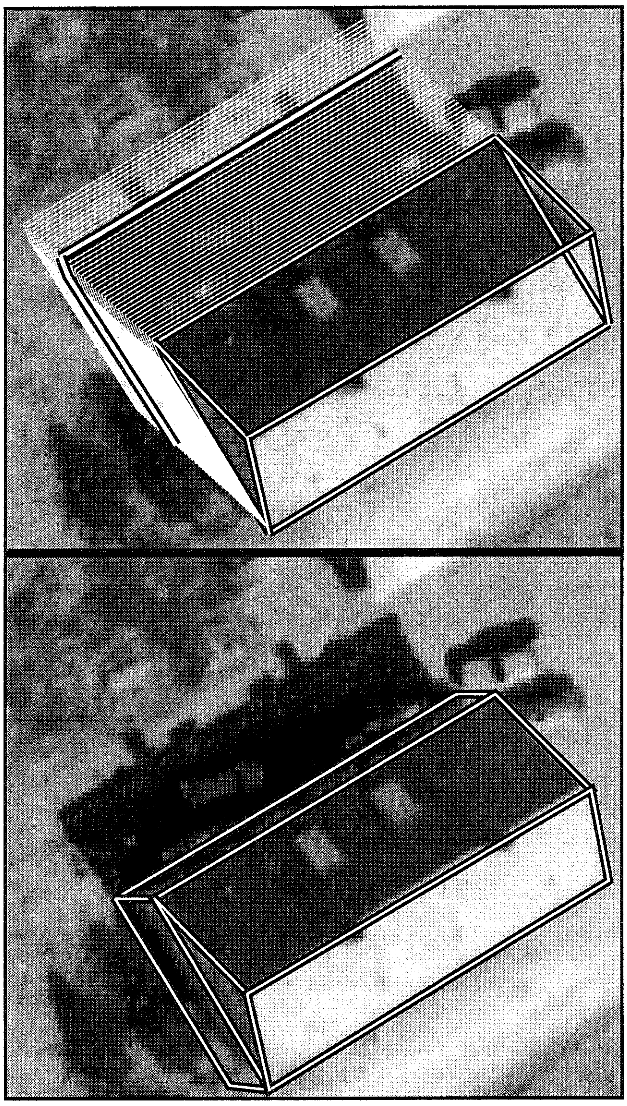

Figure 4: Automated shadow-based vertical extrusion of the triangular prism from Figure 3

ing, due to edge fragmentation along the roof ridge caused by shadow-casting protusions on the roof. However, the triangular faces of the prisms are in close proximity at each vertex, and PIVOT generates a third hypothesis by removing the intermediate points and fitting the new hypothesis to an idealized prism model in object space.

In this case, the rectangular volume supporting the prism does not have enough visible edges to generate a 2–corner, and so PIVOT's first hypothesis generation phase is unable to generate this volume. However, PIVOT looks for visible verticals at corner points of prisms to estimate the height of a possible peaked roof building [McGlone and Shufelt, 1994]; it also uses object-space shadow mensuration to estimate building height. The basic method is reminiscent of earlier work [Irvin and McKeown, 1989], but the approach described here operates in object space rather than image space. PIVOT iteratively extrudes the prism from the ground plane in object space, each time computing the unoccluded cast shadow boundary in object space and projecting it back to image space. This iteration halts when the projected shadow boundary has a good match with the image gradient, with the proper dark-to-light transition. Figure 4 shows the shadow boundaries produced at each iteration; the best match with the image is highlighted.

The building hypothesis mechanisms in PIVOT make heavy use of the photogrammetric camera model, which allows the system to generate geometrically consistent hypotheses, merge partial volumes, extrude triangular prisms by measuring verticals, and use date/time acquisition information in conjunction with the camera model to perform rigorous shadow mensuration. The use of photogrammetric analysis leads to robust performance in difficult image analysis situations where traditional image space vision techniques fail, while still utilizing only a single image.

5 3D BUILDING HYPOTHESIS VERIFICATION

The previous section closed by discussing the use of solar azimuth and elevation information in conjunction with a photogrammetric camera model to perform shadow mensuration of peaked roof building heights. This information is used again, in the hypothesis verification stage of PIVOT, to evaluate the consistency of photometric effects; namely, the expected shadow-to-ground dark-to-light transition, and the change in intensity across faces of buildings with respect to the sun vector.

PIVOT tests shadow-to-ground consistency by computing the median intensities inside two regions of the image, the *shadow region* and the *ground region*. The shadow region is created by projecting all building points to the ground along the sun vector, and then computing the enclosing polygon of these ground points. This polygon is then projected back to image space, and the building boundary polygon is subtracted from the shadow polygon, leaving the shadow region. A fixed-width distance transform is then run on the shadow region, and the building boundary region is subtracted again, leaving a ground region which is adjacent to the shadow region. If the median intensity of the ground region is less than that of the shadow region, the building hypothesis is rejected. Figure 5 illustrates these regions, as well as the key features of the next topic, inter-surface illumination consistency.

Qualitatively, we expect that surfaces which face the sun should be brighter than those which face further away. We implement this idea with a simple illumination model from computer graphics, $I = \alpha + \beta \cos\theta$, where I is the surface intensity, θ is the angle between the surface normal and the sun vector (we consider only surfaces with $\theta < \pi/2$), and α and β are constants which depend on surface material properties. For each surface, I (the median surface intensity) and $\cos\theta$ are known; α and β are unknowns. Thus, for two or more surfaces, a least-squares line fit solves for α and β. Then, if $\beta \leq 0$, we reject the building hypothesis, since its surfaces appear brighter as they face further away from the sun. PIVOT tests the consistency of roof and wall surfaces separately, since roofs are frequently composed of a different surface material than walls.

After rejecting photometrically inconsistent hypotheses, PIVOT computes a sum of scores which measure the goodness of fit to image gradient, the intensity homogeneity of each surface, and the magnitude of the shadow-to-ground transition. All hypotheses are then sorted by score, and PIVOT traverses the sorted list, each time selecting the hypothesis $\mathcal{H}$ in the list with the highest score and removing any other hypotheses in the list which were formed by attachment or extrusion of $\mathcal{H}$, or which form components of $\mathcal{H}$ if $\mathcal{H}$ was

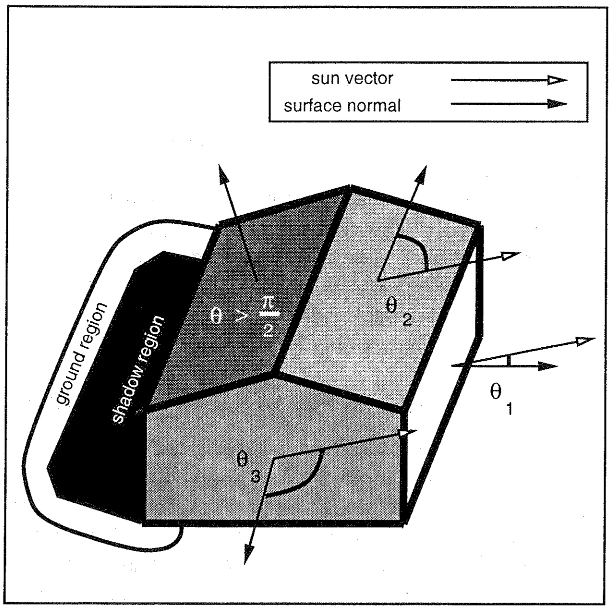

Figure 5: Computing the consistency of inter-surface intensities and the shadow-ground transition

itself formed by attachment or extrusion. Finally, if building hypotheses overlap in object space, the overlapping hypothesis with the best score is kept, and the other overlapping hypotheses are rejected.

Before proceeding to a performance analysis for PIVOT, it is important to note that the qualitative shadow analysis and illumination consistency tests, as well as the object space overlap tests, require the use of a rigorous camera model. The verification of 3D building hypotheses necessitates the ability to accurately project points to object space from image space and vice versa.

6 RESULTS AND ANALYSIS

PIVOT has been tested on 22 images to date, and experimentation continues on a growing body of test scenes. In this section, results are presented for two of these images; the FLAT_L image used in previous sections, and the RADT5WOB image, distributed under the RADIUS research initiative. We use performance evaluation metrics which have been thoroughly described elsewhere [McGlone and Shufelt, 1994]; before proceeding with the results, we first briefly describe this evaluation process.

First, ground-truth site models were compiled using the SiteCity interactive modeling system [Hsieh, 1996], which uses rigorous photogrammetric solutions and semi-automated feature extraction techniques to aid in site model compilation. These ground-truth models are then compared with the PIVOT models in 2D (image space) and 3D (object space). In 2D, models are compared on a pixel-by-pixel basis; in 3D, on a voxel-by-voxel basis (object space is subdivided into cubes $0.5m$ on a side for this evaluation). The ground truth is used to label each pixel/voxel as *building* or *background*. Then, a *true positive (TP)* pixel is one labeled as *building* by both the ground-truth and PIVOT; a *true negative (TN)* is one labeled as *background* by both. A *false positive (FP)* is labeled as *building* by the ground-truth, but as *background* by PIVOT; a *false negative (FN)* is the opposite of an FP.

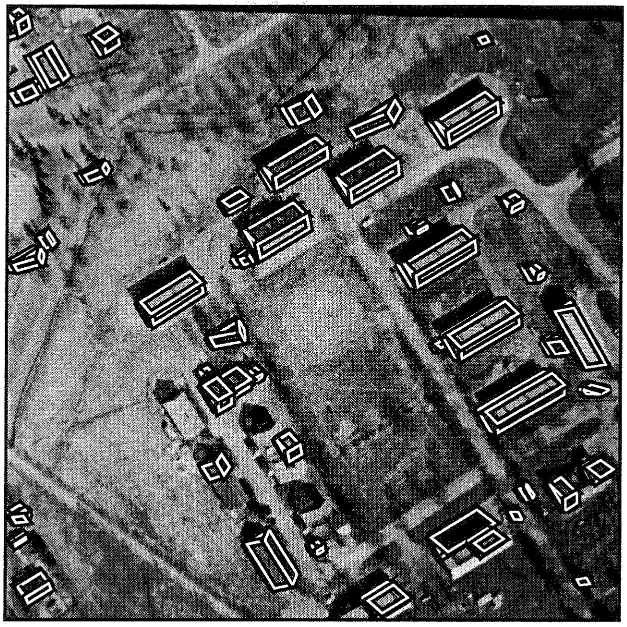

Figure 6: PIVOT results on FLAT_L

evaluation	building detection %	branch factor	miss factor	quality %
2D	67.6	0.56	0.48	49.0
3D	54.9	0.82	0.82	37.9

Table 7: Evaluation results for FLAT_L

The number of TP, FP, TN, and FN pixels/voxels are counted, and then four metrics are computed:

- Building detection percentage: $100 \times TP/(TP + FN)$
- Branching factor: FP/TP
- Miss factor: FN/TP
- Quality percentage: $100 \times TP/(TP + FP + FN)$

Figure 6 shows the final 3D object space building hypotheses produced for FLAT_L, and Table 7 gives the performance statistics for these results. PIVOT hypothesizes several structures where no buildings exist, due to false alignments of edges along vanishing lines; this explains the relatively high branching factors and low quality percentages. Many buildings are well modeled; the exceptions lie in the lower right corner of the image, where edge fragmentation hindered primitive generation. Only one building, the lightly colored peaked roof structure, is completely missed by PIVOT; that building is actually found but discarded because the median intensity of its sunward roof facet is *darker* than the other facet.

Figure 8 shows the final 3D object space building hypotheses produced for RADT5WOB, and Table 9 gives the performance statistics for these results. PIVOT detects at least some portion of every building in the scene, and the high building detection percentage and low miss factor reflect this good performance. The low branching factor and high quality percentage indicate that PIVOT is generating low amounts of false positives; only three of PIVOT's buildings lie entirely on background pixels.

These examples are representative of PIVOT's performance on other images; with the use of photogrammetric techniques,

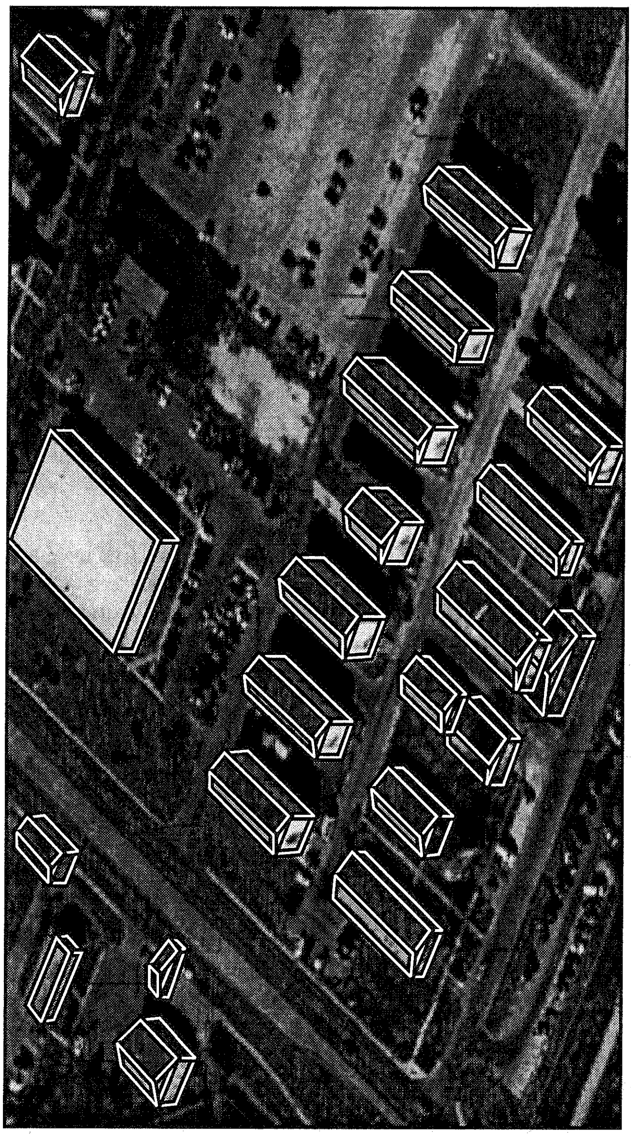

Figure 8: PIVOT results on RADT5WOB

this fully automated monocular system is able to achieve robust performance on images taken from widely differing viewpoints, with large variations in image photometry and scene content.

7 CONCLUSIONS AND FUTURE WORK

Future work on PIVOT is directed along several research avenues, many of which have been mentioned in this paper: a generalized attachment strategy for primitives to handle partial alignments of primitives; extrusion of primitives along all vanishing lines, not just in the vertical direction; the addition of trihedral vertices as a supplementary intermediate feature; the ability to handle severe low-level edge fragmentation; and capabilities for interfacing with semi-automated systems.

While PIVOT is still under development, current results illustrate the power and potential of the thorough integration of photogrammetric methods in automated feature extraction algorithms for aerial image analysis. The combination of computer vision techniques with a rigorous central projection camera model leads to superior building extraction performance on a wide variety of scenes, a necessity for the use of automated systems for the timely construction of highly detailed spatial databases.

evaluation	building detection %	branch factor	miss factor	quality %
2D	90.7	0.24	0.10	74.3
3D	84.4	0.30	0.18	67.1

Table 9: Evaluation results for RADT5WOB

8 ACKNOWLEDGEMENTS

I wish to thank the members of the Digital Mapping Laboratory, particularly Steve Cochran, Yuan Hsieh, Chris McGlone, Dave McKeown, and Michel Roux, for many insightful discussions on issues in automated building extraction.

REFERENCES

[Barnard, 1983] S. Barnard. Interpreting perspective images. *Artificial Intelligence*, 21:435–462, 1983.

[Biederman, 1985] I. Biederman. Human image understanding: Recent research and a theory. *Computer Vision, Graphics, and Image Processing*, 32:29–73, 1985.

[Braun *et al.*, 1995] C. Braun, T. H. Kolbe, F. Lang, W. Schickler, V. Steinhage, A. B. Cremers, W. Förstner, and L. Plümer. Models for photogrammetric building reconstruction. *Computers and Graphics (UK)*, 19(1):109–118, 1995.

[Förstner, 1995] W. Förstner. Mid-level vision processes for automatic building extraction. In *Proceedings: Ascona Workshop on Automatic Extraction of Man-Made Objects from Aerial and Space Images*, Ascona, Switzerland, 1995.

[Fritsch *et al.*, 1994] D. Fritsch, M. Sester, and T. Schenk. Test on image understanding. In *Proceedings: ISPRS Commission III Symposium on Spatial Information from Digital Photogrammetry and Computer Vision, Volume 30, Part 3/1*, pages 243–248, Munich, Germany, 1994.

[Hsieh, 1996] Y. Hsieh. Design and evaluation of a semi-automated site modeling system. In *Proceedings of the ARPA Image Understanding Workshop*, Palm Springs, California, February 1996.

[Irvin and McKeown, 1989] R. B. Irvin and D. M. McKeown. Methods for exploiting the relationship between buildings and their shadows in aerial imagery. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(6):1564–1575, November 1989.

[McGlone and Shufelt, 1994] J. C. McGlone and J. A. Shufelt. Projective and object space geometry for monocular building extraction. In *Proceedings: IEEE Conference on Computer Vision and Pattern Recognition*, pages 54–61, June 1994.

[McKeown and McGlone, 1993] D. M. McKeown and J. C. McGlone. Integration of photogrammetric cues into cartographic feature extraction. In *Proceedings: SPIE Conference on Integrating Photogrammetric Techniques with Scene Analysis and Machine Vision, Volume 1944*, pages 2–15, April 14-15 1993.

[Shufelt, 1996] J. A. Shufelt. Performance evaluation and analysis of vanishing point detection techniques. In *Proceedings of the ARPA Image Understanding Workshop*, Palm Springs, California, February 1996.