Multi-Image Focus of Attention for Rapid Site Model Construction

Robert T. Collins
The Robotics Institute, Carnegie Mellon University,
5000 Forbes Ave, Pittsburgh, PA 15213
http://www.cs.cmu.edu/~rcollins/

Abstract

A multi-image focus of attention mechanism has been developed that can quickly distinguish raised objects like buildings from structured background clutter typical to many aerial image scenarios. The underlying approach is the space-sweep stereo method, in which features from multiple images are backprojected onto a virtual, horizontal plane that is methodically swept through the scene. Backprojected gradient orientations from multiple images are highly correlated when they come from scene locations containing structural edges that are roughly horizontal, like building roofs and terrain; otherwise, they tend to be uniformly distributed. These observations are used to define a structural salience measure that can determine whether a given volume of space contains a statistically significant number of structural edges, without first performing precise reconstruction of those edges. The utility of structural salience for computing focus of attention regions is illustrated on sample data from Ft.Hood, Texas.

1. Introduction

In recent years there has been an enormous increase in the availability of high-resolution aerial imagery from airborne and satellite sensing systems devoted to mapping, reconnaissance, and earth-resource management. Concurrently, there has been a growing need to produce site models that include man-made cartographic features such as buildings and roads. Despite the ever-increasing need for accurate three-dimensional site models, and no shortage of high-resolution imagery from which to build them, manual site model construction remains a very tedious and expensive process. This is clearly a situation where automated IU techniques could have a great impact. However, although several automated building extraction systems have been developed [4, 12, 15, 16, 17, 20, 23], they are typically very slow on cluttered, real-world scenes because they spend an inordinate amount of time generating and testing hypotheses that later turn out to be false.

In applications such as automated target detection, where large sets of imagery must be rapidly screened, the general strategy is to apply a rapid focus of attention algorithm to identify areas that are likely to contain objects of interest. We claim that this rapid separation of wheat from chaff is just what is needed to improve the performance of automated model extraction systems, but that 2D focus of attention methods based on local pixel classification are not well-suited for identifying generic objects such as buildings that are characterized by their underlying 3D structure rather than by a small set of image appearances.

In this paper we describe a 3D focus of attention mechanism that can quickly distinguish raised objects like buildings from structured background clutter typical to many aerial scenarios. The method is based on a novel multi-image matching algorithm that we have developed. The key to using multi-image matching for rapid focus of attention is the observation that we can scan volumes of the scene to determine whether they contain a statistically significant number of structural edges, *without first performing precise reconstruction of those edges.* This observation allows us to compute a salience measure that determines whether or not a scene volume is likely to be occupied by a man-made object. Unlike traditional binocular stereo algorithms, our approach handles sets of images taken from a variety of viewing angles, resolutions and times of day, and works best when many images are available.

2. Related Work

There are two reasons why existing building extractors cannot perform rapid construction of high-quality, geometric site models. The first is that they are not rapid. Most systems are broken into a hypothesis generation phase, followed by a verification phase based on shadows or 3D in-

formation gathered from corroborating images [4, 15, 21]. The main speed problem is in the hypothesis generation phase, where low-level image features like line segments are grouped into closed polygons representing building roofs. This process is inherently slow, due to the combinatorial nature of perceptual grouping. The only way to speed up such systems is to generate fewer hypotheses, or better yet, not even attempt to find hypotheses in areas that don't contain buildings. The second drawback to existing building extractors is that they lack reliability - they miss too many buildings that are there, and hallucinate too many buildings that are not there. In our experience, if the system sensitivity is set high enough to find a good number of buildings, the system will invariably also generate a huge number of false positives that considerably slow down system run-times.

To make extraction algorithms that are faster and more reliable, some researchers have resorted to semi-automated approaches that put a "human in the loop" to tightly constrain where IU techniques are applied and to provide contextual information [11, 13, 14]. For example, if a human is willing to roughly delineate a building roof with a closed polygon, the combinatorial hypothesis generation phase of building extraction is avoided, and the system can focus on the more tractable geometric task of precisely estimating building shape and height from multiple views. However, this "solution" does not satisfy the ultimate goal of a fully automated model extraction system, and even semi-automated modeling would grow tedious over time. Even if the user input could be reduced to a single mouse click on top of each building roof in the scene, it would still be a time-consuming task to process a large, built-up area. Moreover, one mouse click could hardly convey all the contextual information necessary to ensure reliable segmentation of a building's shape and size.

A fully automated system will need a different mechanism to focus attention on buildings. Previously developed focus-of-attention techniques can be divided into into three categories: monocular schemes based on classifying 2D image regions using color or texture, methods based on exploiting high-level knowledge about the relationships between 2D regions, and methods that derive 3D information about the scene:

1) Monocular (2D) color or texture-based FOA methods are not well-suited to extracting man-made objects from grey-scale aerial imagery. Although color-based approaches work well for classifying material types using hyperspectral imagery [7, 10, 19], they are problematic in grey-scale imagery where buildings and roads can appear either lighter or darker than their surroundings. Texture-based methods work best for land-use classification of large areas containing repetitive patterns, such as farm land or parking lots [10].

2) Knowledge of visual context can provide domain-specific FOA in some situations. For example, when searching for small, indistinct objects such as cars in a large aerial image, it can be more effective to first search for larger structures like roads and then look for cars in proximity to the roads [8, 24, 25]. Another example is to use shadow regions, typically dark regions that are easy to extract, to cue the search for nearby buildings [16]. These approaches tend to be based purely on 2D proximity cues, and rely heavily on domain-specific assumptions that aren't likely to transfer to other application areas.

3) Grimson et.al [9] describe a 3D FOA mechanism for object recognition that uses binocular stereo to determine groups of line segments that are near each other in 3D, and therefore likely to belong to the same object. However, they use a traditional line matching algorithm that has a combinatorial complexity too great to be used over the whole image, and thus must first employ a simple 2D FOA mechanism (in this case color) to determine regions of the image in which the stereo matching algorithm should be applied. Another example is the active vision work of Coombs et.al. [5] on zero-disparity filtering. They note that for binocular stereo systems there exists a virtual surface in space called the horopter that corresponds to the set of 3D points that appear to have zero stereo disparity between the left and right images at a particular vergence angle. They use this idea to determine a 3D segmentation of the scene by slowly changing the stereo vergence angle (which effectively sweeps the horopter through space) and grouping pixels that have zero disparity at each depth.

We believe that the use of 3D information is key to distinguishing single raised objects like buildings from the kind of structured background clutter typical to many aerial scenarios. We have developed a novel multi-image stereo algorithm that rapidly derives a coarse 3D scene segmentation by sweeping a virtual planar surface through space and noting locations where several corresponding viewing rays converge. Our approach is most closely related to zero-disparity filtering, but is a non-biological generalization that handles configurations of multiple cameras.

3. Multi-Image Focus of Attention

This section presents a technical overview of our approach to multi-image focus of attention. In Section 3.1 we describe the space-sweep method, a multi-image stereo algorithm that we use to efficiently scan all possible sets of multi-image feature correspondences. This basic algorithm forms the heart of a structure salience measure defined in Section 3.2, which uses edge information from multiple images to measure how likely it is that a given volume of the scene contains a significant amount of 3D structure. Section 3.3 illustrates how this salience measure can be used

to perform multi-image focus of attention for raised structures such as buildings in a set of aerial images of Ft.Hood, Texas. Finally, Section 3.4 describes further metric scene information that can be computed using this approach.

3.1. Space-Sweep Stereo

We have developed a geometric algorithm called the space-sweep stereo method for efficiently bringing local image patches from potential multi-image correspondences into proximity, where they can be tested for compatibility [3]. The method is based on the premise that areas of space where image feature viewing rays (nearly) intersect are likely to be the 3D locations of observed scene features. Consider an arbitrary 3D scene location – it projects into a known 2D location in each image using the known camera projection equations. If features in the local neighborhoods of these image locations are found to be "compatible", in a sense to be made clear later, then there is a high likelihood that those image patches are in correspondance, and that the 3D scene location actually contains a scene feature that is being viewed by each camera. In this way, we can simultaneously determine multi-image feature correspondences and the 3D locations of features in the scene.

We organize the generation of all possible multi-image correspondences with a geometric algorithm that sweeps a virtual plane, partitioned into cells, through the scene along a line perpendicular to the plane. At each position of the plane along the sweeping path, features from each image are backprojected onto the plane, and feature information falling within the same cell is checked for compatibility. Cells containing feature data with a high degree of compatibility are output as the likely locations of 3D scene features, and the image locations that project into that cell are hypothesized to be in correspondence. Organizing the search for multi-image correspondences as a space-sweep in object space leads to an efficient geometric algorithm in terms of both storage space and computation time, and easily generalizes to any number of images.

Efficient Backprojection

Assume the sweeping plane is swept along the Z-axis of the scene, so that the plane equation has the form $Z = z_i$. At each position along the sweep, every cell on the sweeping plane accumulates information from image posititions that project into it. Image data is mapped onto the sweeping plane using the known camera projection equations. For the standard perspective camera model, the transformation that backprojects features from an image onto the plane $Z = z_i$ is a nonlinear 2D projective transformation that can be represented as a 3x3 matrix H_i.

The key to efficiency in the space-sweep method is that, rather than performing a new projective transformation to backproject image features onto each different position of the sweeping plane along its path, it suffices to perform the transformation once for some canonical plane position $Z = z_0$, then update the locations by mapping features between plane $Z = z_0$ and $Z = z_i$ directly. It can be shown [3] that the transformation that updates feature positions between planes $Z = z_0$ and $Z = z_i$ is a linear dilation of the form

$$\begin{aligned} x(z_i) &= \delta\, x(z_0) + (1-\delta)C_x \\ y(z_i) &= \delta\, y(z_0) + (1-\delta)C_y \end{aligned} \tag{1}$$

where $(x(z_0), y(z_0))$ and $(x(z_i), y(z_i))$ are backprojected feature positions on each plane, (C_x, C_y, C_z) is the location of the camera, and $\delta = (z_i - C_z)/(z_0 - C_z)$. This transformation represents an isotropic scaling about the point (C_x, C_y) – all orientations and angles are preserved. Also note that it is a separable transformation, thus updating of feature positions on the sweeping plane can be performed using separable image warping techniques [6]. Special graphics hardware is available for performing image warping of this type very quickly.

Using Gradient Orientation Features

In [3], the space-sweep method was used to backproject and combine Canny edge features [2] across multiple views to reconstruct 3D edges in the scene. For our present focus of attention application, we use a different type of image features – gradient orientations – for two reasons. First, we want a more dense set of feature information than is provided by Canny edges; our goal is to determine occupancy of volumes of space, not to reconstruct edges. Second, we do not want to make premature decisions at the individual image level as to what edge features are important. Very weak gradient information, when found to be consistent over multiple images, is a powerful indicator of the presence of 3D structure. Even in places where there are no structural edges, say in the middle of a roof or a field, there are still a surprising number of low-contrast edges due to surface texture and discoloration that can be picked up by this technique.

Gradient orientation information is incorporated into the space-sweep stereo algorithm as follows:

1. Compute the image gradient for each image $I_i, i = 1, \ldots, n$ using any standard gradient operator, such as Canny or Sobel. For each pixel in I_i we have two gradient components dx and dy.

2. Normalize each gradient vector by its magnitude to get unit gradient components $\sin\phi$ and $\cos\phi$.

$$\text{mag} = \sqrt{dx^2 + dy^2} \tag{2}$$

$$\cos\phi = dx/\text{mag} \tag{3}$$

$$\sin\phi = dy/\text{mag}\,. \tag{4}$$

3. Project unit gradients from each image I_i onto the sweeping plane at elevation z_j. If H_i is the projective transformation that backprojects features from I_i to $Z = z_j$,

then angle ϕ at pixel location (x, y) can be transformed by backprojecting points (x, y) and $(x + \cos\phi, y + \sin\phi)$ onto the sweeping plane.[1] The new backprojected gradient orientation θ can then be computed from this backprojected line segment. Some interpolation may be necessary to fill in gaps in the backprojected gradient image.

4. At each cell location in the sweeping plane $Z = z_j$, combine backprojected orientations $\theta_i, i = 1, \ldots, n$ from all images by forming a resultant vector with components C and S, where

$$C = \sum \cos\theta_i \;\; S = \sum \sin\theta_i \,. \tag{5}$$

3.2. Structure Salience Measure

A measure of structural salience can be defined by noting that backprojected gradient orientations from different images should be highly correlated within sweeping plane cells through which a 3D scene edge passes. This is particularly true for scene edges that are aligned with the sweeping plane, in this case horizontal scene edges along the boundaries of roads and building roofs.

We design a statistical measure of the degree to which a set of gradient orientations are correlated by considering a null hypothesis that all the angles are uniformly distributed. A sufficient statistic for testing uniformity of a set of angles $\{\theta_i | i = 1, \ldots, n\}$ is the length of their vector resultant, namely

$$R = \sqrt{(\sum \cos\theta_i)^2 + (\sum \sin\theta_i)^2}.$$

The problem of computing the probability density function $P(R)$ of R is related to the isotropic random walk on the circle, and has a long history ([18], Section 4.4). We use a formula due to Bennet [1]

$$\Phi(k) = 1 - \int_0^k P(r)\mathrm{dr} = n \int_0^\infty J_0(kt) J_1(t) \left[J_0(t)\right]^{n-1} \mathrm{dt}$$

where n is the number of images (orientation samples), k is any value of the vector resultant length R ranging from 0 to n, and J_0 and J_1 are the standard Bessel functions of order zero and one, respectively ([22], Section 6.5). In practice, this function can be tabulated offline for the number of views one wants to use. Using this formula, the probability that the resultant R takes a value between $0 \leq a < b \leq n$ is computed as $P(a < r < b) = \Phi(a) - \Phi(b)$.

Our salience measure is computed by comparing the theoretical distribution due to uniformity with the observed distribution of resultant lengths over a volume of interest in the scene. If they are in good agreement, then it is unlikely that there is a significant amount of man-made structure within that volume of the scene. Conversely, if they disagree significantly, it is a good indicator of man-made scene structure. The chi-square statistic measures how well two distributions match ([22], Section 14.3). To use it, the observed distribution is collected as a histogram ranging from 0 to n of the resultant lengths from all sweeping plane cells located within the volume of interest. The number of entries in each histogram bucket is then compared with the number that would be expected if the resultants were distributed according to the theoretical distribution $P(R)$ using the chi-square statistic

$$\chi^2 = \sum \frac{(\text{observed}(i) - \text{expected}(i))^2}{\text{expected}(i)} \,.$$

Those scene volumes where the value of χ^2 is high contain a large number of locations in which gradient orientations from multiple images are highly correlated. Volumes where the χ^2 value is low are composed mainly of locations where backprojected gradient information from multiple images is uncorrelated. The chi-square statistic thus provides us with a saliency measure to test whether a given volume in the scene contains a significant amount of scene structure.

By plotting the chi-square statistic computed for a scene patch over a range of elevations, we can derive information on what elevations contain the most scene structure (Figure 1). Typically, when no buildings are present in that scene patch, the plot exhibits only one prominant peak due to features on the terrain. When a building is present, two peaks are seen - one for ground-level structure, and one for edges along the roof.

3.3. Results on Ft.Hood Area

To illustrate use of the structural salience measure to perform focus of attention for building extraction, a 760x740 square meter test area of Ft.Hood, Texas was chosen along with seven nadir and oblique images of the site. Figure 2a shows a 2360x2300 pixel subimage of the area used. In this instance, a prior planar terrain model was available (computed by hand for a previous project). To produce focus of attention regions, salience measures were computed at horizontal grid spacings every 0.5 meters within a volume with a horizonal footprint 30 meters per side and encompassing the range of elevations from 4 to 20 meters above the terrain in that location. The resulting grid of salience measures for the whole 760x740 meter area was smoothed and thresholded to produce the focus of attention regions shown in Figure 2b (shown projected onto the same image). By masking out areas where the salience measure suggests there is no scene structure, over 75% of the image can be discarded, much of it highly cluttered. Clearly, any building extraction algorithm would perform much more efficiently

[1] Recall that it is more efficient to do the projective backprojection once, then perform separable image warping using Equation 1 for subsequent positions of the sweeping plane.

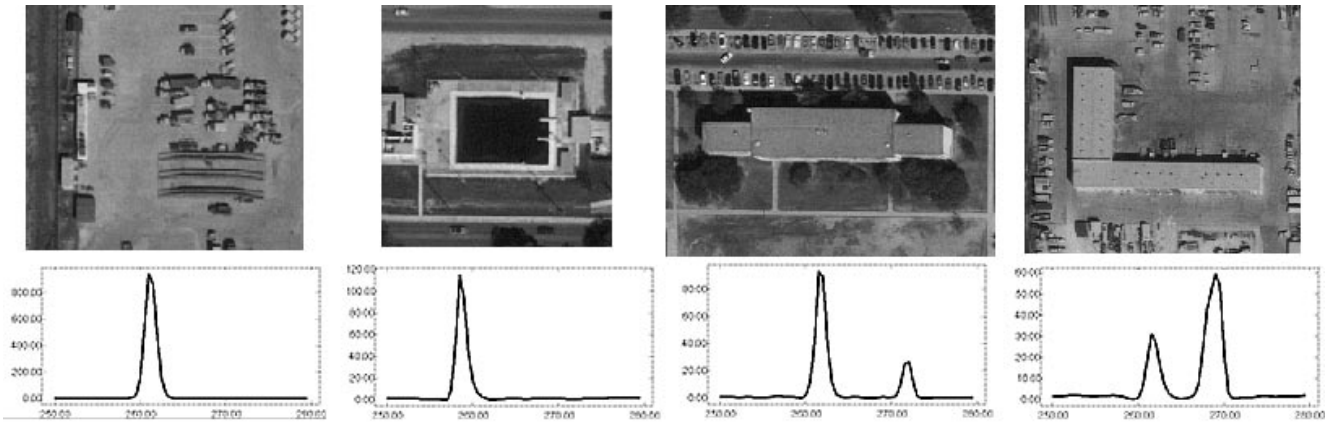

Figure 1. Four scene patches together with a plot of structural salience vs. elevation. (A) Ground clutter. (B) A swimming pool. (C) and (D) Buildings with surrounding terrain.

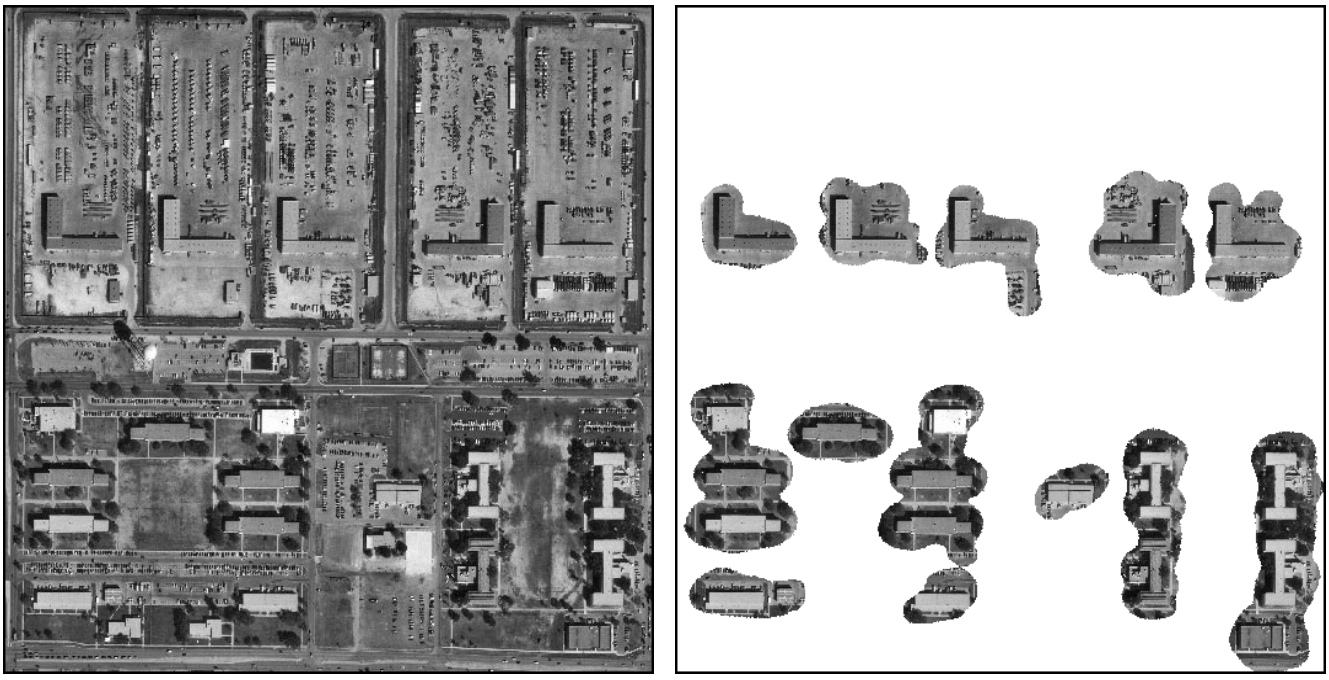

Figure 2. Focus of attention regions derived for a 760x740 meter area of the Ft.Hood, Texas site, using seven nadir and oblique views. Over 75% of this sample image can be discarded, much of it highly cluttered.

and correctly if it were presented with this reduced set of pixels rather than the whole image.

3.4. Computing Additional Information

As argued earlier, these focus of attention regions are a useful product by themselves since they allow subsequent building extraction algorithms to ignore large portions of the scene where there is no useful information. However, we briefly note here some other types of geometric information that are easily extracted using the same underlying mechanism.

Computing an Elevation Map

If we plot the salience measure computed for a horizontal scene patch over a range of elevations, peaks in the plot denote elevations where a relatively large amount of scene structure can be found. By identifying prominant peaks, and choosing the one occuring at the highest elevation, we obtain an estimate of the scene elevation at that location. Repeating this process for each pixel in a grid of horizontal scene locations results in a coarse height map of the scene, as shown in Figure 3. This result highlights the fact that this is a multi-image stereo algorithm. It also shows that, even though we are doing feature-based stereo, by using a dense set of gradient orientation features (rather than a sparse set of line or point features) we can derive a relatively dense elevation map.

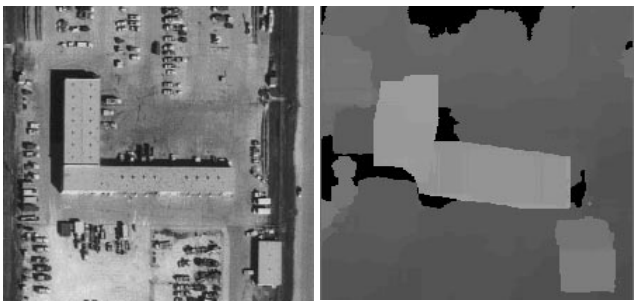

Figure 3. Scene patch along with a coarse elevation map derived using the structural salience measure.

Computing Dominant Orientation

In some building extraction applications, knowing the predominant orientation of the building structure may be helpful. This information is easily computed from the gradient resultant components C and S from Equation 5. For each cell within a volume that has a high structural salience, the orientation of the gradient resultant can be computed as $\omega = \arctan(S/C)$. Forming a histogram of these resultant orientations and identifying peaks allows the dominant orientations of the structure to be quickly identified. A common case is to find two dominant orientations that are orthogonal to each other, since many building roofs have a rectilinear structure.

4. Conclusion and Future Work

We believe our multi-image focus of attention technique will greatly improve the timeliness and quality of the site modeling process. Its benefits are that it: 1) provides three-dimensional focus of attention to identify individual man-made object like buildings that do not have a canonical two-dimensional color or texture, 2) enables full automation by replacing the human guidance needed by current semi-automated extraction algorithms, 3) enables rapid modeling by quickly discarding large, cluttered areas of the scene that have no relevance to the building extraction task, 4) directly benefits existing systems by acting as a preprocessor to explicitly identify image areas that are likely to contain man-made structures, and 5) is a true multi-image approach that combines information from multiple oblique views in a nonpreferential way to simultaneously determine multi-image correspondences and three-dimensional scene structure.

Being multi-image, feature-based stereo matching, our proposed method can continue on to extract precise 3D building models. We plan to pursue this course of action in our future work. Our approach lends itself most naturally to a coarse-to-fine *reconstruction* hierarchy, as opposed to the coarse-to-fine *image resolution* hierarchy used by many algorithms. At the coarsest level, computed most quickly, each building is represented as an enclosed volume of space. The next level of processing determines the building's height and predominant orientation. A third level of processing could finally begin to extract precise 3D locations and orientations of edge features, and link them together to form space curves delineating surface discontinuites, leading to a final level where the the building wireframe is fleshed out with surfaces and entered into the geospatial database. The method can also determine the elevation of terrain around each building, and delineate man-made features on the terrain, such as roads and walkways. All of these steps can be performed using the same underlying multi-image stereo method – space-sweep stereo – which is a rapid matching technique that enables us to make full use of the multi-look imagery provided by current and planned aerial image acquisition systems.

References

[1] W.R.Bennet, "Distribution of the Sum of Randomly Phased Components," *Quarterly of Applied Mathematics*, Vol.5(4), 1948, pp. 385–393.

[2] J.Canny, "A Computational Approach to Edge Detection," *IEEE Pattern Analysis and Machine Intelligence*, Vol.8(6), 1986, pp. 679–698.

[3] R.Collins, "A Space-Sweep Approach to True Multi-Image Matching," *IEEE Computer Vision and Pattern Recognition*, San Francisco, CA, June 1996, pp. 358-363.

[4] R.Collins, Y.Cheng, C.Jaynes, F.Stolle, X.Wang, A.Hanson and E.Riseman, "Site Model Acquisition and Extension from Aerial Images," *International Conference on Computer Vision*, Cambridge, MA, June 1995, pp. 888-893.

[5] D.Coombs, I.Horswill, and P.von Kaenel, "Disparity Filtering: Proximity Detection and Segmentation," *Proceedings Intelligent Robots and Computer Vision XI*, SPIE Vol 1825, 1992, pp.195-206.

[6] K.Fant, "A Nonaliasing, Real-time Spatial Transform Technique," *IEEE Comp. Graphics and Applications*, Vol. 6(1), January 1986, pp. 71-80.

[7] S.J.Ford and D.M.McKeown, "Utilization of Multispectral Imagery for Cartographic Feature Extraction," *DARPA Image Understanding Workshop*, January 1992, pp.805-817.

[8] T.D.Garvey, *Perceptual Strategies for Purposive Vision*, Ph.D. Dissertation, Department of Electrical Engineering, Stanford University, Stanford, CA 1975.

[9] W.E.L.Grimson, G.Klanderman, P.A.O'Donnell, and L.A.Ratan, "An Active Visual Attention System to 'Play Where's Waldo' ", *Arpa Image Understanding Workshop*, Monterey, CA, Nov 1994, pp.1059-1065.

[10] C.Harlow, M.Trivedi, R.Conners and D.Philips, "Scene Analysis of High Resolution Aerial Scenes," *Optical Engineering*, Vol.25(3), March 1986, pp.347-355.

[11] A.Heller, P.Fua, C.Connolly and J.Sargent, "The Site-Model Construction Component of the RADIUS Testbed System," *ARPA Image Understanding Workshop*, Palm Springs CA, pp.345-355.

[12] M.Herman, T.Kanade, and S.Kuroe, "Incremental Acquisition of a Three-Dimensional Scene Model from Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-6, No. 3, May 1984, pp.331–340.

[13] Y.Hsieh, "Design and Evaluation of a Semi-Automated Site Modeling System," *ARPA Image Understanding Workshop*, Palm Springs CA, pp.435–459.

[14] S.Heuel and R.Nevatia, "Including Interaction in an Automated Modeling System," *ARPA Image Understanding Workshop*, Palm Springs CA, pp.429-434.

[15] A.Huertas and R.Nevatia, "Detecting Buildings in Aerial Images," *Computer Vision, Graphics and Image Processing*, Vol. 41(2), 1988, pp.131–152.

[16] R.Irvin and D.McKeown, "Methods for Exploiting the Relationship between Buildings and their Shadows in Aerial Imagery," *IEEE TRans. on Systems, Man and Cybernetics*, Vol.19(6), 1989, pp. 1564–1575.

[17] Y.Liow and T.Pavlidis, "Use of Shadows for Extracting Buildings in Aerial Images," *Computer Vision, Graphics and Image Processing*, Vol. 49, 1990, pp. 242–277.

[18] K.V.Mardia, *Statistics of Directional Data*, Academic Press, New York, 1972.

[19] Matsuyama, "Knowledge-Based Aerial Image Understanding Systems and Expert Systems for Image Processing," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. GE-25(3), May 1987, pp.305-316.

[20] D.M.McKeown, "Toward Automatic Cartographic Feature Extraction," *Mapping and Spatial Modelling for Navigation*, L.F.Pau, Ed., Nato ASI Series, Vol.F 65, Springer-Verlag, 1990.

[21] D.McKeown et.al., "Research in the Automated Analysis of Remotely Sensed Imagery: 1994-1995," *ARPA Image Understanding Workshop*, Palm Springs CA, Feb 1996, pp.215-245.

[22] W.Press, *Numerical Recipes in C*, 2nd Edition, Cambridge University Press, NewYork, 1995.

[23] J.Shufelt and D.McKeown, "Fusion of Moncular Cues to Detect Man-made Structures in Aerial Imagery," *Computer Vision, Graphics and Image Processing*, Vol. 57(3), 1993, pp. 307-330.

[24] T.M.Strat, "Employing Contextual Information in Computer Vision," *DARPA Image Understanding Workshop*, Washington DC, April 1993, pp.217-229.

[25] L.E.Wixson and D.H.Ballard, "Using Intermediate Objects To Improve The Efficiency Of Visual Search," *International Journal of Computer Vision*, V.12, 1994, pp. 209-230.