AERIAL IMAGE MATCHING INCORPORATING OBJECT RECOGNITION

Hongwei Zhu, Research Assistant
Frank L. Scarpace, Professor
Environmental Remote Sensing Center
University of Wisconsin-Madison
Madison, WI 53706
hongweizhu@wisc.edu
scarpace@wisc.edu

ABSTRACT

A stereo image matching algorithm based on areas, multi-scale regions, edges and recognized objects, like buildings and roads is discussed. The algorithm includes five key parts: (1) Multi-scale image segmentation; (2) Object recognition/classification; (3) Matching using area, multi-scale regions, edges and recognized objects; (4) Bare earth surface reduction; (5) and manual editing of the automated matching result. The image is segmented first at different scales. Each scale represents objects at different abstraction levels. Based on the multi-scale segmentation result, some large and prominent objects, mainly buildings, roads and vegetation are recognized/classified. The larger and more prominent recognized objects and high level regions are matched first; then smaller objects and regions at lower levels. The dynamic programming technique is used to get an optimal solution. Properties of objects are also used as constraints in the matching. The result from the above matching is a DSM (Digital Surface Model), i.e., elevations of the tops of the objects. It is reduced to a DEM (Digital Elevation Model) based on the objects/regions characteristics. As expected, the generated DEM is not perfect. Manual editing is applied to fix any remaining problems. The object recognition and matching steps are interleaved. The matching results are used to assist the object recognition; in turn, the newly recognized objects are used to update the matching. The algorithm generates DEMs with higher accuracy than area-based matching alone.

INTRODUCTION

Stereo image matching is used to find conjugate points on overlapping images. This is a very critical part of digital photogrammetry for DEM (Digital Elevation Model) generation. It is also an intensively investigated topic in computer vision, in which it is referred to as a "correspondence problem". In the past decades, numerous researchers have worked on this issue and many different methods have been proposed. But, it still has not been completely resolved. The difficulty comes from many sources. Some of them are radiometric distortion, geometric distortion, occlusion, repetitive pattern and lack of features (Schenk, 1999). Generally, in digital photogrammetry, the published methods can be classified into three categories -- area-based, feature-based and symbolic (relational) (Schenk, 1999). The entities used for matching evolved from gray levels to edges, regions, then to symbolic description. The philosophy behind the evolution from one to another is to improve the uniqueness and globality of the matching entity. In the computer vision community, aside from the above three types, a pixel based method which optimize some cost function globally has been also intensively investigated. Some of the algorithms are based on dynamic programming (Ohta and Kanade, 1985; Birchfield and Tomasi, 1999; Levitin, 2003), graph cut (Veksler, 1999; Kolmogorov and Zabih, 2001, 2002), belief propagation (Felzenszwalb and Huttenlocher, 2004; Sun, Shum and Zheng, 2003; Tappen and Freeman, 2003).

In this research, the proposed matching entities include areas, multi-scale regions, edges and recognized objects, like buildings and roads. The proposed research scheme includes the following key steps:

(1). Multi-scale image segmentation
(2). Object recognition/Classification
(3). Matching using area, multi-scale regions, edges and recognized objects
(4). Bare earth surface reduction
(5). Manual editing of the automated matching result

Steps, (2) and (3) are interleaved. The matching results can be used to assist the object recognition; in turn, the newly recognized objects can be used to update the matching.

METHODOLOGY

The key tasks of this research consist of multi-scale image segmentation, object recognition/ classification, matching, bare earth surface reduction and manual editing. The overall flow diagram of the algorithm is shown in figure1.

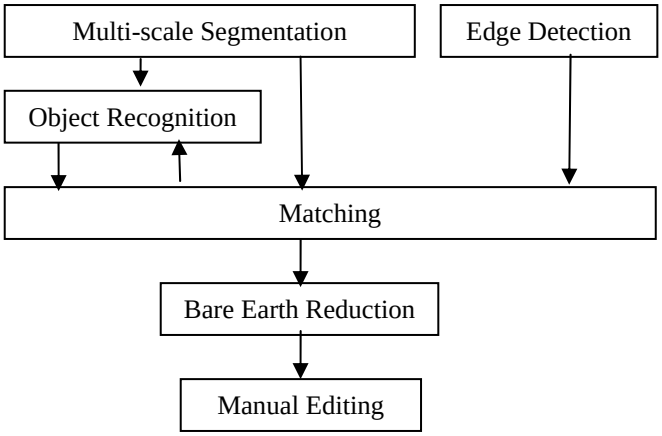

Figure 1. Algorithm Diagram.

Multi-scale Image Segmentation

The multi-scale segmentation algorithm by Baatz (2000) is used. In this algorithm, the user defines a scale parameter TH that controls the average region size of the segmented image. At the beginning, each pixel is a segment, the algorithm then loop over the following 3 steps until converge is reached.

(1) search for contiguous similar segments to merge;

(2) if found, try to merge them and calculate the change of the heterogeneities dH as shown in the following equation:

$$dH = (n_1+n_2)H_{new} - (n_1*H_1 + n_2*H_2)$$

where n_1, n_2 are the number of pixels in regions 1 and 2; H_1, H_2 are their heterogeneities and H_{new} is the heterogeneity of the combined region. If dH < TH, merge the two segments. If there are more unprocessed segments, go to (1);

(3) otherwise, check if there were changes made. If no change, exit; otherwise, start another iteration from (1).

More details of the algorithm can be found in Baatz's paper. The image is segmented at different scales. Each scale represents objects at different abstraction levels. Higher level region contains lower level regions. At higher level, a region might represent a forest; at lower level, a region might be a single tree in the forest.

Object Recognition/Classification

With the multi-scale segmentation result, some objects, like buildings, roads and vegetation areas are recognized or classified. Smaller objects within the identified objects at lower level, like markers on the roads are not recognized, because the goal of this research is matching, not complete reconstruction of objects. But those objects are used for matching. The recognized objects are matched and the matching results are used for recognition as explained in the following paragraphs. These object recognition and matching are executed alternatively to improve the performances of each other.

Based on the reviewed literature and visual analysis of the images, some strategies to recognize/ classify the buildings, roads and vegetation are implemented and summarized below.

Road extraction consists of two closely related parts: road segment extraction and road network construction. In this research, only road segment extraction is studied. The homogeneous regions are considered as road if they are elongated with smooth parallel boundaries and have almost consistent width along the running direction. After the matching result is obtained, the road can also be verified by checking if the matched points form a smooth 3D surface.

Buildings are detected as homogeneous regions with pairs of almost parallel boundaries, and some of the pairs are almost perpendicular to each other. Buildings are also above the surrounding ground and have limited height ranges. The 2D building detection need not be perfect, only the building outline or part of the outline, or even just some cue of the buildings existence is useful. The matching result is used to help detect the full extent of the building by fitting points to 3D planes. The object-oriented surface reduction strategy is applied afterward to reduce the points on building to bare ground.

Vegetations are identified by irregular region boundaries and texture analysis. If color or infra-red images are available, the color or NDVI can also be used. Additionally, the matched point elevations have high variance over the vegetated areas.

Matching

The matching is performed on epipolar images. The epipolar images can be created before or after segmentation and object recognition. The multi-scale region, recognized objects and detected edges (by Canny edge detector) are used for matching. First, the larger and more prominent recognized objects are matched, then the smaller ones. The higher level regions are then matched and followed by regions at lower levels which are inside larger regions to get denser matches. To search for conjugate objects, the objects' types, areas, perimeters, orientations, etc. are compared; for regions, similar properties are used for comparison. The detected edges that do not belong to the objects or regions are then matched to enhance the above matching results. Only those matches with a high confidence are kept. During the matching process, dynamic programming is used to get an optimal solution. The recognized objects are matched first because they are more unique and easier to match, so should be more reliable. Utilizing this strategy, the match is more reliable and efficient. Properties of objects, like "road surface is smooth and building roof is planar" are used as constraints in the matching process.

Figure 2 illustrates the matching strategy with a pair of much simplified images. The original images are segmented and objects, like building, road are recognized or identified. The road in the left image is the largest object, so its conjugate object is searched for first; then buildings H1 and H2. After all the recognized objects are matched, the higher level region R1 is then matched; then the region R2 at lower level. The searching for match of R2 is limited in R1', the conjugate region of R1. Once the conjugate objects and regions are found, matching points can be found easily. As illustrated in figure 2, epipolar line e1 crosses the building H1 and the road at points 1, 2 and 3, 4 respectively. The epioplar line e1' on the right image crosses the conjugate building and road at 1',2 and 3',4' correspondingly. The points 1 and 1', 2 and 2', 3 and 3', 4 and 4' can be considered as conjugate points. Because the objects' or regions' boundaries might not cross the conjugate epipolar lines at exactly the same location on left and right images, some points pairs found might not be real conjugate points, they have to be tested and refined.

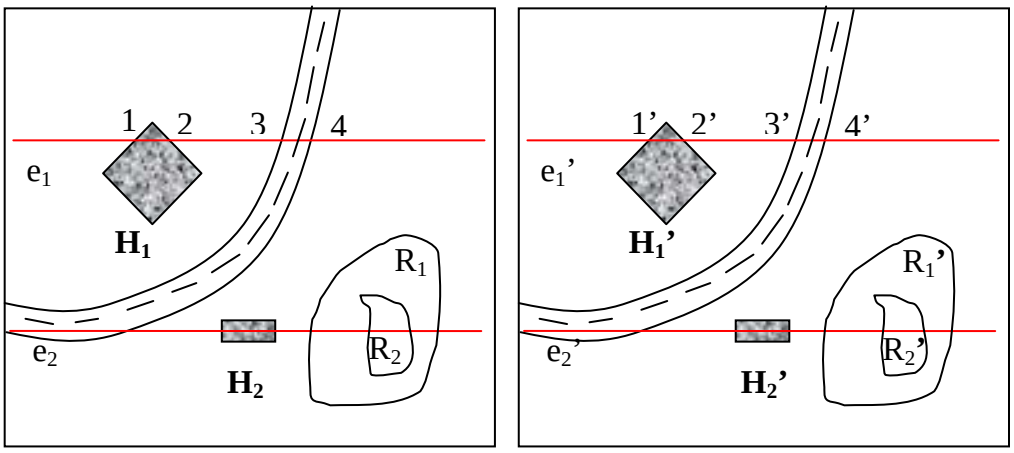

Figure 2. Illustration of the Matching Strategy.

The above described strategy does not necessarily produce dense enough matching points. For any large area that there is no match from the above object/region-based matching, dynamic programming with the above matched points as anchors or area-based match are used to densify the points. Especially, road, building and vegetations areas should have denser points than other places. There are a number of advantages of doing this: (1) for road areas, more details will likely be found and the road can be modeled better; (2) for building areas, more surrounding points can

help to better extract the building outline and model the local bare surface; (3) for vegetation areas, more points on bare ground are expected which means the bare ground reduction is more successful.

LSM (Least Square Matching) is used to achieve sub-pixel matching accuracy when needed.

Bare Earth Surface Reduction

The matching results in generation of a DSM (Digital Surface Model), i.e., elevations of the tops of objects. In photogrammetric applications, the DEM representing the bare ground is the desired product. So, the DSM must be reduced to a DEM. Filtering is the approach most researchers have used to produce a DSM from a DEM. The problem with filtering is that they are based on some models which are not necessarily correct. In addition, they are done blindly, without knowing what is on the ground, let alone the properties of the objects on the ground. This will result in smooth surface with important details smoothed out. Object oriented surface reduction discussed in this paper can overcome some of the problems.

Road surface is expected to extract correctly during match if the road segment was detected. It can be further refined by forcing points on the road to fit onto piecewise smooth surfaces.

Building roof points can be deleted or reduced to bare ground easily if the outline was extracted correctly. The building occupied area can generally be treated as a flat plane. In the absence of good building extraction, (TIN) surfaces can be constructed around the building area; the points at higher plateaus can be considered as non-ground points and reduced. Another approach is to build a histogram of the points' elevations. A histogram with two dominant peaks is anticipated. Separate the peaks at the trough; elevations at the left of the division are considered as bare ground.

Vegetation point reduction is more difficult. For single trees or small bushes, the points above them might be simply deleted. For forested areas, with dense matching, some matched points may be on the ground. This is usually true for photogrammetric projects because the photos are usually taken when the leaves are off. After fitting the on-ground points with a smooth surface, other non-ground points can be deleted or reduced. One approach to identify the on ground points is to find the regions that contain points at a small segmentation scale, and check whether these regions are bare ground. The second approach is to construct an average or least square surface with all matched points, the points above the surface can be classified as non-ground points and deleted. This procedure can be applied again to the remaining points until there is enough evidence that the bare ground surface is reached. The third approach is to start from the surrounding open areas, extend the open area surface to the tree covered area.

Manual Editing of the Matching Result

A manual stereo editing tool was developed in last few years and was applied to NCRST (National Consortium on Remote Sensing in Transportation) (Zhu, etc, 2005; Koncz, etc, 2002) and WBI (Wisconsin Buffer Initial) projects. Some useful editing functions were implemented. Points to be edited can be selected using rubber banding to draw a rectangular region on the display screen or by selecting points to form a polygon to define the region of interest. The elevation of the selected points can then be adjusted by the mouse wheel or set to a user inputted value, or fit to a least square 3D plane of the selected points. The unwanted points can also be deleted.

EXPERIMENTAL DATASET AND RESULT ASSESSMENT

Experimental Dataset

Color and color infrared aerial photos were acquired for University of Wisconsin-Madison campus in May, 1996. The nominal scale of the photographs is 1:5000, and the forwardlap and sidelap of the block are 80% and 40% respectively. The photos were scanned at 15μm, with pixel size on ground 7.5cm by 7.5cm. They were oriented with geodetic and GPS controls.

There were DEM, buildings and other information collected with the Socet Set softcopy program. The compiled DEM accuracy is about 12cm. They are used as reference data. Some GPS points are also collected as check points.

The study focuses at the north part of the campus as indicated by the rectangle in figure 3. There are different terrain types presented in this area: open field, forested hill, residential area with roads and buildings, etc. For each terrain types listed above, two or three stereo models are picked for study.

Figure 3. Study Area: University of Wisconsin-Madison Campus.

Result Assessment

Area-based and this object-recognition integrated matching methods are used to collect two sets of 3D points, one for each method. The data collected using Socet Set are considered as "truth" and are used as reference. The RMSE (Root Mean Square Error) is calculated to evaluate the accuracy of the matching results.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (Z_i^{'} - Z_i)^2}{n}}$$

where n is number of reference points; Z_i is the elevation of the reference point and Z'_i is the linearly interpolated elevation from the matched 3D points at the reference point location (Maune, 2001; Zhu, etc, 2005; Koncz, etc, 2002). To get the linearly interpolated elevation, a TIN (Triangulated Irregular Network) is created from the matched 3D points. Each reference point will be positoned within one 3D triangle when they both are projected to the X-Y plane. The vertical line through the reference point intersects the triangle at one point. The elevation of the intersection point is the linearly interpolated elevation at the reference location.

To assess the accuracy of DEM collected by the discussed method and compare it to those collected by area-based method, RMSE values are calculated for DEMs collected by both methods. There are categorized and overall assessments. For each terrain type -- open field, forested area and residential area with roads and buildings, RMSE values are calculated. There are also overall RMSE values calculated.

To assess the manual editing time improvement from area-based to the object-recognition integrated method, the time spend on manual editing the DEMs collected by these two technologies to achieve similar accuracy are kept tracked and compared.

CONCLUSIONS

We have not yet reached any results at the time this paper was written. The hypotheses of the research are "The DEM generated with the proposed method will have a better accuracy in terms of RMSE than area-based method alone. The manual editing time of the matching results from the proposed method will be less than that from the area-based method, when similar accuracy is achieved".

REFERENCES

Baatz, M., A. Schäpe (2000). Multi-resolution segmentation – an optimization approach for high quality multi-scale image segmentation. In: STROBL, J. et al. (Hrsg.): *Angewandte Geographische Informationsverarbeitung – Beiträge zum AGIT-Symposium Salzburg 2000*, Karlsruhe, Wichmann-Verlag, pp.12-23.

Birchfield, S. and C. Tomasi (1999). Depth discontinuities by pixel-to-pixel stereo. *International Journal of Computer Vision*, *35*(3):269-293.

Felzenszwalb, P. and D. Huttenlocher (2004). Efficient belief propagation for early vision. IEEE Conference on Computer Vision and Pattern Recognition, 2004.

Kolmogorov, V. and R. Zabih (2001). Computing visual correspondence with occlusions using graph cuts. International Conference on Computer Vision, 2001.

Kolmogorov, V. and , R. Zabih (2002). Multi-camera scene reconstruction via graph cuts. European Conference on Computer Vision, 2002.

Koncz, N., H. Zhu, F. L. Scarpace, A. Vonderohe and T. Adams (2002). Comparison of surface models derived by manual, LIDAR and digital photogrammetric techniques for a highway corridor, Proceedings of the 2002 NCRST/TRB/ISPRS/Pecora Conference, Denver, CO, November 8-15, 2002.

Levitin, A. (2003). Introduction to The Design and Analysis of Algorithms. Addison Wesley.

Maune, D., editor (2001). Digital Elevation Model Technologies and Applications: The DEM User Manual.

Ohta, Y. and T. Kanade (1985). Stereo by intra- and inter-scan line search using dynamic programming. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *7*(2):139–154.

Schenk, T. (1999). Digital Photogrammetry, vol. 1. TerraScience.

Sun, J., H. Shum and N. Zheng (2003). Stereo matching using belief propagation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *25*(7):787-800.

Tappen, M. and W. Freeman (2003). Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters. International Conference on Computer Vision, 2003.

Veksler, O. (1999). Efficient graph-based energy minimization methods in computer vision. Ph.D. thesis, Cornell University.

Zhu, H., A. Padmanabhan, N. Koncz, F. L. Scarpace, A. Vonderohe (2005). Automated photogrammetric surface extraction using LiDAR data as first approximations. Proceeding of American Society for Photogrammetry and Remote Sensing Annual Conference, March 7-11, 2005.