APPLYING COMPUTER VISION TECHNIQUES TO PERFORM SEMI-AUTOMATED ANALYTICAL PHOTOGRAMMETRY

*David Nilosek**

Rochester Institute of Technology
54 Lomb Memorial Drive
Chester F. Carlson Center for Imaging Science
drn2369@cis.rit.edu

Carl Salvaggio

Rochester Institute of Technology
54 Lomb Memorial Drive
Chester F. Carlson Center for Imaging Science
salvaggio@cis.rit.edu

ABSTRACT

The purpose of this research is to show how common computer vision techniques can be implemented in such a way that it is possible to automate the process of analytical photogrammetry. This work develops a workflow that generates a sparse three-dimensional point cloud from a bundle of images using SIFT, RANSAC, and a sparse bundle adjustment along with basic photogrammetric methods. It then goes on to show how the output of the sparse reconstruction method can be used to generate denser three-dimensional point clouds that can be facetized and turned into high resolution three-dimensional models. This workflow was successfully tested on a five image dataset taken with RIT's WASP imaging sensor over the Van Lare wastewater treatment plant in Rochester, NY.

1. INTRODUCTION

With recent advances in the field of computer vision, the number of automatic and semi-automatic methods of reconstructing three-dimensional point clouds of objects using multi-view images has grown. Many of these algorithms are focused on using a large database of images collected from many people to reconstruct structures that many people visit, a process often referred to as "phototourism"[1] .

There is an interest in the remote sensing and photogrammetry community to find ways to automatically extract three-dimensional models of scenes. Techniques using LIDAR data along with a sparse three-dimensional point cloud generated using the phototourism technique have been developed [2]. Other techniques have attempted to reconstruct building structure by matching lines and planes across aerial imagery[3]. Finally, other methods have attempted to use remote sensing segmentation/classification algorithms to learn more about the scene to do the reconstruction [4]. The approach of this research is to utilize photogrammetric techniques on remotely-sensed imagery to perform analytical photogrammetry by taking advantage of techniques developed in the computer vision community. This problem poses a challenge to the standard three-dimensional reconstruction process. The number of images overlapping the target of interest is generally small, on the order of two to three overlapping images. This will reduce the number of correspondences and consequently the fidelity of the three-dimensional point cloud. This problem circumvented by augmenting the phototourism process with other computer vision techniques to generate denser correspondences. This involves merging concepts from both the computer vision community and the photogrammetry community to produce an accurate three-dimensional model of a target object in the scene. A workflow has been established that uses nadir imagery taken from the RIT WASP sensor to extract structure from a scene[5].

This workflow is broken up into four separate parts; 1) feature detection and camera pose estimation , 2) sparse three-dimensional reconstruction and optimization, 3) geo-rectification, and 4) dense model extraction. This process exports this model as an OBJ/ODB file for input into the physical simulation environment, such as the DIRSIG environment developed by the Digital Image and Remote Sensing Lab at RIT [6]. Currently all facets on the object can only be assigned a single spectra, however, future work will attempt to use classification/segmentation methods to assign appropriate spectra to different facets on the model.

2. FEATURE DETECTION AND POSE ESTIMATION

The scale invariant feature transform (SIFT) algorithm is currently one of the more popular feature detectors used by the computer vision community to perform image-to-image correspondence [7]. This algorithm is capable of generating thousands of invariant features in an image. Invariant, in this case, means the spatial region around the feature will remain constant if the viewpoint changes. The SIFT algorithm uses difference of Gaussian kernels of varying widths to generate features along edges and at corners in the image. A gradient

*This work is being carried out under funding received from the United States Department of Energy, National Nuclear Security Administration BAA PDP08 Grant Number DE-ARS2-07NA28115.

978-1-4244-9300-5/10/$26.00 ©2010 IEEE

histogram of the region around each feature is calculated. The gradient histogram describes the general orientation of the region around the feature and it is described relative to the orientation and scale of the detected feature. Due to the relative description of the gradient histogram, it can be used as a scale and rotation invariant descriptor for each feature. This descriptor will not change greatly when the viewpoint for the target is changed. This allows for matching to be easily done between two images described by SIFT features. Matching features using SIFT is simply done by finding the closest matching descriptor between two images using a minimum Euclidean distance between the descriptor vectors [7]. This yields the initial correspondence between the images. Figure 1 shows how thousands of points between two images can be matched.

Once the correspondences are found the next step is

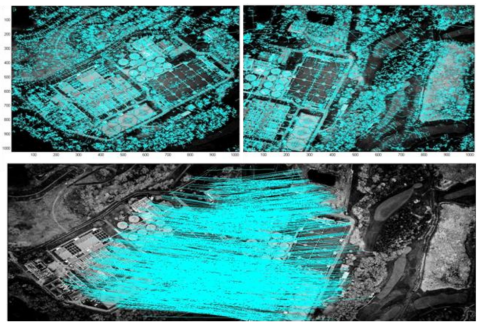

Fig. 1. Thousands of features detected and matched between two images[8]

removal of the bad correspondences and estimation of the epipolar geometry of the system. This is done in one process using the Random Sample Consensus (RANSAC) model fitting approach. RANSAC is an optimization algorithm that repeatedly and randomly selects a number of observations from a large set of data and attempts to determine the best model, excluding outliers[9].

The model that is used in this situation is taken from the computer vision community. For any two images looking at the same object from two different views; a point in one image will correspond to a line in other image, called an epipolar line. This epipolar line could be thought of as an image of the line originating from each point in the first image. This relationship between the two images is described by the fundamental matrix.

$$\mathbf{F}\mathbf{x_1} = \mathbf{l_2} \tag{1}$$

F is the 3x3 fundamental matrix, x_1 is a homogeneous point in image 1, and l_2 is the epipolar line in image 2. Homogenous geometry dictates that the dot product between a point and a line where the point falls on the line must equal zero. So the corresponding point in image 2 is related to the epipolar line in image 2 as

$$\mathbf{x_2^T}\mathbf{l_2} = 0 \tag{2}$$

Substituting equation 1 into equation 2, allows the fundamental correspondence equation to be derived.

$$\mathbf{x_2^T}\mathbf{F}\mathbf{x_1} = 0 \tag{3}$$

Any two matching points, given that the fundamental matrix relationship exists between the two images looking at the same object, must obey this equation[10]. The fundamental correspondence equation is the model fit using RANSAC. The dataset comes from the initial SIFT matches. Correspondences are randomly selected and the fundamental matrix is calculated. The fundamental matrix that fits the dataset best is chosen as the fundamental matrix for the two images. Matches that do not satisfy equation 3, within some threshold, are determined to be false matches and are removed. The output of the SIFT-RANSAC process yields a sparse selection of point correspondences between two images and the fundamental matrix describing the pose of these images relative to each other.

3. SPARSE THREE-DIMENSIONAL RECONSTRUCTION AND OPTIMIZATION

Once a selection of point correspondences has been found, basic photogrammetry can be used to calculate the three-dimensional point for each correspondence. One advantage that aerial images have in this process is that the camera position data is often available for each image, eliminating the need to estimate it from the data provided, thus reducing errors. For the purposes of this research it is assumed that the imagery was taken coincident with accurate inertial measurement data.

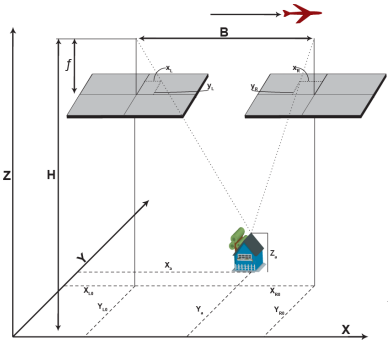

Fig. 2. The model used for calculating the three-dimensional coordinates in basic photogrammetry

Figure 2 depicts the model that is assumed using basic photogrammetry to calculate the three-dimensional point. The X,Y, and Z coordinate can all be calculated based on

known parameters between the two images using the geometry of the model. B is the baseline between the two images, H is the flying height, f is the focal length, and the subscripts l and r refer to the left and right images, respectively[11].

$$X = \frac{Bx_l}{x_l - x_r} \tag{4}$$

$$Y = \frac{By_l}{x_l - x_r} \tag{5}$$

$$Z = H - \frac{Bf}{x_l - x_r} \tag{6}$$

This model assumes the plane is flying flat and level along the x-axis in this coordinate system. It is unlikely that the plane will do this in a real collection so the flight line must be found and the image coordinates have to be transformed relative to the flight line to fit this model. The coordinates also have to be projected onto a virtual focal plane that simulates a nadir-looking image to remove the roll, pitch, and yaw effects of the aircraft. This error correction is done here so that a good estimate of the three dimensional point can be found. Having an accurate estimate of the three dimensional point allows for the second error correction step to be performed with more accuracy.

Once the three-dimensional coordinate estimate is found using basic photogrammetry, the system of point matches, cameras, and coordinates are optimized using a sparse bundle adjustment (SBA). The SBA minimizes the reprojection error across the whole system. The camera can be mathematically described as a three-dimensional to two-dimensional projection matrix:

$$\mathbf{P} = \mathbf{KR}\left[\mathbf{I}\right| - t] \tag{7}$$

Where K is the camera calibration matrix, R is the rotation matrix, and t is the three-dimensional position of the camera [10]. The projection of a three dimensional point to the image plane is then defined as

$$\mathbf{x} = \mathbf{PX} \tag{8}$$

SBA tries to minimize the Euclidean distance between the known feature coordinate and the feature coordinate based on the model:

$$\sum d\left(\mathbf{x}, \mathbf{PX}\right)^2 \tag{9}$$

SBA will minimize the error across the whole scene by adjusting all the parameters including the three-dimensional point coordinates and the camera parameters. This process is used on a bundle of image to image matches in order to minimize the error across the whole scene [12].

4. GEORECTIFICATION

After the error across the three-dimensional points is minimized using SBA, the point cloud is then georectified by exploiting the available positioning data that comes with imagery. The collinearity equations are used to perform the georectification. It is assumed that the position data from the imagery has been applied to the imagery so that each pixel has a UTM coordinate. The three-dimensional points are then projected through the collinearity equations onto the georefereced image. This is used as a mapping so that each three-dimensional point can be referenced to a universal transverse mercator (UTM) geographic coordinate. This process can be accomplished using Equation 8 and projecting the points through the camera model. Both processes provide the same results.

5. DENSE THREE-DIMENSIONAL RECONSTRUCTION

Each of the previous steps are part of a workflow that generates a sparse georeferenced three-dimensional point cloud and a fundamental matrix describing the matching images. Those steps are repeated over a bundle of images to generate a number of sparse point clouds. In these image bundles a "base image" will be defined as the image which all the other images in the bundle overlap. These regions of overlap can be easily determined using the camera parameters and the sparse bundles adjustment. A user may be interested in generating a higher resolution three-dimensional point cloud of a target within the sparse point cloud. This can be done within the framework of this workflow.

From Equation 1, it is known that a point-to-line correspondence can be found using the fundamental matrix. This matrix is known at the end of the sparse point cloud extraction process, so it is possible to generate epipolar lines between images. Figure 3 shows this relationship between three images.

Fig. 3. The point to epipolar line correspondence between three images

Using the fundamental matrix, the correspondence problem can then be reduced to a search along a single line. With this point-to-line correspondence, a region around the point in the first image is cross-correlated along the line in the other images until the best match is found. This can be done for every point in the image, however, this is very computationally expensive. In this workflow, the user selects a region of interest (ROI) over a target that they wants to generate a

denser point cloud for, and passes the ROI along with the corresponding fundamental matrix to this process. The ROI must be within a region of overlap between the images. The dense correspondence process is done for every point in the ROI. The dense three-dimensional point cloud extraction follows exactly the same approach as the sparse point cloud extraction, using basic photogrammetry. The dense point clouds can also be put through the sparse bundle adjustment and then georectified and inserted into the sparse point cloud.

Having a dense three-dimensional point cloud leads to the option of being able to generate a dense facetized three-dimensional model. Currently the point cloud is facetized using a basic Delauney triangulation, the facetized model can then be output in any format. This, however, is not the best way to attempt to generate a CAD-like high resolution model of a target. Work is currently in progress to implement "smart" plane and line fitting which is automated or semi-automated. These processes have typically been implemented on LIDAR data. [13, 14] The dense three-dimensional point cloud is comparable to LIDAR point clouds which leads to the presumption that these methods may work in generating accurate CAD-like models from dense three-dimensional point clouds.

6. RESULTS

This workflow was tested on a dataset flown by the RIT WASP sensor over the Van Lare wastewater treatment plant in Rochester, New York. The sparse point cloud extraction process was run on this bundle of five images. The center image of the bundle is the base image. Figure 4 shows the four generated point clouds after SBA had been applied. It also shows the whole point cloud projected down onto the base image for georectification.

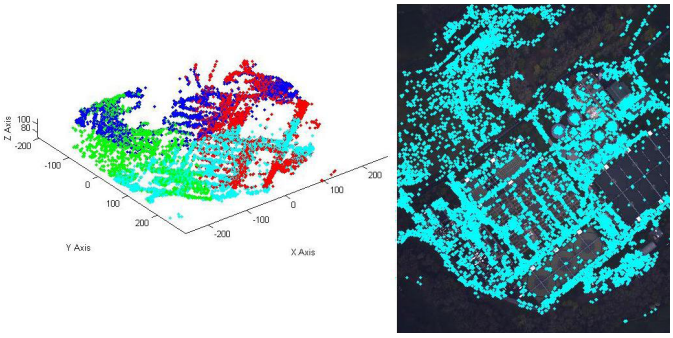

Fig. 4. The four generated point clouds after sparse bundle adjustment as well as the point clouds projected onto the base image for georectification.

Figure 3 shows a selection of three sub-images from this dataset that were used in the generation of a dense point cloud. Figure 5 shows the output of the dense point cloud process.

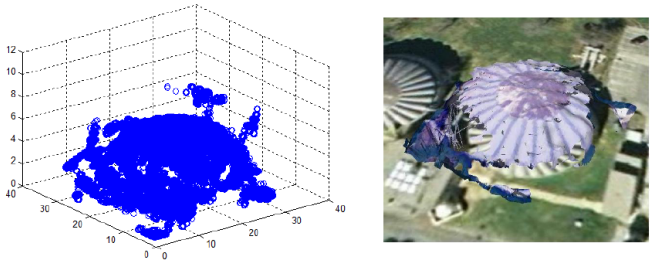

Fig. 5. The dense point cloud generated from figure 3

7. CONCLUSION

This work demonstrated that it is possible to use a well-defined computer vision methodology along with basic photogrammetric principles to develop a semi-automated workflow to extract sparse and dense point clouds from a set of aerial images. One additional area under study is implementing better methods of turning the dense three-dimensional point cloud into a high resolution CAD-like model. Algorithms like "smart-boxes" will tested on the dataset to see if the dense point cloud is dense enough to work with algorithms that are intended for use with LIDAR data [14]. Computer vision and photogrammetry are two fields of study that do not often come together to help each other. This work is putting forth the idea that these two areas could benefit a great deal from the individual strengths in the automation of analytical photogrammetry.

8. REFERENCES

[1] N. Snavely, S.M. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3D," in *ACM SIGGRAPH 2006 Papers*. ACM, 2006, pp. 835–846.

[2] I. Stamos, L. Liu, C. Chen, G. Wolberg, G. Yu, and S. Zokai, "Integrating automated range registration with multiview geometry for the photorealistic modeling of large-scale scenes," *International Journal of Computer Vision*, vol. 78, no. 2, pp. 237–260, 2008.

[3] C. Baillard, C. Schmid, A. Zisserman, and A. Fitzgibbon, "Automatic line matching and 3D reconstruction of buildings from multiple views," *International Archives of Photogrammetry and Remote Sensing*, vol. 32, no. 3; SECT 2W5, pp. 69–80, 1999.

[4] C. Baillard and H. Maītre, "3-D reconstruction of urban scenes from aerial stereo imagery: a focusing strategy," *Computer Vision and Image Understanding*, vol. 76, no. 3, pp. 244–258, 1999.

[5] "http://www.cis.rit.edu/ lias/wasp/," .

[6] "http://dirsig.org," .

[7] D.G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[8] ASPRS, *Airborne Synthetic Scene Generation (Aerosynth)*, Sam Amtpmop. Texas, November 2009.

[9] M Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[10] R. Hartley and A. Zisserman, *Multiple View geometry in Computer Vision*, Cambride University Press, 2nd edition, 2004.

[11] B. A. DeWitt and P. R. Wolf, *Elements of Photogrammetry (with Applications in GIS)*, McGraw-Hill Higher Education, 3rd edition, 2000.

[12] M. Lourakis and Argyros A., *The Design and Implementation of a Generic Sparse Bundle Adjustment Software Package Based on the Levenberg-Marquardt Algorithm.*, Institute of Computer Science, 4th edition, 2004.

[13] S Lach, S. Brown, and J. Kerekes, "Semi-automated dirsig scene modeling from 3d lidar and passive imaging sources," SPIE, 2007, vol. 6214.

[14] L. Nan, A. Sharf, H. Zhang, D. Cohen-Or, and B. Chen, "Smartboxes for interactive urban reconstruction," in *ACM SIGGRAPH 2010 papers*. ACM, 2010, pp. 1–10.