# AIRBORNE SYNTHETIC SCENE GENERATION (AEROSYNTH)

**Karl Walli**, Lt Col, USAF-AFIT/CI
**Dave Nilosek**, MS Student
**John Schott**, PhD
**Carl Salvaggio**, PhD
Center for Imaging Science
Rochester Institute of Technology
Rochester, NY 14623

## ABSTRACT

Automated synthetic scene generation is now becoming feasible with calibrated camera remote sensing. This paper implements computer vision techniques that have recently become popular to extract ”structure from motion” (SfM) of a calibrated camera with respect to a target. This process is similar to Microsoft’s popular ”PhotoSynth” technique (PhotoSynth09), but, blends photogrammetric with computer vision techniques and applies it to geographic scenes imaged from an airborne platform. Additionally, it will be augmented with new features to increase the fidelity of the 3D structure for realistic scene modeling. This includes the generation of both sparse and dense point clouds useful for synthetic macro/micro-scene reconstruction.

Although, the quest for computer vision has been an active area of research for decades, it has recently experienced a renaissance due to a few significant breakthroughs. This paper will review the developments in mathematical formalism, robust automated point extraction, and efficient sparse matrix algorithm implementation that have fomented the capability to retrieve 3D structure from multiple aerial images of the same target and apply it to geographical scene modeling.

Scenes are reconstructed on both a macro and a micro scale. The macro scene reconstruction implements the scale invariant feature transform to establish initial correspondence, then extracts a scene coordinate estimate using photogrammetric techniques. The estimates along with calibrated camera information are fed through a sparse bundle adjustment to extract refined scene coordinates. The micro scale reconstruction uses a denser correspondence done on specific targets using the epipolar geometry derived in the macro method.

The seeds of computer vision were actually planted by photogrammetrists over 40 years ago, through the development of “space resectioning” and “bundle adjustment” techniques. But it is only the parallel breakthroughs, in the previously mentioned areas that have finally allowed the dream of rudimentary computer vision to be fulfilled in an efficient and robust fashion. Both areas will benefit from the application of these advancements to geographical synthetic scene modeling. This paper will explore the process the authors refer to as Airborne Synthetic Scene Generation (AeroSynth).

Key words: Structure from motion, bundle adjustment, multi-view imaging, scene synthesis, computer vision.

## AEROSYNTH INTRODUCTION

Recovering 3D structure from 2D images requires only that the scene is imaged from two different viewing geometries and that the same features can be accurately identified. Figure 1, depicts a site of interest imaged from multiple views using an airborne sensor; here the point of interest is the top of a smokestack that will be imaged with the effects of parallax displacing it with respect to other features within the scene. This parallax displacement effect has been used for decades within the photogrammetry community to recover the 3D structure within a scene. Unfortunately, robust automated techniques to match similar features within a scene have been fairly elusive until very recent breakthroughs in the area of computer vision.

## RECOVERING SPARSE STRUCTURE FROM IMAGES

The key to automatically recovering 3D structure from an imaged scene is to identify reliable invariant features, match these features from images with diverse angular views of that scene and then generate accurate mathematical relationships to relate the images. This information can then be utilized in concert with the camera external and internal orientation parameters to derive scene structure that is defined within the World Coordinate System (WCS) of choice.

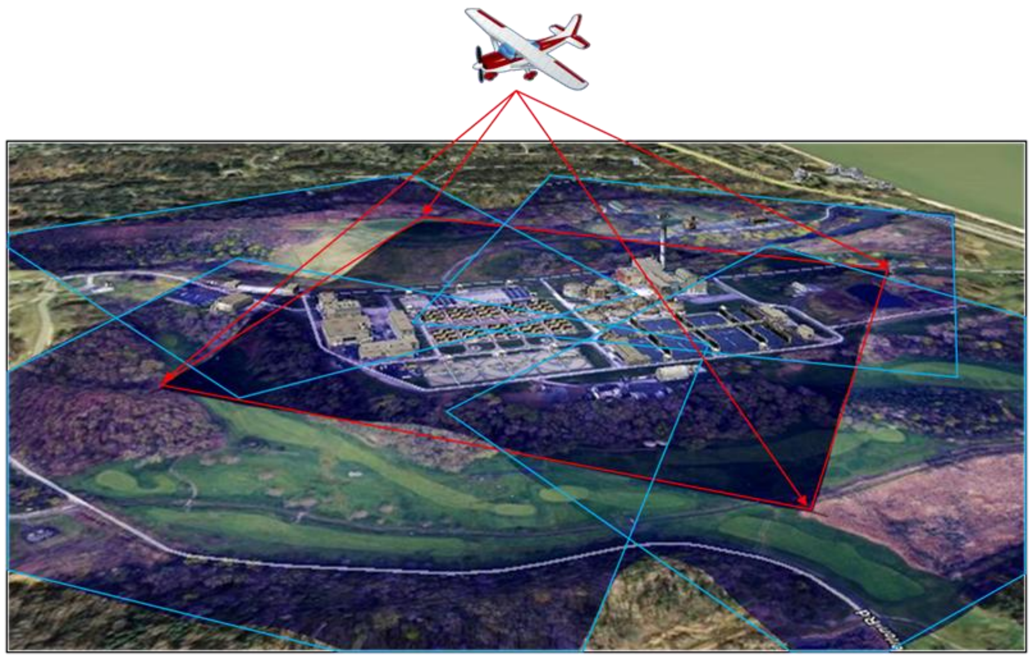


**Figure 1 – Example showing the angular diversity required to recover 3D Terrain from Airborne Imagery.**

## Airborne Dataset

For this study, the working imagery was obtained from the Rochester Institute of Technology, Center for Imaging Science's (RIT/CIS), Wildfire Airborne Sensing Program (WASP) multimodal sensor suite. This sensor provides 4kx4k Visible Near Infrared (VNIR) and 640x512 Short Wave Infrared (SWIR), Mid-Wave Infrared (MWIR), and Long Wave Infrared (LWIR) images. Google Earth (GE) was utilized as the GIS visualization tool, with a detailed model of the Frank E. VanLare Water Treatment Plant (Pictometry, 2008) embedded within the standard satellite imagery and 30[m] terrain elevation maps (Figure 1 & Figure 4). Figure 1 shows the region of overlap (outlined in red) of 5 WASP images where the site of interest is contained in the central (base) image.

## Invariant Feature Detection and Matching

The Scale Invariant Feature Transform (SIFT) operator, proposed by David Lowe in 1999 (Lowe, 2004), has become a "gold standard" in 2D image registration due to its ability to robustly identify large quantities of semi-invariant feature within images. The SIFT technique can consistently isolate thousands of potential invariant features within an arbitrary image as seen in Figure 2. This is extremely useful when attempting to create sparse structure from matched point correspondences, since any matching features can then be processed to obtain the 3D structure of the imaged scene. In addition, more recent independent testing has confirmed that the SIFT feature detector, and its variants, perform better under varying image conditions than other current feature extraction techniques (Moreels & Perona, 2006) & (Mikolajczyk & Schmid, 2005).

The SIFT algorithm utilizes a Difference of Gaussian edge detector of varying widths to isolate features and define a gradient mapping around them. These gradient maps are then compared for similarity in another image and matches result from the most likely invariant feature pairs. Once potential matches are found, outliers usually need to be culled based on the requisite epipolar relationships that must exist between two images of the same scene. This has always been challenging in the past due to the effects of parallax, but, this can now be robustly addressed using techniques highlighted in the next section.

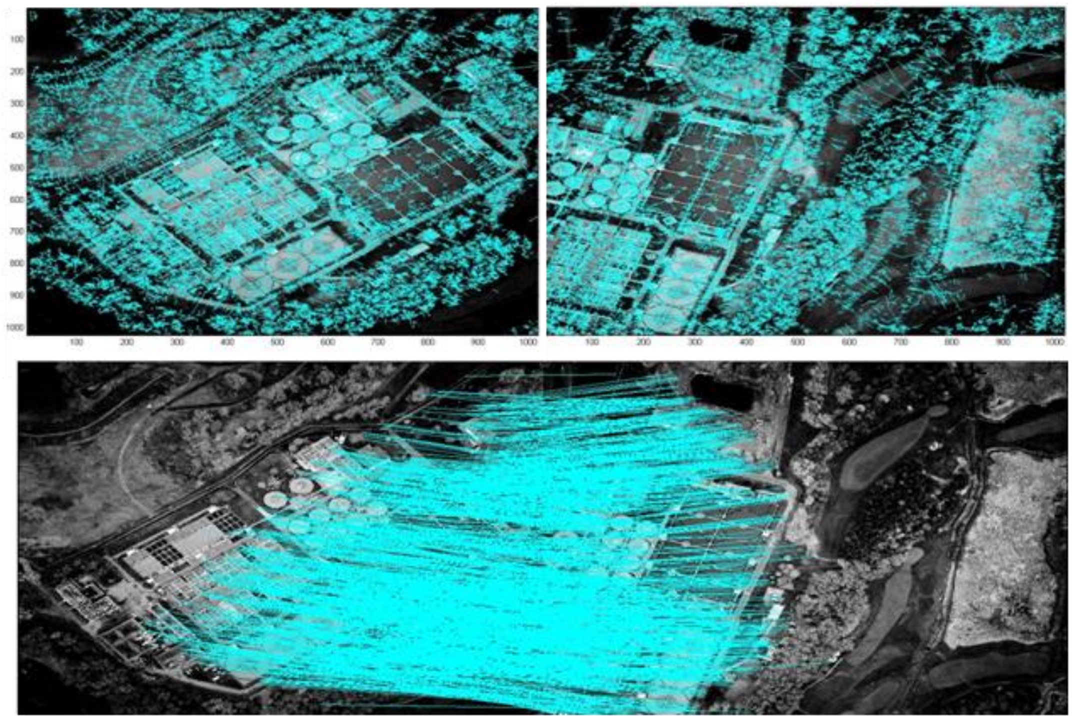


**Figure 2 - Thousands of invariant keypoints generated and matched using the SIFT algorithm.**

## Outlier Removal

In order to successfully remove erroneous matches derived using the SIFT algorithm, the potential match set will be processed using the RANdom Sample Consensus (RANSAC) technique in conjunction with the Fundamental Matrix relationship between images of the same scene (Figure 3). RANSAC has proven to be a robust technique for outlier removal, even in the presence of large numbers of incorrect matches (Hartley & Zisserman, 2004). Also, because it is not necessary to test all the sets of points for a solution, it can be efficiently utilized with techniques like SIFT that provide large numbers of automated matches.

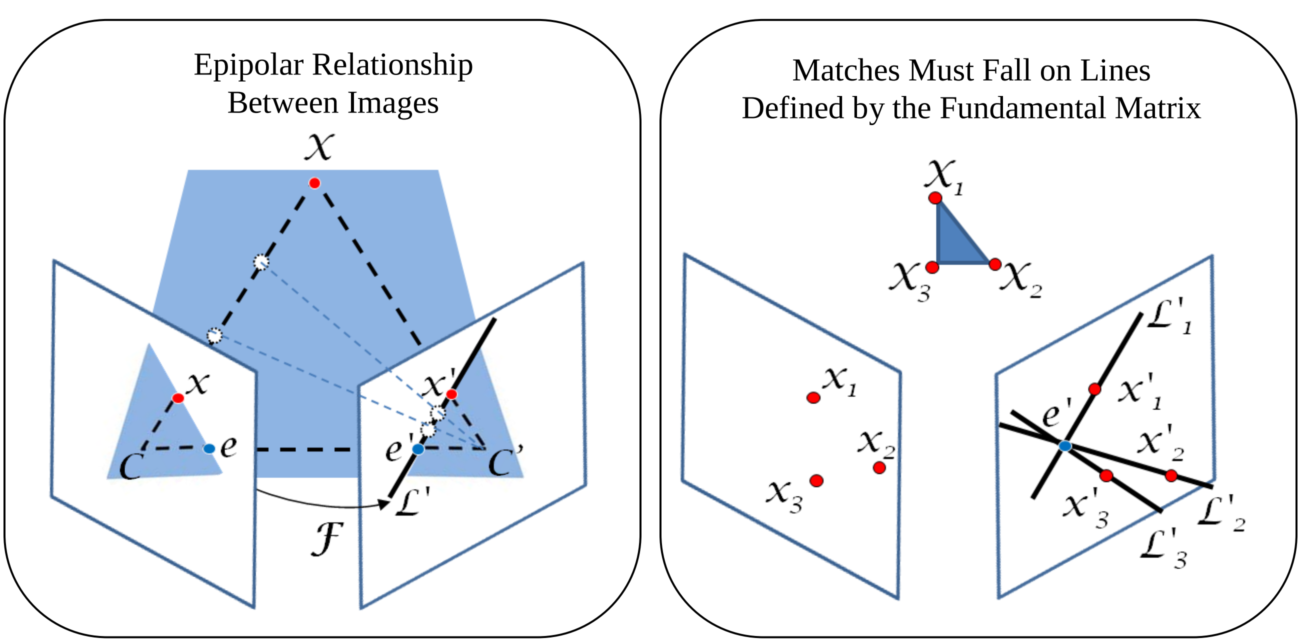



**Figure 3 – Depiction of the Fundamental Matrix constraint between two images which is used for outlier removal.**

In the diagram above (Figure 3), the Fundamental Matrix ***F*** dictates that for a given 3D scene point ***X***, a ray must pass from the camera center ***C*** (a focal length behind the image plane) through the image location ***x*** and this ray will be imaged by the camera ***C'*** as an epipolar line ***l'***, passing from the image of the same model point ***x'*** to that cameras epipole ***e'***. The epipole is the image of the other camera center (which may be off the image entirely).

Anyone that has worked for any length of time with automatic image registration can attest to the challenging issues parallax can cause when relating features. The limitation of utilizing a 2D Projective Homography to relate imagery with large degrees elevation difference between acquisition stations, can be somewhat addressed through the use of the Fundamental Matrix relationship. This relationship constrains the matches to an epipolar line even under extreme parallax situations and can be simply formalized in a mathematical manner as shown below (Hartley & Zisserman, 2004).

*Fundamental Matrix*

$$Fx = l' \quad \textbf{(1)}$$

and so, $\boldsymbol{x'^T F}$ must be in the left null-space of $\boldsymbol{x}$ and $\boldsymbol{Fx}$ must be in the right null-space of $\boldsymbol{x'^T}$.

*Fundamental Null Space*

$$x'^T F x = 0 \quad \textbf{(2)}$$

So, for a given point ***x***, the preliminary match point must lie along the epipolar line ***l'*** in order for it to be a valid match. So, the proposed feature matches that do not fit this epipolar constraint are probably bad matches.

Once the initial matched point set has been obtained using the automated SIFT technique, it is usually necessary to test for these bad matches or "outliers". The RANSAC algorithm (Fischler & Bolles, 1981) can be utilized to iteratively take a random sample of the matches to create a Fundamental Matrix relationship between the images. Once this is done, the veracity of that relationship can be tested by comparing the number of resulting inliers against a statistically relevant number of additional tests. The Fundamental Matrix that produces the most match point inliers is then accepted as the best mathematical model and any outliers to this model are then removed.

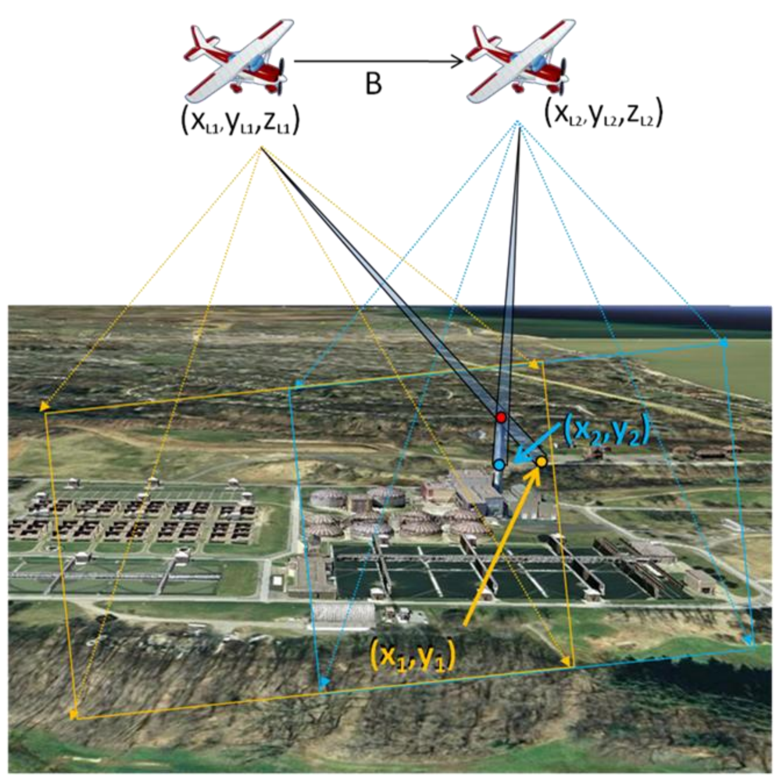



**Figure 4 – Graphic showing two collection stations of an airborne sensor utilized to recover 3D Structure.**

## Initial Estimate of Sparse Structure

The initial estimation technique that is utilized to derive the 3D scene structure utilizes a simple approach that is augmented for more general situations by compensating for the aircraft motion and implementing coordinate system conversions. This basic process can be visualized in Figure 4 and the following equations (DeWitt & Wolf, 2000) can be utilized to derive 3D structure once these corrections have been accomplished. The flying height of the initial sensor location is represented by $\boldsymbol{T_{z1}}$, the baseline distance between sensor locations is $\boldsymbol{B}$, the pixel distance between matches is $\boldsymbol{p_i}$, and each image location is described as $[\boldsymbol{x_{1i}}, \boldsymbol{y_{1i}}]$ and $[\boldsymbol{x_{2i}}, \boldsymbol{y_{2i}}]$.

| | | |
|---|---|---|
| *Focal Plane Distance* | $p_i = x_{1i} - x_{2i}$ | **(3)** |
| *X-location Relative* | $X_i = \dfrac{B * x_{1i}}{p_i}$ | **(4)** |
| *Y-location Relative* | $Y_i = \dfrac{B * y_{1i}}{p_i}$ | **(5)** |
| *Z-location World Coord* | $Z_i = T_{z1} - \dfrac{B * f}{p_i}$ | **(6)** |

Figure 5 depicts the corrections that are required for any deviation of the flight line from the coordinate axis of the images and the pitch, yaw, and role of the aircraft. Initial coordinate conversions are required to align the image planes with the flight path and compensate for heading and yaw. Unless the acquisition platform is capable of acquiring perfectly nadir imaging on a routine basis, it is necessary to rectify the image or image correspondences to enable proper linear 3D structure estimation. The approach the author has taken to accomplish this is to back-project the image correspondences onto a virtual focal plane that is located at the focal length (f), but, is situated perpendicular to the earth's surface as depicted in Figure 5B.

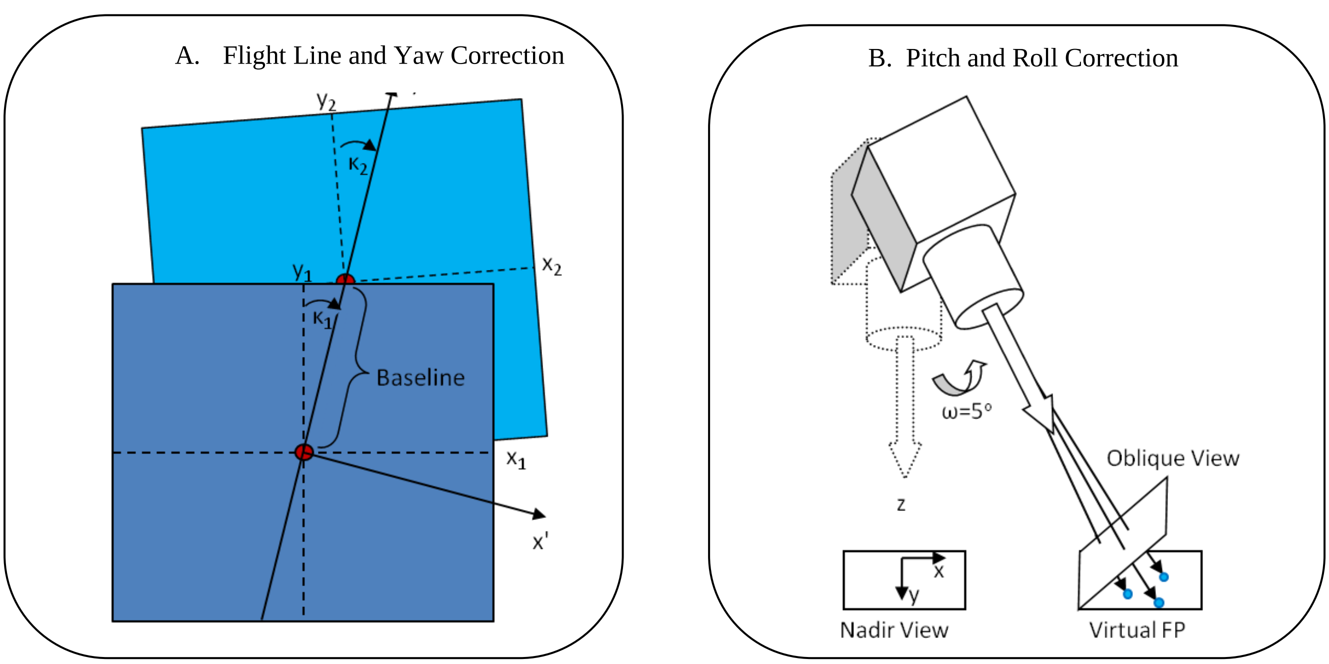



**Figure 5 – Corrections are required to compensate for flight line orientation and aircraft pitch, yaw, and roll.**

It is important to note that the height estimate ($\boldsymbol{Z_i}$) is dependent on the ratio of the Baseline ($\boldsymbol{B}$) to the pixel distance ($\boldsymbol{p_i}$) of the matches projected onto the virtual focal plane. This ratio can be corrected to one that is aligned with the flight line by performing a coordinate system conversion to the base image plane and then compensating for the relative Baseline distance (Equation (7)). Finally, the corrected image plane distance can be calculated using Equation (8). Here, the offset from the flight line is represented by $\boldsymbol{K}$ and $\boldsymbol{T_{xi}}$ and $\boldsymbol{T_{yi}}$ are respectively the Longitude and Latitude of the camera centers.

*Baseline Distance Correction*

$$B = \sqrt{|cosK|(T_{x2} - T_{x1})^2 + |sinK|\left(T_{y2} - T_{y1}\right)^2} \quad \textbf{(7)}$$

*Image Distance Correction*

$$p_i = \sqrt{(x_{2i} - x_{1i})^2 + (y_{2i} - y_{1i})^2} \quad \textbf{(8)}$$

The initial results can be viewed with their respective camera stations in Figure 6, where nearly 20,000 individual point correspondences were automatically recovered from 5 matching images (4 image pairs) to produce a Sparse Point Cloud (SPC) representation of the scene. Note that here the results are still in a relative (meter based) coordinate system centered on the base camera location.

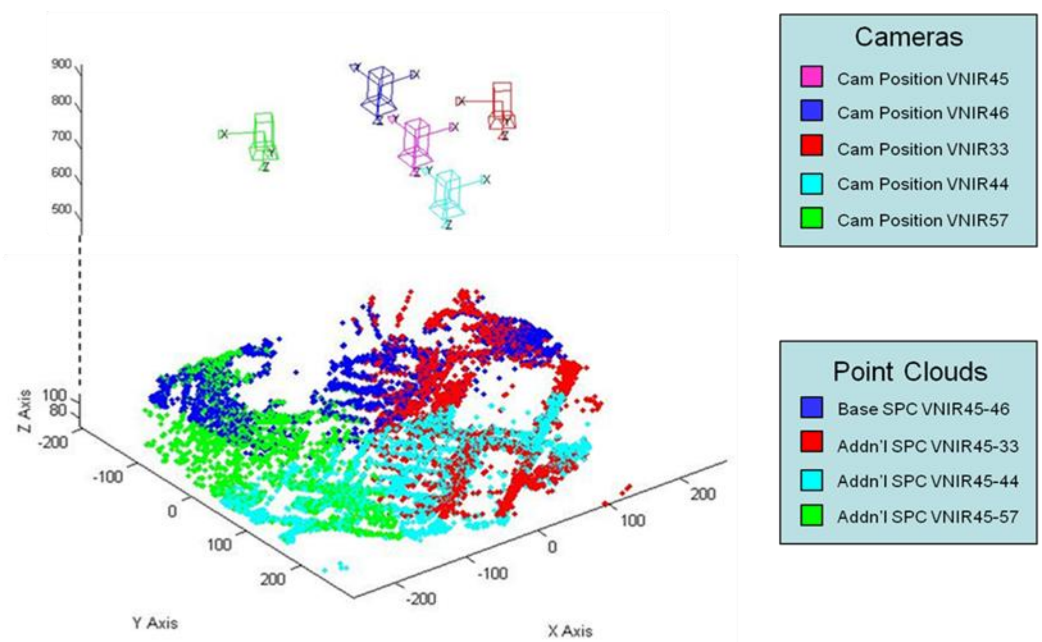



**Figure 6 – The initial estimates of the four individual SPC's can be seen compared to the camera locations.**

### Non-Linear Optimization of Sparse Structure

Many of the problems presented in this research cannot be solved by linear methods alone. In these cases, it is necessary to apply non-linear estimation techniques to provide accurate solutions. Such real world problems as the resectioning of images to models and the Bundle Adjustment (BA) of multiple images, to reconstruct 3D structure, both require nonlinear minimization solutions. In fact, for BA, these solutions often depend on calculating the interaction of several thousand variables simultaneously. Due to its stability and speed of convergence, the Levenberg–Marquardt Algorithm (LMA) is currently one of the most popular approaches which is routinely used to solve these challenging problems.

When utilizing LMA, the computational challenge is to minimize a given cost function. For applications such as resectioning and BA, this cost function is defined as the sum of the squared error between image points (actual data) and projected 3D model points (predicted values) dictated by the current set of parameter. The minimization function takes advantage of the relationship between the estimated 3D structure ($\overline{\boldsymbol{X}}_i$) and its 2D projection onto the image plane ($\overline{\boldsymbol{x}}_i$) as mathematically formalized below (Hartley & Zisserman, 2004).

*Projection Matrix Simplified*

$$\bar{\boldsymbol{x}}_i = \boldsymbol{P}\bar{\boldsymbol{X}}_i \tag{9}$$

*Projection Matrix Expanded*

$$\begin{bmatrix} x_1 & x_2 & \cdots x_i \\ y_1 & y_2 & \cdots y_i \\ 1 & 1 & \cdots 1 \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \\ P_{31} & P_{32} & P_{33} & P_{34} \end{bmatrix} \begin{bmatrix} X_1 & X_2 & \cdots X_i \\ Y_1 & Y_2 & \cdots Y_i \\ Z_1 & Z_1 & \cdots Z_i \\ 1 & 1 & \cdots 1 \end{bmatrix} \tag{10}$$

The Projection Matrix (***P***) can then be utilized directly for minimization since it incorporates the cameras internal calibration parameters (***K***) , and external orientation (***R)*** and position (***t***). This minimization equation then takes the following form (Equations (11) and (12)).

*Projection Minimization Function*

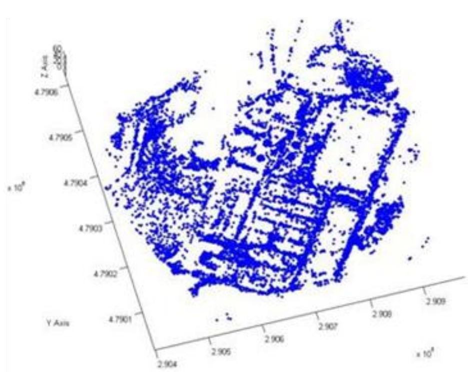
$$\sum_i d(\bar{x}_i, P\bar{X}_i)^2 \tag{11}$$

*Expanded Minimization Function*

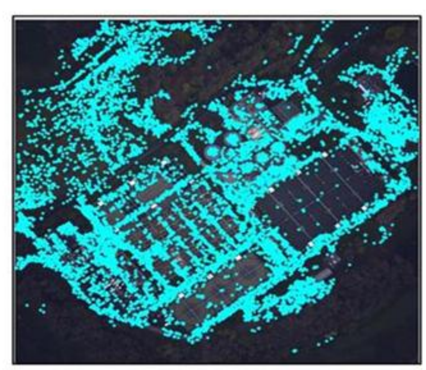
$$\sum_{i=1}^{n} \left\| \bar{x}_i - \hat{X}_i(K, R, t, \bar{X}_i) \right\|^2 \tag{12}$$

The Sparse Bundle Adjustment (SBA) algorithm of Lourakis and Argyros (Lourakis & Argyros, 2004) is optimized for speed and efficiency. It can easily optimize against several camera variables and the structure of tens of thousands of 3D points simultaneously to produce a sparse image bundle that is mutually self-consistent. However, as with any engineering code, it requires specific formatting for the input variables and special care when preparing the camera IOPs and EOPs. The next section addresses this topic in order to ensure that accurate global coordinates can be obtained after utilizing this SBA minimization.

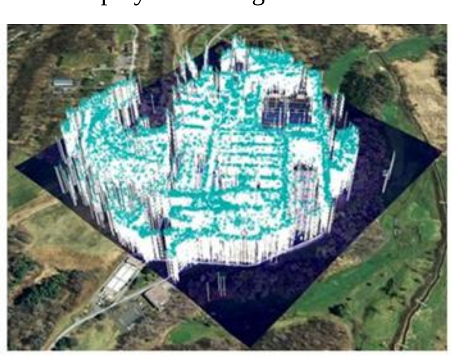
A. Final SPC in global UTM.



B. Results Projected back onto Base Image.



C. SPC displayed in Google Earth.



D. SPC converted into faceted mesh model.

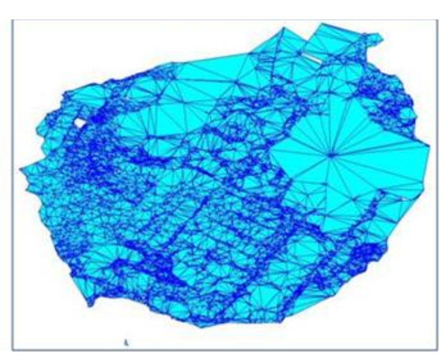


**Figure 7 – Example results of the Sparse Bundle Adjustment processing on the Sparse Point Cloud.**

### Relating the Results to World Coordinate System

Since the results of the SBA process minimize against a relative coordinate system anchored on the base camera position, it can be difficult to determine the absolute locations of the 3D points even though there is good self consistency between the camera locations and the SPC. In order to recover the absolute location of the3D points, the collinearity equations (Equations (13) and (14)) were utilized to re-project the 3D points back into the base image locations of the initial feature matches as seen in Figure 7B.

*Collinearity Eq X-component World Coord.*

$$X - X_L = (Z - Z_L)\left[\frac{m_{11}(x - x_0) + m_{21}(y - y_0) + m_{31}(-f)}{m_{13}(x - x_0) + m_{23}(y - y_0) + m_{33}(-f)}\right] \qquad \textbf{(13)}$$

*Collinearity Eq Y-component World Coord.*

$$Y - Y_L = (Z - Z_L)\left[\frac{m_{12}(x - x_0) + m_{22}(y - y_0) + m_{32}(-f)}{m_{13}(x - x_0) + m_{23}(y - y_0) + m_{33}(-f)}\right] \qquad \textbf{(14)}$$

In this case, only the minimized depth parameter ($\boldsymbol{Z_i}$) retained its absolute coordinate value and so could be utilized with the camera locations to determine the world coordinate Latitude ($\boldsymbol{Y_i}$) and Longitude ($\boldsymbol{X_i}$) values.

## RECOVERING DENSE STRUCTURE FROM IMAGES

The key to recovering a Dense Point Cloud (DPC) from matching images lies in the ability to relate the images on a pixel-to-pixel level (Nilosek & Walli, 2009). This is the transition point between the macro and micro scene reconstruction, and the micro process needs certain information derived in the macro process. At this point in the process each image is related to a base image of the scene through a fundamental matrix derived using the RANSAC process. The macro process has also derived the regions of overlap for each image with respect to the base image. Each fundamental matrix and region of overlap are passed off to the micro process. Ideally this process would relate every pixel in every overlapping image to the base image however due to computing power restrictions, examples in this paper focus on specific targets inside the regions of overlap.

### Dense Correspondence - Relating Images at the Pixel Level

The utility of the Fundamental Matrix for outlier match removal has been shown, now this matrix will be used to help derive a dense set of matches between overlapping regions. Using this matrix and equation (1) for every point in the base image an epipolar line that contains the corresponding point can be found in each other overlapping image. Figure 8 shows how epipolar lines are found in different overlapping regions from a single point in one image for three different images.

**Figure 8 – Left: Target with single point chosen. Middle/Right: Corresponding epipolar lines.**

This property of the Fundamental Matrix reduces the correspondence search to a 1 dimensional search along epipolar lines. The images are rectified so that the epipolar lines are along the horizontal then a normalized cross correlation is computed on a small area selected around the single point in the base image. The maximum response from the normalized cross correlation is chosen as the match. This is done for every pixel over the entire area which results in a very dense correspondence between the multiple views. The estimate of the dense structure follows the same pipeline as estimating the sparse structure. First basic photogrammetry is used to extract an initial estimate of the structure. Then the camera parameters, initial estimate of the structure and correspondences are used in minimizing the reprojection error between all the images using the SBA method. The collinearity equations can also be used to place the dense structure in the world coordinate system. The dense structure is also mapped with an image of the target. Figure 9 shows the initial estimate of the structure then the final product after all the processes.

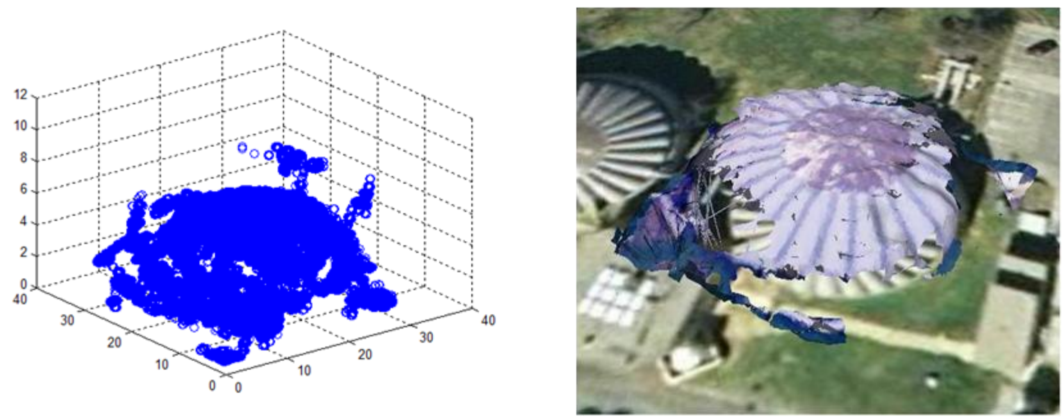



**Figure 9 – Left: Initial estimate of the structure of the dense point cloud from three images. Right: Result after SBA, world coordinate mapping and image texturing.**

Once the dense structure of a specific target has been acquired it is added to the sparse structure. Figure 10 shows the dense structure incorporated into the sparse structure overlaid on a map. Also on this map are hand generated CAD models of the same structures. Based on the CAD model the dense structure is not too far off from the correct structure. One very clear issue stands out when working with only nadir imagery, and that is that it is very difficult to reconstruct the sides of objects. Oblique imagery can be used to view the sides of objects however, the fairly severe projective transforms that relate oblique images together provide its own set of correspondence problems.

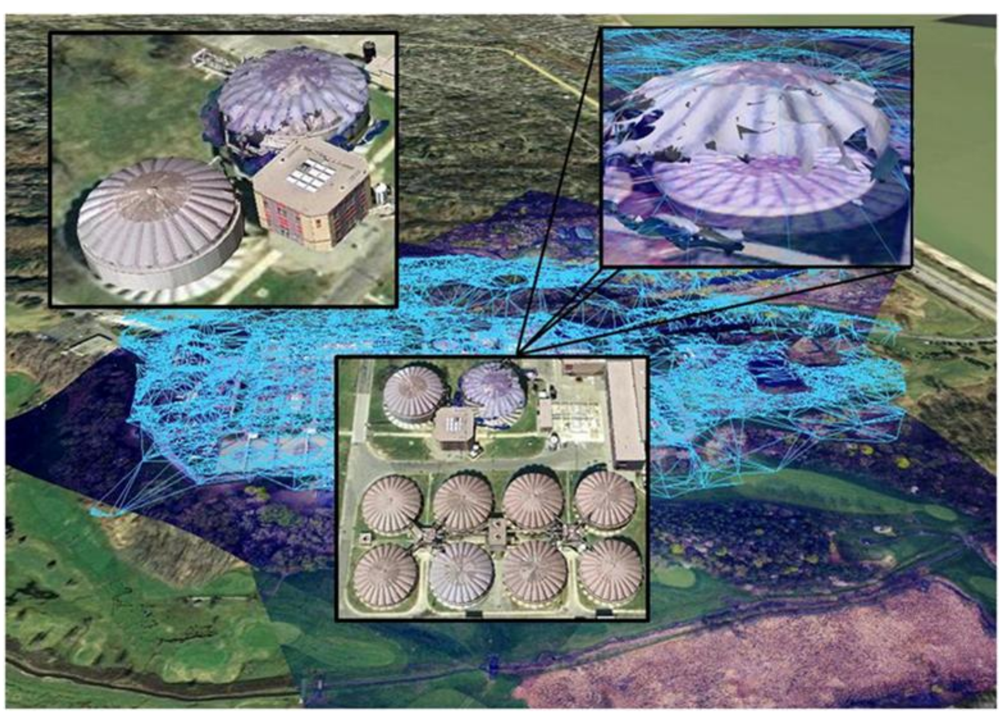


**Figure 10 – Resulting 3D structure recovered from three overlapping images using Dense Point Correspondences (Model provided by Pictometry International Corporation and embedded within Google Earth).**

### Matching Oblique Images using ASIFT – Maximizing Angular Diversity

Recently an algorithm has been developed that attempts to describe features as projectively invariant. This algorithm is called Affine Scale Invariant Feature Transform (Morel & Yu, 2009). This algorithm builds off of the original SIFT by taking the initial images and simulating rotations along both the x and y axis. It essentially performs many SIFT operations over these simulated images in order to find the best matching rotation between the images in order to remove it. Once the initial matching is found using ASIFT, the same RANSAC process using the Fundamental Matrix as the fitting model can be used to weed out the outliers found with ASIFT. Figure 11 shows an example of matching points using ASIFT and then RANSAC.

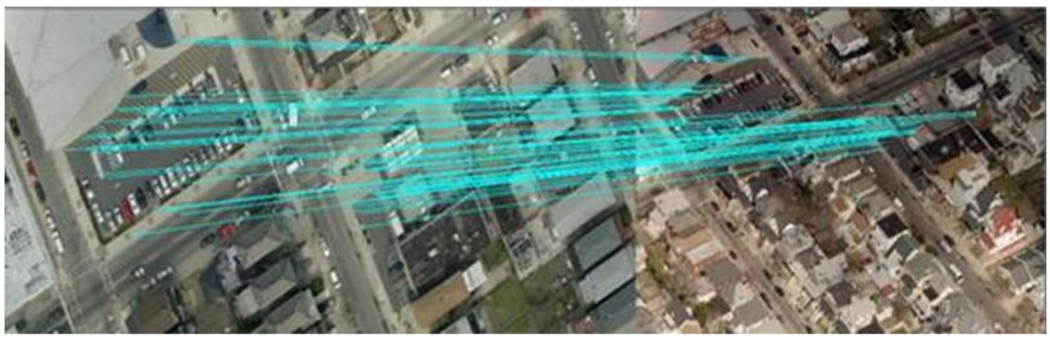


**Figure 11 – Matching between a nadir and oblique images using ASIFT and then RANSAC with the Fundamental Matrix as the fitting model (Images courtesy Pictometry International Corp).**

The next step is to utilize the SPC, resulting fundamental matrices and regions of overlap to extract a DPC of a target area within the scene. Since a projective transformation can greatly impair the normalized cross-correlation method of point matching, a separate approach may be required for dealing with images that capture significant angular diversity of a target.

### Growing a Depth Map from the Sparse Correspondence

Since an accurate sparse representation of the structure of the scene has already been derived, this structure can be utilized as a good starting point to 'grow' a dense matching between images. (Goesele, Snavely, Curless, Hoppe, & Seitz, 2007). A dense matching is generated around each sparse match using an optimization method that minimizes the normalized pixel intensity difference between each overlapping image with respect to the base image. Here each projected SPC location is utilized as an initial seed and the matched image locations are slowly grown from the pixels surrounding these points. In this way a dense correspondence mapping can be obtained between images by constraining the epipolar line search space.

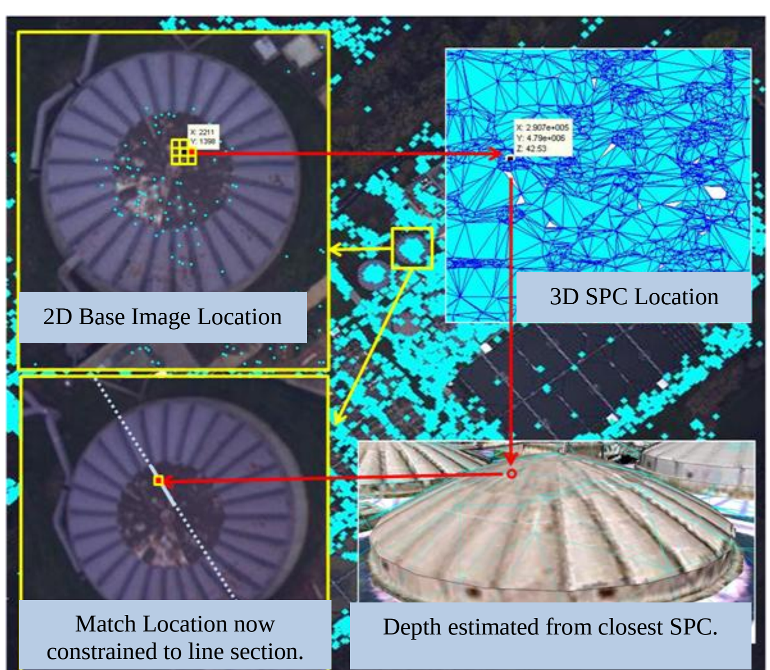



**Figure 12 – Growing 3D depth maps based on the initial SPC results and epipolar relationships.**

## AEROSYNTH SUMMARY

Due to the fast growth in the computer vision arena, regarding SfM techniques, it is fruitful for the photogrammetry community to keep abreast and apply these techniques to the area of remote sensing. The authors' AeroSynth technique for recovering 3D structure from images is a blend of the both photogrammetric and computer vision approaches. It utilizes the automatic feature isolation/matching, epipolar relationships and SBA of the computer vision community and combines it with the linear 3D point estimation and collinearity relationships of photogrammetry. As a result, the image bundle, SPC, and DPC that is produced can be related to the WCS and directly injected into GIS applications for automatic analysis and comparison to existing archival data.

## REFERENCES


DeWitt, B. A., & Wolf, P. R. (2000). *Elements of Photogrammetry (with Applications in GIS)* (3rd ed.). McGraw-Hill Higher Education.

Fischler, M., & Bolles, R. (1981). Random Sample Consensus: A Paradigm for Model Fitting with applications to Image Analysis and Automated Cartography. *Communications of the ACM, Volume24, Issue 6* (pp. 381-395). New York: ACM.

Goesele, M., Snavely, N., Curless, B., Hoppe, H., & Seitz, S. (2007). Multi-View Stereo for Community Photo Collections. *2007 IEEE 11th International Conference on Computer Vision*, *1-6*, pp. 825-832. Rio de Janeiro, Brazil.

Hartley, R. I., & Zisserman, A. (2004). *Multiple View Geometry in Computer Vision* (Second ed.). Cambridge University Press, ISBN: 0521540518.

Lourakis, M., & Argyros, A. (2004). *The Design and Implementation of a Generic Sparse Bundle Adjustment Software Package Based on the Levenberg-Marquardt Algorithm.* Institute of Computer Science - FORTH.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *INTERNATIONAL JOURNAL OF COMPUTER VISION* , *60*, 91-110.

Mikolajczyk, K., & Schmid, C. (2005). A Performance Evaluation of Local Descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.* , 1615-1630.

Moreels, P., & Perona, P. (2006). Evaluation of Features Detectors and Descriptors based on 3D Objects. *International Journal of Computer Vision* , 263-284.

Morel, J. M., & Yu, G. (2009). ASIFT: A New Framework for Fully Affine Invariant Image Comparison. *SIAM Journal on Imaging Sciences* , *2* (2).

Nilosek, D., & Walli, K. (2009). AeroSynth: Aerial Scene Synthesis from Images. *SIGGRAPH*, (Poster Session). New Orleans, LA.

Pictometry, C. I. (2008). *Pictometry Homepage*. Retrieved July 15, 2009, from Pictometry Website: http://www.pictometry.com/home/home.shtml