A Comparison of Stereo and Multiview 3-D Reconstruction Using Cross-Sensor Satellite Imagery

Ozge C. Ozcanli, Yi Dong, Joseph L. Mundy
Vision Systems Inc.
Providence, RI
ozge@visionsystemsinc.com

Helen Webb, Riad Hammoud, Victor Tom
BAE Systems
Burlington, MA

Abstract

High-resolution and accurate Digital Elevation Model (DEM) generation from satellite imagery is a challenging problem. In this work, a stereo 3-D reconstruction framework is outlined that is applicable to nonstereoscopic satellite image pairs that may be captured by different satellites. The orthographic height maps given by stereo reconstruction are compared to height maps given by a multiview approach based on Probabilistic Volumetric Representation (PVR). Height map qualities are measured in comparison to manually prepared ground-truth height maps in three sites from different parts of the world with urban, semi-urban and rural features. The results along with strengths and weaknesses of the two techniques are summarized.

1. Introduction

The number of high-resolution[1], commercial satellite images available for Earth monitoring is rapidly expanding. There are numerous applications exploiting this imagery from mapping for Geographic Information System (GIS) purposes to scientific analysis of ice sheet dynamics. In this paper, derivation of high resolution Digital Elevation Models (DEM) using satellite imagery is investigated. DEMs are important by-products of Earth imaging and high-quality, dynamic DEM generation utilizing the daily stream of worldwide satellite images would be very beneficial to higher level exploitation of the imagery.

Traditionally, stereoscopic images are used for DEM generation and such images are collected specially and on demand by the image vendors. The archive of nonstereoscopic images is much larger however and expanding more rapidly. Methods that can utilize two or more nonstereoscopic images with some overlap area possibly captured by different satellites to reconstruct a DEM present an exciting new opportunity to exploit the existing archives. In this paper, two such techniques that are applicable to nonstereoscopic images for DEM generation are compared in terms of reconstruction accuracy. The first method is a stereo reconstruction approach similar to NASA's Ames Stereo Pipeline [1,2,3]. The second method is a lesser known multiview probabilistic volumetric reconstruction (PVR) algorithm applied to satellite imagery [4,5]. Section 2 outlines the stereo reconstruction technique used in this paper and also explains the differences to Ames Stereo Pipeline. The major difference is that, the technique used in this paper does not require sparse feature correspondences to be computed in the two images and manually verified as in [1] thanks to the application of an automated geolocation correction algorithm [6] prior to reconstruction. Automatic geolocation correction is briefly described in Section 1.1 and it georegisters all the overlapping images of a site to a common geodetic coordinate frame by correcting the errors in their metadata. Such large scale and automated georegistration makes PVR, the multiview reconstruction technique of [5] applicable as well since PVR performs a dense triangulation of 3-D surfaces by shooting rays from all the pixels in the images. Thus, for this algorithm to work, the rays need to be precisely aligned at the correct volumetric elements, *voxel*s, necessitating the relative georegistration step. Section 3 gives a brief overview of PVR technique.

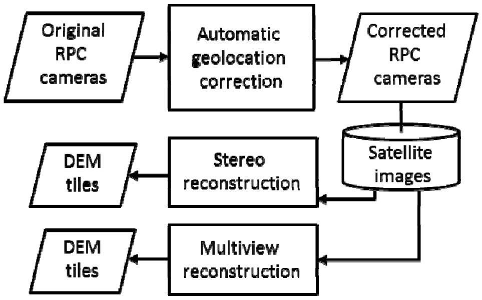


Figure 1: The process flow diagram of the DEM generation pipeline used in this paper. Prior to reconstruction, Rational Polynomial Coefficient (RPC) camera model of the satellite images are bias-corrected as explained in Section 1.1.

The resulting DEM tiles are compared to manually prepared height maps on three different sites. The evaluation methodology is explained in Section 1.3 and the results are summarized in Section 4. In summary, the

[1] Ground Sampling Distances (GSD) of 1 meter or better.

major contributions of this paper are: i) a fully automated multiview stereo reconstruction pipeline that can compute and fuse results from hundreds of cross-sensor nonstereoscopic images and ii) a quantitative analysis of the DEM accuracy in comparison to the true multiview approach of PVR algorithm using the same set of input images.

1.1. Automatic Geolocation Correction

Both of the 3-D reconstruction methods used in this paper are applicable to nonstereoscopic images, possibly cross-sensor images, only after the errors in the geolocation of the images reported in the metadata are corrected. For this purpose, a recent framework [6] is used which takes advantage of the fact that many overlapping images can be retrieved from the archive for a given region of interest. For the experiment sites of this paper, more than 200 images collected by GeoEye-1, Worldview-1, WorldView-2 and Quickbird satellites are acquired. GeoEye-1 has the best geolocation accuracy of 3 meter 90% Circular Error (CE) on the ground [7] and Quickbird is the worst with 23 meters CE90 [8]. A seed group is selected from the satellite with the best geolocation accuracy; GeoEye-1 in our experiments. This seed group is bundle adjusted using automatically computed tie points and further refined using edge features. In this context, bundle adjustment means adjusting corresponding image to ground rays so that they intersect at a single 3-D point. After seed group's bundle adjustment, all the remaining images are grouped such that groups of three or more images are formed where two images are from the corrected seed group and the rest are uncorrected. These groups of three or more images are also bundle-adjusted and refined using edge features; but this time in a way that does not change the seed pair's corrected metadata. This establishes the coordinate frame onto which the other images in the group are aligned. In this way, all the images in the collection are georegistered to a common coordinate frame where triangulation can be done for 3-D reconstruction purposes.

Each satellite image is equipped with a Rational Polynomial Coefficient (RPC) camera model that is specially developed for satellite imagery [9] and delivered in the metadata of each image. This model combines intrinsic and extrinsic calibration of the platform in one set of polynomial equations and enables the projection of 3-D points given as (latitude, longitude, elevation) to 2-D image pixels, (u, v). The function can be written as $\mathrm{u} = \mathrm{F}_{RPC}(\mathrm{lat}, \mathrm{lon}, \mathrm{elev})\mathrm{u}_\mathrm{s} + \mathrm{u}_0$ and $\mathrm{v} = \mathrm{F}_{RPC}(\mathrm{lat}, \mathrm{lon}, \mathrm{elev})\mathrm{v}_\mathrm{s} + \mathrm{v}_0$ for some scaling parameters $\mathrm{u}_\mathrm{s}, \mathrm{v}_\mathrm{s}$, some offset parameters $\mathrm{u}_0, \mathrm{v}_0$ and for a high order polynomial function F_{RPC} with 80 coefficients. Since the satellite camera is far from the imaged surfaces (typically ~500 km), the rays for the individual pixels are almost parallel to each other. Thus geopositioning errors can be corrected during bundle adjustment by small translations in the image domain where one translation is used to correct an entire image of size 50 km at max. This type of correction specifically computes a correction offset, $(\Delta\mathrm{u}_0, \Delta\mathrm{v}_0)$, termed as bias correction offset in [10]. Depending on the geopositioning error on the ground and the resolution of the image, the correction offsets in the image domain are on the order of 5-50 pixels. For the images used in the experiments of this paper, the worst case correction offsets ranged from 5 pixels, for GeoEye-1 imagery with ~0.5 meter GSD, up to 30 pixels in radius for Quickbird imagery with ~1 meter GSD.

1.2. Pre-processing of satellite imagery

After geolocation correction, all the satellite images (PAN bands) are calibrated from raw digital pixel values to radiance and then to top-of-atmosphere reflectance using calibration metadata [11,12]. Dark pixel subtraction is applied to remove haze and in the case of stereo approach in Section 2, histogram equalization is run to adjust dynamic ranges in addition to radiance correction. The texture classifier of [13] is adapted to detect clouds automatically if visible in the region of interest.

1.3. Evaluation Methodology

In this paper, three sites are chosen to compare the two DEM generation techniques where more than 200 satellite images from GeoEye-1, WorldView-1, WorldView-2 and Quickbird satellites are available per site. Since no laser-illuminated detection and ranging (LIDAR) data is available, ground-truth DEMs were prepared manually for each site. Figure 9-10-11 show example ground-truth DEMs. The ground-truth DEMs are generated given two or more satellite images with corrected RPC cameras by interactive construction of 3-D surfaces. DEM surface height is estimated by boundary alignment in multiple images. For this purpose, mostly buildings but also possibly other horizontal surfaces such as roads, pavements, etc. that are visible in all the images are first labeled as polygonal outlines in the first image. Then given one of the polygons, its height in 3-D is adjusted manually such that when it is projected onto other images using their RPC cameras, the polygon aligns accurately with the corresponding structure in the other images. The set of 3-D polygonal outlines created in this manner are projected onto a map-aligned[2] orthographic image using RPC camera of one of the images. Then the polygons are filled with the absolute height value of the structure outlined with that polygon to produce the ground-truth DEM. It is very labor intensive to generate these ground-truth DEMs as many finely detailed structures need to be

[2] The image is aligned with North and East directions in y and x axes.

labeled, and sometimes the exact boundaries are ambiguous in the images. For all the sites, a random selection of a subset of the man-made structures with varying heights is labeled, but deemed sufficient for quantitative performance analysis.

Given a ground truth DEM and a reconstructed DEM of a region, the pixel-wise height differences to the ground-truth DEMs are measured. The percentage of all the pixels that have a specified difference in meters with the ground-truth height values are reported as plots shown in Figure 9.

1.4. Related Work

DEM generation from aerial imagery is a well-studied field with the major approach being stereo reconstruction from one or more pairs of stereoscopic images. The latter case is called multiview stereo and the individual reconstructions of each available pair is fused in various ways either in 2-D or 3-D. The second approach is the direct use of all available views of a scene for reconstruction in 3-D which is termed as true multiview. It is beyond the scope of this paper to review all the approaches here, but [4] is a good survey paper that lists major contributions to stereo and multiview approaches. In [4], the authors list full automation as the major challenge of the state of the art, and for satellite imagery, the manual steps are usually during bundle adjustment or selection/acquisition of favorable images. In this paper, two *fully-automated* reconstruction pipelines that are applicable to *cross-sensor* and *nonstereoscopic* satellite images are compared. The first one is a multiview stereo (Section 2) and second one is a true multiview (Section 3) approach. Both algorithms are capable of using as many images as available for a given site after automated relative geo-registration (bundle adjustment) of [6] is applied. The multiview stereo approach of Section 2 uses the semi-global matching algorithm of [14] to generate disparity maps as well as its orthographic fusion algorithm of the maps via simple median filtering. However, the affine rectification of satellite images to generate scan line correspondences is an idea borrowed from [3]. The difference to [3] is the use of fully-automated geolocation correction, a priori to reconstruction, so that the accurate alignment induced by RPC camera models can then be exploited to generate a sufficient number of *accurate* correspondences for affine rectification.

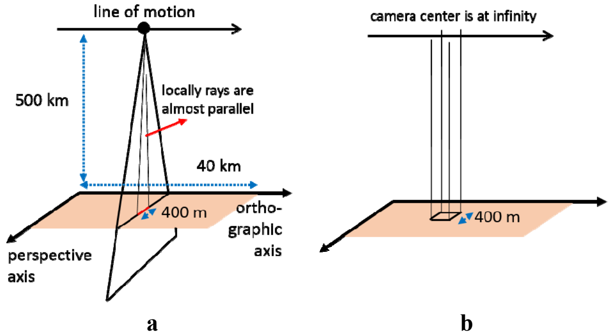


Figure 2: **a**) Pushbroom camera model of satellite imagery **b**) A local fit as an affine camera model where the camera center is at infinity.

2. Multiview Stereo Reconstruction

In computer vision, stereo reconstruction is a well-studied problem when the imaging geometry of the stereo pair is given by perspective cameras. If the cameras are perspective then the epipolar geometry can be used to *rectify* the images [15]. Rectification is a procedure where a 2-D rigid transform is applied to both images so that the corresponding epipolar lines become horizontal. In this rectified form, the problem of solving for the correspondence of pixels in the left and right images transforms into a 1-D problem of solving for the correspondence along the scan lines of the images. Many robust algorithms [16] exist that efficiently solve for the correspondence along the scan lines to generate a *disparity map*. However, in the case of satellite imagery, the imaging geometry is given by a *pushbroom* camera, Figure 2a, for which the epipolar curves are not linear [17]. However, observe from Figure 2a that locally the rays are well approximated by parallel lines due to the high altitude of the camera. Thus, an *affine camera model* can be used as an approximate projection model for *local* image patches of a large satellite image. Note that due to the perspective axis of pushbroom model, Figure 2a, the orientation of the rays change slightly at different parts of the satellite image and it becomes necessary to fit a different affine approximation at different parts of the image. The important consequence of affine approximation is that the epipolar curves given by affine camera models are linear and thus local image patches can now be rectified. In [17] it has been theoretically proven that epipolar curves of pushbroom cameras can indeed be well approximated by piece-wise linear models as given by local affine camera fits.

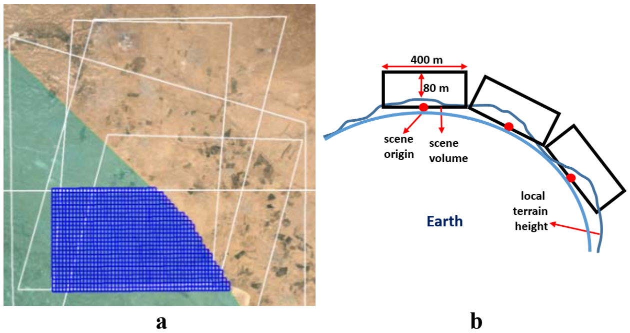


Figure 3: a) 400 m x 400 m x 80+ m scene boxes shown in blue in the intersection area of five satellite images with outlines shown in white. b) An illustration of the side views of 3-D scene boxes tiling the Earth surface. Earth curvature is exaggerated for demonstration purposes.

To generate local image patches in a systematic way for many satellite images, the region of interest to be reconstructed on Earth surface is first tiled with 3-D scene boxes. Each scene box has a local Euclidean coordinate system with an origin specified in WGS84 coordinates on Earth surface, Figure 3b. The scene has a flat plane at the base elevation; however, the sizes of the scene boxes are kept small and thus they are good approximations to the curved Earth surface. The size of the surface area of the scene boxes are 400x400 m^2 in our experiments, and many scenes are generated side by side with shifted origin points. The base elevation and heights of the volumes are initialized using Advanced Spaceborne Thermal Emission and Reflection Radiometer Digital Elevation Model (ASTER DEM) tiles of the area with 80 meter margin added on top of the highest terrain elevation to account for buildings[3]. The 3-D scene boxes bound the elevation variation in each 400x400 m^2 area and provide the local coordinate system for affine camera fitting. Specifically, the box origin has coordinates (0,0,0) and for any 3-D point $X = (x, y, z)$ in the local coordinate system, a corresponding geodetic coordinate $\overline{X} = (\text{lat}, \text{lon}, \text{elev})$ can be computed. The 3-D scene boxes also provide a mechanism to crop smaller patches from the satellite images that observe the same scene. For this purpose, the eight corners of the scene box are projected onto the image using the RPC model and the image is cropped using the smallest 2-D box that contains all the projected corners.

The key insight in this paper is that once the RPC cameras of two images are georegistered to each other through bias correction as explained in Section 1.1, they can be used to compute precise 3-D to 2-D correspondences. Specifically, given a 3-D scene box, a set of 3-D points, $\{(X_i, \overline{X}_i)\}$ can be initialized in this box *randomly*. Then given the RPC camera of an image, the corresponding set of 2-D image points, $\{(u_i, v_i)\}$, can be computed using $u_i = F_{RPC}(\overline{X}_i)u_s + u_0$ and $v_i = F_{RPC}(\overline{X}_i)v_s + v_0$. An affine camera model, P_A, is then fit using the 3-D to 2-D correspondence set $\{X_i \leftrightarrow (u_i, v_i)\}$ via the standard algorithm given in [15].

Given two cropped satellite images, I and I′, of a scene, rectification entails computation of two homography matrices H and H′ that rotate the two images so that the *corresponding* epipolar lines are horizontal and the images are minimally distorted. We use the same affine epipolar rectification technique of Ames Stereo Pipeline as described in [18] to compute H and H′ using an affine fundamental matrix, F_A. In [18], F_A is computed numerically from image correspondences whereas in our case it is derived analytically [15] using the locally fitted affine camera models, P_A and $P_A{}'$. Specifically:

$$F_A = [E']_\times P_A{}' P_A^+$$

where P_A^+ is the pseudo-inverse of P_A, E' is the epipole defined by $E' = P_A{}'C$ and C is the center of P_A. C is a homogeneous point at infinity defined using the principal ray, (r_x, r_y, r_z), of P_A and it is defined as $C = [r_x \;\; r_y \;\; r_z \;\; 0]^T$. Note that after matrix multiplications F_A is in the following form:

$$F_A = \begin{bmatrix} 0 & 0 & a \\ 0 & 0 & b \\ c & d & e \end{bmatrix}$$

and the epipole points in the left and right images can also be written as $E = (-d, c, 0)^T$ and $E' = (-b, a, 0)^T$. Using the derivation of [18], the rotation matrices that make the epipolar lines horizontal can be written as:

$$R = \begin{bmatrix} E_0/\|E\| & E_1/\|E\| \\ -E_1/\|E\| & E_0/\|E\| \end{bmatrix} \quad R' = \begin{bmatrix} E'_0/\|E'\| & E'_1/\|E'\| \\ -E'_1/\|E'\| & E'_0/\|E'\| \end{bmatrix}$$

For the rotated images RI and R′I′, the epipolar lines are horizontal; however, they are not in corresponding scan lines yet. The images need to be scaled and shifted for this purpose. At this point, it is necessary to use some correspondence points to solve for the optimal scale, shift and skew parameters that minimally distort the images. Also stereo matching algorithms work best if the points on the ground plane have zero disparity. Thus the same method of generating random correspondences via RPC models is used. Specifically, given the 3-D scene box, a set of 3-D points, $\{(X_i, \overline{X}_i)\}$, are initialized such that $z = 0$, i.e. they are on the local ground plane[4]. Then these points are projected onto local image patches to generate image correspondences: $\{x_i = (u_i, v_i) \leftrightarrow x'_i = (u'_i, v'_i)\}$ where $u_i = F_{RPC}(\overline{X}_i)u_s + u_0$, $v_i = F_{RPC}(\overline{X}_i)v_s + v_0$ and $u_i = F'_{RPC}(\overline{X}_i)u'_s + u'_0$, $v_i = F'_{RPC}(\overline{X}_i)v'_s + v'_0$. The problem is then formulated as a linear set of equations in the form $Ax = b$ to solve for the optimal transformation parameters. Specifically, first the translation and scale for y axis of the

[3] The height of the scene volume is denoted as 80+ meter in Figure 7a as it changes with respect to the terrain.

[4] Note that the absolute elevation of the ground plane of the scene box was determined using ASTER DEM.

image is solved by setting up:

$$A_{i,:} = [(R'x'_i)_1 \quad 1]$$
$$b_i = (Rx_i)_1$$

such that $A\begin{bmatrix} y_s \\ y_0 \end{bmatrix} = b$. Second, the scaling, skew and translation for x axis of the image is solved by setting up:

$$A_{i,:} = \left[\begin{bmatrix} 1 & 0 & 0 \\ 0 & y_s & y_0 \\ 0 & 0 & 1 \end{bmatrix} R'x'_i\right]$$
$$b_i = (Rx_i)_0$$

such that $A\begin{bmatrix} x_s \\ x_{skew} \\ x_0 \end{bmatrix} = b$. Then the final rectification homography matrices become:

$$H = \begin{bmatrix} 1 & -x_{skew}/2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} R$$
$$H' = \begin{bmatrix} x_s & x_{skew}/2 & x_0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & y_s & y_0 \\ 0 & 0 & 1 \end{bmatrix} R'$$

An example rectified pair of images as HI and $H'I'$ are shown in Figure 4a-b. Observe from Figure 4a-b that the scan lines of the images indeed correspond.

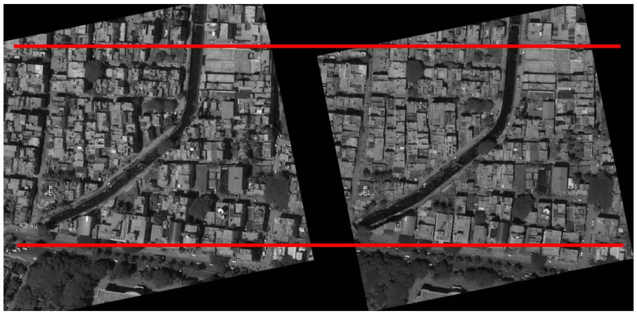


Figure 4: **a-b)** The rectified images using local affine approximations. The red lines mark two example scan lines to demonstrate their correspondence.

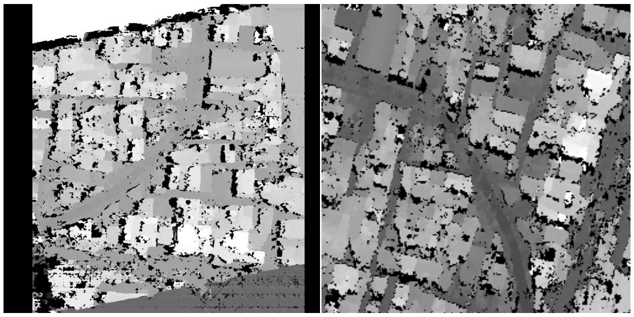


Figure 5: **a)** The disparity map for the pair in Figure 4. **b**) The orthographic height map created using the disparity map in (**a**). The map has 1 meter GSD and size 400x400.

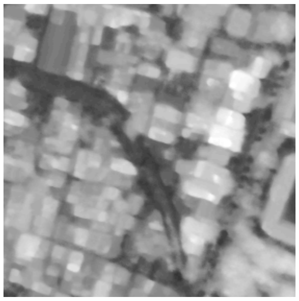
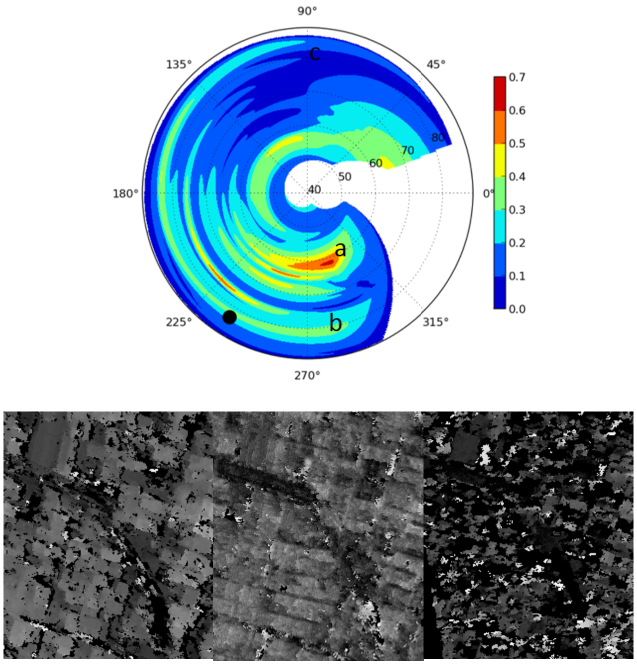


Figure 6: Median height map from over 5000 rectified pairs of 73 satellite images (using only PAN bands with GSDs 1 meter or better).

Given rectified image pairs as in Figure 4a-b, any stereo matching algorithm that generates a disparity map for the pixel correspondences along the scan lines is applicable. In this paper, the well-known semi-global matching (SGM) algorithm of [14] is used. This algorithm enforces a local continuity constraint to generate regularized disparity maps and it has been shown to produce good results for aerial imagery [14]. Figure 5a shows the disparity map for the example in Figure 4. Given the pixel to pixel correspondences via the disparity map, it is possible to reconstruct an orthographic height map by triangulating the 3-D points. Since back-projection using RPC camera models entails costly optimization [10], a reverse method is used in this paper. Specifically, all possible 3-D points in the 3-D scene box with 1 meter postings in x, y and z coordinates is traversed. For each 3-D point, it is projected onto both left and right rectified images and the consistency of the projections' disparity is measured with the value in the disparity image given by SGM algorithm. For each column in 3-D, i.e. given (x, y), the z value with the most consistent disparity is selected as the height of that column. Thus, a map-aligned orthographic height map is created as shown in Figure 5b. Observe from Figure 5b that the buildings, river, streets and even some trees are recovered fine albeit with numerous gaps.

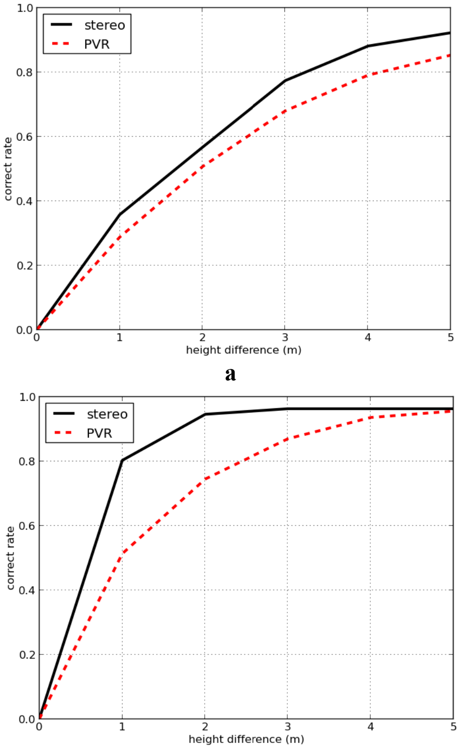
Given N images that are geolocation corrected, it is possible to create $N(N-1)$ pairs and generate a DEM using the described approach. In this paper, all these DEMs are combined simply by taking the median height at each pixel value disregarding the gaps. An example final map is shown in Figure 6a. The final height map has no gaps and the relative heights of the buildings and streets seem to be accurately recovered even given the challenging dense urban scene structure in Bangalore, India.



a **b** **c**

Figure 7: Top row: The variation of DEM accuracy with respect to the viewing angle combination of the second image generated from over 5000 pairs in Bangalore site. In the plot, azimuth angle changes from 0° to 360° and elevation changes from 40° to 90° with 90° corresponding to a nadir image. The bar on the right shows the color code for the percentage of pixels with 3 meter difference to the ground-truth DEM. The first image's elevation and azimuth angles are marked as a black dot and it is collected by GeoEye-1. Bottom row: a) shows the best DEM for a pair with second image collected by Quickbird satellite, b) an average result by WorldView2 and c) the worst view angle configuration with the second image by Quickbird. The view angles of a,b and c are marked on the plot.

The combined result is impressive given that the paired images are collected by possibly different satellites and at very different times. Obviously, the quality of DEMs from individual pairs differ quite a bit. The sun-angle and view-angle combinations of the paired images, illumination changes due to atmospheric effects and local surface properties determine the quality. The effect of image viewing angles, $(\text{elevation}, \text{azimuth})$ as reported in the metadata, on the accuracy of DEM is shown in Figure 7 for the Bangalore site. The ground-truth height map is used to measure the percentage of pixels with 3 meters difference to the ground-truth value. For the best pair in Figure 7a, roughly 70% of the pixels differed by 3 meters or less from the ground truth. Observe from the plot that it is possible to single out regions in the plot that result in good quality DEMs. For example, there seems to be a large range of azimuth angles for the second image that generate high quality DEMs when its elevation is the same as the first image. Observe that when the angular configuration is not ideal, then the DEM is very poor with lots of gaps as in Figure 7c. More research is needed to discover rules in selection of the pairs pertaining to relative view and sun angle combinations in order to avoid the combinatorial explosion in the number of possible pairs. However, in this paper, all possible $\mathrm{N(N-1)}$ pairs given N images of a site are used to produce a combined DEM.



a

b

Figure 8: **a)** The correct rate of pixels with height value differences of 1 to 5 meter to values in Figure 10d for the site in Figure 10. **b)** The same plot for the site in Figure 11.

3. Multiview Reconstruction using PVR

A completely different approach to produce a DEM using N images is through multiview triangulation directly in 3-D. In this paper, the probabilistic volumetric representation (PVR) of [4,5] is chosen as the multiview approach to be compared to the stereo approach outlined in Section 2. PVR algorithm represents the 3-D volume of a scene using a regular grid of cubic volume elements, *voxels*, e.g. of size 1 m^3, and computes a surface occlusion probability and a surface appearance model for each voxel in the volume. The rays from the available images are cast into the volume using the images' camera models and the surface existence probabilities are updated simultaneously with their appearance models using the appearance of the rays. It is shown that the algorithm converges to the correct 3-D surface model as more images are used to update the model [4]. The critical assumption is that the

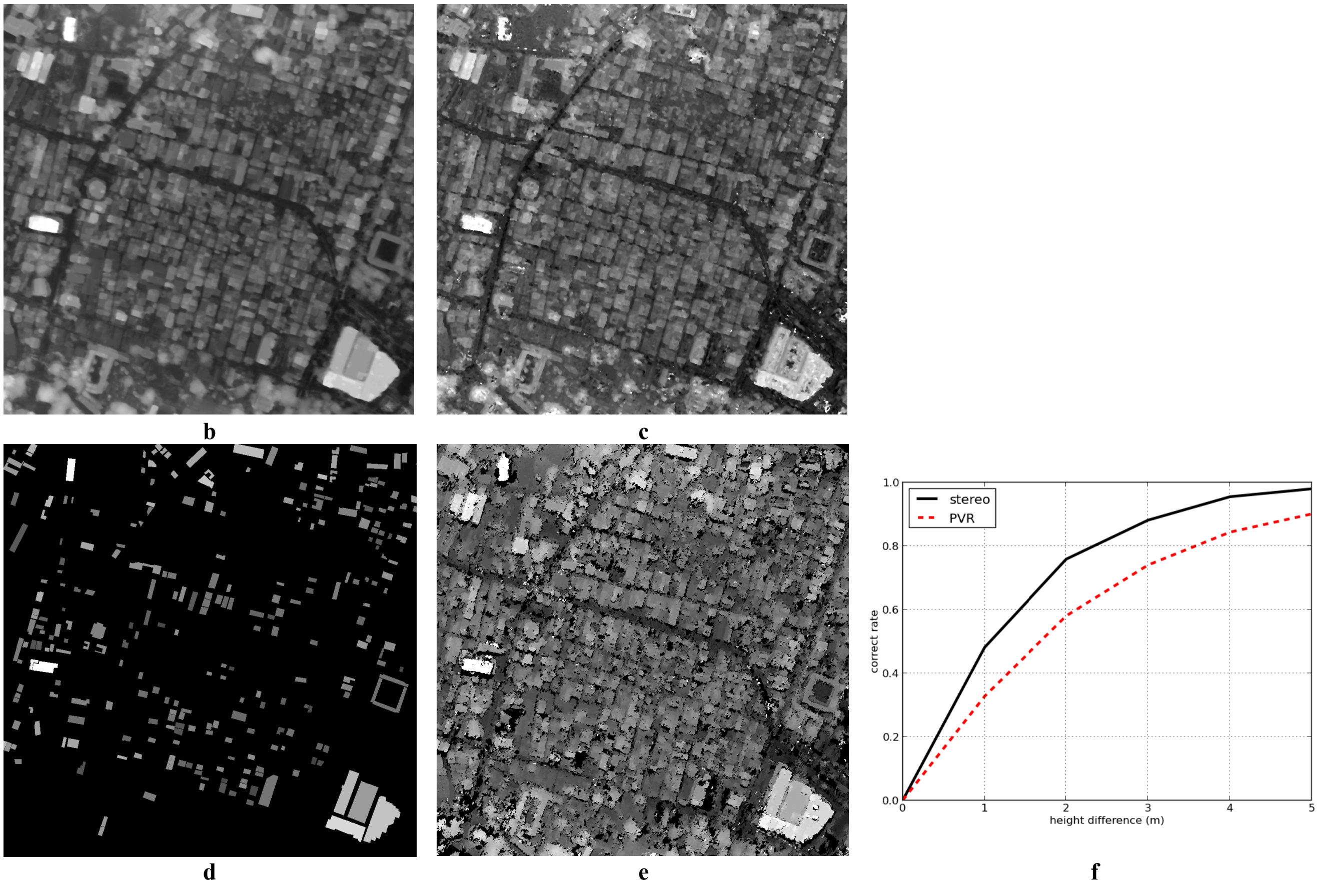


Figure 9: Dense urban site from Bangalore, India, $N = 73$. **b)** Multiview stereo result from 5256 pairs. **c)** PVR result. **d)** Manually prepared ground-truth DEM. **e)** The stereo result from the best pair based on correct rate with respect to (**d**). **e)** The rate of pixels with height value differences of 1 to 5 meter to values in (**d**). The rate is calculated using only non-black pixels in (**d**).

geopositioning errors of the input images are *relatively* corrected such that the triangulation errors can be absorbed within a voxel. The geolocation correction step outlined in Section 1.1 assures that the images are relatively correct and thus the method becomes applicable given a large set of images. For each experimental site, all the available images are used to update a 3-D model using PVR. The map-aligned orthographic height maps are created from the 3-D model by projecting the height value of the first *visible* voxel along each ray of the orthographic camera. The visibility of each voxel along a ray is computed from the surface existence probabilities of the voxels [4].

4. Results

Three sites with dense urban, semi-urban and rural characteristics are chosen from Bangalore, India; Sydney, Australia and Canberra, Australia. Figures 9-11 show a crop from the sites along with the DEMs given by the stereo and multiview approaches as well as the manually generated ground-truth DEM. For each site a plot of the percentage of pixels (y axis) with a height difference of 1 to 5 meters (x axis) to the ground-truth DEM is shown as well, Figure 9d and Figure 8a-b. Overall, multiview stereo approach generates better quality DEMs with 84% of the pixels within 3 meter difference to the ground-truth for the very challenging urban site. Observe that multiview stereo generates fewer gaps especially on the building roof tops compared to PVR algorithm. Also observe from the rural site of Figure 11, the large homogenously textured areas are much better reconstructed by the stereo approach. The major reason for this superiority is the multidirectional continuity constraint enforced by SGM algorithm during disparity computation. PVR treats each voxel independently during reconstruction and no such regularization is enforced. The second reason is the mutual information based matching cost of SGM being largely immune to illumination differences of surfaces viewed in different images whereas PVR models the absolute values of the surface appearances. On the other hand, PVR reconstructs a sharper geometry and has $\mathrm{O(N)}$ complexity given N images, whereas multiview stereo algorithm has $\mathrm{O(N^2)}$ complexity.

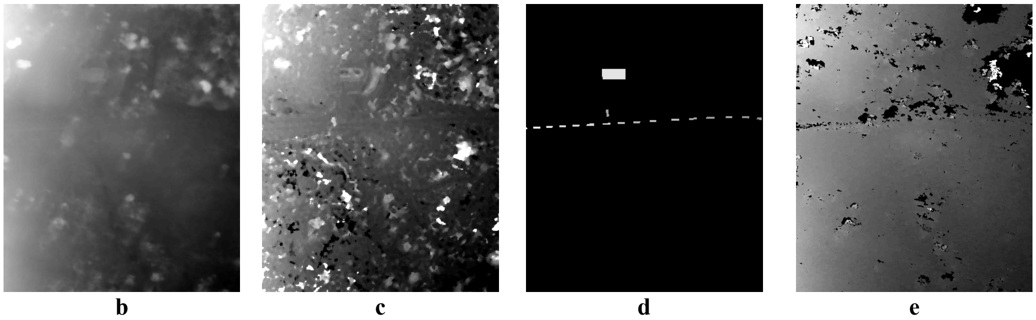


Figure 10: Rural site from Canberra, Australia, $N = 149$. **b-c-d-e** as in Figure 9.

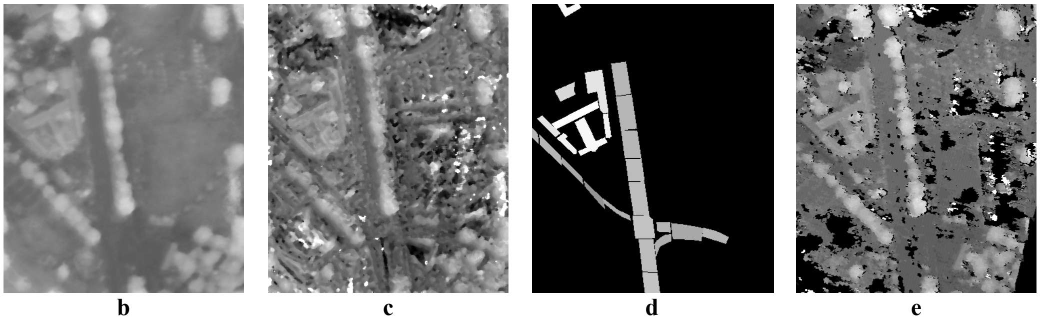


Figure 11: Semi-urban site from Sydney, Australia, $N = 51$, **b-c-d-e** as in Figure 9.

5. Conclusions

In this paper, two fully automatic algorithms that generate DEMs from a large set of multiview and nonstereoscopic satellite images collected by different satellites are evaluated based on their reconstruction accuracy given manually prepared DEMs. The multiview stereo algorithm is found to generate more accurate DEMs compared to PVR algorithm on the urban and rural sites used in this paper. Note that more experimentation with LIDAR data is needed to quantify the absolute accuracy of the elevations given by both algorithms. There is a lot of room for improvement for both reconstruction algorithms. Multiview stereo algorithm would benefit from a 3-D fusion approach rather than the simplistic median value based composition of the height maps. PVR algorithm clearly falls behind due to its well-known shortcoming for reconstruction of homogenously textured areas and a 3-D regularization of the reconstructed surfaces is required. A third avenue of improvement would be a combined approach to quickly initialize a surface with the stereo algorithm using a subset of images and then update with the rest via PVR to sharpen the surfaces.

Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory, contract FA8650-12-C-7211. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. **Disclaimer**: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

References

[1] "The Ames Stereo Pipeline," Intelligent Robotics Group, NASA Ames Research Center, 2014. [Online]. http://byss.arc.nasa.gov/stereopipeline/binaries/asp_book-2.4.2.pdf

[2] M. J. Broxton and L. J. Edwards, "The Ames Stereo Pipeline: Automated 3D Surface Reconstruction from Orbital Imagery," in *Lunar and Planetary Science Conference 39*, vol. Abstract #2419.

[3] Z. M. Moratto, M. J. Broxton, R. A. Beyer, M. Lundy, and K. Husmann, "Ames Stereo Pipeline, NASA's Open Source Automated Stereogrammetry Software," in *Lunar and Planetary Science Conference 41*, vol. Abstract #2364.

[4] Thomas Pollard and Joseph L. Mundy, "Change Detection in a 3-D World," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2007.

[5] Thomas Pollard, Ibrahim Eden, Joseph L. Mundy, and David B. Cooper, "A Volumetric Approach to Change Detection in Satellite Images," *ASPRS*

Photogrammetric Engineering & Remote Sensing Journal, vol. 76, no. 7, pp. 817-831, 2010.

[6] Ozge C. Ozcanli et al., "Automatic Geolocation Correction of Satellite Imagery," in *Proc. of IEEE Conf. of Computer Vision Pattern Recognition (CVPR) Workshops*, 2014.

[7] GeoEye1. https://www.digitalglobe.com/sites/default/files/DG_GeoEye1.pdf.

[8] Quickbird. https://www.digitalglobe.com/sites/default/files/QuickBird-DS-QB-Prod.pdf.

[9] G. Dial and J. Grodecki, "RPC replacacement camera models," in *Proceedings of the ASPRS 2005 Annual Conference*, 2005.

[10] Jacek Grodecki and Gene Dial, "Block Adjustment of High-resolution satellite images described by Rational Polynomials," *Photogrammetric Engineering and Remote Sensing*, vol. 69, pp. 59-68, 2003.

[11] "Radiometric use of WorldView-2 Imagery," DigitalGlobe,.

[12] Nancy E. Podger, William Colwell, and Martin Tay, "GeoEye-1 Radiance at Aperture," GeoEye1, 2011. [Online]. https://apollomapping.com/wp-content/user_uploads/2011/09/GeoEye1_Radiance_at_Aperture.pdf

[13] M. Varma and A. Zisserman, "A Statistical Approach to Texture Classification from Single Images," *International Journal of Computer Vision*, vol. 62, no. 1, pp. 61-81, 2005.

[14] Heiko Hirschmuller, "Stereo Processing by Semi-Global Matching and Mutual Information," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328-341, 2008.

[15] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed.: Cambridge University Press, 2000.

[16] Middlebury Dataset and Evaluations. [Online]. vision.middlebury.edu/stereo/

[17] Taejung Kim, "A Study on the Epipolarity of Linear Pushbroom Images," *Photogrammetric Engineering & Remote Sensing*, vol. 66, no. 8, pp. 961-966, 2000.

[18] Zack Moratto. (2013) Satellite Epipolar Rectification. [Online]. http://lunokhod.org/?p=1206#ref

[19] Michel Morgan, Kyung-Ok Kim, Soo Jeong, and Ayman Habib, "Epipolar Resampling of Space-born Linear Array Scanner Scenes Using Parallel Projection," *Photogrammetric Engineering & Remote Sensing*, vol. 72, no. 11, pp. 1255-1263, 2006.

[20] Thomas Pollard and Joseph L. Mundy, "Change Detection in a 3-D World," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2007.

[21] Przemyslaw Musialski et al., "A survey of urban reconstruction," *Computer Graphics Forum*, vol. 32, no. 6, pp. 146-177, 2013.

[22] Clive S. Fraser and Harry B. Hanley, "Bias compensation in rational functions for IKONOS satellite imagery," *Photogrammetric Engineering and Remote Sensing*, vol. 69, pp. 53-58, 2003.

[23] Pablo d'Angelo and Peter Reinartz, "DSM based orientation of large stereo satellite image blocks.," *Int. Arch. Photogrammetry and Remote Sensing Spatial Inf. Sci*, vol. 39, pp. 209-214, 2012.

[24] Jaehong Oh, Charles Toth, and Dorota Grejner-Brzezinska, "Automatic Georeferencing of Aerial Images Using High-resolution stereo satellite images," in *ASPRS Annual Conference*, 2010.

[25] Mark D. Pritt and Kevin J. LaTourette, "Automated Georegistration of Motion Imagery," in *Applied Imagery Pattern Recognition Workshop (AIPR)*, 2011.

[26] Victor Tom, G. K. Wallace, and G. J. Wolfe, "Image registration by a statistical method," in *Proc. SPIE Applications of Digital Image Processing VI*, vol. 432, 1983.

[27] Edward M. Mikhail, James S. Bethel, and J. C. McGlone, *Introduction to Modern Photogrammetry.*: John Wiley & Sons, Inc., 2001.