Geometric-constrained multi-view image matching method based on semi-global optimization

Wenhao Zhao\textsuperscript{a,b}, Li Yan\textsuperscript{a} and Yunsheng Zhang\textsuperscript{c}  
\textsuperscript{a}School of Geodesy and Geomatics, Wuhan University, Wuhan, China; \textsuperscript{b}National Geomatics Center of China, Beijing, China; \textsuperscript{c}School of Geosciences and Info-Physics, Central South University, Changsha, China

**ABSTRACT**  
Targeting at a reliable image matching of multiple remote sensing images for the generation of digital surface models, this paper presents a geometric-constrained multi-view image matching method, based on an energy minimization framework. By employing a geometrical constraint, the cost value of the energy function was calculated from multiple images, and the cost value was aggregated in an image space using a semi-global optimization approach. A homography transform parameter calculation method is proposed for fast calculation of projection pixel on each image when calculation cost values. It is based on the known interior orientation parameters, exterior orientation parameters, and a given elevation value. For an efficient and reliable processing of multiple remote sensing images, the proposed matching method was performed via a coarse-to-fine strategy through image pyramid. Three sets of airborne remote sensing images were used to evaluate the performance of the proposed method. Results reveal that the multi-view image matching can improve matching reliability. Moreover, the experimental results show that the proposed method performs better than traditional methods.

1. Introduction

The generation of digital surface models (DSMs) from airborne or spaceborne remote sensing images via a reliable and dense matching is an essential and difficult task in digital photogrammetry and remote sensing (Zhu et al. 2005; Zhang and Gruen 2006; Zhu, Wu, and Tian 2007; Martina and Christian 2012; Wu, Zhang, and Zhu 2012). Reliable dense matching has received vast attentions in the last 40 years and tremendous efforts have been made to improve image matching performance (Otto and Chau 1989; Furukawa and Ponce 2010; Bulatow, Wernerus, and Heipke 2011; Steve and Arko 2012; Knaptitsch et al. 2017). Several state-of-the-art dense matching methods have been implemented in currently commercial digital photogrammetric software systems for producing DSMs (Zhang et al. 2007; Lemaire 2008; Match-T DSM 2017). However, image matching is still a hindrance and numerous human interactions are required for the generation of DSMs with a favorable quality (Gruen 2012).

Several image constraints have been used for a reliable image matching. For example, in computer vision community, the image matching based on global constraint has obtained great achievements, such as the matching method based on graph cut and belief propagation (Boykov, Veksler, and Zabih 2001; Felzenszwalb and Huttenlocher 2006). These methods transfer matching task into an energy minimization-based optimization procedure, which considers both smooth and occlusion assumptions. However, most of them focus on short-baseline stereo pair with little image distortion and cannot adapt to the processing of large size airborne or spaceborne remote sensing images. With the recent developments in multiple digital aerial cameras, highly overlapped images can be obtained without any additional cost, and a single physical object can be observed on more than 10 images (Leberl et al. 2010). These advances have encouraged research on multi-baseline image matching methods. Employing multiple images in image matching exploits redundant information and can alleviate the matching ambiguity caused by occlusions, surface discontinuities, and other problems. Several methods consider using the redundant information from multiple images to improve the matching result. Zhu et al. (2010) presented a triangulation-constrained multiple image matching method for improving the stereo matching reliability, but this method can only match distinctive interest point. Zhang et al. (2007) and Hirschmuller (2008) combined 3D information from separate multiple stereo pairs to obtain integrated 3D information. However, multiple image information is ignored when dealing with different stereo pairs.

**CONTACT** Yunsheng Zhang zhangys@csu.edu.cn

© 2018 Wuhan University. Published by Taylor & Francis Group. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
This paper presents an energy-based image matching method incorporating geometrical constraint, which uses multiple image information to calculate the matching cost. A semi-global optimization approach was employed to obtain a dense matching result for DSM generation. In Section 2, a review on image matching is given. The proposed method is presented in detail in Section 3. In Section 4, multiple aerial remote sensing images were employed for the experimental analysis and quantitative evaluation of the developed method. Finally, concluding remarks are presented and discussed.

2. Related work

2.1. Local vs. global

Traditional stereo matching method for remote sensing images is based on the correlation coefficient method, such as the normalized correlation coefficient (NCC) presented in (Helava 1978) and the least square matching proposed in (Gruen 1985), which depends on the selection of an appropriate window to calculate the correlation values. On the one hand, if the window size is small and does not contain enough intensity variations, a poor match result will be obtained. On the other hand, if the window size is large, the regions where the depth of the scene (i.e. disparity) varies significantly may be covered that will also lead to a poor match result. In both situations, the maximum (Max) of correlation may not represent correct matches owing to the geometrical distortions among stereo images. For alleviating such matching ambiguity problem, a number of researchers have presented varieties of correlation method, such as the adaptive window (Kanade and Okutomi 1994), shift window (Bobick and Intille 1998), multiple window (Hirschmüller 2001), and adaptive support-weight approach that considers color similarity and geometrical proximity (Yoon and Kweon 2006). All these methods are fast and can improve the matching reliability to some extent compared with the original correlation method. However, the matching result of local image matching method based on local information has no or weak relation to its neighborhood, so that the global reliability of the result cannot be ensured.

Deriving 3D information from 2D images is an ill-posed problem. A number of unreliable results cannot be avoided when only using local information for matching as mentioned earlier. Nevertheless, the large portion of matching result is right and only small portion of matching result is inconsistent with remote sensing images. If global consistency is considered, the wrong matching result can be reduced, improving matching reliability to a significant extent. In processing remote sensing images, image matching based on relaxation algorithm and dynamic programming are two kinds of typical global matching methods (Zhang et al. 2000; Alobeid, Jacobsen, and Heipke 2010). The image matching method based on dynamic programming can obtain optimized matching result on epipolar lines but ignores the relationship among epipolar lines. For addressing this problem, Alobeid, Jacobsen, and Heipke (2010) performed a vertical median filtering process to reduce the streaking problem inherited in the dynamic programming-based matching method, and thus, the artifacts in the final result can be removed. Hirschmüller (2008) performed dynamic programming in multiple directional and obtained not only epipolar lines but also several prominent results. The global image matching via relaxation based on the assumption of the local smooth of terrain or object utilizes contextual information to reduce local ambiguity and achieve global consistency. The smooth consideration used in this strategy can lead to a number of artifacts when facing surface discontinuities. For tackling this problem, Zhang and Gruen (2006) explicitly constructed edges to model surface details and constrained the relaxation process to ensure matching reliability.

To deal with the ill-posed problem in stereo matching, Terzopoulos (1986) proposed a regulation method for transferring the ill-posed problem to a well-defined problem. This strategy depends on the reduction of defined energy function via several optimization methods, and thus, global consistency matching result can be finally obtained. However, these methods are time consuming and cannot be adapted in the practical processing of airborne and spaceborne remote sensing images (Alobeid, Jacobsen, and Heipke 2010).

2.2. Single stereo vs. multi-baseline

Traditional aerial mapping is based on film photography, which often depends on a single-stereo matching method to obtain dense matching result and derive DSMs. In a number of limited situations, only stereo pair image can be obtained, such as the rover servicing on the Mars surface, which is a fixed-baseline stereo camera installed on a rover for obtaining stereo image and generating a surface model of Mars for navigation (Wu, Zhang, and Zhu 2011). The researchers from computer vision field pay more attention to a single-stereo matching method. After the launch of stereo benchmark by (Scharstein and Szeliski 2002), a large number of image matching methods based on single stereo are proposed. However, most of them require parameter tuning to perform better on benchmark data-sets, and the performance of them on remote sensing images has not been well studied.

Zhu et al. (2005) presented an image matching method based on self-adaptive triangulations for the stereo pair of airborne remote sensing images. Triangulations are dynamically updated along with the matching process by inserting newly matched points into triangulations. The dynamic updating of triangulations is adaptive to image textures by propagating the matching from good textural areas to that of difficult areas. Wu, Zhang, and
Zhu (2011, 2012) extended the matching method to a wide-baseline image and incorporated edge features to process poor textural stereo image matching.

However, if only a single stereo is employed for image matching, the blunder rate from automatically matching can be rather high, and removing such blunders will result in a number of holes in the final result. When digital aerial images become possible during the last decade, including ADS40/80, DMC, SWDC, and Ultracam-X, an object point can be viewed on more images. Thus, efforts are changed to use the redundant information from multiple images to generate DSM. In addition, higher measurement accuracy can be achieved through the spatial intersection of more than two imaging rays. For matching highly overlapped images, stereo matching methods are naturally extended to a multi-baseline model. One effort is to combine DSM from a single stereo matching method to produce an integrated DSM (Bignon 2003; Zhang et al. 2007; Hirschmuller 2008; Lemaire 2008; Yan et al. 2016). Zhang et al. (2007) presented a matching algorithm for choosing a better stereo pair for matching based on terrain slope, orientation, and sensor geometry. Elevations from different stereo pairs and methods (back matching) are used to detect blunders to improve the reliability of the matching result. The Match-T DSM (2017) selects the best suited image pairs based on the analysis of the DSM slope to perform stereo pair image matching. For several areas with poor texture, the Match-T DSM tries up to 20 stereo pair combinations in per matching unit. After that, the Match-T DSM employs a true 3D filtering to combine all the DSMs from stereo pairs to produce a merged DSM. Bignon (2003) stated that once images are loaded into the Pixel Factory, algorithms create hundreds of stereo pairs and distribute the computations over available computers to parallelize and speed up the automatic stereo matching. Multi-correlation is performed between along-track and side-lap stereo pairs. After stereo matching, all results are merged to form a complete DSM. Yan et al. (2016) proposed a graph network to select matching stereo pairs and fusing point clouds. Another effort is to match all available images simultaneously to reduce blunders in matching result. Zhang and Gruen (2006) presented an image matching method for digital elevation model generation using three-view IKONOS images. This method calculates the average of the NCC value between the search and the reference images to form a strong reliability indicator and adopts the relaxation method considering edge feature as a constraint to obtain a reliable matching result. Furukawa and Ponce (2010) presented three stages of multi-baseline matching method, namely, patch-based multi-view stereo (PMVS). In this method, seed points are matched by feature point matching first. Then, the seed points are propagated to its neighborhoods. Finally, a filtering process is performed. The propagation and filtering process are repeated for several times. PMVS has been used to process the unmanned aerial vehicle remote sensing image, and its accuracy is high if the texture of images is good (Steve and Arko 2012).

2.3. Image space vs. object space

Typical image matching using disparity image representation is performed in image space. An excellent review on such kind of methods can be found in Scharstein and Szeliskiand’s stereo matching benchmark evaluation homepage (Scharstein and Szeliski 2002; EVAL 2017). A number of other matching methods performing directly on the original stereo image have been proposed (Lhuillier and Long Quan 2002; Lee et al. 2003; Radhika et al. 2007). These methods explicitly use epipolar constraints to restrict the corresponding search range and a number of segment information to restrict the search range further, such as triangulation constraints or edge features (Wu, Zhang, and Zhu 2012).

When multiple images are available, a number of matching methods are used to match all images simultaneously in an object space, such as (Gruen and Baltzavias 1988; Zebedin et al. 2006). Zebedin et al. (2006) adopted a multiple image matching method called plane sweeping to obtain the coarse level of DSM first. Then, the DSM was improved and densified using an iterative and hierarchical multi-view matching algorithm.

Another kind of object-spaced based image matching method based on volume optimization has been proposed, such as the voxel coloring/space carving approaches (Slabaugh et al. 2004). These methods initialize the reconstructed object with coarse volume grids and then apply visibility calculation and photo consistency checking to remove the impossible grids. Finally, all visible and photo consistency voxels are retained, which are used to model the scene or render new viewpoint scene. However, most of the voxel-based matching methods do not consider the relation among voxels in spite of their visibility. Recently, Hernandez, Torr, and Cipolla (2007) introduced the Markov random field into voxel coloring/space carving-based matching method, which considered not only the visibility and photo consistency, but also the smooth assumption among voxels, and then used a graph cut to optimize volume.

3. Geometric-constrained multi-view image matching based on semi-global optimization

3.1. Overview of the method

The framework of the proposed method in this paper is shown in Figure 1. The main steps are highlighted in Figure 1, which include the following:

1. Cost value calculation considering geometrical constraints. In this step, a base image is selected at first. Then, a multi-view census transform is
3.2. Calculation of cost values considering a multi-view geometrical constraint

A large number of studies have been conducted on the cost function performance evaluation for image matching. For remote sensing images, the commonly used method is the NCC method. However, selecting a proper window for the NCC value calculation is difficult. Based on a comparative work on the cost functions for stereo matching presented in (Hirschmuller and Scharstein 2009), the census transform proposed in (Zabih and Woodfill 1994) is a good option for image matching, which is a full invariant to illumination change and robust to discontinuity. Census transform uses intensity differences within a pixel's neighborhood to determine the bit string representing that pixel. An example is presented in Figure 2, where the central pixel is 13; the top-left is less than the center, so that it was transformed to 0; the bottom-right is larger than the center, so that it was transformed to 1; and the window was transformed to a bit string as 00011110. The costs were calculated using the Hamming distance of two census transformed pixels in a given correlation window, which counts only the number of positions at which the corresponding symbols differ. In the matching procedure, one image was selected as the base image, while the others served as the match images.

For the cost values that were calculated on multiple images, a geometric model was used to link all available images. For airborne remote sensing images, the collinearity condition equations derived from the interior orientation (IO) and exterior orientation (EO) parameters were used to model the relationship between the 2D image space and the 3D object space. Based on the geometric model, multiple images can be simultaneously matched. The epipolar constraint was implicitly integrated to restrict the search range. The cost value was computed with respect to the elevation value in the object space rather than the disparity in the image space as used in traditional stereo matching to include all available images for cost value calculation (Scharstein and Szeliski 2002). Hence, the Hamming distance functions of all individual stereo image pairs can be integrated in a single framework.

Taking the point $p_0$ on the base image shown in Figure 3 as an example, the image ray from the perspective center $S_0$ to $p_0$ intersects with the elevation plane (defined by the elevation value $Z$) at the point $A$ in the object space. The point $A$ was back projected to all match images, and the possible corresponding points $p_i$ ($i = 1, 2, \ldots, n$) were then obtained. Census transform was performed in a square window around these possible corresponding points. For balancing the reliability and calculation speed, a $5 \times 5$ square window was used. The

![Figure 1. Flowchart of the proposed geometric-constrained multi-view image matching method.](image1)

![Figure 2. Illustration of a census transform.](image2)
similarity between the census transform of a fixed window was measured using the Hamming distance. For \( n + 1 \) multiple images, one image was selected as the base image, while the other images were treated as the match images. Then, \( n \) individual Hamming distances were calculated. An average distance \( C(p, Z_p) \) was used as the cost value.

\[
C(p, Z_p) = \frac{1}{n} \sum_{i=1}^{n} C_0(p, Z_p) \quad (1)
\]

For a given object or scene, the elevation value \( Z \) will be limited in a range \((Z_{\text{min}}, Z_{\text{max}})\). For a numerical solution, the elevation range was divided into \( k \) equal sections. The interval \( dZ \) was determined according to the ground sampling distance of the images, ensuring that two successive elevations can be back projected to two nearby continuous pixels. Thus, each \( Z \) value can be calculated by \( Z_i = Z_{\text{min}} + i \times dZ \), where \( i \in (0, k) \). For each pixel on the base image, a cost cube will be calculated associated to all possible elevation value \( Z \).

If IO and EO parameters are given, a projection matrix can be easily deduced from a collineation equation (Hartley and Zisserman 2004). Figure 3 shows that if the projection matrix for image \( I_0 \) is \( P \), then the projection matrix for image \( I_i \) is \( P_i \). The point \((x, y)\) in image \( I_0 \) can be mapped to \((x', y')\) in image \( I_i \) with the help of the elevation plane defined by the elevation \( Z \). The mapping process can be performed using Equation (2) directly based on homography transformation parameters, which were derived from the projection matrix and the given elevation value \( Z \).

\[
\begin{bmatrix}
  x' \\
  y' \\
  1
\end{bmatrix} = H \begin{bmatrix}
  x \\
  y \\
  1
\end{bmatrix}, \quad \text{where} \quad H = \begin{bmatrix}
  P'_{11} & P'_{12} & P'_{13} & Z_{\text{min}} \\
  P'_{21} & P'_{22} & P'_{23} & Z_{\text{min}} + \Delta Z \\
  P'_{31} & P'_{32} & P'_{33} & Z_{\text{max}}
\end{bmatrix}
\]

For each possible elevation value \( Z \), the homography transform parameters were firstly calculated based on Equation (2). The corresponding projection of the point \( P_0 \), as shown in Figure 3, was directly calculated via the pre-calculated homography transform parameters. The cost value of the current value \( Z \) was then calculated. After processing all possible \( Z \) values, an initial cost cube was obtained.

### 3.3. Semi-global optimization for elevation selection

After calculating the cost for each \( Z \) value of a given pixel, the base image was used as the cost aggregation space. Similar to the disparity space image (DSI), which is commonly used in stereo matching methods (Scharstein and Szeliski 2002), an elevation space image (ZSI) is defined in this paper. The ZSI associated with the base image is defined by substituting the gray value of the base image with the elevation value. Therefore, the matching cost and the smoothness constraints are expressed by defining the energy \( E(Z) \), which depends on the ZSI, as follow:

\[
E(Z) = \sum_p C(p, Z_p) + \sum_{q \in I_p} \lambda_1|Z_p - Z_q| = \Delta Z
\]

In Equation (3), the first term is the sum of the matching costs related to the elevation value \( Z \). The latter two terms penalize the discontinuities with penalty factors \( \lambda_1 \) and \( \lambda_2 \), which have small or large elevation differences within a neighborhood \( q \) of the pixel \( p \). This minimization approximation was realized by aggregating \( S(p, Z) \) of path wise costs into a cost cube similar to the semi-global matching method proposed in (Hirschmuller 2008), as follow:

\[
S(p, Z) = \sum_r L_r(p, Z) \quad (4)
\]

where \( L_r(p, Z) \) in Equation (5) represents the cost of pixel \( p \) with elevation \( Z \) along one direction \( r \) on the base image, which is described by following equation:

\[
L_r(p, Z) = C(p, Z) + \min_r L_r(p-r, Z) + \Delta Z + \lambda_1,
\]

\[
\min_r L_r(p-r, Z + \Delta Z) + \lambda_1,
\]

\[
\min_l L_r(p-r, i \times \Delta Z) + \lambda_2)
\]

\[
- \min L_r(p-r, i \times \Delta Z)
\]

This regularization term function favors the planar and sloped surfaces but still allows larger elevation to jump.
in the direction of cost aggregation. After aggregating the cost cube on the base image, the elevation at each pixel was selected as the elevation value corresponding to the minimum cost of the cost cube.

3.4. Propagation through pyramid images

Given the large size of images, a coarse-to-fine propagation based on an image pyramid is a common strategy used in processing remote sensing images (Zhang and Gruen 2006). Similarly, we employed a coarse-to-fine method for a fast matching and reliable propagation. Each pyramid level was generated by multiplying a Gaussian kernel, which reduces the resolution by a factor of two. The pyramid level number is a predefined value that is either inputted by the user or can be determined based on the given elevation range covered by the images. For all the experiments in this paper, the pyramid level was set manually.

The match processing starts with an initial matching at the top level of the pyramid image. At the start matching level, the possible elevation value is set between the Max ($Z_{\text{max}}$) and the minimum elevation values ($Z_{\text{min}}$). However, as the image resolution reduces, the interval of the elevation value $dZ$ will increase accordingly. Consequently, the number of the possible elevation value is reduced, which will reduce the cost value and optimization time cost and save the memory cost. At each pyramid level, except for the top level of the image pyramid, an intermediate elevation image is obtained based on a match processing. This intermediate elevation image is used to provide the approximations in the subsequent lower level of the pyramid levels. After the initial matching, the Max and the minimum elevation values for each pixel on the base image will be determined by the intermediate elevation image. Then, a narrow search range is determined around the approximations. Take Figure 4 as an example, for each pixel in the intermediate level of pyramid image, the height value $Z_i$ is obtained from the last level of pyramid image, then the value of $Z_{\text{max}}$ and $Z_{\text{min}}$ will be obtained by $Z_i \pm s \cdot dZ$, respectively ($s$ is empirically set to 12 in this paper). Thus, while the matching procedure is propagating through the image pyramid, the elevation image computed from the higher level of the image pyramid is successively refined at the lower level. Finally, a dense and accurate elevation value is obtained.

However, owing to the problems caused by occlusion, repetitive patterns, texture-less area, and so on, multiple images are also integrated for cost calculation and semi-global optimization methods are used to select the elevation value. Consequently, a number of blunders may exist. Therefore, an automatic blunder detection procedure was carried out at each pyramid level to remove several possible wrong matches and provide a reliable intermediate elevation image to guide the subsequent match processing. For blunder detection, another image was selected as base image by changing the role of one match image and the original defined base image. In each level of the image pyramid, the elevation image was calculated for the alternative selected base images using the proposed method. A consistency check was then employed to eliminate several possible wrong matches. Figure 5 shows that for a generated 3D point $A$, its projection on the two selected base images $b_0$ and $b_1$ was $(x_0, x_1)$. For a generated 3D point $A'$, its projection on the two selected base images $b_0$ and $b_1$ was $(x'_0, x'_1)$. If $x_0$ and $x'_0$ is in the same pixel, $x_1$ should be very near to $x'_1$. Considering the orientation parameter error, the distance $d$ between $x_1$ and $x'_1$ should be less than a given threshold (one pixel in all the experiments in the paper). Otherwise, the matching result will be inconsistent and has to be removed from the result.

After the eliminating process, a number of holes existed in the intermediate elevation images, so that an interpolation procedure was employed to interpolate the elevation value from the nearby reliable elevation value. For balancing time cost and reliability, the interpolation method based on an inverse distance to a power was used. For each value to be interpolated, the neighborhood was detected first. Then, the weighted elevation value was determined, and the weight was calculated according to the distance from the current point to the selected neighbor point.

![Figure 4. Illustration of dynamically determined height range.](image-url)

![Figure 5. Illustration of the filtering process based on consistency checking.](image-url)
4. Experimental results and analysis

Two sets of experiments were carried out using images with different texture to evaluate the proposed method. The first experimental set was designed to evaluate the contribution of multiple images to the result. Aerial remote sensing images with related airborne laser scanning (ALS) data as ground truth data were used and the accuracy was evaluated. The second experiment employed a set of UltraCam-X (UCX) airborne remote sensing images and a set of nadir airborne remote sensing images from an aerial oblique system with typical textures and scenes for further experiments. For a comparison, three methods were applied, such as the result from the proposed geometric-constrained multi-view image matching method (called “gcSGM”) and other two methods, of which one is the PMVS (the PMVS2 downloaded from http://www.di.ens.fr/pmvs/ was used) and the other is based on a local method using the NCC calculated from multiple images as cost function and a winner-take-all (WTA) strategy for selecting the best elevation value.

Results were compared to reference data based on the DSM derived from laser scanning data or check points manually measured to evaluate the accuracy of the DSM generated using the proposed method quantitatively. The generation result was compared to the reference data. The elevation (vertical) root-mean-squared error (RMSE) value and the Max value were computed. The elevation (vertical) RMSE statistic can be computed using the following equation:

\[
\text{RMSE}_z = \sqrt{\frac{\sum_{j=1}^{m} (Z_{\text{DSM},j} - Z_{\text{REF},j})^2}{n}}
\]

where \(Z_{\text{DSM},j}\) is the elevation value derived using the proposed method, \(Z_{\text{REF},j}\) is the elevation value of jth check points from the reference data, and \(n\) is the number employed for RMSE computation. A low RMSE indicates a high quality of the generated DSM.

4.1. Experiment on image numbers

A series of digital aerial remote sensing images with accurate IO and EO parameters were downloaded from the ISPRS webpage (http://www.itc.nl/ISPRS_WGIII4/tests_datasets.html, accessed on 4 January 2016) for experimental analysis. A sub area with image size of 1200 x 1200 pixels was cropped from the image, as illustrated in Figure 6. In addition, the ALS data associated with these images is provided, which was used as the reference data. For evaluation of the effect of image number on matching result, images were gradually added into the proposed matching process. In Figure 6, Images 1 and 2 (indicated by 1–2 in the following tables and figures) were used to carry out the proposed method. Images 1, 2, and 3 (indicated by 1–3) were used to carry out the experiment. The proposed method was performed on Images 1, 2, 3, and 4 (indicated by 1–4).

After the proposed matching method is performed on the experimental images, 3D points were obtained. The 3D points were interpolated to DSM for visually and quantitatively comparison. DSM is illustrated as a surface map (Figure 7(a), (c), and (e)) and a shaded-relief map (Figure 7(b), (d), and (f)). By analysis of Figure 7(a), (c), and (e), the feature is sharper when more images are employed for matching, especially for the gable roof in Ellipses 1, 2, and 3. According to the corresponding shaded-relief map in Figure 7(b), (d), and (f), the DSM is visually better when more images are added for matching.

For the quantitative evaluation of the image matching results, the associated ALS data were used to calculate the RMSE and the Max elevation difference. The detailed comparative results are listed in Table 1.

According to Table 1, when all images are used in calculating the cost values and seeking the right elevation value, the RMSE is 0.479 m. This value is significantly less than that acquired when only two images are used for matching. When all images are involved for matching, the Max number of successful matches is obtained. The Max difference is reduced when all the four images are used for matching. The RMSE is 0.525 m when only two images are employed for matching, and 700,992 points are obtained. After the third image is added, the RMSE was reduced to 0.512 m, and 702,392 points were obtained. This trend reveals that the increase of the image number is helpful to improve the matching accuracy and success ratio.

4.2. Results with UCX airborne remote sensing images

Unlike the airborne remote sensing images with low buildings used in the above experiment, a set of UCX airborne remote sensing images were used for experimental
analysis, as shown in Figure 8. The images contain several tall buildings. These images were oriented by digital photogrammetric workstation VirtuoZo via bundle adjustment with ground control points, so that accurate IO and EO parameters were obtained. For comparison, the PMVS and the "NCC–WTA" methods were performed on these images. Owing to the limitation of the computer RAM, the images were downsampled for four times, with the size of 2355 × 3608 pixels. For a fair comparison, the same target image and area was selected for PMVS, and the other parameters were used as default as suggested by the authors of PMVS.

After point clouds are produced from the three methods directly, they were interpolated to DSM. As shown in Figure 9, the surface maps are in the left column and the shaded-relief maps are in the right column. To illustrate the result with more details, we marked six areas in Figure 9, named UCX-Area 1, UCX-Area 2, UCX-Area 3, UCX-Area 4, UCX-Area 5, and UCX-Area 6, and then we zoomed in the UCX-Areas 3, 4, 5, and 6 (as shown in Figure 10). Based on the final DSM results in Figure 9, the PMVS and the proposed method (gcSGM) produce smoother DSM, while the DSM derived from the proposed method has more detailed features. For example, as shown in UCX-Area 3 in Figure 9(b) and UCX-Area 5 in Figure 9(c), the proposed method can produce the model of vegetation clearly. Comparing UCX-Area 4 in Figure 9(b) with UCX-Area 6 in Figure 9(c), two separated buildings are linked together after interpolation for the 3D points derived from the PMVS method. On the contrary, a number of gross errors exist in the DSM using the NCC method with a WTA strategy, as shown in UCX-Areas 1 and 2 of Figure 9(a). As shown in Figure 10, the proposed method can obtain the best DSM result.

For a quantitative evaluation, 196 evenly distributed check points were manually measured on VirtuoZo to evaluate the DSMs generated using the three methods. The RMSE and the Max values were calculated, and the results are listed in Table 2. This table shows that...
values are relative high for this set of experiment images, which indicates that the resolution of image influences the accuracy of the final DSM result.

4.3. Results with oblique air-borne remote sensing images

Four images, as shown in Figure 11, were used for further experimental analysis. The images are a subset of images captured using a five-camera aerial oblique system. All the employed images were obtained from a nadir camera. The images include some contents related to buildings. For a comparison, the PMVS and the NCC
The method can produce DSM with more sharp features. Comparing Ob-Area 2 in Figure 12(d) with Ob-Area 4 in Figure 12(f), the proposed method obtains more detailed information on the area with vegetation. As shown in Figure 13, the proposed method can obtain better DSM results.

After matching processing is, respectively, performed with the three methods, the obtained 3D point cloud data were interpolated to produce DSM. As shown in Figure 12, the left column presents the surface maps and the right corresponds to the shaded-relief maps. To illustrate the result with more details, we marked four areas in Figure 12, named Ob-Area 1, Ob-Area 2, Ob-Area 3, and Ob-Area 4, and then we zoomed in them (as shown in Figure 13). Based on the DSM, the models produced by the PMVS and the proposed method are smoother. According to Ob-Area 1 in Figure 12(d) and the corresponding Ob-Areas 3 in Figure 12(f), the proposed method can produce DSM with more sharp features. Comparing Ob-Area 2 in Figure 12(d) with Ob-Area 4 in Figure 12(f), the proposed method obtains more detailed information on the area with vegetation. As shown in Figure 13, the proposed method can obtain better DSM results.

No control point was found for orienting the oblique remote sensing image data sets to an absolute world coordinate system, thus 161 evenly-distributed check points were manually selected and alternatively used to evaluate the accuracy of the disparity in image space. The disparity map was produced by the back projection of the DSM map to the two selected base images. The RMSE and the Max of the disparity difference are listed in Table 3.

According to Table 3, the proposed method obtained the maximum number of 3D point. The RMSE value of the disparity difference on the check points was 2.484.
pixels, which is less than the results produced by the PMVS and “NCC–WTA” methods. For the NCC method with WTA, a number of gross errors were found in the DSM, and its RMSE and the Max of disparity value are relative larger than the other two methods. Less 3D point cloud data were produced by the PMVS. However, its DSM is smooth, except for a small portion of defect. In addition, less detail was produced by this method owing to the small number of 3D points. Results show that the proposed method performs better than the other two methods.

5. Conclusions

This paper presented a multi-view geometric-constrained matching method for the DSM generation based on a semi-global optimization, which can match multiple images at the same time. Experimental analyses using multiple airborne remote sensing images with typical textures convey that multiple images are helpful to improve the quality of stereo matching. The developed census transformation can be used to measure the similarity of multiple images simultaneously. Using an energy framework based on the stereo matching that considers smooth assumption and discontinuity, the proposed method can automatically match images with densely populated high buildings suffering with severe surface discontinuity problems. Furthermore, the proposed method can produce prominent DSM result with sharp features.

In the near future, we will employ multiple satellite images for the implementation of the proposed matching in an object space, and the proposed method will be optimized using parallel computing for fast process and large number of images. Another future work is to employ an image warp processing during the matching propagation through the pyramids of image to simultaneously match images from large different views.

Funding

This work was supported by the National Key Research and Development Program of China [grant number 2017YFC0803802]; and the National Natural Science Foundation of China [grant number 41771486].

Notes on contributors

Wenhao Zhao is a PhD candidate in School of Geodesy and Geomatics, Wuhan University. His research interests include image matching and DEM generalization.

Li Yan is a professor in School of Geodesy and Geomatics, Wuhan University. He has authored or co-authored more than 100 scientific articles in journals, books, and conference proceedings. His work focuses on mapping with satellite images, aerial oblique images processing, and autonomous positioning and navigation.

Yunsheng Zhang is an associate professor in Photogrammetry and Remote Sensing at the School of Geosciences and Info-Physics, Central South University, Changsha, China. He has authored or co-authored more than 30 scientific articles in journals, books, and conference proceedings. His work focuses on UAV, aerial oblique image and Lidar data processing, change detection.

ORCID

Yunsheng Zhang http://orcid.org/0000-0002-2779-2015

References

Alobeid, A., K. Jacobsen, and C. Heipke. 2010. “Comparison of Matching Algorithms for DSM Generation in Urban Areas from Ikonos Imagery.” Photogrammetric Engineering & Remote Sensing 76 (9): 1041–1050.

Bignone, F. 2003. “Processing of Stereo Scanner: From Stereo Plotter to Pixel Factory.” Paper presented at the 49th Photogrammetric Week, Stuttgart, Germany, September 1–5.

Boblick, A., and S. Intille. 1998. “Large Occlusion Stereo.” International Journal of Computer Vision 33 (3): 181–200.

Boykov, Y., O. Veksler, and R. Zabih. 2001. “Fast Approximate Energy Minimization via Graph Cuts.” IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (11): 1222–1239.

Bulatov, D., P. Wernerus, and C. Heipke. 2011. “Multi-View Dense Matching Supported by Triangular Meshes.” ISPRS Journal of Photogrammetry and Remote Sensing 66 (6): 907–918.

EVAL. 2017. Accessed October 22, 2017. http://vision.middlebury.edu/stereo/eval/

Felzenszwalb, P., and D. Huttenlocher. 2006. “Efficient Belief Propagation for Early Vision.” International Journal of Computer Vision 70 (1): 41–54.

Furukawa, Y., and J. Ponce. 2010. “Accurate, Dense, and Robust Multiview Stereopsis.” IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (8): 1362–1376.

Gruen, A. 1985. “Adaptive Least Squares Correlation: A Powerful Image Match.” South African Journal of Photogrammetry, Remote Sensing and Cartography 14 (3): 175–187.

Gruen, A. 2012. “Development and Status of Image Matching in Photogrammetry.” The Photogrammetric Record 27 (137): 36–57.

Gruen, A., and E. Baltsavias. 1988. “Geometrically Constrained Multiphoto Matching.” Photogrammetric Engineering & Remote Sensing 54 (5): 633–641.

Hartley, R., and A. Zisserman. 2004. Multiple View Geometry in Computer Vision. New York: Cambrige University Press.

Helava, U. V. 1978. “Digital Correlation in Photogrammetric Instruments.” Photogrammetria 34: 19–41.

Hernandez, C., P. H. S. Torr, and R. Cipolla. 2007. “Multiview Stereo via Volumetric Graph-Cuts and Occlusion Robust Photo-Consistency.” IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (12): 2241–2246.

Hirschmüller, H. 2001. “Improvements in Real-Time Correlation-Based Stereo Vision.” Paper presented at the IEEE Workshop on Stereo and Multi-Baseline Vision, Kauai, Hawaii, USA, December 9–10, 141–148.
Hirschmuller, H. 2008. "Stereo Processing by Semiglobal Matching and Mutual Information." IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2): 328–341.

Hirschmuller, H., and D. Scharstein. 2009. "Evaluation of Stereo Matching Costs on Images with Radiometric Differences." IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (9): 1582–1599.

Kanade, T., and M. O. Okutomi. 1994. "A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment." IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (9): 920–932.

Knapitsch, A., J. Park, Q. Zhou, and V. Kolton. 2017. “Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction.” ACM Transactions on Graphics (TOG) 36 (4). doi: 10.1145/3072959.3073599

Leberl, F., A. Irschara, T. Pock, P. Meixner, M. Gruber, S. Scholz, and A. Wiechert. 2010. "Point Clouds: Lidar versus 3D Vision." Photogrammetric Engineering & Remote Sensing 76 (10): 1123–1134.

Lee, H., T. Kim, W. Park, and H. Lee. 2003. "Extraction of Digital Elevation Models from Satellite Stereo Images through Stereo Matching Based on Epipolarity and Scene Geometry." Image and Vision Computing 21 (9): 789–796.

Lemaire, C. 2008. “Aspects of the DSM Production with High Resolution Images.” International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 37(B): 1143–1146.

Lhuillier, M., and L. Long Quan. 2002. "Match Propagation for Image-Based Modeling and Rendering." IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (8): 1140–1146.

Martina, L., and G. Christian. 2012. “Accuracy Assessment of Digital Surface Models Based on Worldview-2 and ADS80 Stereo Remote Sensing Data.” Sensors 12 (5): 6347–6368.

Match-T DSM. 2017. Accessed October 27, 2017. https://geospatial.trimble.com/products-and-solutions/inpho

Otto, G., and T. Chau. 1989. "Region-Growing' Algorithm for Matching of Terrain Images." Image and Vision Computing 7 (2): 83–94.

Radhika, V., B. Kartikeyan, B. Krishna, S. Chowdhury, and P. Srivastava. 2007. "Robust Stereo Image Matching for Spaceborne Imagery." IEEE Transactions on Geoscience and Remote Sensing 45 (9): 2993–3000.

Scharstein, D., and R. Szeliski. 2002. "A Taxonomy and Approaches for Computing Visual Correspondence." International Journal of Computer Vision 47 (1–3): 7–42.

Zabih, R., and J. Woodfill. 1994. "Non-Parametric Local Transforms for Computing Visual Correspondence." Paper presented at the European Conference of Computer Vision, Stockholm, Sweden, May 2–6.

Zededin, L., A. Klaus, B. Gruber-Geymayer, and K. Karner. 2006. "Towards 3D Map Generation from Digital Aerial Images." ISPRS Journal of Photogrammetry and Remote Sensing 60 (6): 413–427.

Zhang, L., and A. Gruen. 2006. “Multi-Image Matching for DSM Generation from IKONOS Imagery.” ISPRS Journal of Photogrammetry and Remote Sensing 60 (3): 195–211.

Zhang, B., S. Miller, S. Walker, and K. Devencena. 2007. "Next Generation Automatic Terrain Extraction Using Microsoft Ultracam Imagery." Paper presented at the ASPRS 2007 Annual Conference, Tampa, FL, USA, May 7–11.

Zhang, Z., J. Zhang, M. Liao, and L. Zhang. 2000. “Automatic Registration of Multisource Imagery Based on Global Image Matching.” Photogrammetric Engineering & Remote Sensing 66 (5): 625–629.

Zhu, Q., B. Wu, and Y. Tian. 2007. "Propagation Strategies for Stereo Image Matching Based on the Dynamic Triangle Constraint." ISPRS Journal of Photogrammetry and Remote Sensing 62 (4): 295–308.

Zhu, Q., Y. Zhang, B. Wu, and Y. Zhang. 2010. “Multiple Close-Range Image Matching Based on a Self-Adaptive Triangle Constraint.” The Photogrammetric Record 25 (132): 437–453.

Zhu, Q., J. Zhao, H. Lin, and J. Y. Gong. 2005. "Triangulation of Well-Defined Points as a Constraint for Reliable Image Matching.” Photogrammetric Engineering & Remote Sensing 71 (9): 1063–1069.