A Triangulation-based Hierarchical Image Matching Method for Wide-Baseline Images

Bo Wu, Yunsheng Zhang, and Qing Zhu

Abstract

This paper presents a triangulation-based hierarchical image matching method for wide-baseline images. The method includes the following three steps: (a) image orientation by incorporating the SIFT algorithm with the RANSAC approach, (b) feature matching based on the self-adaptive triangle constraint, which includes point-to-point matching and subsequent point-to-area matching, and (c) triangulation constrained dense matching based on the previous matched results. Two new constraints, the triangulation-based disparity constraint and triangulation-based gradient orientation constraint, are developed to alleviate the matching ambiguity for wide-baseline images. A triangulation based affine-adaptive cross-correlation is developed to help find correct matches even in the image regions with large perspective distortions. Experiments using Mars ground wide-baseline images and terrestrial wide-baseline images revealed that the proposed method is capable of generating reliable and dense matching results for terrain mapping and surface reconstruction from the wide-baseline images.

Introduction

Most methods for terrain mapping or surface reconstruction from ground stereo vision are based on hard-baseline (or fixed-baseline) stereo imaging systems, in which stereo cameras are mounted on a rigid camera bar with a stereo base generally from several centimeters to half a meter (Li *et al.*, 2004 and 2007; Di and Li, 2007; Wu, 2006). For hard-baseline stereo vision, the image matching is relatively easy. This is because the disparities and local pattern distortions on the stereo images are relatively small, which allows for limiting the matching search area using simple constraints such as the epipolar geometry and obtaining matching results using the standard similarity functions such as the Normalized Cross-Correlation (NCC) method (Helava, 1978). However, hard-baseline stereo vision can only be used to map the nearby terrain surrounding the imaging systems (for example, up to 100 m for a stereo imaging system with a hard baseline of 30 cm) due to the fact that the depth estimation errors from stereo vision are directly proportional to the square of the distance from the camera to the targets and inversely proportional to the baseline (Wu, 2006; Olson and Abi-Rached, 2009). However, mapping distant terrain (several hundred meters from the camera) is sometimes required, for example, mapping tasks in remote and unreachable environments such as deserts, polar areas, and areas after natural disasters. In these cases, wide-baseline stereo vision is employed. For wide-baseline stereo vision, the imaging system takes images of the same scene from different locations using the same camera to form a wide baseline (a few meters to dozens of meters), which enables mapping of distant targets as far as a few hundred meters from the cameras with an acceptable accuracy (Olson *et al.*, 2003; Olson and Abi-Rached, 2005 and 2009; Di and Li, 2007).

The wide-baseline stereo vision improves the accuracy of the depth for distant terrain; however, it introduces significant difficulties for stereo matching between the stereo images. Wide-baseline stereo vision is difficult for two reasons. First, there is poor knowledge regarding the relative position and orientation information between the images taken at the two ends of the baseline, compared to the situation of the hard baseline case. Second, the change in perspective makes stereo matching difficult since the image textures of the same objects change significantly in the two images due to different viewpoints. Figure 1 shows examples of wide-baseline images from a wide-baseline mapping task in Duck Bay area of Victoria Crater on Mars. The images were taken by the Opportunity rover of the Mars Exploration Rover (MER) 2003 mission and were downloaded from NASA's Planetary Data System (PDS). The wide baseline for the images is about 5.5 m. Figure 1a shows the sand dunes close to the center of Victoria Crater. The mapping area is about 300 m from the camera location. Repetitive and homogeneous textures are commonly seen in the sand dune areas. Figure 1b shows the crater wall, which is about 60 m from the camera location. Distinct occlusions and surface discontinuities can be found in the image pairs. Figure 1c shows the slope area on the crater wall, which is about 20 m from the camera location. As the observed distance is relatively short, the image textures change significantly due to the different viewpoints.

Bo Wu is with the Department of Land Surveying & Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (lsbowu@polyu.edu.hk).

Yunsheng Zhang is with the State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, P.R. China, and formerly at the Hong Kong Polytechnic University.

Qing Zhu is with the State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, P.R. China.

Photogrammetric Engineering & Remote Sensing
Vol. 77, No. 7, July 2011, pp. 695–708.

0099-1112/11/7707–0695/$3.00/0

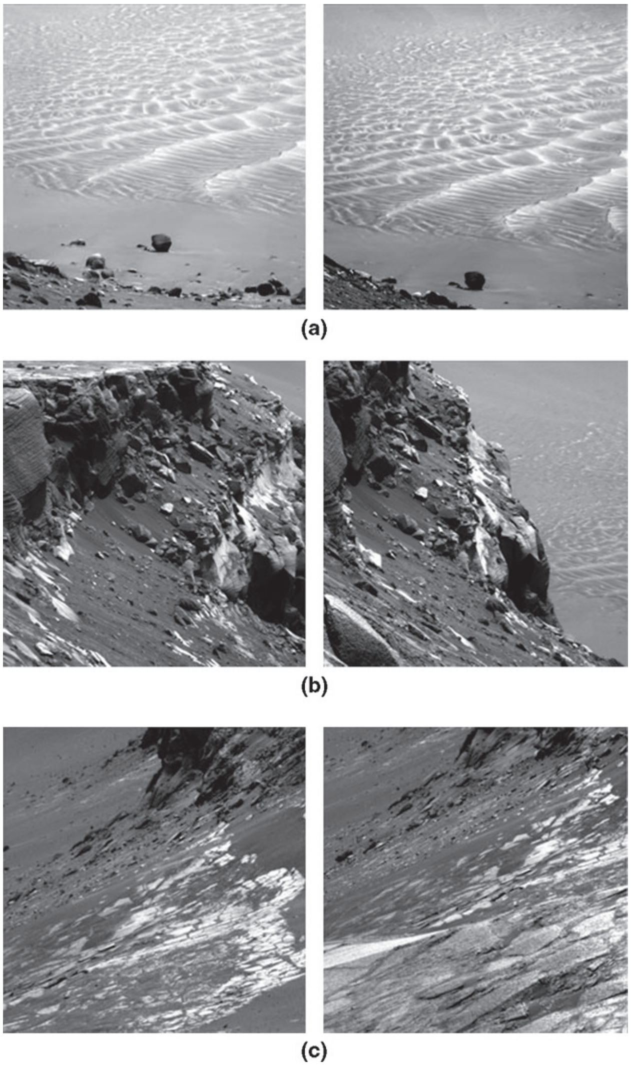

Figure 1. Stereo wide-baseline images at Duck Bay area of Victoria Crater on Mars downloaded from PDS: (a) data set 1, (b) data set 2, and (c) data set 3. In each dataset, the left image was acquired at the first camera location and the right image at the second location at a distance of about 5.5 m.

Wide-baseline stereo vision is not commonly used in photogrammetry and computer vision for terrain mapping and surface reconstruction due to the above difficulties. Olson *et al.* (2003) presented an endeavor for this task. They proposed a framework to process the Mars ground wide-baseline images to generate density disparity maps. The method depends on the initial orientation parameters of the images and uses very sparse feature-matching results to determine the motion of the images. The matching is based on a maximum likelihood function. Zhu *et al.* (2005) presented a new image matching method based on a self-adaptive triangle constraint, which was used for stereo aerial image matching (Wu, 2006; Zhu *et al.*, 2007a) and close-range terrestrial image matching with a hard baseline and it proved able to produce reliable matching results. However, for wide-baseline images, it is much more difficult to obtain reliable matching results due to the difficulties addressed above. Therefore, this paper presents a triangulation-based hierarchical image matching method to improve the reliability and automation of wide-baseline image matching based on the previous work (Wu, 2006; Zhu *et al.*, 2005 and 2007a).

After providing a literature review on how to improve the reliability of wide-baseline image matching, a triangulation-based hierarchical image matching method for wide-baseline images is presented in detail. Three data sets of wide-baseline images (Figure 1) collected on Mars are employed for experimental analysis. Two pairs of terrestrial wide-baseline images and the associated lidar point cloud data are used for quantitative evaluation of the developed method. Finally, concluding remarks are presented and discussed.

Related Work

Image matching is a challenging and often ill-posed problem, especially for wide-baseline images. During the past decades, a few endeavours have been devoted in the field of photogrammetry and computer vision to improve the reliability, automation, and efficiency of wide-baseline image matching which can be generally categorized into two classes based on the matching primitives. One is feature-based matching and the other is area-based matching.

1. *Feature-based Matching*: The most well-known method for image matching between images with large perspective or scale changes is the scale invariant feature transform (SIFT) method (Lowe, 1999 and 2004). SIFT combines a scale invariant interest point detector and a descriptor based on the gradient distribution in the detected local regions. The interest points are detected based on local 3D extrema in the scale-space pyramid built with difference-of-Gaussian (DOG) filters, which is invariant over a wider set of transformations, especially scale change (Lowe, 1999). In the SIFT descriptor, each interest point is characterized by a vector with 128 unsigned eight-bit numbers generated from a local region, which defines the multi-scale gradient orientation histogram. The similarity is measured by comparing the two vectors associated with the two matching points (Lowe, 2004). The SIFT descriptor provides robustness against errors caused by orientation issues and small geometric distortions. However, it can only detect blob-like interest points (Mikolajczk and Schmid, 2004) and produce relative sparse matching results (Zhu *et al.*, 2007b). Lingua (2009) analyzed the SIFT method for photogrammetric applications and determined that SIFT is a good method for automatic tie point extraction and coarse DSM (Digital Surface Model) generation. After the Wenchuan earthquake in China on 12 May 2008, many aerial images were collected without regular flight tracks and camera orientations. The relative orientation cannot be precisely carried out using traditional methods under these circumstances. The SIFT method was adopted for down-sampled images to obtain coarse but robust relative orientation parameters, which was very important for further processing (Zhang *et al.*, 2009). In addition to SIFT, other scale and affine invariant feature detection and matching methods have been presented in the past (Tuytelaars and Van Gool, 2000; Matas *et al.*, 2004; Mikolajczyk and Schmid, 2004 and 2005). Automatic image orientation becomes possible even with close range images under disorder (Snavely *et al.*, 2008). However, the matching results from these methods are relative sparse and cannot provide sufficient correspondence for detailed terrain mapping.
2. *Area-based Matching*: Area-based matching usually works directly on local image windows, and it can acquire dense correspondences (Lhuillie and Quan, 2002). However, for wide-baseline images, the conventional area-based image matching cannot obtain accurate and dense matching results due to the image perspective distortions, for example, caused by a wide baseline. Megyesi and Chetverikov (2004) presented an affine matching method for wide-baseline

images that accounts for local affine distortion and propagates the best matching affine parameters on each surface until a surface discontinuity is reached. Kannala and Brandt (2007) reported a matching propagation strategy to obtain quasi-dense matching results from wide-baseline images. It depends on an intensity moment to adapt the current estimates of the local affine transformation. This iterative estimation is very time consuming. Olson and Abi-Rached (2005 and 2009) used a maximum likelihood method to obtain a density map from terrain images. This method is based on time-consuming global optimization. Strecha *et al.* (2003) presented a multi-view wide-baseline stereo system for the reconstruction of precise 3D models. Based on a few sparse set of initial depth estimates, an algorithm was developed to propagate these initial depth estimates by an inhomogeneous time diffusion process, which is guided by a properly weighted matching energy that takes into account the matching to all views. Tola *et al.* (2008 and 2010) developed a method entitled DAISY for wide-baseline image matching, which is similar to the SIFT method. It has to calculate a high dimensional descriptor for every pixel and is thus time consuming.

Feature-based image matching obtains robust but sparse matching results, while area-based matching can obtain dense matching results but the matching reliability may depend on the texture conditions of the images. Therefore, this paper presents a triangulation-based hierarchical image matching method for wide-baseline images that incorporates the merits of both the feature-based matching and the area-based matching methods and produces reliable and dense matching results with high efficiency and automation. The details are described in the following sections.

Triangulation-based Hierarchical Image Matching for Wide-Baseline Images

The triangulation-based hierarchical image matching method firstly employs a SIFT algorithm and the RANSAC approach to obtain a few robust correspondences on the wide-baseline images, and then calculates the relative orientation parameters of the images. The robust correspondences obtained are then used to generate initial Delaunay triangulations. Then, interest point matching is carried out based on the self-adaptive triangle constraints (Wu, 2006; Zhu *et al.*, 2005 and 2007a). The interest point matching starts by detecting interest points using a Harris-Laplace detector (Zhu *et al.*, 2007b) within a pair of triangles in the initial triangulations, then matches these interest points under the triangle constraint, and obtains a pair of corresponding points with maximum reliability. After that, the method inserts the newly matched corresponding points into the triangulations and updates the triangulations dynamically, then handles the next pair of triangles and repeats the same process until the termination conditions (the triangles are small enough or cannot match successfully for at least one pair of points) of the matching propagation are met. Because the most distinctive point is always successfully matched first, the dynamic updating of triangulations is just the process of self-adaptive matching propagation. This local geometry constraint of triangles can adapt to the changes in image texture automatically and will finally produce more reliable matching results (Wu, 2006; Zhu *et al.*, 2005 and 2007a). After interest point matching, the same process is repeated to match the remaining interest points in one image with all the pixels in another image. Then, dense matching is performed based on the triangulations generated from the previous matched points and finally matching results are obtained. This hierarchical image matching strategy incorporates the merits of both the feature-based matching, and the area-based matching and has the capability of generating reliable and dense matching results efficiently. The framework of the method is illustrated in Figure 2.

Based on the previous work on self-adaptive triangle-constrained image matching (Wu, 2006; Zhu *et al.*, 2005 and 2007a), this paper highlights the following three innovations particularly developed for the purpose of matching wide-baseline images: (a) robust image orientation enabling wide-baseline image matching without knowing any prior information such as the image orientation parameters,
(b) triangulation-based disparity constraint and triangulation-based gradient orientation constraint, which help alleviate the matching ambiguity, and (c) triangulation-based affine-adaptive cross-correlation (TAACC), which enables finding correct matches even in the image regions with large perspective distortions. The following sections will describe the details of the method.

Image Orientation

Since the wide-baseline images are taken at different locations, the illumination environment may change and the contrast may be different for the image pairs. Therefore, an image preprocessing (image enhancement) is performed before the image orientation. The Wallis filter (Pratt, 1991) is employed to enhance and sharpen the texture patterns and increase the signal-to-noise ratio. After enhancement, the texture details are enriched in both low- and high-level contrast regions which will be helpful for the subsequent processing.

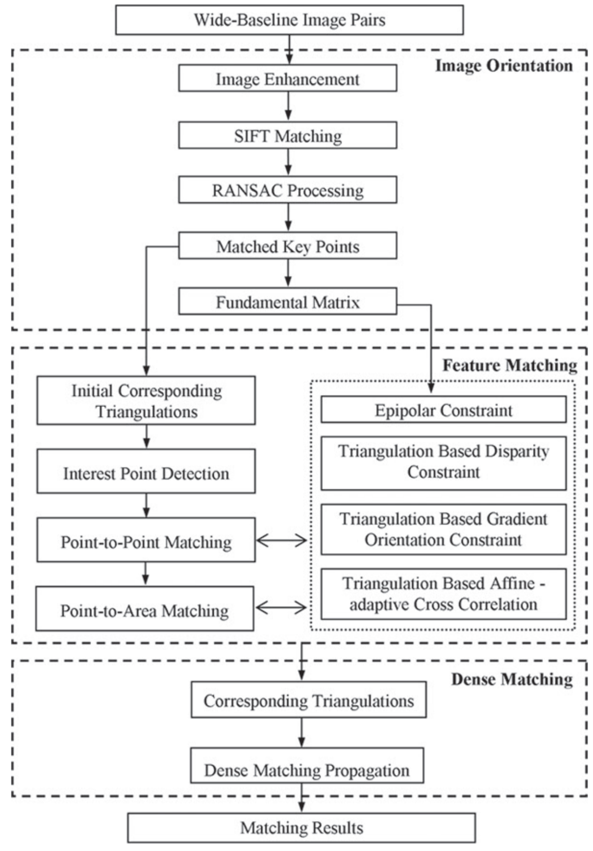
Wide-Baseline Image Pairs
Image Orientation
Image Enhancement
SIFT Matching
RANSAC Processing
Matched Key Points
Fundamental Matrix
Feature Matching
Initial Corresponding Triangulations
Interest Point Detection
Point-to-Point Matching
Point-to-Area Matching
Epipolar Constraint
Triangulation Based Disparity Constraint
Triangulation Based Gradient Orientation Constraint
Triangulation Based Affine - adaptive Cross Correlation
Dense Matching
Corresponding Triangulations
Dense Matching Propagation
Matching Results

Figure 2. Framework of the triangulation-based hierarchical image matching method.

Image orientation has been studied thoroughly in photogrammetry and computer vision (Zhang, 1998; Stewénius *et al.*, 2006). It is the process to solve the relative relation or orientation between two images of the same scene from which the epipolar geometry can be determined. This allows matching adjacent points in multiple images without knowing anything about the position of the camera. The fundamental matrix, describing the relation between two overlapping images, is a compact way of representing the epipolar geometry of the two images. In this paper, image orientation for wide-baseline images is achieved using the matched key points generated by the SIFT algorithm (Lowe, 1999 and 2004). A RANSAC approach (Fischler and Bolles, 1981) is used to remove the possible mismatches. A least squares algorithm (Hartley and Zisserman, 2003) is employed to calculate the fundamental matrix.

SIFT Matching

As mentioned previously, the SIFT algorithm (Lowe, 1999 and 2004) is a good solution to identify corresponding points on images with large perspective or scale changes. Therefore, the SIFT algorithm is used to find some key corresponding points on the wide-baseline images. Figure 3a, 3c, and 3e show the SIFT matching results for the experimental data sets illustrated in Figure 1 in which the white lines indicate the matched point pairs between the image pairs. From the matching results, there are some "messy"lines that may indicate mismatches. These mismatches must be removed for the calculation of an accurate fundamental matrix. The RANSAC approach as described in the following section is used for this purpose.

RANSAC Approach

RANSAC (Fischler and Bolles, 1981) is an algorithm used to derive a usable model from a set of data that contains both inliers and outliers. The RANSAC algorithm starts by randomly selecting an instance of the data, which the algorithm supposes is a part of the inliers. A model is then built assuming that the chosen instance is an exact instance of the model. This model is then used to determine how much of the remaining data fits the model by determining whether each instance of the data fits reasonably well to the model, i.e., seeing if it is an inlier. This is used as a criterion to determine the best model which has the largest number of inliers. This process is repeated for every instance of the data to find the overall best model. This model is then recalculated using all of its inliers, instead of just using the single data instance, to produce a more accurate model.

In this case the outliers come from the possible mismatches from the SIFT matching results. RANSAC is adopted to exclude the outliers (Hartley and Zisserman, 2003). At first, seven samples from the SIFT-matching results are selected randomly, and a fundamental matrix is then

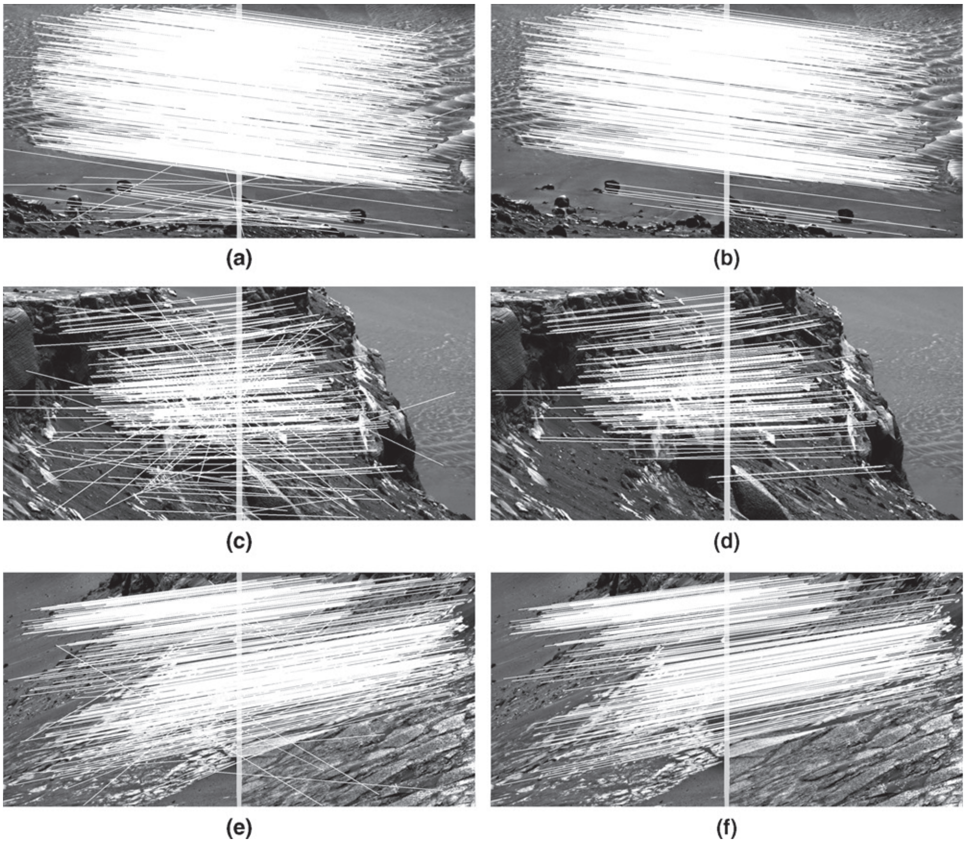

Figure 3. SIFT matching and outlier removal using the RANSAC approach on the wide-baseline images: (a) SIFT matching results for data set 1, (b) results after outlier removal for data set 1, (c) SIFT matching results for data set 2, (d) results after outlier removal for data set 2, (e) SIFT matching results for data set 3, and (f) results after outlier removal for data set 3.

calculated using these samples. After that, the number of inliers is calculated that are consistent with the previously calculated fundamental matrix. The algorithm then repeats the previous step and chooses the fundamental matrix with the largest number of inliers and the outliers are finally removed. Figure 3b, 3d, and 3f show the results after RANSAC processing, in which the mismatches are effectively removed.

Calculation of the Fundamental Matrix

Because the SIFT algorithm mainly detects blob-like features (Mikolajczk and Schmid, 2004), there may be shifts existing between the blob centers and their accurate corresponding point locations. Therefore, a least squares process (Gruen, 1985) is employed here to refine the SIFT matching results. After that, the fundamental matrix is calculated by using a least squares algorithm (Hartley and Zisserman, 2003) using the successfully matched key points.

To evaluate the accuracy of the image orientation results, all the matched key points are divided randomly into two groups. One group is used as control points to calculate the fundamental matrix, and the other group is used as checkpoints to calculate the residual r according to the following equation (Hartley and Zisserman, 2003):

$$r = \frac{\sum_{i=1}^{N}\sqrt{d(x_i, Fx_i')^2 + d(x_i', Fx_i)^2}}{N} \qquad (1)$$

where x_i,x_i' are a pair of matched points, and N is the number of matched points used for checkpoints; F is the fundamental matrix calculated from the control points; F' is the transpose of F, $d(x_i, Fx_i')$, and $d(x_i', Fx_i)$ give the distance from the point to its corresponding epipolar line determined by the fundamental matrix.

Table 1 displays the image orientation results for all three data sets, and shows that the residuals are about 0.2 pixels. This proves the good performance of the proposed image orientation method.

TABLE 1. IMAGE ORIENTATION RESULTS FOR THE THREE DATA SETS

Data Set	Number of Control Points	Number of Check Points	Residual (pixel)
Data Set 1	532	532	0.206
Data Set 2	112	112	0.156
Data Set 3	199	200	0.131

Triangulation-based Feature Matching

The triangulation-based feature matching includes two steps. The first step is a point-to-point matching that only matches the interest points detected in both images. The second step is a point-to-area matching based on the previous point-to-point matching results which uses the remaining interest points in one image and searches for their correspondence using all the pixels in another image.

Triangulation-based Point-to-Point Matching

Before the point-to-point matching, interesting points need to be detected. A Harris-Laplace detector (Mikolajczk and Schmid, 2004; Zhu *et al.*, 2007b) is used here instead of using the SIFT method. This is because SIFT mainly detects blob-like points (Mikolajczk and Schmid, 2004) while the significant points such as corners and highly textured points may not be able to be successfully detected, and this disadvantage is critical to the subsequent terrain mapping and reconstruction. The Harris-Laplace detector (Mikolajczk and Schmid, 2004; Zhu *et al.*, 2007b) responds to corners and highly textured points. It can detect interest points also invariant to scale, computing a multi-scale representation for the Harris interest point detector and then selecting points at which a local measure (the Laplacian) is maximal over scales.

In addition to the epipolar constraint determined by the previous image orientation process and the adaptive triangle constraint, this paper investigates the following new constraints and strategies particularly designed for wide-baseline image matching.

(1) Triangulation-based Disparity Constraint

After image orientation processing, there are a few matched key points generated from the SIFT method. These matched key points are then used to generate a pair of initial Delaunay triangulations that serve as constraints for the subsequent image matching. Each triangle in the triangulations is considered as a local smooth area, and the disparity of the corresponding points within the triangle area should have some relationships with the disparity of the triangle vertex (Pollard *et al.,* 1986; Zhu *et al.*, 2005).

In Figure 4, triangle *abc* and *a′b′c′* are a pair of corresponding triangles in the left (Figure 4a) and right image (Figure 4b), respectively; *a* and *a′* are a pair of triangle vertices, and *f* is a point close to *a* ; *f* and *f′* are a pair corresponding points. Under the assumption of local smooth constraint, the disparity gradient of *a* and *f* should satisfy the following equation:

$$|\rho_f - \rho_a| \leq rDist_{fa} \qquad (2)$$

where ρ_f and ρ_a are the disparity of f and a, respectively, $Dist_{fa}$ is the distance from f to a; r is a predefined value within the range of (0.66, 1) which is given by Pollard *et al.* (1986).

In other words, if f and its corresponding point f' meet the disparity gradient constraint, f' should be located in a circle with the center of $f + \rho_a$ and the radius of $rDist_{fa}$.

For wide-baseline images taken on the ground, the disparities in different image regions change differently with respect to the distance from the targets to the camera locations. The farther from the camera, the smaller the disparity, which results in the upper part of the image generally having a smaller disparity than the bottom part. Figure 5 shows the disparity changes for data set 3. A line drawn from top to bottom in Figure 5a indicates a profile from far to close regions. The corresponding profile is identified in Figure 5b, from which increasing parallax can be found from top to bottom.

Based on the above analysis, the r value in Equation 2 has been reconsidered to be adaptive to the wide-baseline images. The following heuristic values are suggested based

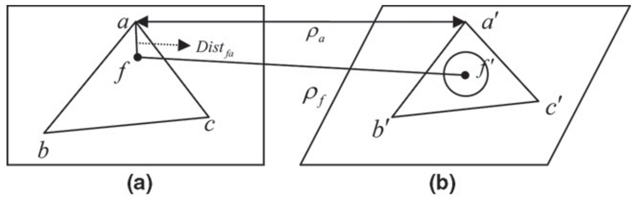

Figure 4. Disparity gradient: (a) left image, and (b) right image.

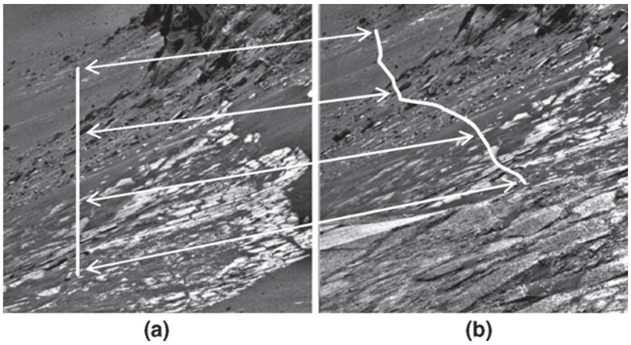

Figure 5. Disparity changes from top to bottom on data set 3: (a) left image, and (b) right image.

on the experiments using images with moderate wide baseline of a few meters:

$$r = \begin{cases} 0.66 & \text{if } (y < 0.33 \times \textit{height}) \\ 0.75 & \text{if } (0.33 \times \textit{height} \le y < 0.66 \times \textit{height}) \\ 1 & \text{if } (y \ge 0.66 \times \textit{height}) \end{cases} \qquad (3)$$

where y is the y-coordinate of the detected interest point in the image space and *height* means the image height.

(2) Triangulation-based Gradient Orientation Constraint
Another new constraint, triangulation-based gradient orientation constraint, is used in this method that is based on an assumption that the gradient orientation between correspondences should be similar to each other within a local area in the image pair. Figure 6 shows an enlarged region on data set 3 where the corresponding triangulations generated from the matched corresponding points are shown on the image. Each pair of corresponding vertexes in the triangulations has an index number, and their gradient orientations are marked with arrows starting from the vertexes. From Figure 6, similar gradient orientations can be found for all the corresponding vertexes in the left (Figure 6a) image and the right (Figure 6b) image.

For each pair of corresponding triangles, the differentiations of the gradient orientations for the three vertexes can be calculated and a median differentiation can be derived. Figure 7 shows the statistics of the median differentiation of the gradient orientation for all the corresponding triangles in the initial corresponding triangulations for data set 3 generated from the matched points in image orientation. As shown in Figure 7, the majority of the median differentiations are close to zero, which means the majority of the gradient orientations for the corresponding triangle vertexes are consistent. Only a very few are different, which may be related to significant image texture changes in a local region or possible mismatches.

Therefore, the median differentiation of gradient orientations for each pair of corresponding triangles is used to help find correct matches within this pair of corresponding triangles. In the image matching process, the variance of the median differentiation δ for all the corresponding triangles in the initial corresponding triangulations is calculated, and 3δ is used as a threshold in the subsequent image matching for the gradient orientation constraint.

(3) Triangulation-based Affine-adaptive Cross Correlation (TAACC)
In the original self-adaptive triangle constrained image matching method (Wu, 2006; Zhu *et al.*, 2005 and 2007a), an NCC (Normalized Cross-Correlation) method (Helava, 1978; Lhuillie and Quan, 2002) was used to measure the similarity of the corresponding points. However, the traditional NCC is based on rectangular correlation windows and is not invariant to rotation and scale changes. Therefore, this paper developed a triangulation-based affine-adaptive cross correlation (TAACC) for similarity measurement in which the correlation windows of the matching points are warped before the calculation of their similarity according to the affine transformation parameters propagated from the surrounding triangles.

As illustrated in Figure 8, for each pair of corresponding triangles Δabc and $\Delta a'b'c'$, the affine transformation parameters are calculated by using itself and its adjacent three triangles Δeac,Δagb,Δbfc. Six pairs of corresponding points *a-a', b-b', c-c', e-e', f-f'*, and *g-g'* are used to calculate the affine transformation parameters of the triangle pair Δabc and $\Delta a'b'c'$. Assuming *a", b"*, and *c"* are the locations calculated using the derived affine transformation parameters from *a, b,* and *c*. The residuals d_a, d_b, and d_c can be calculated by comparing the distances between *a"-a', b"-b', c"-c'*. The mean value of the residuals $\varepsilon_d = (d_a + d_b + d_c)/3$ for the three vertexes can be calculated. If the mean residual ε_d is less than a predefined threshold (e.g., three pixels), the affine distortion is assumed to be able to be compensated by the calculated affine transformation parameters. Then, a warp process for the correlation windows is performed to compensate the rotation and scale changes. For efficient calculating, only the four corners of the correlation window in the searched image are calculated using the affine

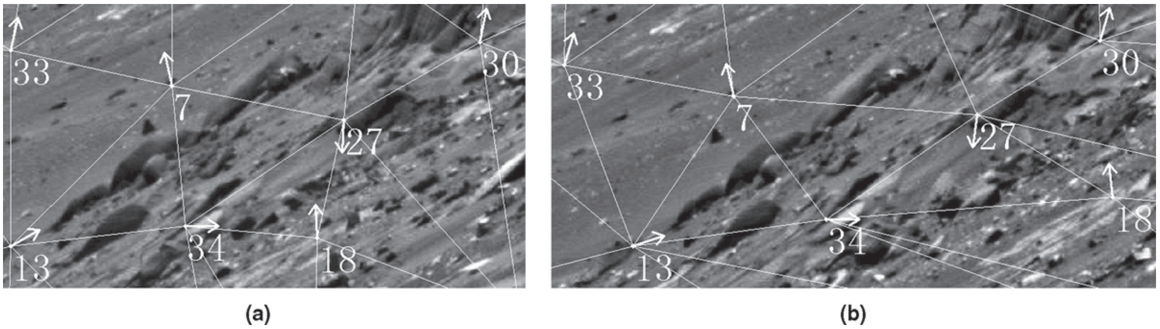

Figure 6. Gradient orientation of the corresponding vertex for data set 3: (a) left image, and (b) right image.

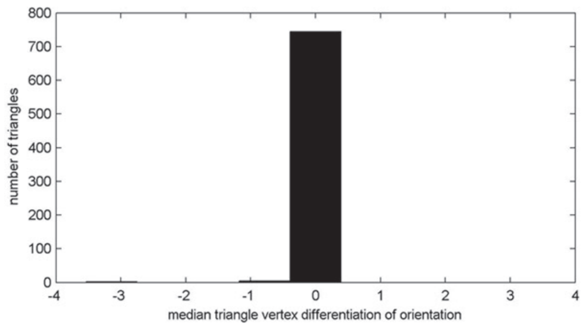

Figure 7. Histogram of median differentiation of the gradient orientation for all the corresponding triangles in the initial corresponding triangulations for data set 3.

transformation parameters. The image coordinates of the rest of the pixels in the correlation window are interpolated by a bilinear interpolation method; their gray values are interpolated from the search images again using the bilinear interpolation method.

Figure 9 shows the detailed results of TAACC. In Figure 9, the matched points are marked using crosses. The small rectangle in Figure 9b is the correlation window on the left image for point 1. If there is no such a process of using TAACC, the correlation window for its corresponding point 1′ on the right image (Figure 9c) will be the same with the one on the left image (Figure 9b). After using TAACC, a distorted rectangle is calculated from the affine transformation parameters as illustrated in Figure 9c. For the corresponding points 1 and 1′, the correlation value without considering the affine transformation is 0.85. After warping the correlation windows using the triangulation-based affine transformation parameters, the correlation value increases to 0.94. Another example is shown in Figure 9d and 9e. The small rectangle in Figure 9d is the correlation window for point 2, and the distorted rectangle in Figure 9e is the correlation window of its corresponding point 2′, which is calculated from the affine transformation parameters. For the corresponding points 2 and 2′, the correlation value without considering the affine transformation is 0.72. After warping the correlation windows using the triangulation-based affine transformation parameters, the correlation value increases to 0.95.

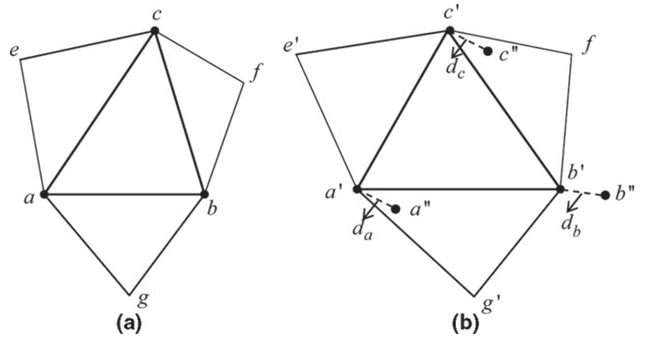

Figure 8. Affine distortion estimation based on corresponding triangulations: (a) left triangulation, and (b) right triangulation.

However, sometimes the affine distortion cannot be compensated by the affine transformation parameters as described above, i.e., the mean residual ε_d is larger than a predefined threshold. In this case, the correlation window in the searched image is determined using an alternative method similar to the method presented by Megyesi and Chetverikov (2004) and Xu *et al.* (2009) which is an iterative process to scale and rotate the correlation window step by step until the correlation values stop increasing. This method provides more accurate correlation windows and is ideal for the local regions with significant affine distortions; however, it is extremely time consuming. Therefore, this paper uses a combination of the triangulation-based affine estimation and the iterative affine estimation to improve both the matching reliability and efficiency. According to the experiments using the three data sets, a majority of the affine distortions can be estimated using the triangulation-based method and only 2 percent, 12.4 percent, and 12.0 percent of the feature matching are performed using the iterative method for data sets 1, 2, and 3, respectively.

The actual image matching propagation takes place as follows. Assuming the left image is the reference image and the right image is the searched image, the algorithm chooses one interest point with maximum interest strength (Zhu *et al.,* 2007a) in a selected triangle on the reference image. After that, potential correspondences are obtained under the triangle constraint, triangle-based disparity constraint, triangle-based gradient orientation constraint, and epipolar constraint. If there is no corresponding point found under the constraints, turn to the next interest point. If multiple potential corresponding points are obtained, triangulation-based affine estimation or iterative affine estimation is used to calculate the affine transformation parameters. The matching scores are obtained by warping the correlation window in the reference images using the calculated affine parameters. The corresponding point with the highest matching score is chosen as the matching hypothesis for the left-to-right matching. And then, a right-to-left matching with the same process is carried out. If the matching result from the left-to-right image is consistent with the matching from the right-to-left image, the algorithm accepts this matching result. Otherwise, the algorithm ignores the matching result and turns to the next interest point. The same process is repeated and newly matched corresponding points are inserted in the corresponding triangulations until the termination conditions of the propagation are met. Details about the implementation of the self-adaptive triangle-constrained image matching method can be found in Wu (2006) and Zhu *et al.* (2005 and 2007a).

Plate 1 shows the point-to-point matching results for data set 1, 2, and 3. As seen in Plate 1, the matched corresponding points are visually accurate but sparse. Disparity maps interpolated from the matched points are used to evaluate the performance of the image matching as illustrated in Plate 1b, 1d, and 1f. Bright values represent larger disparities (discard the black background). Significant brightness changes within a local region in the disparity maps may indicate possible mismatches except those situations where there are stand-alone objects existing in the image, such as the rocks close to the bottom in Plate 1a. The disparity maps are relatively smooth but apparently they do not provide enough details.

Triangulation-Based Point-to-Area Matching

For wide-baseline images, the repeatability rate of the interest points (Zhu *et al.*, 2007b) detected in the stereo pairs may be low due to the large perspective changes. This will result in

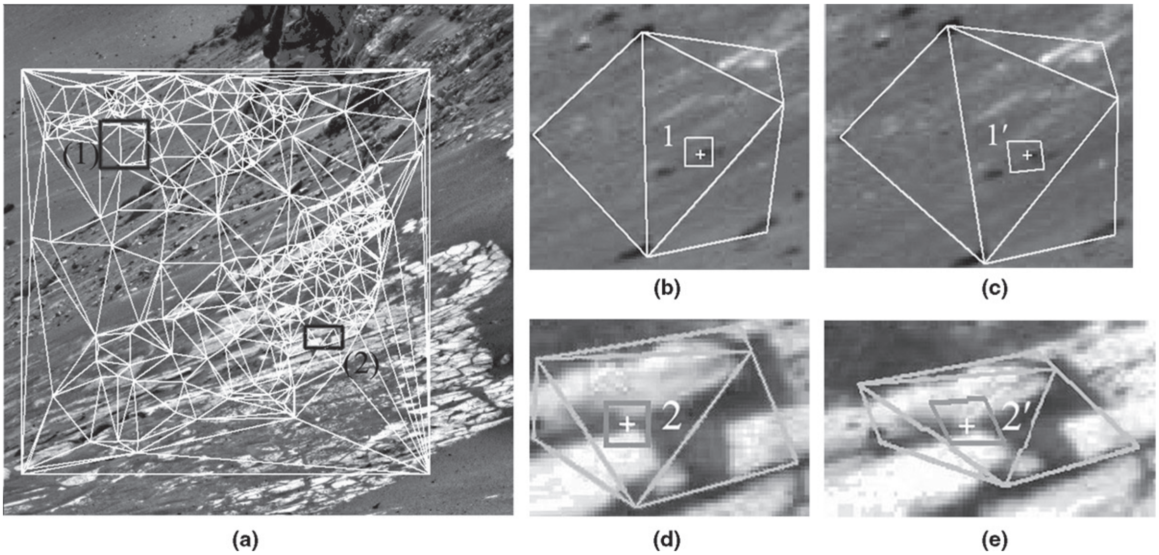

Figure 9. Results of the triangulation-based affine-adaptive cross correlation on data set 3: (a) a reference image with labeled study areas, (b) left image for the study area 1, (c) right image for the study area 1, (d) left image for the study area 2, and (e) right image for the study area 2.

sparse matching results if only matching the detected interest points in the stereo images. Therefore, this paper presents a point-to-area matching based on the previous point-to-point matching results. The point-to-area matching uses the remaining interest points (after the point-to-point matching) in one image and searches for their correspondence in all the pixels in a local area in another image.

In this case, the corresponding triangulations generated from the previous point-to-point matching results are denser than the initial corresponding triangulations for point-to-point matching, and they provide stronger constraints for point-to-area matching. The matching propagation is similar to the previous point-to-point matching. For an interest point in a triangle on the reference image (e.g., the left image), the algorithm searches for matching candidates in all the pixels along the epipolar line inside the corresponding triangle on the searched image (e.g., the right image). Again, the triangle constraint, triangle-based disparity constraint, triangle-based gradient orientation constraint, and epipolar constraint are employed to help find correct matches. The triangulation-based affine estimation or the iterative affine estimation is used to define the correlation window on the searched image. For all the pixels that satisfy the previously mentioned constraints, their matching scores are calculated based on the warped correlation window and then a matching-score curve is obtained. If the highest matching score is larger than a predefined threshold (e.g., 0.8) and the ratio of the highest matching score to the second highest matching score is larger than a predefined threshold (e.g., 1.25), the correspondence is considered to be a matching hypothesis. And if the hypothesis passes the right-to-left consistency check, the correspondence is accepted as a correct matching. Then, the newly matched points are inserted into the triangulations. The matching propagation will be terminated after all the detected interest points in the reference image are examined.

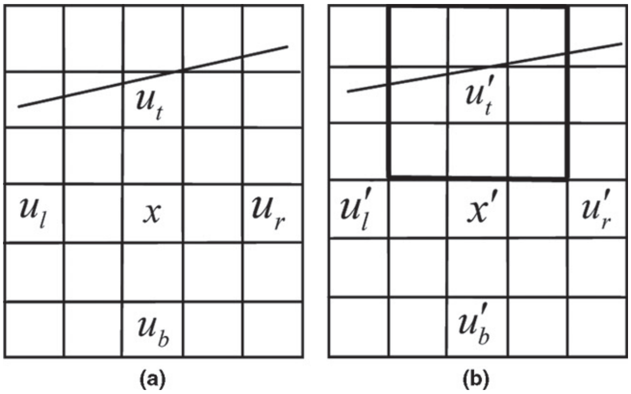

Figure 10. Dense matching propagation: (a) neighborhood of *x* in the left image, and (b) neighborhood of *x′* in the right image.

Plate 2 shows the point-to-area matching results for data set 1, 2, and 3. As can be seen from Plate 2, the matched points are much denser than the previous point-to-point matching. The disparity maps are relatively smooth and more details can be found.

As shown in Plate 2, there are still some regions without correspondences after the point-to-area matching. Therefore, a triangulation-based dense matching is presented to make further dense matches based on the previous matching results.

Triangulation-based Dense Matching

The triangulation-based dense matching employs an affine matching propagation strategy inspired by the popular dense matching methods (Otto and Chau, 1989; Lhuillier and Quan, 2002; Megyesi and Chetverikov, 2004), which is a "the best the first" strategy. In this method, "the best" has two types of meaning, in which one is to select the most reliable seeds to guide subsequent matching and

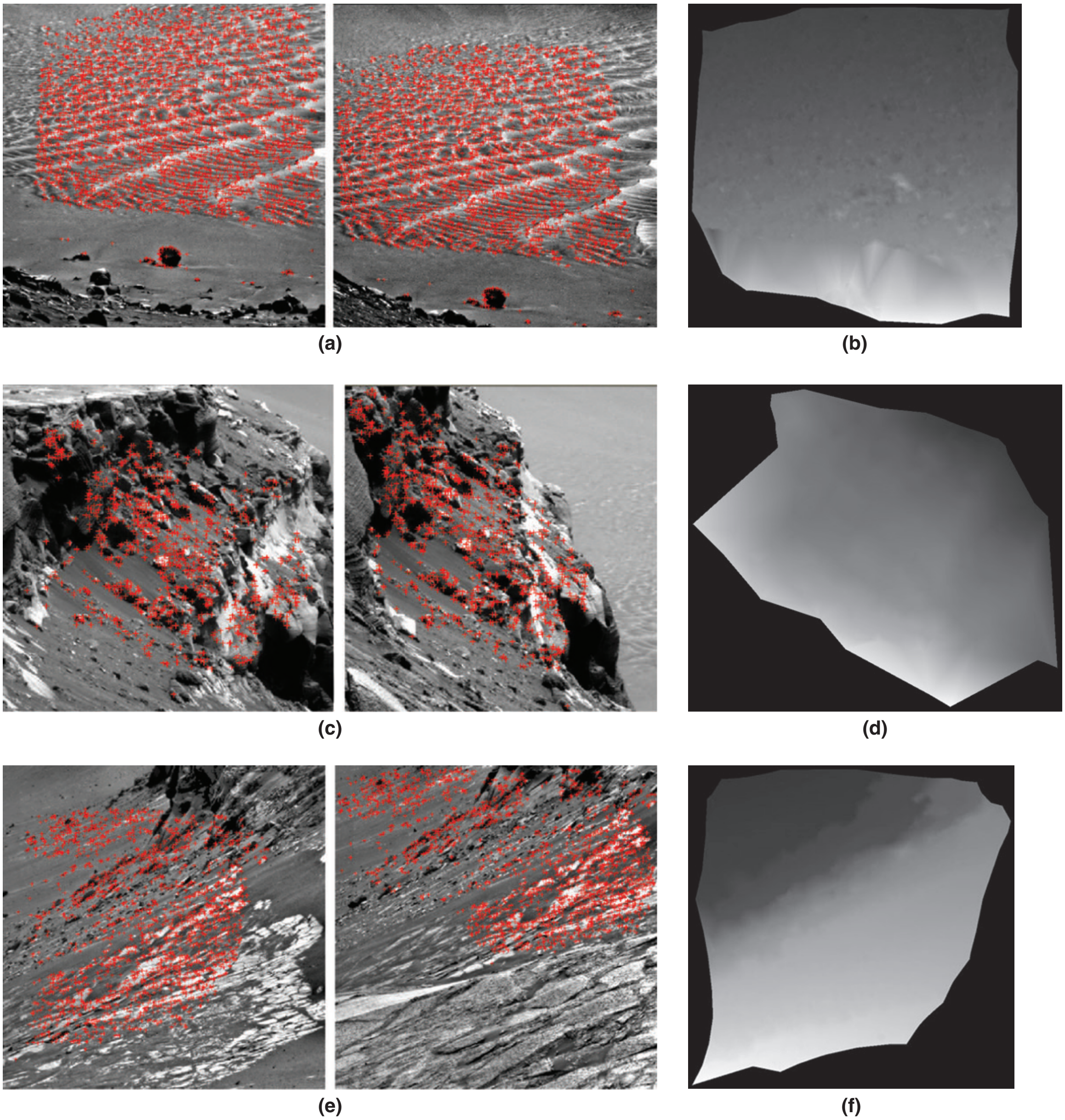

Plate 1. Point-to-point matching results. (a) Matched interest points for data set 1, (b) disparity map for data set 1, (c) matched interest points for data set 2, (d) disparity map for data set 2, (e) matched interest points for data set 3, and (f) disparity map for data set 3.

another is to select the easiest part in the image to match first and then propagate the matching to the relatively hard parts.

The matching results (triangle vertexes) from the previous feature matching are used as seeds in the dense matching. The matching starts from the top part of the image and propagates to the bottom part. This is because the disparity is generally smaller for the top part and larger for the bottom part for wide-baseline images, which means matching is relatively easier for the top part compared to the bottom part. In the actual process, the matching starts from the image top and propagates to the bottom row by row. For each row in the image, the seed point with maximum matching score as recorded in the previous matching process is selected first for matching and propagates the matching to its neighborhood. The seeds are stored in a seed list, which is a heap data structure enabling fast selecting and incremental adding of new matched points as seeds.

Figure 10 shows the dense matching propagation process. In Figure 10, x and x' are a pair of seed points. The four-connected neighbor pixels of x in the left image are u_t,u_r,u_b,u_l, and the matching candidates of the four neighbor pixels in the right image can be forecasted by the parallax of the seed as marked by u'_t,u'_r,u'_b,u'_l. The four neighborhood pixels u_t,u_r,u_b,u_l are first checked to see whether already matched or not. If not, the corresponding point of u is searched in a 3 × 3 window around u'. The correlation window is warped using the affine

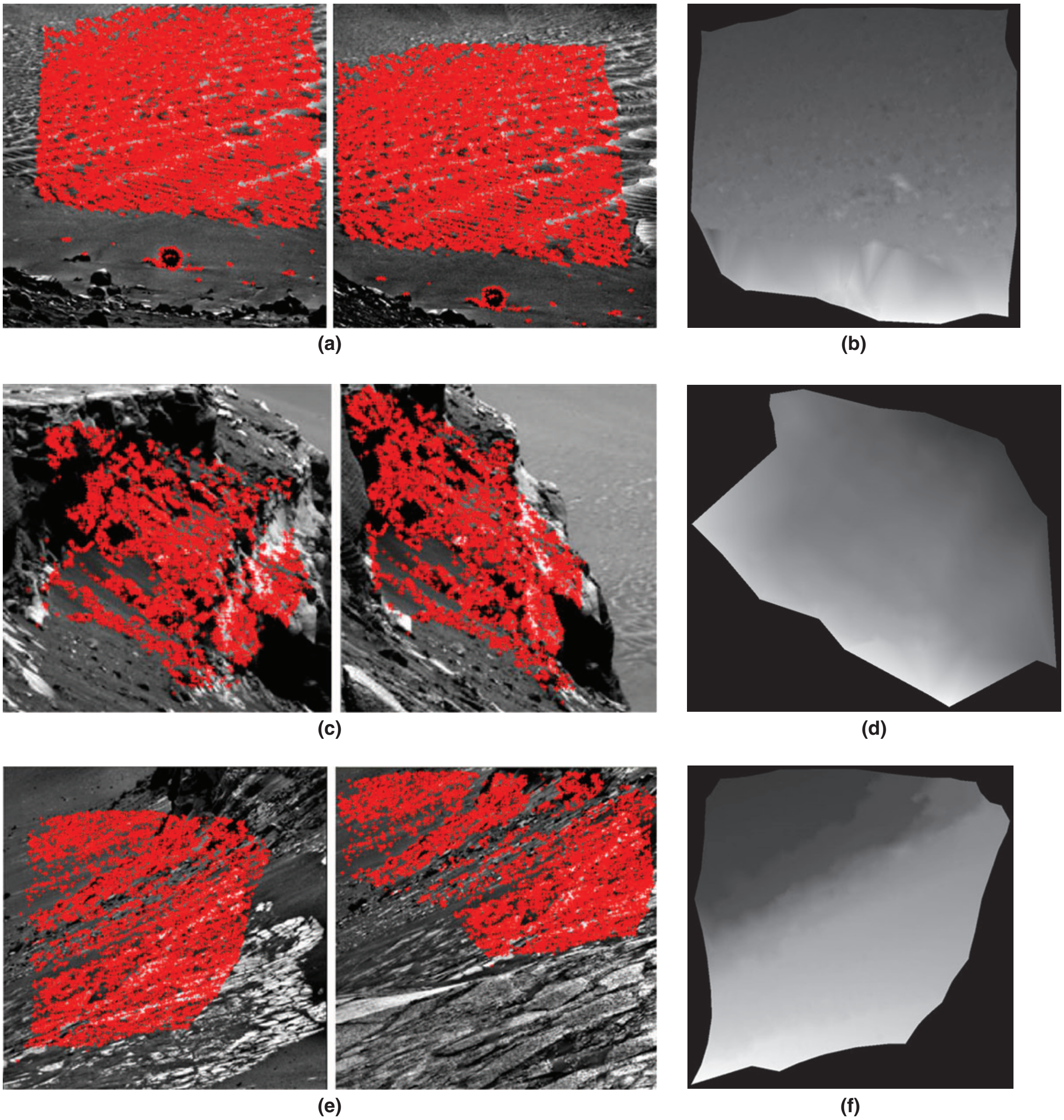

Plate 2. Point-to-area matching results: (a) matched points for data set 1, (b) disparity map for data set 1, (c) matched points for data set 2, (d) disparity map for data set 2, (e) matched points for data set 3, and (f) disparity map for data set 3.

transformation parameters calculated from the triangle surrounding the current point. If the matching score is larger than a predefined threshold (e.g., 0.8), then the matching is accepted. The newly matched points are then inserted into the seed list. If all the four neighbor pixels of the seed have been processed, the current seed will be removed from the seed list. After all the seeds in the seed list are processed, the matching propagation is terminated.

Plate 3 shows the dense matching results for data sets 1, 2, and 3. As can be seen from Plate 3, more points are matched and more details can be found in the disparity maps. The disparity maps also show smooth brightness, which proves the good performance of the proposed method. Also, it can be noted from Plate 3 that there are very few small regions in the images that do not have successfully matched points at the end. This is mainly due to the following two reasons. First, there is insufficient texture in those regions to correctly match the pixels. Second, those regions that are visible in one image may not be visible in the other image due to occlusion problems. It should be noted that it is very difficult even for human eyes to find correct matches in those regions.

Intensive Experimental Analysis

To intensively evaluate the developed method, two standard data sets: Herz-Jesu-P8 and Fountain-p11

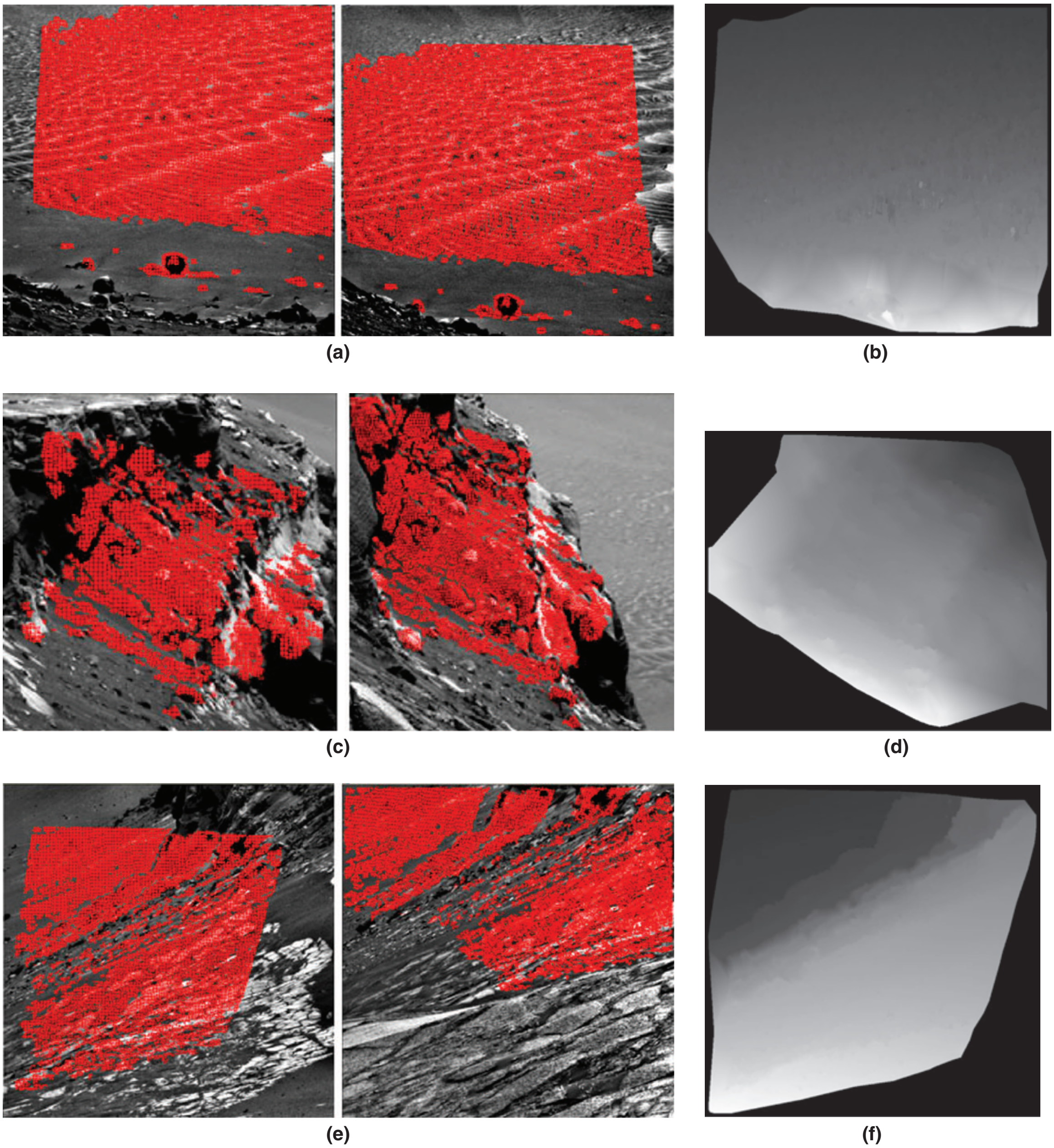

Plate 3. Dense matching results: (a) matched points for data set 1, (b) disparity map for data set 1, (c) matched points for data set 2, (d) disparity map for data set 2, (e) matched points for data set 3, and (f) disparity map for data set 3.

were downloaded from the Computer Vision Laboratory of EPFL in Switzerland (http://cvlab.epfl.ch/~strecha/multiview/denseMVS.html). Each of the data set includes a pairs of terrestrial wide-baseline images and the associated lidar point cloud data. The lidar data are well aligned with the images and it will be used as ground truth to evaluate the performance of the image matching. Figure 11 shows the Herz-Jesu-P8 and the Fountain-11 data sets. The interior orientation (IO) and exterior orientation (EO) parameters of the stereo images are known. The radial distortions for the images have been corrected. The Herz-Jesu-P8 data set has a wide-baseline of 8.47 m, and the distance from the camera to the scene is about 14 m. The Fountain-p11 data set has a wide-baseline of 6.92 m, and the distance from the camera to the scene is about 8.2 m.

Image matching was performed to process the wide-baseline images in the Herz-Jesu-P8 and the Fountain-p11 data sets. 3D points were derived based on the matching results using the associated IO and EO parameters. To compare the 3D points derived from the wide-baseline images with the lidar data, the lidar points were firstly back-projected onto the image pairs using the IO and EO parameters of the images, and only those lidar points whose back-projected points overlapped with the matched points from the wide-baseline images were selected for further comparison. Then, the RMSE (root mean squared error) and

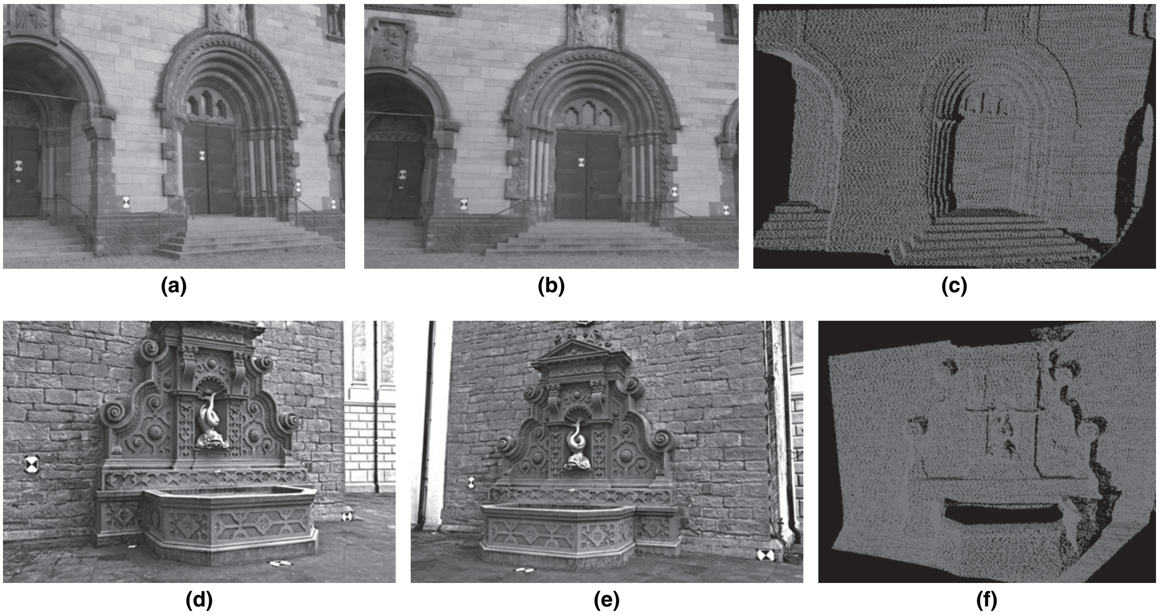

Figure 11. The Herz-Jesu-P8 and the Fountain-p11 data sets: (a) left image of the Herz-Jesu-P8, (b) right image of the Herz-Jesu-P8, (c) 3D view of the associated lidar data for Herz-Jesu-P8, (d) left image of the Fountain-p11, (e) right image of the Fountain-p11, and (f) 3D view of the associated lidar data for Fountain-p11.

maximum value of the difference between the 3D points derived from the wide-baseline images and the lidar points were calculated as indicators of the performance of image matching. The image matching method presented in this paper (TAACC) was compared with other two image matching methods for the two experimental data sets. One is DAISY (Tola *et al.*, 2008 and 2010), which is a method for wide-baseline image matching. The other is NCC (Lhuillie and Quan, 2002), which is a traditional dense matching method. The source code for DAISY was downloaded from the author's website (http://cvlab.epfl.ch/~tola/daisy.html). Epipolar constraint and a winner-take-all matching strategy were employed to determine the corresponding points using DAISY. NCC was implemented based on the principles presented in Lhuillie and Quan (2002). The results are shown in Table 2 and Figure 12.

For the Herz-Jesu-P8 data set, there are 342,053 points derived from the wide-baseline image matching results based on TAACC. For DAISY and NCC methods, there are 338,833 and 336,341 points obtained, respectively. TAACC produces more matched points than the other two methods. The RMSE and maximum value of the differences between the 3D points derived from image matching and the lidar points is 4.2 cm and 58.7 cm for TAACC, respectively. They are better than those of DAISY and NCC method, which indicates the very good performance of the proposed TAACC method. Similar results can be found for the Fountain-p11 data set.

Figure 12 plots the percentage of correctly calculated depth against the error threshold setting to a fraction of the scene's depth range for the two data sets. For example, for the Herz-Jesu-P8 data set, about 82 percent of 3D points generated from the NCC matching results have differences of less than 0.5 percent of the scene's depth range (7 cm) compared with the lidar points. While for the DAISY and TAACC method, the percentage is 87.5 percent and 90 percent under the same condition, respectively.

In the experiments, the proposed method also behaves efficiently. The method is implemented using C++, and the processing time (from image orientation to the final dense matching results) is about three minutes for the two data sets using a 2.5GHZ CPU machine.

Table 2. Experimental Results

Data Set	Method	Matched Points	Lidar Points Used for Comparison	RMSE	Maximum
Herz-Jesu-P8	TAACC	342053	261643	4.2 cm	58.7 cm
	DAISY	338833	255648	7.1 cm	79.0 cm
	NCC	336341	253022	9.9 cm	112.1 cm
Fountain-p11	TAACC	591155	324084	6.2 cm	68.6 cm
	DAISY	513796	271537	7.5 cm	82.0 cm
	NCC	503745	294885	11.0 cm	127.2 cm

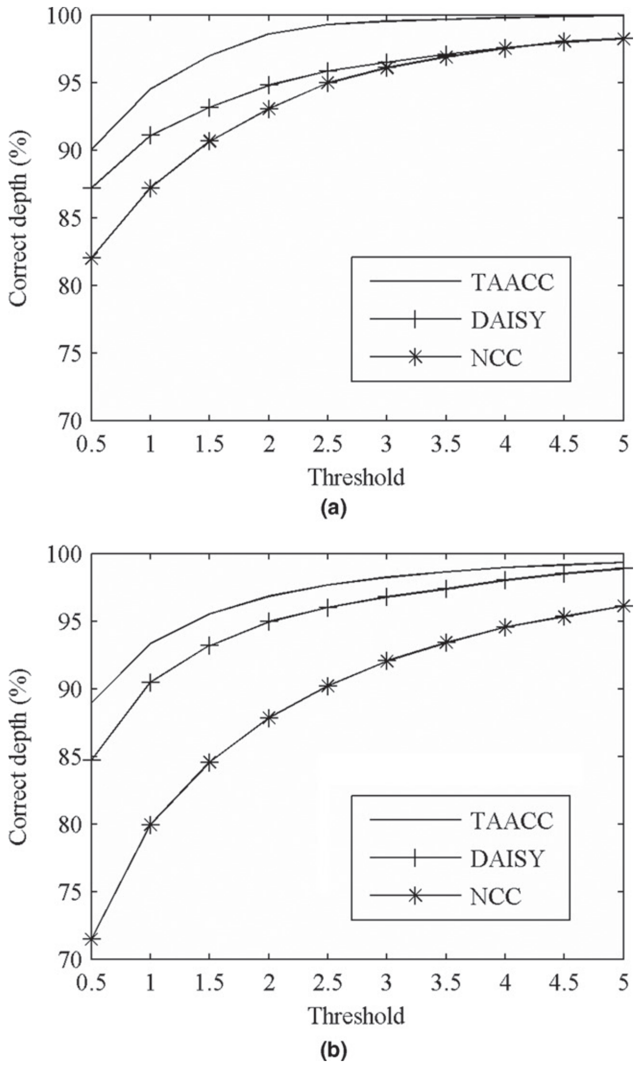

Figure 12. The correctly estimated depth against the error: (a) results for the Herz-Jesu-P8 data set, and (b) results for the Fountain-p11 data set.

Conclusions and Discussion

This paper presented a triangulation-based hierarchical image matching method for wide-baseline images. The experiment analyses using actual wide-baseline images conveyed the following conclusions:

1. The SIFT algorithm incorporated with the RANSAC approach can provide reliable but sparse correspondences, which is ideal for image orientation of the wide-baseline images. This also enables reliable wide-baseline image matching without knowing any prior information about the images.
2. The triangulation-based disparity constraint and triangulation-based gradient orientation constraint can help alleviate the matching ambiguity, particularly for wide-baseline images.
3. The triangulation-based affine-adaptive cross-correlation enables correct matches to be found on wide-baseline images even in the local regions with large perspective distortions.
4. The proposed triangulation-based hierarchical image matching strategy incorporating the merits of both feature-based matching and area-based matching with the capability of generating reliable and dense matching results efficiently is ideal for terrain mapping or surface reconstruction from wide-baseline images.

Acknowledgments

The work described in this paper was supported by a grant from the Research Grants Council of Hong Kong (Project No: PolyU 5312/10E) and a grant from the Start-up Research Fund Program of the Hong Kong Polytechnic University. The research was also supported by the National Basic Research Program of China/973 Program (Project No: 2010CB731801) and National Natural Science Foundation of China (Project No: 40871212).

References

Di, K., and R. Li, 2007. Topographic mapping capability analysis of mars exploration rover 2003 mission imagery, *Proceedings of the 5th International Symposium on Mobile Mapping Technology*, 28–31 May, Padua, Italy.

Fischler, M.A., and R.C. Bolles, 1981. Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography, *Communications of the ACM*, 24(6):381–395.

Gruen, A., 1985. Adaptive least squares correlation: A powerful image matching, *South African Journal of Photogrammetry, Remote Sensing and Cartography,* 14(3):175–187.

Hartley, R., and A. Zisserman, 2003. *Multiple View Geometry in Computer Vision*, Cambridge University Press, Cambridge, UK, 672 p.

Helava, U.V., 1978. Digital correlation in photogrammetric instruments, *Photogrammetria*, 34:19–41.

Kannala, J., and S.S. Brandt, 2007. Quasi-dense wide baseline matching using match propagation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* Minneapolis, Minnesota, pp. 1–8.

Lhuillier, M., and L. Quan, 2002. Match propagation for image-based modeling and rendering, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 24(8):1140–1146.

Li, R., K. Di, L.H. Matthies, R.E. Arvidson, W.M. Folkner, and B.A. Archinal, 2004. Rover localization and landing-site mapping technology for the 2003 Mars exploration rover mission, *Photogrammetric Engineering & Remote Sensing*, 70(1):77–90.

Li, R., K. Di, A.B. Howard, L. Matthies, J. Wang, and S. Agarwal, 2007. Rock modeling and matching for autonomous long-range mars rover localization, *Journal of Field Robotics*, 24(3):187–203.

Lingua, A., D. Marenchino, and F. Nex, 2009. Performance analysis of the SIFT operator for automatic feature extraction and matching in photogrammetric applications, *Sensors*, 9(5):3745–3766.

Lowe, D.G., 1999. Object recognition from local scale-invariant features, *Proceedings of the International Conference on Computer Vision,* Corfu, Greece, pp. 1150–1157.

Lowe, D.G., 2004. Distinctive image features from scale-invariant key points, *International Journal of Computer Vision,* 60(2):91–110.

Matas, J., O. Chum, M. Urbana, and T. Pajdlaa, 2004. Robust wide-baseline stereo from maximally stable extremal regions, *Image and Vision Computing,* 22(10):761–767.

Megyesi, Z., and D. Chetverikov, 2004. Affine propagation for surface reconstruction in wide baseline stereo, *Proceedings of the 17th International Conference on Pattern Recognition*, Cambridge, UK, pp. 76–79.

Mikolajczyk, K., and C. Schmid, 2004. Scale and affine invariant interest point detectors, *International Journal of Computer Vision,* 60(1):63–86.

Mikolajczyk, K., and C. Schmid, 2005. A performance evaluation of local descriptors, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 27(10):1615–1630.

Olson, C.F., H. Abi-Rached, M. Ye, and J.P. Hendrich, 2003. Wide-baseline stereo vision for Mars rovers, *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems,* Las Vegas, Nevada, pp. 1302–1307.

Olson, C.F., and H. Abi-Rached, 2005. Wide-baseline stereo experiments in natural terrain, *Proceedings of the 12th International Conference on Advanced Robotics*, Seattle, Washington, pp. 376–383.

Olson, C.F., and H. Abi-Rached, 2009. Wide-baseline stereo vision for terrain mapping, *Machine Vision and Applications*, doi:10.1007/s00138-009-0188-9.

Otto, G.P., and T.K. Chau, 1989. A region-growing algorithm for matching of terrain images, *Image and Vision Computing*, 7(2):83–94.

Pollard, S., J. Porrill, J. Mayhew, and J. Frisby, 1986. Disparity gradient, Lipschitz continuity, and computing binocular correspondence, *Proceedings of Robotics Research: The Third International Symposium* (O.D. Faugeras and G.Giralt, editors), Gouvieux, France, pp. 19–26.

Pratt, W.K., 1991. *Digital Image Processing*, John Wiley & Sons, New York, 503 p.

Snavely, N., S.M. Seitz, and R. Szeliski, 2008. Modeling the world from internet photo collections, *International Journal of Computer Vision,* 80(2):189–210.

Stewénius, H., C. Engels, and D. Nistér, 2006. Recent developments on direct relative orientation, *ISPRS Journal of Photogrammetry and Remote Sensing,* 60(4):284–294.

Strecha, C., T. Tuytelaars, and L. Van Gool, 2003. Dense matching of multiple wide-baseline views, *Proceedings of the IEEE Conference on Computer Vision*, Nice, France, pp.1194–1201.

Tola, E., V. Lepetit, and P. Fua, 2008. A fast local descriptor for dense matching, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* Alaska, pp. 1–8.

Tola, E., V. Lepetit, and P. Fua, 2010. DAISY: An efficient dense descriptor applied to wide-baseline stereo, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815–830.

Tuytelaars, T., and L. Van Gool, 2000. Wide baseline stereo matching based on local, affinely invariant regions, *Proceedings of the Eleventh British Machine Vision Conference,* 11–14 September, Bristol, England.

Wu, B., 2006. *A Reliable Stereo Image Matching Method Based on the Self-adaptive Triangle Constraint*, Ph.D. dissertation, Wuhan University, China, 110 p.

Xu, Z., F. Zhang, F. Sun, and Z. Hu, 2009. Quasi-dense matching by neighborhood transfer for fish-eye images, *Acta Automatica Sinica,* 35(9):1159–1167.

Zhang, Z., 1998. Determining the epipolar geometry and its uncertainty: A review, *International Journal of Computer Vision,* 27(2):161–198.

Zhang, Z., Y. Zhang, T. Ke, and D. Guo, 2009. Photogrammetry for first response in Wenchuan earthquake, *Photogrammetric Engineering & Remote Sensing,* 75(5):510–513.

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.

Zhu, Q., B. Wu, and Y. Tian, 2007a. 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., B. Wu, and N. Wan, 2007b. A filtering strategy for interest point detecting to improve repeatability and information content, *Photogrammetric Engineering & Remote Sensing*, 73(5):547–553.

(Received 19 July 2010; accepted 24 September 2010; final version 27 October 2010)