


Article

Template Matching for Wide-Baseline Panoramic Images from a Vehicle-Borne Multi-Camera Rig

Shunping Ji ^{1,2,*} , Dawen Yu ¹, Yong Hong ³ and Meng Lu ⁴

¹ School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; yudawen@whu.edu.cn

² Beijing Advanced Innovation Center for Imaging Technology, Capital Normal University, Beijing 100001, China

³ Ledor Spatial Information Technology Corporation, Wuhan 430000, China; hongyong@ledor.com.cn

⁴ Department of Physical Geography, Faculty of Geoscience, Utrecht University, Princetonlaan 8, 3584 CB Utrecht, The Netherlands; m.lu@uu.nl

* Correspondence: jishunping@whu.edu.cn; Tel.: +86-135-5405-7323

Received: 3 May 2018; Accepted: 18 June 2018; Published: 21 June 2018



Abstract: Automatic detection and locating of objects such as poles, traffic signs, and building corners in street scenes captured from a mobile mapping system has many applications. Template matching is a technique that could automatically recognise the counterparts or correspondents of an object from multi-view images. In this study, we aim at finding correspondents of an object from wide baseline panoramic images with large geometric deformations from sphere projection and significant systematic errors from multi-camera rig geometry. Firstly, we deduce the camera model and epipolar model of a multi-camera rig system. Then, epipolar errors are analysed to determine the search area for pixelwise matching. A low-cost laser scanner is optionally used to constrain the depth of an object. Lastly, several classic feature descriptors are introduced to template matching and evaluated on the multi-view panoramic image dataset. We propose a template matching method combining a fast variation of a scale-invariant feature transform (SIFT) descriptor. Our method experimentally achieved the best performance in terms of accuracy and efficiency comparing to other feature descriptors and the most recent robust template matching methods.

Keywords: template matching; panoramic camera; mobile mapping system; feature descriptors

1. Introduction

A ground mobile mapping system (MMS) mounted with multiple sensors such as a mono/stereo camera, panoramic camera, laser scanner, and GPS/INS (global positioning system/inertial navigation system) has shown a wide range of implications in city planning, 2D/3D map making, traffic surveillance, and autonomous cars [1–4]. A challenge is to effectively locate objects of interest, such as manhole covers, telegraph poles, building corners, and so on, from multi-view images acquainted by an MMS, for constructing and maintaining georeferenced datasets of certain objects [5,6]. An intuitive solution is to automatically recognize the designated object instances with high precision and recall rate, then calculate their geolocation by GPS/INS and triangulation according to specific points in objects (for example, the geometric center of a manhole cover or the base of a pole). However, this solution can hardly be approached in practice with the difficulty of accurately extracting and segmenting objects. To construct a highly accurate object dataset for traffic maps or 3D city maps, manual recognition and measuring are still necessary.

In this paper, we automatically identify correspondences from multi-view images given a user-specified object in one of the images. With the correspondences, 3D world coordinates of the object

could be calculated using GPS/INS. The strategy can greatly reduce the manual work of selecting at least two correspondent points from stereo or multi-view images. One technology that has been used for specific object recognition is called template matching [7,8]. In contrast to a common sparse stereo matching, which typically extracts arbitrary features from images and then matches them [9,10] (see Figure 1a), template matching finds corresponding objects or image patches given a template. Figure 1b shows a classic case in early film photogrammetry where four fiducial marks are matched to the corresponding template for precise interior orientation. In this study, we aim to automatically locate a user-specified object in multi-view panoramic images through template matching (Figure 1c).

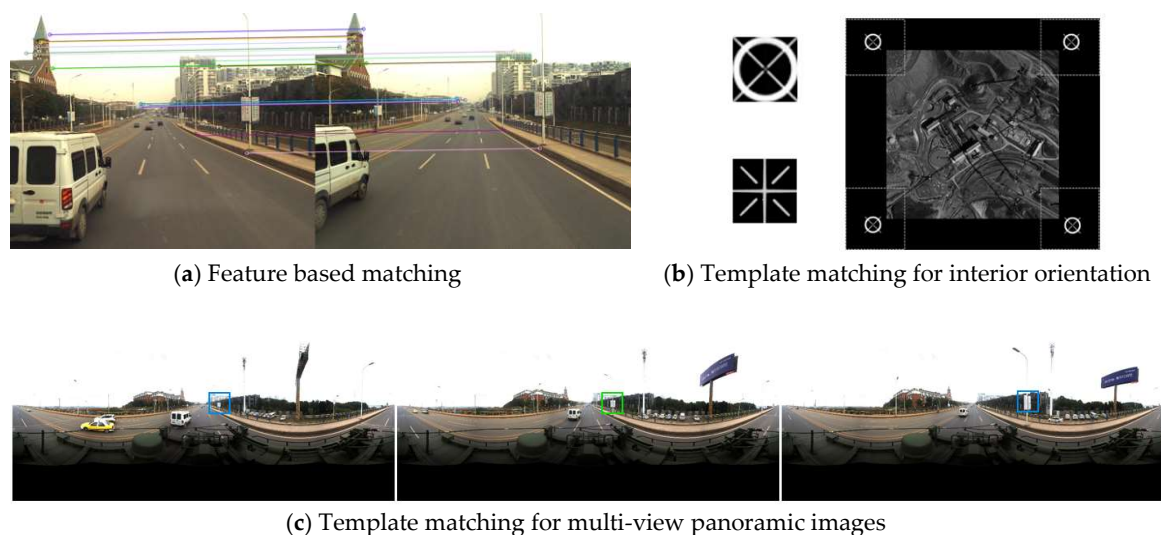


Figure 1. Some examples of image matching. (a) shows the feature-based matching method that firstly extracts arbitrary features (denoted by small circles) from images and then matches them (denoted by lines); (b) shows a traditional template matching case that uses the corresponding template patch to locate four fiducial marks in an aerial image; (c) shows the template matching case in our study. Given an object or patch (denoted by green box) in one image, the corresponding objects (denoted by blue box) with significant distortion should be retrieved from multi-view panoramic images.

The main advantage of a panoramic camera over a mono/stereo camera is that it covers a 360° scene at one shot. Its sensor models and applications have been studied for years. The spherical perspective [11], global description of omni-directional images [12], and rotation and scale invariance of image patches [13] of a spherical camera have been studied theoretically. However, no physical spherical camera has been produced due to the challenging manufacturing process. One type of manufactured panoramic camera, called a catadioptric camera, utilizes a paraboloid mirror to reflect light from all around a scene into a lens for light collecting. Geyer and Daniilidis gave detailed projective geometry for a catadioptric panoramic sensor [14]. Barreto and Araujo studied the geometry of the catadioptric projection of lines and its use in calibration [15]. Another type of panoramic camera uses a high-speed rotating CCD to collect a 360°-cylinder scene [16]. The third and the most popular panoramic camera used in an MMS is called multi-camera rig system, which consists of a series of fisheye lenses. Shi et al. proposed a rigorous model for a multi-camera rig and integrated the model into a GPS-supported simultaneous localization and mapping (SLAM) [17]. Kaess and Dellaert used a multi-camera rig for a SLAM with an ideal spherical model [18]. Ji et al. compared the difference between the ideal model and the rigorous model in indoor and outdoor scenes [19].

Although a popular substitute for the ideal omnidirectional camera, the multi-camera rig is limited by the current manufacturing techniques: the projection center of each fisheye lens hardly overlaps and the optical axis of each lens seldom lies in the same plane; the separate fisheye images are stitched to a panoramic image with inevitable stitching errors. These technical limitations restrain the applications

of 3D street scene reconstruction and high-precision surveying. However, the multi-camera rig could be used to measure objects of interest with an accuracy of decimeter level. In [19], the localization error has been analyzed, and experimentally proved to be less than 0.2 m. In this paper, we leverage multi-view panoramic images captured from a multi-camera rig to locate user-specified objects and focus the work on the challenges of applying template matching to detect distorted objects from panoramic images.

Three problems of using a multi-camera rig in template matching should be addressed. First, the wide baseline (e.g., 8 m) of a multi-camera causes significant perspective distortion between multi-view images. The distortions could beat some classic matching methods using absolute intensity difference or normalized correlation of image patches. Secondly, a panoramic camera introduces further geometric deformation. One part of the deformation is caused by the projection from a 3D sphere to a 2D image, which makes the epipolar no longer a line but a trigonometric curve. The other part is the stitching error raised from the camera model of a multi-camera rig. The stitching error cannot be eliminated without knowing the depth of the scene. An object farther away from the panoramic projection sphere (e.g., 20 m radius) has larger errors as its corresponding points in the adjacent fisheye images could not be overlapped with each other. Third, compared to an arbitrary feature matching where correspondences are compared in feature space, a template matching requires pixelwise searching in a given range in the original image. This indicates that the setting of an adequate search range may affect the efficiency and success rate, as a larger search range brings more false positives.

Considering these restrictions, our work focuses on two aspects. One is to investigate the epipolar geometry of the multi-camera rig and the epipolar accuracy that could be achieved. Shi et al. proposed the stereo geometry of a multi-camera rig and pointed out that the epipolar line is the intersection of the panoramic sphere and a 2D plane roughly passes through camera center [17]. Ji et al. further analyzed the accuracy of the epipolar according to correspondent points in the stereo pairs [19]. However, these studies do not explicitly provide the formula of the epipolar model of a multi-camera rig.

The other goal of our work is to develop an optimal feature descriptor or a robust template matching method that could match wide baseline panoramic images. The most used technology in template matching is normalized correlation with pixel values (intensity) [20]. However, correlation could be defeated by any geometric distortion more than a shift. To make the descriptor (or similarity measure) more robust, the strategy used in the feature matching can be utilized. Although an object selected by a user (typically an image patch) does not satisfy the definition of an empirical feature, it can be described by any feature descriptor. For example, CENSUS is a simple and popular feature descriptor used in stereo matching, which consists of a sequence vector with binary elements comparing between a point of interest and its neighborhood [21]. The descriptor used in the SURF feature consists of Haar wavelet responses in horizontal and vertical directions and the absolute values of these responses [9]. The SIFT descriptor is a 128-dimensional bin-based gradient vector [10].

Besides utilizing feature descriptors in template matching, several recent articles examined other strategies for template matching in challenging situations. Itamar et al. proposed a similarity measure named deformable diversity similarity for template matching based on properties of the nearest neighborhood [22]. Shaul et al. used a measurement called Best-Buddies similarity to make template matching robust against complex geometric deformations and high levels of outliers [23]. Wisarut et al. presented an adaptive template matching based on the sum of squared difference to enhance the matching quality in object tracking and reduce high computational cost [24]. Wu et al. also proposed a method to speed-up template matching and decrease the computational costs of conventional methods [25]. Yoo et al. presented a histogram-based template matching method for the situation of large-scale differences between target and template images [26]. To deal with heavy appearance variations, Sun et al. proposed a multiple template method to track fast motion by generating virtual templates that are affinely transformed from the original one [27]. Korman et al. also proposed a fast method for template matching considering 2D affine transformations [28]. Hong et al.

proposed a compact orientation templates (DOT) representation with a fast partial occlusion handling approach [29]. Generally, an ideal template matching method should be effective and efficient. We propose an accelerated SIFT descriptor for template matching with improved effectivity and efficiency comparing to the most recent template matching methods that are considered robust.

The main contribution of the paper is to thoroughly investigate the challenging template matching on wide baseline multi-view panoramic images for the first time, and to propose a high performance matching method. In Section 2, we review the camera and stereo geometry of a multi-camera rig and deduce the explicit epipolar model under a sphere projection. In Section 3, we analyze the epipolar errors to constrain the searching range for a pixel-wise matching. Several feature descriptors are introduced to the template matching on multi-view panoramic images, and an efficient matching strategy that significantly alleviates computational costs is proposed. The experiments in Section 4 are designed to evaluate the accuracy and efficiency of these matching methods. The discussion and conclusion are given in Sections 5 and 6 respectively.

2. Geometry of A Multi-Camera Rig System

2.1. Ideal Panoramic Camera Model

An ideal spherical panoramic model represents the collinearity of an arbitrary 3D point P with coordinate vector X_p , the corresponding panoramic point u with coordinate vector $X = [x', y', z']$, and the panoramic center S (Figure 1a). In Equation (1), R and T are the rotation matrix and translation vector, respectively, and λ is the scale difference between the panoramic and world coordinate systems. In Equation (2), X is restricted on the surface of a sphere with radius r .

$$\lambda X = R^T(X_p - T) \quad (1)$$

$$x'^2 + y'^2 + z'^2 = r^2 \quad (2)$$

Coordinate X can be obtained from a 2D image point $x = [x, y]$. In Equation (3), φ_h and φ_v are the horizontal angle with the range $[-\pi, \pi]$ and the elevation angle with the range $[-0.5\pi, 0.5\pi]$ respectively. w and h are the width and height of the panoramic image respectively. Equation (4) calculates sphere coordinate X using a right-hand coordinate system.

$$\varphi_h = \frac{\pi(2x - w)}{w}; \varphi_v = \frac{\pi(h - 2y)}{2h} \quad (3)$$

$$\begin{aligned} x' &= r \cos(\varphi_v) \sin(\varphi_h) \\ y' &= r \cos(\varphi_v) \cos(\varphi_h) \\ z' &= r \sin(\varphi_v) \end{aligned} \quad (4)$$

2.2. Rigorous Panoramic Camera Model

As is shown in Figure 2b, a multi-camera rig is composed of five separate fish-eye lenses. The rigorous camera model is therefore a transformation from a world object P' to the corresponding projection point u_c on the fish-eye image (the solid line). For convenience, the fish-eye image coordinates are usually firstly transformed to a virtual plane camera coordinates by choosing a fisheye camera model [30] and a calibration model [31]; we denote the transformation as K_c . Then, according to the rotation R_c and translation T_c between a fisheye camera and the virtual panoramic camera, Equation (5) describes a fisheye image point u_c with a coordinate x_c projected to the corresponding panoramic camera point u with a coordinate X . K_c, R_c, T_c are typically fixed values with pre-calibration, and k is the scale factor between the ideal plane and the panoramic sphere, which can be calculated by combining Equations (2) and (5).

$$X = kR_cK_cx_c + T_c \quad (5)$$

The third step is to associate the obtained panoramic coordinate X with its world coordinate X_p . According to the solid line in Figure 2b that passes C , u , and P' , a more rigorous model could be constructed as Equation (6) where the projection centre lies in the fisheye camera.

$$T_c + \lambda(X - T_c) = R^T(X_p - T) \tag{6}$$

Neither (1) nor (6) could fully compensate the geometric stitching errors caused by a multi-camera rig system. However, the projection centres of the separate fish-eye cameras C and the panoramic centre S are very close, and the angle between SP and CP' is very small, which could ensure that by using (6) the projection errors of an object, within a certain distance to the projection sphere, are limited to less than one pixel. Note that every fisheye camera has its own translation parameter T_c to the panoramic camera.

In this paper, the rigorous camera model (5) and (6) is used for relative orientation and local bundle adjustment of panoramic image series for recovering the geometric relations between the adjacent multi-view images. The ideal camera model (1) is utilized to generate epipolar curve and to evaluate its accuracy, as the object is measured on a panoramic image with 360° covering.

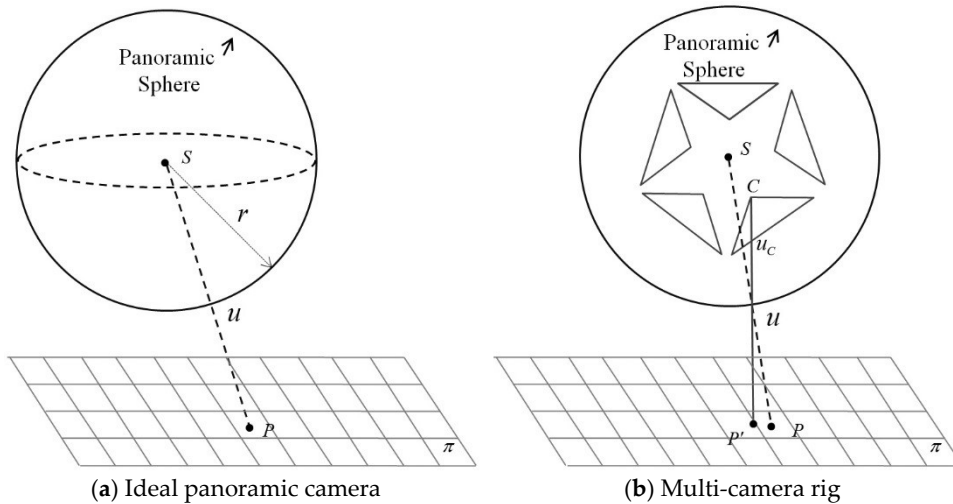


Figure 2. The ideal panoramic camera (a) where S , u , and p are collinear, and the multi-camera rig (b) where in fact C , u , and p' are collinear and C and S don't overlap.

2.3. Ideal Epipolar of Panoramic Stereo

Both the ideal and rigorous multi-camera model have their corresponding epipolar geometry. As points are matched in the stitched panoramic images other than separate fisheye images, only the epipolar of an ideal panoramic stereo is utilized.

We set the baseline $B = [B_x \ B_y \ B_z]$ and the corresponding rays $X_1 = [X_1 \ Y_1 \ Z_1]$ and $X_2 = [X_2 \ Y_2 \ Z_2]$ in panoramic coordinates. $X_2' = RX_2$ is the ray that has been translated to the coordinates of the left camera by a rotation R . Then we have

$$B \cdot (X_1 \times X_2') = 0 \tag{7}$$

Equation (7) can be expanded by the third line of the determinant to obtain Equation (8), where a , b , and c are determined by X_1 and R and $a = B_y Z_1 - B_z Y_1$, $b = B_z X_1 - B_x Z_1$, $c = T_x Y_1 - T_y X_1$. Combined with Equation (2), the epipolar line of ideal panoramic stereo images is a large circle through the panoramic camera centre.

$$aX_2 + bY_2 + cZ_2 = 0 \tag{8}$$

Finally, we take Equations (3) and (4) into (8):

$$a \sin\left(\frac{2\pi}{w}x\right) + b \cos\left(\frac{2\pi}{w}x\right) - \text{carctan}\left(\frac{\pi}{h}y\right) = 0 \quad (9)$$

where (x, y) is pixel coordinate and w and h are the width and height of the image, respectively, indicating a trigonometric curve epipolar line of a panoramic stereo.

3. Multi-View Template Matching for Panoramic Images

3.1. Pre-Processing of Feature-Based Matching and Local Bundle Adjustment

The exterior orientation parameters obtained by the GPS/INS system could be utilized to recover epipolar geometry. However, we found the errors in GPS/INS observations could cause bias in images up to a dozen pixels. We execute relative orientation and local bundle adjustment to obtain the relative poses between adjacent images and use them to generate more accurate epipolar curves.

Firstly, SIFT features [10] are extracted from fisheye images and then matched in stereo pairs with a GPU boosted strategy [32]. Then, by transferring these matching candidates to panoramic coordinates with Equation (5), RANSAC [33] is embedded into Equation (6) for outlier elimination. Third, the corresponding features in adjacent stereo pairs are used to recover the scale shift between them. Every 3D feature could vote a scale and the median value is used. In this study, we use only triple view geometry for panoramic epipolar generation and template matching, as we used a wide baseline up to 8 m. Some matching errors may remain in the triple view images after scale fusion. We execute bundle adjustment embedded with weight decaying methods for the three images according to Equation (6) to ensure all of the false correspondences are eliminated. Finally, we obtain the accurate relative poses between the multi-view panoramic images.

3.2. Error Estimation of Panoramic Epipolar

Based on Equation (9), we draw the epipolar line on a panoramic image with sphere projection. In Figure 3, we evenly select points in the image and calculate the bias of the epipolar line, that is, the minimum distance between a true correspondent point and the epipolar line. According to Table 1, the average bias is 8.29 pixels and the maximum bias is 35 pixels. Generally, nearer points exhibit larger errors, as observed in Figure 3. We set a search range of 40 pixels in the orthogonal direction to the epipolar line to ensure every candidate is in the search area.

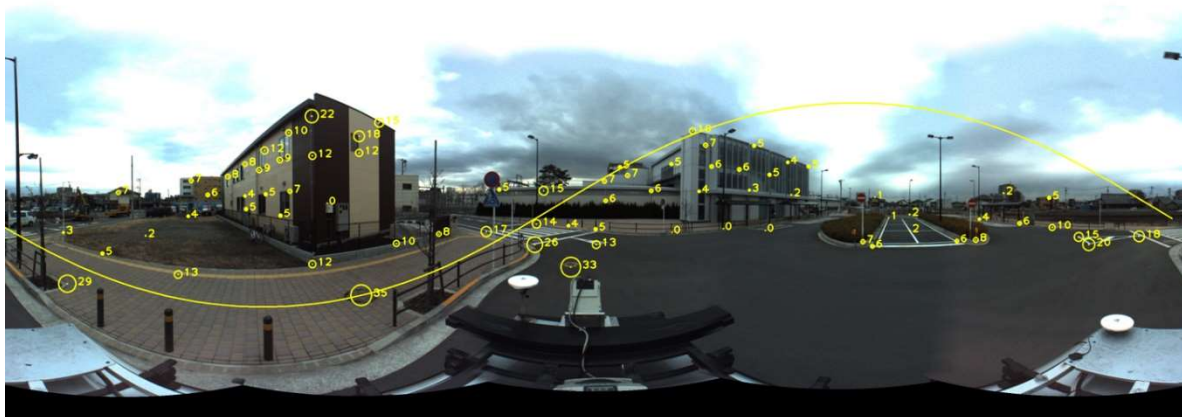


Figure 3. Epipolar of a panoramic image with sphere projection (yellow curve). Points in circles are selected for checking epipolar errors and bigger circles indicate larger errors (numbers beside the circles).

Table 1. Checking epipolar bias in a panoramic image with known correspondent points.

Total Points	Minimum	Maximum	Average	<10 Pixel	10~20	>20 Pixel
76	0	35.0	8.29	53 (69.74%)	18 (23.68%)	5 (6.58%)

The search range along the epipolar line is determined in two ways. In one case, we set the depth of an object to be measured between 0.5~100 m without any support from a depth sensor. A distance farther than 100 m could cause rapid degeneration of localization accuracy and is ignored.

In the other case, a laser scanner is considered available for collecting sparse point cloud. The point cloud has been georeferenced by GPS/INS data and the pre-calibration parameters between the laser scanner and the GPS/INS system. We project 3D laser points to a virtual panoramic image according to the pose of the camera (obtained by GPS/INS data and the pre-calibration parameters between the GPS/INS system and the camera) and the sensor model (1). As shown in Figure 4, the point cloud is sparse and sparser in larger depth. To make maximum use of the data, we set a search window for the given object point. The median value in this window is chosen as the depth of the object. In this paper, the length of the window is set to 20 pixels. Finally, the error range of a depth value d is set to $d \pm n$, where n is empirically set to 2 m.

In both cases, the object with the assumed maximum and minimum depth is projected to the panoramic image respectively to be the two endpoints of an epipolar curve segment, for constraining the search range along the epipolar.



Figure 4. A virtual panoramic depth map (right) generated from 3D point cloud (left) using the same pose of the corresponding panoramic image (middle).

Figure 5 shows the workflow of the depth map supported searching strategy for multi-view panoramic image matching. Given an object point in the n -th image, epipolar lines are generated respectively in the first stereo pair and the second pair. The depth of the object point is retrieved from the corresponding depth map. The search range is determined (the red curved boxes in the second row) based on the range of depth value and the bias of the epipolar line. Within the search range, the optimal candidates are retrieved by pixelwise matching with a given similarity measurement. At last, the multi-view intersection is carried out for locating and error elimination.

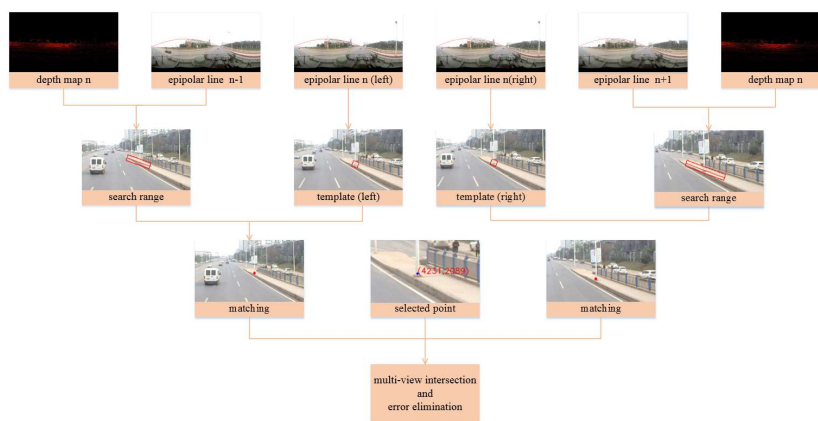


Figure 5. Depth map supported multi-view matching. The left and right stereos are matched separately and in multi-view intersection, mistakenly matched candidates could be observed and eliminated.

3.3. Template Matching with Different Feature Descriptors

Pixel-wise template matching within a given search area depends largely on a robust descriptor used for representing the reference and target image patches. The descriptor is then compared with a normalized correlation (or other similarity measure) and the highest correlation coefficient (or minimum distance) corresponds to the optimal candidate. In this section, we review several commonly used descriptors in template matching and feature matching, and propose an efficient strategy for calculating SIFT descriptors and their similarity.

Regardless of what kind of feature descriptor is used, scale changes that could not be completely compensated will affect the matching results. We resample images to the same scale according to the ratio of distances between the 3D object and the two camera stations. In Figure 6a,b, *depth1* and *depth2* are distances of an interested point measured in different camera stations. If depth map is unavailable, we use empirical values of 0.8, 1, and 1.2 as scale factors and sample only the reference image patch. For those features as pixel intensity that would fail due to a rotation, we dynamically align and resample the patches to be matched to the tangential direction of the epipolar line (Figure 6c,d).

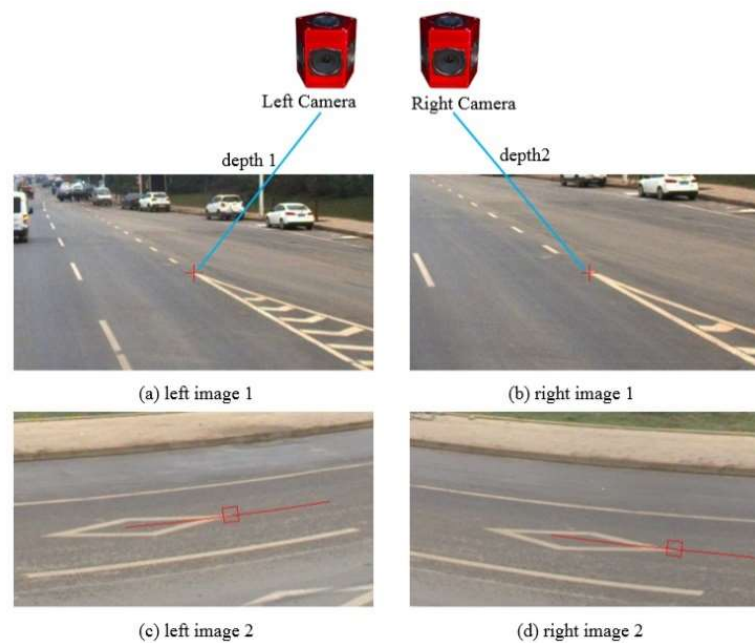


Figure 6. The preprocessing for scale and rotation. To reduce the scale changes between stereo image patches, we can resample (a) based on the ratio $\text{depth1}/\text{depth2}$ or resample (b) based on $\text{depth2}/\text{depth1}$. As for rotation, both of the patch windows could be aligned to the tangent direction of the current point, as shown on (c,d).

In this paper, descriptors listed below are evaluated in the template matching of wide baseline multi-view panoramic images.

Pixel values in dual space. Intensity of pixels is the most used feature in template matching. Except for intensity, we also utilize the phase information by transforming the image patch to the frequency domain with Fourier transformation [20]. The patches in the dual image and frequency space are used to calculate correlation coefficients, and the largest one is considered an optimal candidate.

CENSUS descriptor. By comparing the intensity of a pixel with its neighbor pixels within a 21×21 patch window after epipolar alignment, a CENSUS descriptor is an ordered sequence vector with its elements the binary number for representing “bigger” or “smaller”. The similarity between two CENSUS descriptors is calculated by the normalized Hamming distance.

HOG (Histogram of Oriented Gradient) descriptor [34]. Depending on the quality of the reference and target images, Gamma normalization is selectively executed first. Then, within the 16×16 pixels block, the gradient magnitude and orientation of each point are computed, and the block is divided into 2×2 cells and aligned to the dominant gradient orientation. Afterwards, a tri-linear interpolation is used to calculate 9 orientation bins for each cell. At last, the L2-Hys (Lowe-style clipped L2 norm) scheme is used to normalize the 36-dimensional vector of the block. The similarity depends on the normalized distance of two HOG vectors.

SURF descriptor. After aligned to the tangent direction of the epipolar line, a 20×20 patch centered on the current point is split into 4×4 sub-regions, and a four-dimensional feature vector is computed in each sub-region with 5×5 pixels inside. The four elements of the vector are the sums of Haar wavelet responses in horizontal and vertical directions respectively, as well as the sums of the absolute values of these responses. The normalized distance between the two SURF descriptors (i.e., the two 64-dimensional vectors), is used as the similarity measure.

SIFT descriptor. First, the gradient magnitude and orientation of each point within the 16×16 patch window are computed and the patch is aligned to the main orientation. The Gaussian kernel is then utilized to weight points by the distance to the center point. Second, the gradient magnitude and orientation of each pixel in the 4×4 sub-window are accumulated into an orientation histogram with eight directions to make up a local eight-dimensional vector. A complete SIFT descriptor consists of a 128-dimensional vector of 16 sub-regions after a maximum suppression of larger gradients. The normalized distance between the two SIFT vectors is used as the similarity measure.

Accelerated SIFT descriptor. As the calculation of the pixel-wise SIFT descriptor could be very slow for template matching especially with large search range, we developed a method that dramatically reduces the calculation load of the SIFT descriptor (referred in this study as accelerated SIFT descriptor). The SIFT descriptor consists of 4×4 components each of which includes 4×4 pixels and is described by an eight-dimensional vector. Hence, we can calculate this vector for every pixel via its 4×4 region in advance (2×2 region for simplicity in Figure 7a). In the search range, we finally produce a map with the elements of the vector. This guarantees the calculation of the SIFT descriptor performed only once and we only need to read the descriptor of a candidate from memory instead of re-calculation. Then, a Gaussian window is applied to weight pixels with distance to the current search point. For orientation bias, we rotate the reference image patch instead of the searched patch. To achieve this, we firstly calculate the tangent direction of the epipolar line d_i of the current candidate point i (as every point on the panoramic epipolar has its own tangent direction as shown in Figure 7b), and then obtain the direction difference between d_i and the tangent direction of the reference epipolar line d_0 . The reference patch is rotated with $d_0 - d_i$ degree to compensate the angle bias (Figure 7c). The calculation of the direction difference is not very rigorous and could introduce tiny direction bias however it is very slight and could be ignored for the tremendous efficiency improvement.

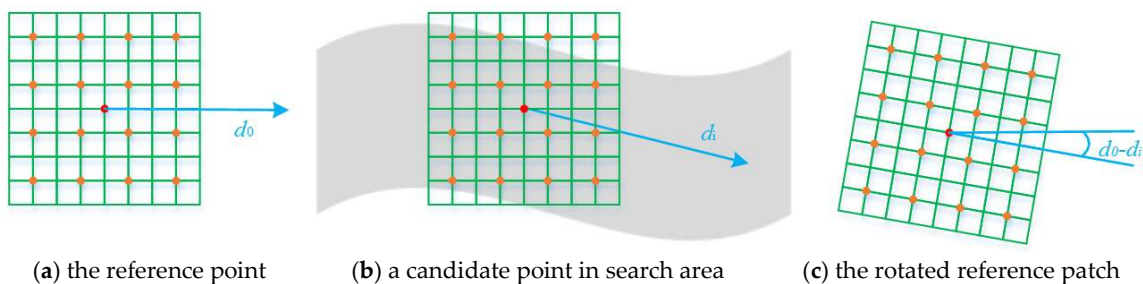


Figure 7. The computation of the accelerated SIFT descriptor. In (a), a SIFT point is described by 16 8-D vectors, which can be stored in the 16 orange points. Therefore, for every pixel in search area (b) we can calculate and store the 8-D vectors only once. When a point is to be matched, its SIFT descriptor is easily retrieved. In (c) the reference descriptor is rotated according to the difference of the two epipolar directions.

4. Experiments and Results

4.1. Test Design

To test the multi-view template matching on panoramic images, we prepared two datasets under different situations. The first dataset consisting of 498 images was captured in 2012 using a PGR's Ladybug3 camera [35] in Kashiwa city, Japan. The camera consists of six fixed fisheye lenses, each of which has a maximum 1616×1232 pixel resolution and 24-bit RGB colour resolution (see Figure 8). The focal length of the fisheye camera is 3.3 mm, and the radius of the panoramic sphere is set to 20 m. A dual-frequency GPS receiver and an IMU (inertial measurement unit) are mounted on the car for georeferencing with CORS RTK technology and GPS/IMU combined filtering (embedded in an Applanix POS/AV system). The processed GPS/IMU data is leveraged to recover the poses of camera, and the latter is then used to calculate the 3D world coordinates of matched points through triangulation and to estimate the search area. A baseline of 2 m interval between adjacent images is used for testing. The search range of an object depth is set to 0.5~100 m.

The other dataset with 168 images was captured in 2017 using a car-mounted Ladybug5 camera in Wuhan city, China. The maximum resolution of the fisheye lens is improved to 2448×2048 pixels. Except for GPS/IMU, a laser scanner was utilized to collect a sparse depth map (see Figure 4). Due to the measurement errors of GPS/IMU and the calibration errors among GPS, camera, and laser scanner, the bias between the depth map and the panoramic image is up to 0.6 m. A baseline of 8 m interval between adjacent images is used for test. The search range along the epipolar line is set to the object's depth range $d \pm 2$ m, where d is the median pixel value within the 20×20 window centered at the interested point in the depth map.

In pre-processing, SIFT feature matching, tie point extraction, and local bundle adjustment were sequentially carried out on the two datasets to obtain the epipolar geometry of the multi-view panoramic images. Hundreds of interested points were then manually selected in one view image to evaluate the performance of different template matching methods.



Figure 8. Six fisheye lenses and the stitched panoramic images captured by the Ladybug 3 camera.

4.2. Results of the Wuhan Data (with Depth Map)

Table 2 shows the performances of the different descriptors embedded in template matching on 80 points selected from road signs, covers, poles, railings, and so on. Although there exist significant geometric distortions, the dual-space intensity based matching could obtain a satisfactory match rate with our preprocessing strategy, that is, alignment to epipolar direction and scale normalization according to depth ratio. Among the feature descriptors, the CENSUS descriptor performs the worst and SIFT the best. Our accelerated SIFT (shorted as AccSIFT) descriptor got the same results as SIFT except for one point. The SURF descriptor performed similar to the dual-space intensity and better than the HOG descriptor.

As to efficiency, more complex descriptors result in less efficiency, and the SIFT descriptor is much lower than the dual-space intensity in execution. Our AccSIFT descriptor got almost the same efficiency as the dual-space intensity and improved 7.9 times compared to the original SIFT descriptor.

Table 2. Comparison of different descriptors on the Wuhan data.

Methods	Match Rate	Time (s)
Intensity	72/80	0.202
CENSUS	65/80	0.119
HOG	69/80	0.344
SURF	73/80	0.631
SIFT	78/80	1.539
AccSIFT	77/80	0.195

Figure 9 shows some vision examples applied with different descriptors. The main distortion in Figure 9a is the scaling distortion; both the dual-space intensity and CENSUS failed to match on one view image while SURF, HOG, and AccSIFT (SIFT) can find all the correct correspondences in multi-view images. It could also be observed that the search range along the epipolar varies largely from different views. Figure 9b shows a challenging situation where the depth of the object is unavailable due to the sparsity of the depth map. A very long search range should therefore be covered. However, except one match error occurred in the CENCUS and HOG methods, other methods correctly found the correspondences. Figure 9c shows an example demonstrating the distortion caused by different view angles. The dual-space intensity and HOG methods failed but all the other methods succeeded.

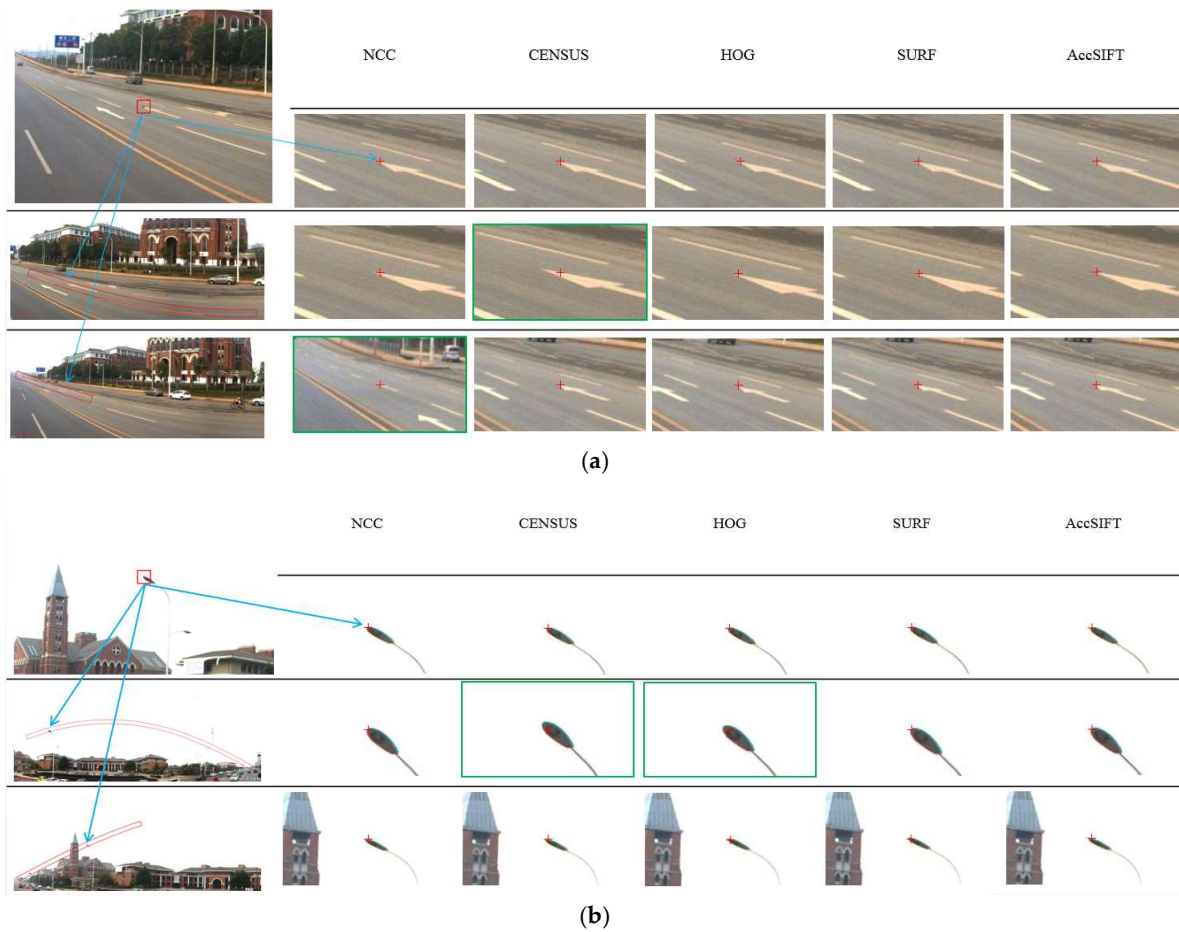


Figure 9. Cont.

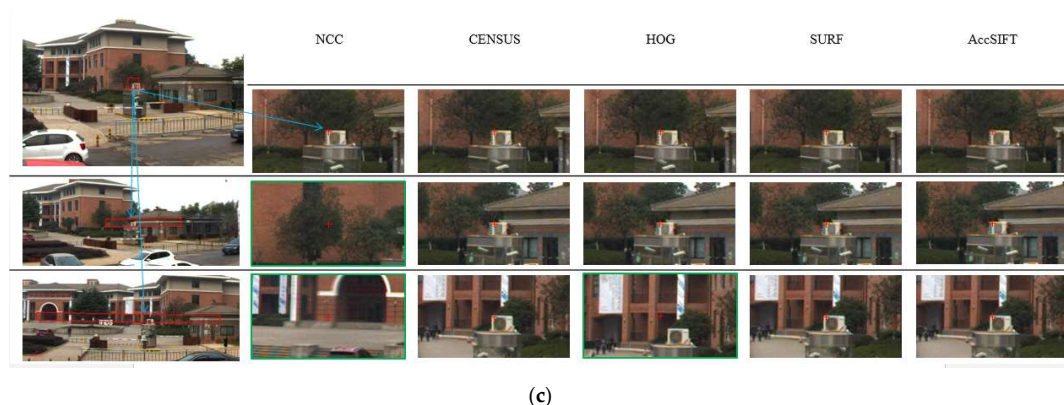


Figure 9. Performances of template matching with different descriptors on the Wuhan data. From left to right: reference image and the two adjacent search image, the results of the NCC, CENSUS, HOG, SURF and AccSIFT descriptors respectively.

4.3. Results of the Kashiwa Data (without Depth Map)

Table 3 shows the performances of template matching with different descriptors on the Kashiwa data where 120 object points were selected for evaluation. The CENSUS descriptor got the worst matching rate and the dual-space intensity and SURF performed almost the same. The SIFT and AccSIFT descriptors performed the best and the second best respectively. As the depth of an object is unknown, a larger search range has to be covered compared to the Wuhan data. This also caused the match rate slightly dropped. For example, the match rate of SIFT dropped from 97.5% to 93.3%. The efficiency dropped correspondingly. It took 4.6 s to match a point when using the SIFT descriptor. Our AccSIFT descriptor improved the efficiency by 450% comparing to the original version, and is even faster than the dual-space intensity.

Table 3. Comparison of the different descriptors on the Kashiwa data.

Methods	Match Rate	Average Time (s)
Intensity	101/120	1.259
CENSUS	91/120	0.572
HOG	96/120	1.797
SURF	103/120	2.849
SIFT	112/120	4.646
Acc SIFT	107/120	1.042

Figure 10 shows some examples of different descriptors applied on the Kashiwa data. In Figure 10a, a point in a far street lamp made all of the methods failed except the dual-space intensity. The results represent a challenging case as the wrongly predicted positions vary largely. It was also a rare case that all of the feature-based descriptors failed to find the correct match. This might be interpreted as meaning that the complicated SIFT or SURF descriptor could simulate more complex distortions and therefore created ambiguity (false positive). Figure 10b shows the large search range and the changes of scale made the intensity, HOG and CENSUS descriptors failed in matching a point on a cover. In Figure 10c the changes of view sight at a corner resulted one matching error of the SURF and HOG descriptors.

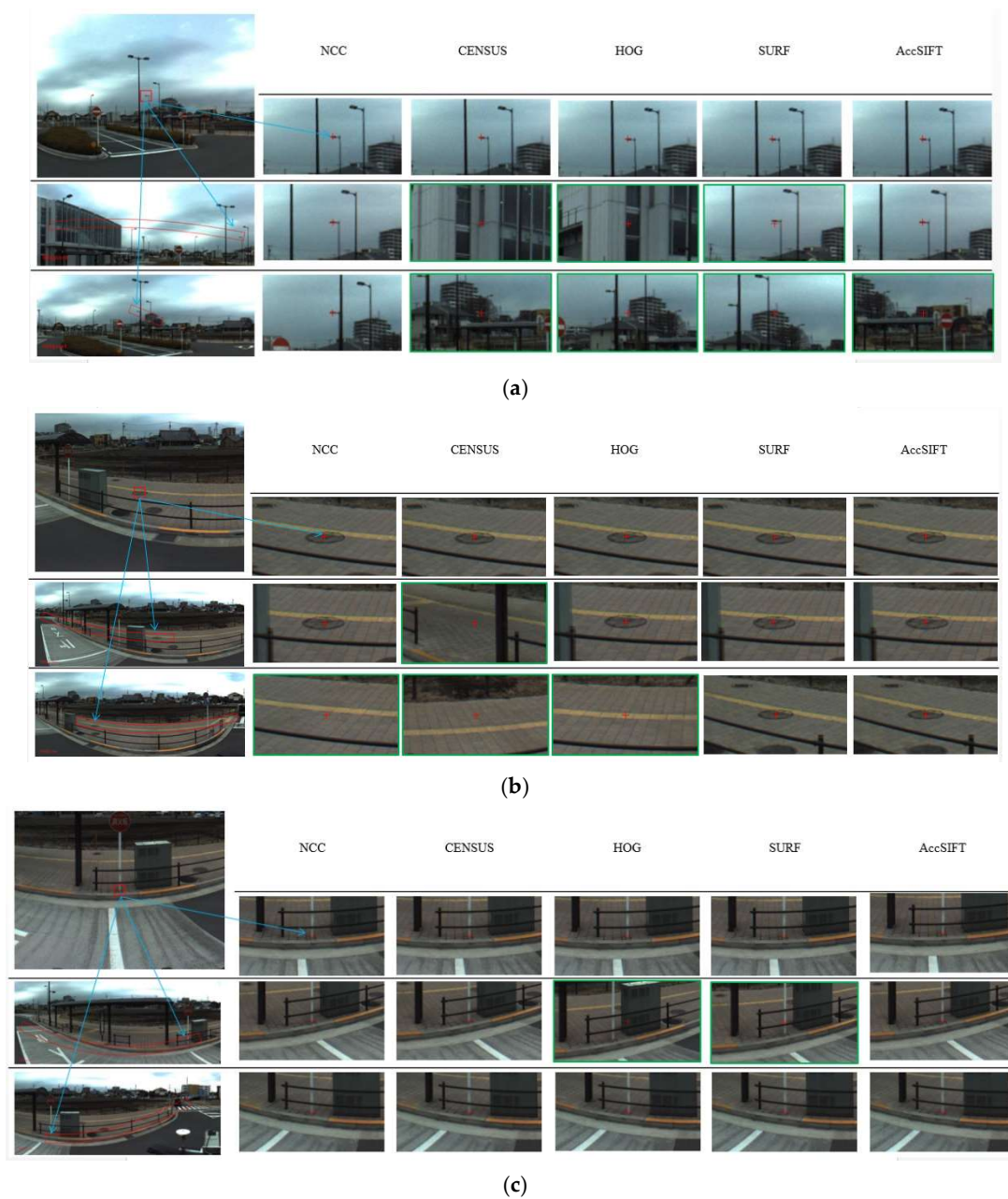


Figure 10. Performances of template matching with different descriptors on the Kashiwa data. From left to right: reference image and the two adjacent search image, the results of the NCC, CENSUS, HOG, SURF and AccSIFT descriptors respectively.

5. Discussion

5.1. The Difference between the AccSIFT and SIFT Descriptors

Considering the processing efficiency, our AccSIFT descriptor was designed slightly differently from the original version. We align the two corresponding patches by rotating an angle that equals the difference between the two tangent directions of the stereo epipolar lines, instead of rotating them respectively to the epipolar directions. It may introduce a slight bias especially when the angle difference becomes larger, and leads to a slight underperformance of our method comparing to the original version. On the Wuhan data, the matching rate dropped 1.25%; on the Kashiwa data, it dropped 4.1%. Figure 11 shows the case in which the AccSIFT descriptor failed while the SIFT

descriptor could find one correct match. It could be observed that large rotations of background buildings caused difficulty in matching. However, the efficiency of the AccSIFT descriptor gained improvement of 790% and 450% against the SIFT descriptor on the two datasets respectively.

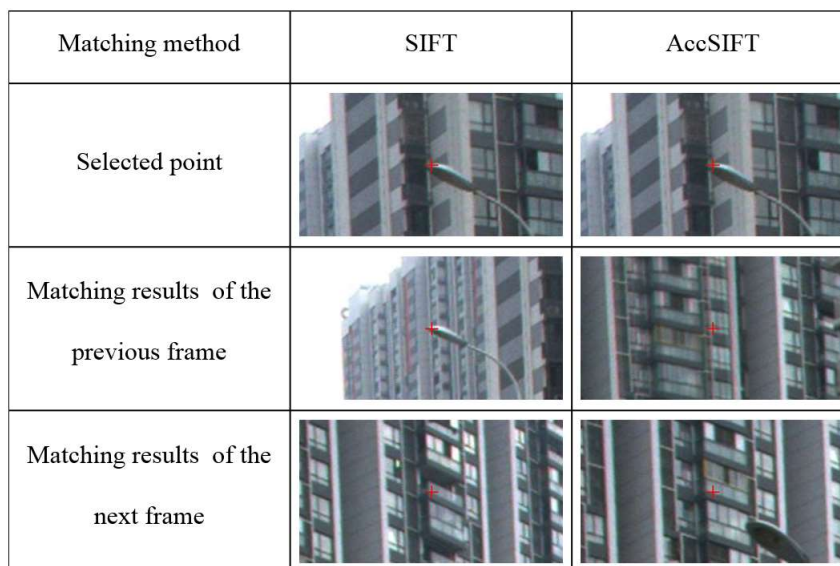


Figure 11. Comparison of the SIFT and the AccSIFT descriptors on matching of the Wuhan images.

5.2. Comparison to Most Recent Studies on Template Matching

We compare our AccSIFT matching method with two most recent studies in robust template matching for large geometric distortions. One method proposed a deformable diversity similarity (short as DDIS) [22]; the other method introduced a similarity called best-buddies similarity (shorted as BBS) [23]. A color based version (other than depth) is applied in this case (shorted as DDIS-C and BBS-C). To be fair, we cropped the corresponding search range (a bounding box) of every interested point from the Wuhan images and masked the background with black in the bounding box (see Figure 12). From Table 4, we could observe our method performing better than the two methods in match rate, and the BBS-C is the worst. As to efficiency, our method and DDIS-C performed almost the same and it is faster than the BBS-C. Figure 12 shows four examples of template matching on the Wuhan data. In all the challenging cases, our method correctly found the correspondences in both of the views, whereas the DDIS-C and BBS-C failed to one view image or the both.

It should be noted that, our AccSIFT descriptor is specially designed for template matching with approximate epipolar constraints (commonly could be obtained via feature matching or GPS/IMU data) and therefore encoded with better rotation invariance. The other two methods depend on the claimed rotation-invariant similarity measure without these easily accessed epipolar constraints.

Table 4. Comparison to most recent studies on the Wuhan data.

Methods	Correct Rate	Average Time (s)
AccSIFT	77/80	0.195
DDIS-C	74/80	0.187
BBS-C	63/80	0.325

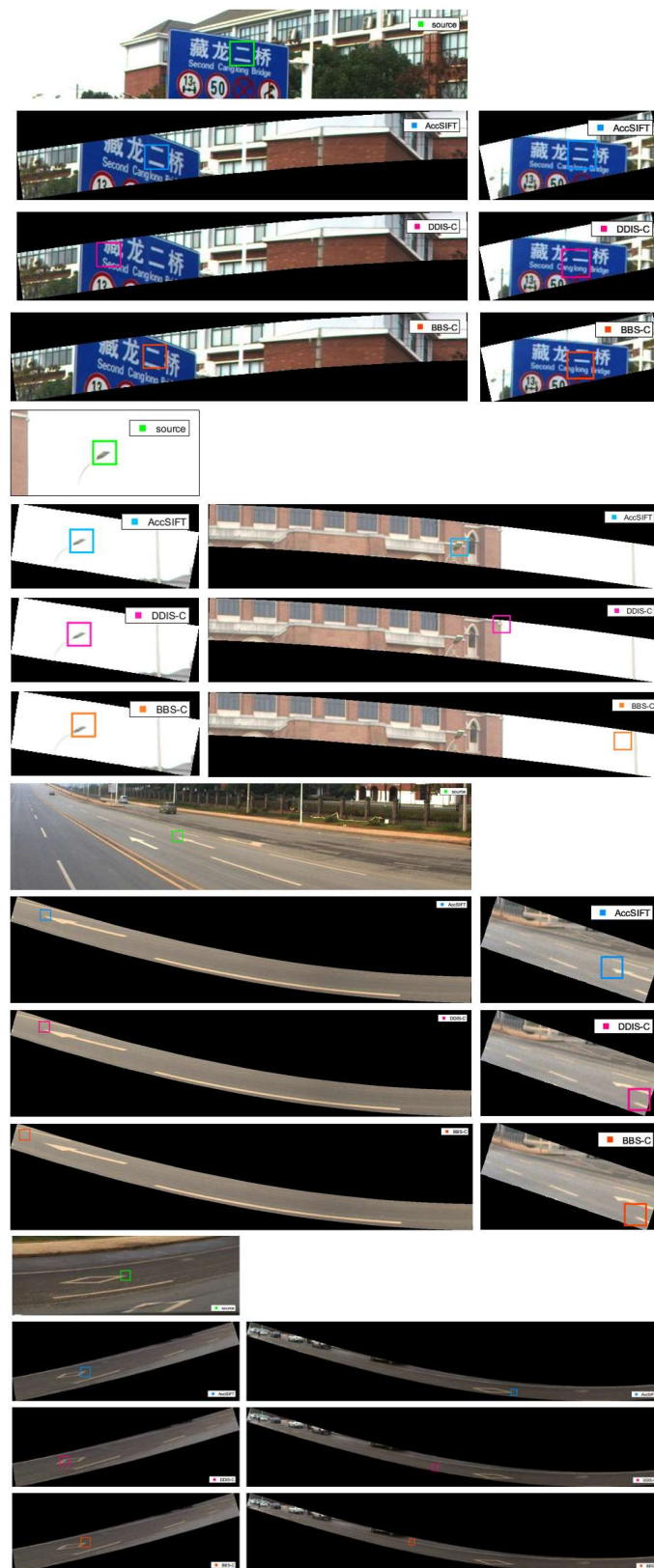


Figure 12. The performances of the three template matching methods on the Wuhan data. The green box in the top images indicates the reference to be matched; the left image patches are cropped from the last frame panoramic image and the right cropped from the next frame. The results of our method are denoted by blue box; the results of DDIS-C and BBS-C are denoted by pink and red box respectively.

6. Conclusions

We studied template matching on wide-baseline multi-view panoramic images from a vehicle-mounted multi-camera rig. Firstly, the epipolar equation of the panoramic stereo is deduced and the epipolar accuracy is analyzed. We then thoroughly evaluate the performances of different feature descriptors on this template matching case, and propose a fast version of a SIFT descriptor, which reached the best performance in terms of efficiency and accuracy. We also showed our method improved from the two most current studies in robust template matching of multi-view panoramic images.

Author Contributions: S.J. led the research and wrote the paper; D.Y. executed most of the experiments; Y.H. and M.L. are involved in research design and manuscript editing.

Funding: This work was supported by the National Natural Science Foundation of China (41471288) and the National key research and development plan of China (2017YFB0503001).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yang, B.; Fang, L.; Li, J. Semi-automated extraction and delineation of 3d roads of street scene from mobile laser scanning point clouds. *ISPRS-J. Photogramm. Remote Sens.* **2013**, *79*, 80–93. [[CrossRef](#)]
2. Paparoditis, N.; Papelard, J.-P.; Cannelle, B.; Devaux, A.; Soheilian, B.; David, N.; Houzay, E. Stereopolis II: A multi-purpose and multi-sensor 3d mobile mapping system for street visualisation and 3d metrology. *Rev Fr. Photogramm. Télédétec.* **2012**, *200*, 69–79.
3. Corso, N.; Zakhor, A. Indoor localization algorithms for an ambulatory human operated 3d mobile mapping system. *Remote Sens.* **2013**, *5*, 6611–6646. [[CrossRef](#)]
4. El-Sheimy, N.; Schwarz, K. Navigating urban areas by VISAT—A mobile mapping system integrating GPS/INS/digital cameras for GIS applications. *Navigation* **1998**, *45*, 275–285. [[CrossRef](#)]
5. Jaakkola, A.; Hyypä, J.; Kukko, A.; Yu, X.; Kaartinen, H.; Lehtomäki, M.; Lin, Y. A low-cost multi-sensoral mobile mapping system and its feasibility for tree measurements. *ISPRS-J. Photogramm. Remote Sens.* **2010**, *65*, 514–522. [[CrossRef](#)]
6. Kim, G.H.; Sohn, H.G.; Song, Y.S. Road infrastructure data acquisition using a vehicle-based mobile mapping system. *Comput.-Aided Civ. Infrastruct. Eng.* **2006**, *21*, 346–356. [[CrossRef](#)]
7. Briechele, K.; Hanebeck, U.D. Template matching using fast normalized cross correlation. *Proc. SPIE* **2001**, *4387*, doi:10.1117/12.421129. [[CrossRef](#)]
8. Brunelli, R. *Template Matching Techniques in Computer Vision: Theory and Practice*, 1st ed.; John Wiley & Sons: Hoboken, NJ, USA, 2009.
9. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-up robust features (SURF). *J. Comput. Vis. Image Underst.* **2004**, *110*, 346–359. [[CrossRef](#)]
10. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]
11. Mei, C.; Benhimane, S.; Malis, E.; Rives, P. Efficient homography-based tracking and 3-D reconstruction for single-viewpoint sensors. *IEEE Trans. Robot.* **2008**, *24*, 1352–1364. [[CrossRef](#)]
12. Paya, L.; Fernandez, L.; Gil, A.; Reinoso, O. Map building and Monte Carlo localization using global appearance of omnidirectional images. *Sensors* **2010**, *10*, 11468–11497. [[CrossRef](#)] [[PubMed](#)]
13. Gutierrez, D.; Rituerto, A.; Montiel, J.M.M.; Guerrero, J.J. Adapting a real-time monocular visual SLAM from conventional to omnidirectional cameras. In Proceedings of the 11th OMNIVIS in IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 343–350.
14. Geyer, C.; Daniilidis, K. Catadioptric projective geometry. *Int. J. Comput. Vis.* **2001**, *45*, 223–243. [[CrossRef](#)]
15. Barreto, J.P.; Araujo, H. Geometric properties of central catadioptric line images and their application in calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1327–1333. [[CrossRef](#)] [[PubMed](#)]
16. Parian, J.A.; Gruen, A. Sensor modeling, self-calibration and accuracy testing of panoramic cameras and laser scanners. *ISPRS-J. Photogramm. Remote Sens.* **2010**, *65*, 60–76. [[CrossRef](#)]

17. Shi, Y.; Ji, S.; Shi, Z.; Duan, Y.; Shibasaki, R. Gps-supported visual slam with a rigorous sensor model for a panoramic camera in outdoor environments. *Sensors* **2012**, *13*, 119–136. [CrossRef] [PubMed]
18. Kaess, M.; Dellaert, F. Probabilistic structure matching for visual SLAM with a multi-camera rig. *Comput. Vis. Image Underst.* **2010**, *114*, 286–296. [CrossRef]
19. Ji, S.; Shi, Y.; Shi, Z.; Bao, A.; Li, J.; Yuan, X.; Duan, Y.; Shibasaki, R. Comparison of two panoramic sensor models for precise 3d measurements. *Photogramm. Eng. Remote Sens.* **2014**, *80*, 229–238. [CrossRef]
20. Lewis, J.P. Fast normalized cross-correlation. In Proceedings of the Vision Interface, Quebec, QC, Canada, 15–19 June 1995; pp. 120–123.
21. Zabih, R.; Woodfill, J. Non-parametric local transforms for computing visual correspondence. In Proceedings of the 3rd European Conference on Computer Vision, Stockholm, Sweden, 2–6 May 1994; pp. 151–158.
22. Talmi, I.; Mechrez, R.; Zelnik-Manor, L. Template matching with deformable diversity similarity. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1311–1319.
23. Dekel, T.; Oron, S.; Rubinstein, M.; Avidan, S.; Freeman, W.T. Best-buddies similarity for robust template matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2021–2029.
24. Chantara, W.; Mun, J.-H.; Shin, D.-W.; Ho, Y.-S. Object tracking using adaptive template matching. *IEEE Trans. Smart Process. Comput.* **2015**, *4*, 1–9. [CrossRef]
25. Wu, T.; Toet, A. Speed-up template matching through integral image based weak classifiers. *J. Pattern Recognit. Res.* **2014**, *1*, 1–12.
26. Yoo, J.; Hwang, S.S.; Kim, S.D.; Ki, M.S.; Cha, J. Scale-invariant template matching using histogram of dominant gradients. *Pattern Recognit.* **2014**, *47*, 3006–3018. [CrossRef]
27. Sun, J.; He, F.Z.; Chen, Y.L.; Chen, X. A multiple template approach for robust tracking of fast motion target. *Appl. Math. Ser. B* **2016**, *31*, 177–197. [CrossRef]
28. Korman, S.; Reichman, D.; Tsur, G.; Avidan, S. Fast-Match: Fast Affine Template Matching. *Int. J. Comput. Vis.* **2017**, *121*, 1–15. [CrossRef]
29. Hong, C.; Zhu, J.; Yu, J.; Cheng, J.; Chen, X. Realtime and robust object matching with a large number of templates. *Multimed. Tools Appl.* **2016**, *75*, 1459–1480. [CrossRef]
30. Schneider, D.; Schwalbe, E.; Maas, H.-G. Validation of geometric models for fisheye lenses. *ISPRS-J. Photogramm. Remote Sens.* **2009**, *64*, 259–266. [CrossRef]
31. Kannala, J.; Brandt, S.S. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1335–1340. [CrossRef] [PubMed]
32. Sinha, S.N.; Frahm, J.M.; Pollefeys, M.; Yakup Genc, Y. GPU-based video feature tracking and matching, EDGE 2006. In Proceedings of the Workshop on Edge Computing Using New Commodity Architectures, Chapel Hill, NC, USA, 23–24 May 2006.
33. Fischler, M.A.; Bolles, R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **1981**, *24*, 381–395. [CrossRef]
34. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; pp. 886–893.
35. Introduction of a PGR's Ladybug3 Camera. Available online: <https://www.ptgrey.com/ladybug3-360-degree-firewire-spherical-camera-systems> (accessed on 1 May 2018).

