IPSJ Transactions on Computer Vision and Applications Vol.7 84–88 (July 2015)

[DOI: 10.2197/ipsjtcva.7.84]

Express Paper

Robust Feature Matching for Distorted Projection by Spherical Cameras

Hajime Taira[1,a)] Yuki Inoue[1,b)] Akihiko Torii[1,c)] Masatoshi Okutomi[1,d)]

Received: March 13, 2015, Accepted: April 20, 2015, Released: July 27, 2015

Abstract: In this work, we proposes a simple yet effective method for improving performance of local feature matching among equirectangular cylindrical images, which brings more stable and complete 3D reconstruction by incremental SfM. The key idea is to exiplictly generate synthesized images by rotating the spherical panoramic images and to detect and describe features only from the less distroted area in the rectified panoramic images. We demonstrate that the proposed method is advantageous for both rotational and translational camera motions compared with the standard methods on the synthetic data. We also demonstrate that the proposed feature matching is beneficial for incremental SfM through the experiments on the Pittsburgh Reserach dataset.

Keywords: feature matching, omnidirectional vision, Structure from Motion

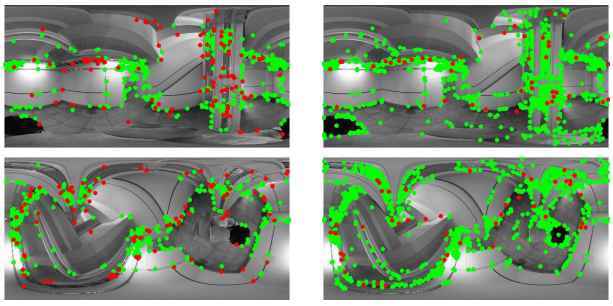
1. Introduction

3D reconstruction from imagery is no longer an infeasible task after great successes of incremental/sequential Structure-from-Motion (SfM). Incremental SfM softwares such as Bundler [14] and VisualSFM [18] have made it possible for the average users to build great 3D models using standard perspective images alone. However, to achieve stable and high quality reconstruction, these softwares require input images to have many common view fields. As a result of this the users need to capture many images (with narrow field of view cameras). One way to reduce the number of images required is to use spherical imaging systems such as Point Grey Ladybug or RICOH THETA. These cameras allow the user to capture images with large amounts of common views with little effort. One drawback of these cameras is the large distortions in the captured images as a result of the spherical projection.

This distortion makes feature matching difficult because the popular local feature detectors and descriptors are designed up to affine invariance [9], [10] (**Fig. 1**). The loss of local feature matching performance is fatal for incremental SfM since all the geometric estimation (and RANSACing) relies on correspondences obtained from local feature matching.

In this work, we propose a simple yet effective method for improving local feature matching among spherical panoramic images, namely, equirectangular images. The key idea is to explicitly synthesize images by rotating the spherical coordinate and to compute features only on the less distorted area in the rectified panorama.

[1] Department of Mechanical and Control Engineering, Graduate School of Science and Engineering, Tokyo Institute of Technology, Meguro, Tokyo 152–8550, Japan
[a)] htaira@ok.ctrl.titech.ac.jp
[b)] yinoue@ok.ctrl.titech.ac.jp
[c)] torii@ctrl.titech.ac.jp
[d)] mxo@ctrl.titech.ac.jp



(a) SIFT (b) Proposed

Fig. 1 Feature matching between equirectangular images. The equirectangular image (top) is matched to the image after the rotational motion (bottom) by a standard SIFT matching (a) and the proposed method (b). In both cases, correct and incorrect matches w.r.t. the ground truth correspondences are shown in green and red dots, respectively. The proposed feature matching (b) has 1,226 correct matches which are more than twice of that obtained by the standard SIFT (a) (535 matches).

Related work. SIFT (DoG keypoint and SIFT decription) [9] is the most well known representation of local features (patches) being invariant to scale and rotation changes. SIFT is often favoured in the large-scale SfM [1], [14], [18] because discriminability is more critical than repeatability when many images of particular landmarks/scenes are given.

MSER and Harris/Hessian-Affine [10] are other popular features that are invariant under an affine transformation. ASIFT [11] is another technique to perform feature matching robustly against affine deformations. These features work well on standard images with perspective projections but cannot handle larger distortions caused by spherical cylindrical projections.

Using multiple perspective images generated by reprojection of cylindrical panorama is a simple approach to reduce the projection distortion [16]. Contrary to the simpleness, the details are not fully investigated since quite a few parameters (focal lengths, image sizes) have to be carefully chosen for improving the pefor-

mance of matching. Besides this, rectification technics of omnidirectional stereo pairs are proposed in Refs. [4] and [7].

SIFT on the sphere (spherical SIFT) [3] takes into account the deformation caused by spherical map projections. It detects and describes SIFT-like local features on the sphere and therefore, gives invariance for changes induced by pure rotation. However, the benefit for the changes by translational motion were not thoroughly demonstrated. Our method is similar to ASIFT and spherical SIFT as it takes into account the deformations according to the spherical projections as described in next sections.

Contributions. We propose a new method for matching features between images in equirectangular projection. The method is based on simple image rectification by rotating spherical image coordinates. The method can be used as a preprocessing step for standard feature computation.

2. Feature Detection and Description with Panoramic Image Rectification

In this section, we first describe the properties of image distortion induced by spherical panoramic projection (Section 2.1). We next introduce the proposed method of feature detection and description with panoramic image rectification (Section 2.2). Finally, we describe the algorithm for matching the features between panoramic images (Section 2.3).

2.1 Properties of Image Distortion on Spherical Cameras

This section formulates the problems on image distortions by spherical cameras associated with the camera motion. We focus on the equirectuangular images captured by spherical cameras but the same approach can be applied to any wide FoV cameras such as omnidirectional and fisheye cameras.

An equirectangular image taken by a spherical camera is represented as

$$u = c(\theta + \pi), \quad v = c(\phi + \pi/2) \tag{1}$$

where u and v are the coordinates on the image (**Fig. 2** (b)), θ and ϕ denote the longitude and latitude in the spherical coordinates (Fig. 2 (a)), and c is a constant defined by the image width w and height $h = w/2$ such that $c = w/2\pi = h/\pi$ (Fig. 2). This equirectangular projection has the property that the lines of longitude and latitude are evenly spaced. Therefore, if we map circular regions (discs) of same size on the sphere to an equirectangular image plane, the discs close to the equators are less distorted but the ones close to the poles are heavily distorted (**Fig. 3**). In other word, the image region close to the equator is similar to a perspective projection whereas the region close to the poles is severely elongated. The region of interest looks very different in two images when features move from the equator to the pole as a result of camera rotation and/or translation. This reduces the matchability of local features.

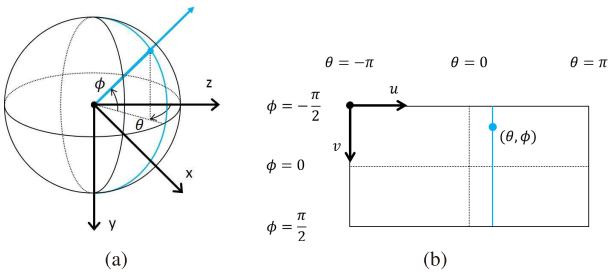


Fig. 2 Imaging with equirectangular projection. (a) Spherical coordinates. (b) Image coordinates of equirectangular projection image.

We tackle this problem by introducing a simple algorithm:

(1) We generate a set of equirectangular images by rotating the camera (spherical) coordinate system along x-axis (**Fig. 4**). We detect and describe features only from weakly distorted region in the rectified images. We call this process "Panoramic image rectification." (Section 2.2).

(2) When matching features between a pair of images, we perform step (1) for both images and use nearest neighbor search with Lowe's ratio test [9] (Section 2.3).

We next describe each step in more detail.

2.2 Panoramic Image Rectification

We generate a set of n equirectangular images using the relationship between a point on the sphere and a point on the equirectangular image in Eq. (1). The rectification is achieved by rotating the spherical coordinates (Fig. 2 (a)) and updating θ and ϕ in Eq. (1).

We generate n rectified images by rotating around an axis

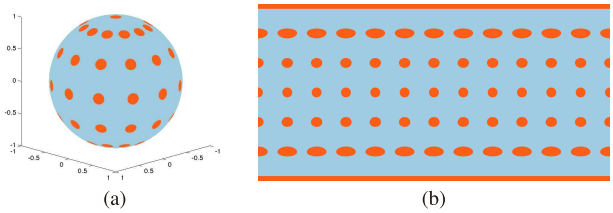


Fig. 3 Image distortion by equirectangular projection. The orange discs have the same size on the sphere (a) but mapping to the equirectantular image severely distorts them (b).

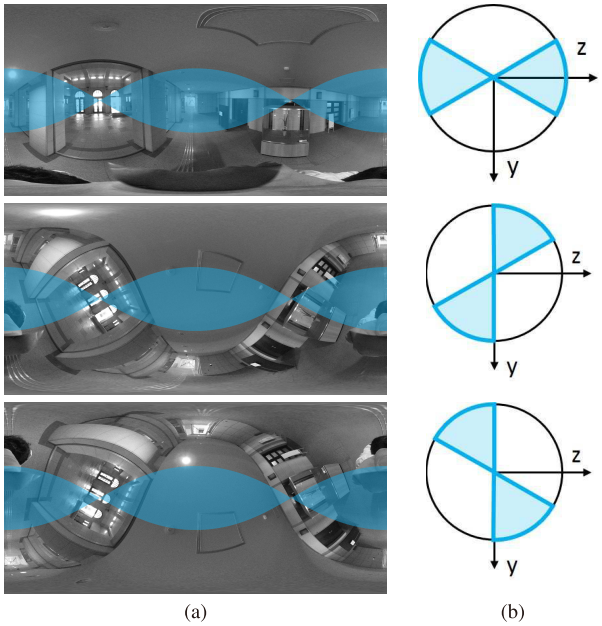
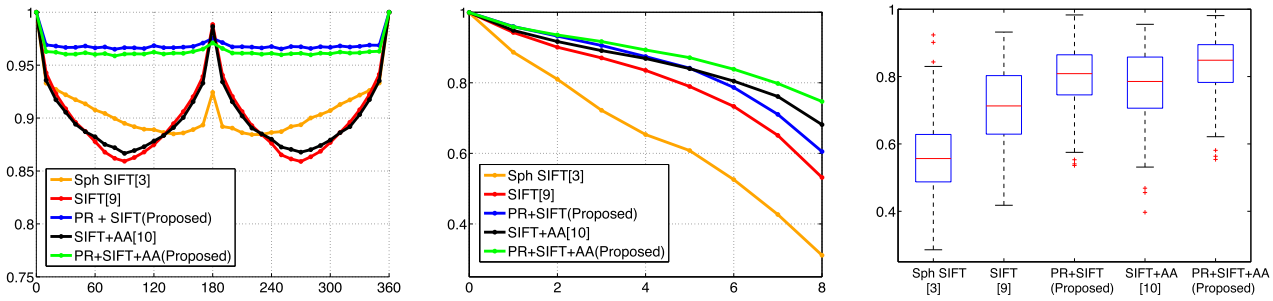


Fig. 4 Example of panoramic image rectification. The rectified equirectangular images (a) obtained by rotation of 0°, 60°, and 120° w.r.t. x-axis. We detect features from masked regions (cyan) in each rectified image. The masked regions on the rectified image correspond to the colored region (cyan) on the spherical coordinates (b).



(a) Camera rotation (b) Camera translation (c) Random camera motion

Fig. 5 Experimental result of feature matching. Graphs show the fraction of matching precision on y-axis for each rotation angle (a) or distance of cameras (b) on x-axis. The boxplot shows the statistics (y-axis) of the fraction of matching precision for each method (x-axis).

pointing at the equator (x-axis throughout the experiments) by the angle $\alpha = m\pi/n$ for $m = 0, 1, \ldots, n-1$. By doing so, all the features will eventually appear nearby the equator of the equirectangular image plane in the range of $\beta = \pi/2n$. We call this n a division number. Figure 4 (a) shows the original equirectangular image (top), the rectified image generated by rotating the spherical coordinates along x-axis at $\alpha = 60°$ (middle) and $\alpha = 120°$ (bottom). It is visually clear that the square object on the ceiling is less distorted on the rectified images.

We next define masks for detecting features in the rectified images. We detect features in the region D of each rectified equirectangular image;

$$\left\{ (\theta, \phi)^\top \in D \mid -\beta < \arctan(\tan\phi / \cos\theta) \leq \beta \right\} \tag{2}$$

where $\beta = \pi/2n$ is the angle determined by the division number n as described above. This "*wave-like*" mask is capable for finding the features from the entire sphere without duplication, as illustrated in Fig. 4 (b). Furthermore, the active regions (size of mask) for feature detection can be controlled using a single parameter n. Finally, we assemble the features detected from the region D for every rectification angle α by back-rotating the keypoint positions to the original coordinate system. Note that the keypoints are detected strictly in the masked areas to exclude duplicated detections but the masks are not used for descriptions of local patches to isolate effects by mask boundaries.

2.3 Feature Matching with Panoramic Image Rectification

Feature matching for a pair of images is fairly simple. For each image, we detect and describe features using the panoramic image rectification described in Section 2.2. Although not a necessary condition, it is reasonable to use the same division number n for both images. Empirically, we set the division number n as 6 in whole experiments. We match the features using the descriptors by searching nearest neighbors followed by Lowe's ratio test (we use the threshold of 0.7 throughout the experiments). Note that any post-processing to refine the feature matches can also be applied such as RANSAC or its variants [12].

3. Experiments

In this section we describe the experimental validation of our method. First, we evaluate performance of the proposed method using DoG keypoint detectors and SIFT descriptors (Section 3.1). We compare it on the synthetic data with the state-of-the-art baseline methods. To inspect the benefits in detail, we separately evaluate the performances for pure rotation and translation. We next evaluate the performance when combined with affine adaptation. Further, we evaluate the performences while variating the division number n and discuss the computational overhead. Finally, we evaluate the benefit of the proposed method in the incremental SfM by reconstructing camera poses on the Pittsburgh Research dataset (Section 3.2).

3.1 Evaluation for Feature Matching

Implementation details. We use 50 equirectangular images of 2,896 × 1,448 pixels. 25 of them are indoor and outdoor scenes captured by ourselves using a commercial spherical camera, RICOH THETA [*1] and 25 images randomly chosen from the Google Pittsburgh Research Data Set [*2]. Our method is implemented with MATLAB using VLfeat library [17].

Evaluation protocol. For the matches $(\mathbf{y}, \mathbf{y}')$ obtained by the nearest neighbor search followed by the ratio test, we define an evaluation function $f(\mathbf{y}, \mathbf{y}')$ with threshold T. We deem the features as correctly matched if $f(\mathbf{y}, \mathbf{y}') < T$. We then measure the performance by *precision* defined as

$$precision = \frac{\#\,correct\ matches}{\#\,feature\ matches} \tag{3}$$

Baseline methods. We compare the proposed method with SIFT [9] and spherical SIFT [3]. We use VLfeat for SIFT, and the implementation of the original authors for spherical SIFT.

Evaluation for pure rotation. We first simulate a set of motion by rotating the camera around z-axis at the interval of 10°. Notice that this axis differs to the x-axis used for the proposed panoramic image rectification. For each rotation angle, we generate 50 pairs of images by coupling the original image and the image after the rotation. Using these pairs of images, we evaluate the matching precision. To measure the correctness of the match we define f as $f_{rot}(\mathbf{y}, \mathbf{y}') = \angle(\mathbf{y}, \mathbf{R}^{-1}(\mathbf{y}'))$ where $\mathbf{R}$ is the rotation matrix obtained from the ground truth rotation. If $f(\mathbf{y}, \mathbf{y}') < 0.1°$, we deem the features as correctly matched.

Figure 5 (a) show the results of matching precision, respectively, for SIFT (SIFT, red), spherical SIFT (Sph SIFT, orange), and SIFT with the proposed panoramic rectification

[*1] https://theta360.com
[*2] Provided and copyrighted by Google.

(PR+SIFT(Proposed), blue). Our method maintains a high precision regardless of the rotation angles. whereas the baseline methods drops in performance when the appearance changes by rotation are large.

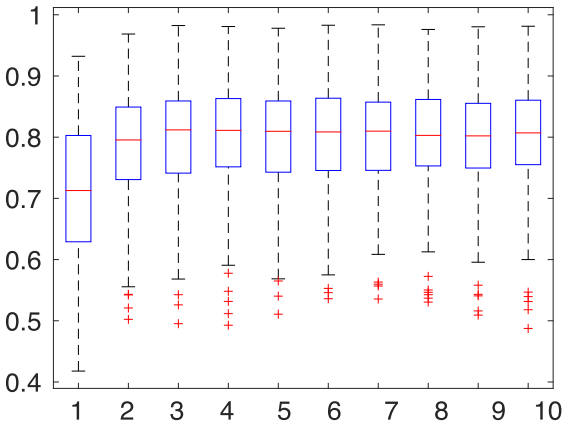


Fig. 6 Experimental result of the proposed matching with random camera motion. The boxplot shows the statistics (y-axis) of the fraction of matching precision for several division number (x-axis).

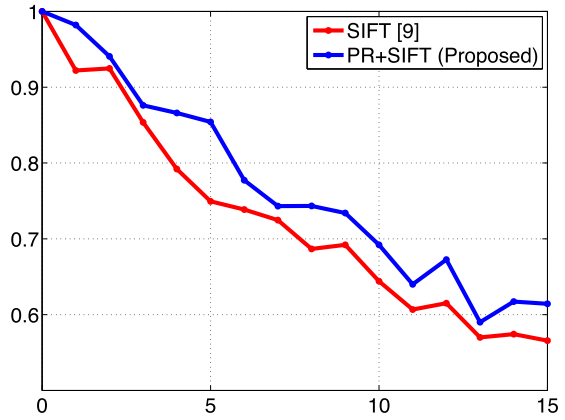


Fig. 7 A comparison of the average fraction of cameras reconstructed (y-axis) when SIFT (red) and the proposed method (PR+SIFT) (blue) are used in incremental SfM. x-axis indicates the number of frames skipped from the original sequence.

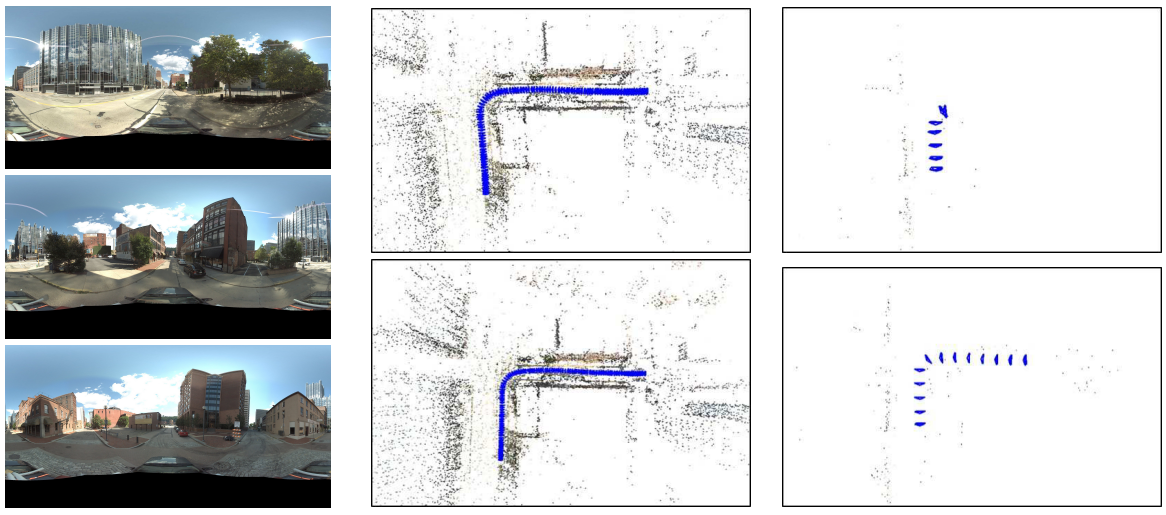
Evaluation for translation. We first prepare a cube of 10^3 to render perspective cutouts taken from indoor and outdoor scenes. We simulate a set of translation by moving the camera from the initial position $(5, 5, 1)$ to $(5, 5, 9)$ at the interval of 1. We used the softwares of Refs. [5] and [6] for rendering pictures on cubes and getting ground truth 3D points.

For each step of translation, we test 50 pairs of images by coupling the image at the initial position and the image after the motion. We evaluate precision with camera translation in the same way as camera rotation. To measure the correctness, we define f as $f_{trans}(\mathbf{y}, \mathbf{y}') = \|\mathbf{Y}(\mathbf{y}) - \mathbf{Y}'(\mathbf{y}')\|_2$ where $\mathbf{Y}$ and $\mathbf{Y}'$ are the 3D points associated to $\mathbf{y}$ and $\mathbf{y}'$ obtained from the ground truth values. We deem the matching is correct if $f_{trans}(\mathbf{y}, \mathbf{y}') < \sqrt{0.1}$.

The results in Fig. 5 (b) show that the feature matching with camera translation is more challenging for all the methods. Spherical SIFT (Sph SIFT, orange) performs worse compared to the standard SIFT (SIFT, red) and our method (PR+SIFT(Proposed), blue). Similar to the rotation results, our method maintains a higher precision than the baseline methods as the proposed method removes artifacts caused by a distorted projection.

Next, for evaluation with random camera motion, we generate 100 pairs of images by generating 2 random rotations and translations for 50 scenes, then match them with the initial panorama where the camera position is $(5, 5, 5)$ with zero rotation. The results in Fig. 5 (c) shows that the proposed method is the most effective (PR+SIFT) in comparison with baseline methods (sph SIFT and SIFT).

Affine adaptation. We also evaluate the performance of the proposed method when combined with affine adaptation [10], [17]. Figure 5 (a) shows that feature matching combined with affine adaptation (SIFT+AA, black) have no superiority to baseline SIFT (SIFT, red) at pure camera rotation. This means that affine adaptation is not able to deal with distorted projection. Figure 5 (b) and Fig. 5 (c) show the results of matching precision in camera translation. Affine adaptation constantly outperforms



(a) (b) (c)

Fig. 8 Experimental results with SfM. Examples of input equirectangular images (a) used for incremental SfM. The SfM results reconstructed with standard SIFT (top) and the proposed method with panoramic image rectification (bottom) using all the frames (b) and every 7 frames (c).

all the baseline methods when the camera is translated. Notice that the proposed method combined with affine adaptation (PR+SIFT+AA(Proposed), green) gives the best performance and the difference is significant on larger motions.

Impact of division number n. We evaluate the effect of division number on matching. We calculate the precision of the SIFT matching combined with the prposed panoramic rectification parametrized by the division number n for 1 to 10. Image pairs are generated randomly in the same manner as the evaluation with random camera motion. **Figure 6** shows that the matching with the division number 2 already results higher precision than the baseline SIFT matching ($n = 1$) and and the gains in performace saturate $n = 4$.

Computational overhead. As the proposed method requires to explicitly generate rectified images, the computational cost for the rectified image generation linearly increases by its number. On the other hand, in principle, the features should be described only the detected features in the masked area of each rectified image. Therefore, the computational overhead for detection and description is marginal.

3.2 Evaluation for Incremental SfM

We evaluate the performance of the proposed feature matching when used as a component of incremental SfM. We generated 50 sets of 100 consecutive frames randomly chosen from the 8,999 equirectangular panoramic images of 3,328 × 1,664 pixels in the Google Pittsburgh Research Data Set. We implemented an incremental SfM pipeline similar to Refs. [13], [15] and evaluate the performance by measuring the number of cameras recovered in each set.

Figure 7 shows the fraction of average number of recovered cameras over the input cameras (y-axis). To compare the performance on different baseline lengths, we skip k frames in each subset (x-axis) and run the SfM. The results clearly show that the incremental SfM combined with the proposed method (PR+SIFT(Proposed), blue) gives consistently better performance than the one with the standard SIFT matching (SIFT, red). **Figure 8** shows examples of 3D reconstruction. The blue cones in Fig. 8 (b) and Fig. 8 (c) represent the camera poses estimated by SfM. Our method can reconstruct all the cameras even when 7 frames are skipped whereas the SfM with the standard SIFT matching fails to reconstruct at the corner of camera path.

4. Conclusion

We proposed a new method for robust feature matching for distorted spherical panoramic images which is simple yet effective and can be combined with any other recent features, e.g., Refs. [2], [8]. We have demonstrated its superiority through the experiments on feature matching and on incremental SfM.

Acknowledgments This work was partially supported by the Grants-in-Aid for Scientific Research (no. 25240025, 15H05313) from the Japan Society for the Promotion of Science.

References

[1] Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M. and Szeliski, R.: Building rome in a day, *Comm. ACM*, Vol.54, No.10, pp.105–112 (2011).
[2] Ambai, M. and Yoshida, Y.: CARD: Compact And Real-time Descriptors, *ICCV*, pp.97–104 (2011).
[3] Cruz-Mota, J., Bogdanova, I., Paquier, B., Bierlaire, M. and Thiran, J.-P.: Scale invariant feature transform on the sphere: Theory and applications, *IJCV*, Vol.98, No.2, pp.217–241 (2012).
[4] Geyer, C. and Daniilidis, K.: Conformal rectification of omnidirectional stereo pairs, *Conference on Computer Vision and Pattern Recognition Workshop, 2003, CVPRW'03*, Vol.7, pp.73–73, IEEE (2003).
[5] Hassner, T.: Viewing real-world faces in 3D, *ICCV*, pp.3607–3614 (2013).
[6] Hassner, T., Assif, L. and Wolf, L.: When standard RANSAC is not enough: Cross-media visual matching with hypothesis relevancy, *Machine Vision and Applications*, Vol.25, No.4, pp.971–983 (2014).
[7] Heller, J. and Pajdla, T.: Stereographic rectification of omnidirectional stereo pairs, *CVPR*, pp.1414–1421 (2009).
[8] Leutenegger, S., Chli, M. and Siegwart, R.Y.: BRISK: Binary Robust Invariant Scalable Keypoints, *ICCV*, pp.2548–2555, Los Alamitos, CA, USA, IEEE Computer Society (2011).
[9] Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints, *IJCV*, Vol.60, No.2, pp.91–110 (2004).
[10] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T. and Gool, L.V.: A Comparison of Affine Region Detectors, *IJCV*, Vol.65, No.1-2, pp.43–72 (2005).
[11] Morel, J.-M. and Yu, G.: ASIFT: A new framework for fully affine invariant image comparison, *SIAM Journal on Imaging Sciences*, Vol.2, No.2, pp.438–469 (2009).
[12] Raguram, R., Chum, O., Pollefeys, M., Matas, J. and Frahm, J.: USAC: A Universal Framework for Random Sample Consensus, *PAMI*, Vol.35, No.8, pp.2022–2038 (2013).
[13] Sato, T., Pajdla, T. and Yokoya, N.: Epipolar geometry estimation for wide-baseline omnidirectional Street View images, *2011 IEEE International Conference on Computer Vision Workshops* (*ICCV Workshops*), pp.56–63 (2011).
[14] Snavely, N., Seitz, S. and Szeliski, R.: Modeling the World from Internet Photo Collections, *IJCV*, Vol.80, No.2, pp.189–210 (2008).
[15] Torii, A., Havlena, M. and Pajdla, T.: From google street view to 3D city models, *OMNIVIS*, pp.2188–2195 (2009).
[16] Torii, A., Sivic, J., Pajdla, T. and Okutomi, M.: Visual place recognition with repetitive structures, *CVPR*, pp.883–890 (2013).
[17] Vedaldi, A. and Fulkerson, B.: VLFeat: An Open and Portable Library of Computer Vision Algorithms (2008), available from ⟨http://www.vlfeat.org/⟩.
[18] Wu, C.: VisualSFM: A visual structure from motion system (2011), available from ⟨http://ccwu.me/vsfm/⟩.

(Communicated by *Takeshi Oishi*)