Robust 3D Reconstruction with Omni-directional Camera based on Structure from Motion

Mengcheng Song, Hiroshi Watanabe
Waseda University
3-14-9, Okubo, Shinjuku-ku, Tokyo, Japan
songmengcheng@akane.waseda.jp,hiroshi.watanabe@waseda.jp

Junich Hara
RICOH Company, Ltd.,
810 Shimo-Imaizumi, Ebina, Kanagawa, Japan
wave@nts.ricoh.co.jp

***Abstract*—3D scenes reconstruction with omni-directional camera is of utmost significance as the camera can capture the 360 degrees scene. It dramatically improves the performance of Structure from Motion (SfM) since the tracked interest points will not miss with a long base-line and sharp rotation. This paper proposes a system with preprocessing and modification based on conventional SfM pipeline for equirectangular images produced by omni-directional camera. The results show that proposed system can well estimate the accurate ego-motion and sparse 3D structure of a synthetic scene as well as a real-world scene by solving the problem of distortion and lower quality at certain area of the equirectangular images.**

Keywords—Omni-directional camera; ego-motion; 3D structure ; structure from motion;

I. INTRODUCTION

3D scene reconstruction and ego-motion estimation are of fundamental importance for robot vision, video stabilization, automotive, augmented reality, etc. In last two decades, it has been widely researched for conventional perspective cameras with narrow field-of-view (FOV). It cannot deal with the situation that the tracked points may be missed when long baseline and sharp rotation happened among the captured images.

Using omni-directional camera instead of pin-hole camera has significant advantage in applications involved with image correspondences. Reference [1] demonstrated that larger FoV could increase the precision of pose estimation and increase the robustness due to the larger visual overlap between image pairs. However, they used the projection model for fisheye and catadioptric camera, which could not be directly applied to equirectangular image.

At present, with the appearance of Ricoh Theta [2], which is a portable omni-directional camera covering FOV of 360 degrees, the focus of video field has been changed to omni-directional vision. Since the equirectangular image produced by Ricoh theta is stitched with two fish-eye images, it will introduce two problems: 1) Distortion on equirectangular projection makes correct feature detection/matching become difficult when we directly apply conventional feature detectors/descriptors which are not invariant to the radial and tangential distortion. 2) If we take into account that the fisheye image is compressed increasingly as moving away from the center, and the stitching of two fisheye images is conducted by the outermost area of the circles which leads to the problem of lower quality (lower details) at certain area in produced equirectangular image.

In this work, we propose a system with preprocessing steps and modifications on the conventional SfM pipeline to solve the problems described above.

II. PROPOSED STRUCTURE FROM MOTION SYSTEM

The modified SfM system for equirectangular images captured by omni-directional cameras with similar system of Ricoh theta is shown in Fig.1. In the following sections, we explain the theory associated with main steps of proposed approach in detail.

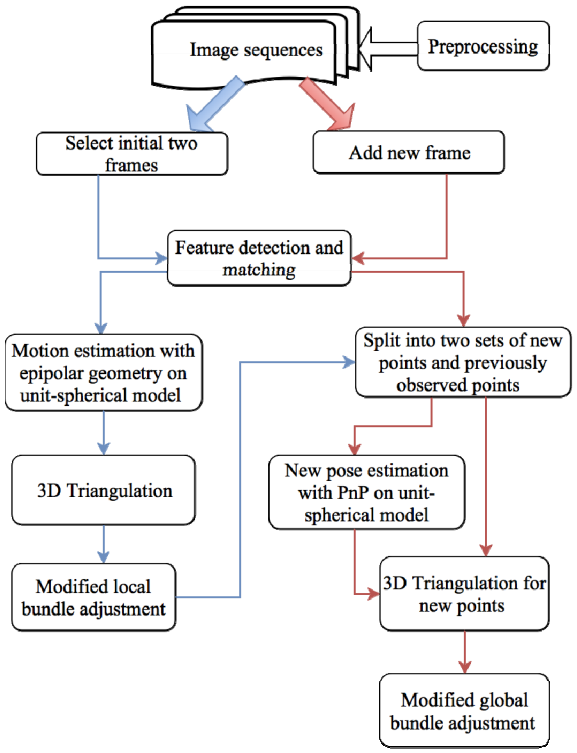


Fig. 1. Proposed SfM system

978-1-5386-2615-3/18/$31.00 ©2018 IEEE

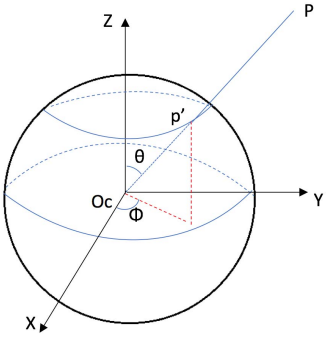


Fig. 2. Unit spherical model

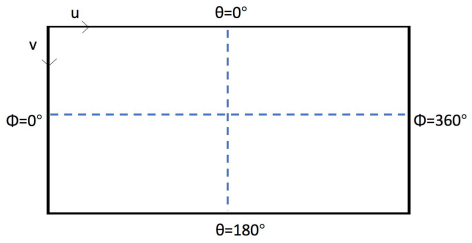


Fig. 3. Equirectangular projection

A. Unit Spherical Model

As depicted in Fig.2, suppose there is a 3D point P_i with respect to the coordinate system of the i-th camera. Let $P_i = [X, Y, Z]^T$, it is projected to p´ on the surface of the unit sphere, then we have the relationship $p' = P_i / \| P_i \|$, and $p' = [\sin\theta \cos\phi,\ \sin\theta \sin\phi,\ \cos\theta\ ]^T$ according to the unit spherical coordinates. Omnidirectional camera like Ricoh theta, can produce equirectangular image shown in Fig.3. Equirectangular projection is a longitude-latitude projection which unfold and flatten the 3D spherical image to 2D pixel domain, then we have the corresponding point $p = [u, v]^T$ related to p´ by:

$$\theta = \pi * v/height$$
$$\phi = 2\pi * u/width \quad (1)$$

where height and width stand for the height and width of resolution of source image respectively.

B. Preprocessing and Correspondence of Feature Points

Cubic mapping can project the omnidirectional image to six patches same as the images captured by perspective camera with different angle of view, it has been used in computer graphics for a long time and recently used in preprocessing for compression of equirectangular images [3]. By this method, the distortion which makes motion vector estimation not suitable can be solved. Thus, we convert our source images into cube map for improving performance of feature matching.

For tackling the problem of lower quality at certain area introduced by stitching and interpolation when obtaining the equirectangular image, we design a system presented in Fig.4. Instead of capturing once at each position, we obtain two

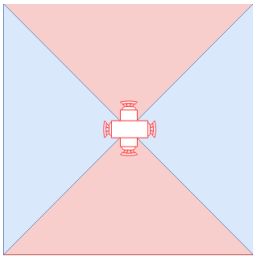


Fig. 4. Rotated system for preprocessing of merging

images with pure rotation at each spot and calculate the rotated angles via detecting most robust matched point by increasing the ratio of distance between pairs of matched points described in [4]. For these two images with pure rotation, the difference in locations of corresponding feature points only occur along the u axis. We can simply use the pixel distance of u to rotate one of them, and merge it with the other.

After preprocessing steps described above, we use affine invariant methods to conduct correspondence of feature points. ASIFT [5] apply the affine transformation space sampling technique to improve SIFT and aims to achieve affine invariance to detect and match more robust pairs of corresponding points between the images captured with different angle of view. As depicted in Fig.5,6, we find that feature correspondence between equirectangular images benefit a lot from combining cube mapping with affine invariant detector and descriptor, since six faces in cube could be considered as perspective images captured with different angle of view.

C. Two-view Structure From Motion

From initially selected first two frames, we only have the information of corresponding points coordinates in preprocessed cubic images to estimate relative pose. This is a typical 2D-2D motion estimation problem solved by epipolar constraint followed by 3D triangulation described in [6]. Here we briefly review this method and introduce the adjustment for dealing with equirectangular images. Epipolar constraint is described as function:

$$p_2^T K^{-T} E K^{-1} p_1 = 0 \quad (2)$$
$$E = [T]_\times R \quad (3)$$

where, K is intrinsic matrix which describes the projection model from 3D domain to 2D pixel domain. E stands for the Essential matrix. T and R stands for translation vector and rotation matrix respectively, and $[\ ]_\times$ is an operator for cross product.

We conduct the eight-point method [6] on 3D unit-spherical domain instead of 2D pixel domain since we have already known the 3D unit vector of each points in equirectangular image. Furthermore, the error calculated by the pixel distance between re-projected point with measured point will be adjusted to fit the 3D spherical domain. We use angular error for error measurement, which will be used in RANSAC [7] to extract best Essential matrix as well as remove outliers. After that, translation vector T and rotation matrix R will be decomposed from E.

After we get the relative pose between two images, the relative projection matrix [R T] can be composed. Then we adjust 3D triangulation problem using unit spherical coordinates p′ as observed points, then the linear triangulation equation can be derived as:

$$\lambda p' = RP_w + T \tag{4}$$

$$\begin{bmatrix} [p_1']_\times & [R_1T_1] \\ \vdots & \vdots \\ [p_i']_\times & [R_iT_i] \end{bmatrix} \begin{bmatrix} P_w \\ 1 \end{bmatrix} = 0 \tag{5}$$

where P_w is the real 3D coordinate respect to the world coordinate system, i represents the i-th camera which can observe that point and λ is an unknown scale factor. P_w can easily be calculated from (5) using singular value decomposition.

To solve the ambiguity of R and T, we triangulate all of the corresponding points then check the depth of those points positive or not as in [6]. This is performed to select the only combination of R and T with the constraint that all of the observed points should be in front of the pin-hole camera. However, it cannot be applied for omni-directional camera, since there is no constraint for enforcing the observed points to be in front of omni-directional camera. Instead, we know that reconstructed 3D points and corresponding measured points on sphere surface should have the same direction. Therefore, we use this property to vote for best R and T.

D. Multi-view Structure From Motion

For adding new images into the structure, we first detect and match feature points between new image and previous one, then split these points into two sets: first set maintains those points observed before, second set contains the new points. We can use the corresponding points of the first set to calculate projection matrix [R T] by adjusting the conventional 6 points PnP [6] with unit spherical coordinates. After we get the pose (R and T) of new frame, we can reconstruct the points of the second set in the same pattern as before.

Bundle adjustment is a method for nonlinear refinement of large scale reconstructed 3D points and poses parameters of all the frames. Conventional SBA [8] uses error of pixel distance between re-projected points and measured points in 2D pixel domain. Here, we adjust the SBA by using the angular error as explained before. This step is conducted each time when we align new frames to the graph.

III. Experiments and Results

In the experiment, we implement the modified structure from motion system in python for omnidirectional camera Ricoh theta S to reconstruct sparse 3D scene points and recover relative poses. We conduct this experiment in both synthetic scene and real-world scene. For building synthetic scene, we use the model provided by [1], and render 3 equirectangular images with the resolution of 2160x1080 pixels from different poses by Blender [9]. To verify our proposed approach, we set the Ricoh theta S on tripod, and capture 3 pairs of images with different poses in corridor scene, each pair include two frames with pure rotation for preprocessing of merging. We use high quality mode to produce equirectangular images with resolution of 5376x2688 pixels.

Detection and matching of feature points are implemented with OpenCV [10]. We compare the performance of feature correspondence using SIFT and Affine SIFT, on both synthetic and real-world images. Number of matched feature points with Lowe ratio described in [4] and correct matched points after removing outliers with RANSAC are compared in raw equirectangular images and preprocessed images. We set angular error of threshold as 0.087 rad for RANSAC to remove outliers. Due to the limited space to show our results, we present the comparison between raw image and cubic mapping in synthetic data as depicted in TABLE I. and Fig.5,6. We then present the benefit of merging of images in real world data as shown in TABLE II. and Fig 7. The results show that with the combination of preprocessing and affine variant SIFT, the number of matched as well as correct matched points after removing outliers are increased dramatically. Additionally, we notice that RANSAC gives faster and better results for estimating essential matrix E by the proposed approach, which gives more accurate pose estimation. Here we use Euler angles (rotated angles around X,Y,Z axis) to represent rotation matrix R, and set the relative translation vector T with unit length due to the scalar problem in monocular structure from motion.

Fig.8 and TABLE III. show the results of sparse 3D reconstruction and poses estimation of synthetic scene. Our proposed system works well with equirectangular images.

IV. Conclusion

We proposed a modified SfM system with unit-spherical model for omnidirectional images. Point correspondences are improved by preprocessing steps including cubic mapping and pure rotation merging to address problems of distortion and lower quality at certain area. Additionally, combined with ASIFT, the number of correct matched feature points can be dramatically increased. Thus, it could be applied to other applications involved with feature correspondences in omni-directional images. For future work, we will implement the dense 3D reconstruction, and test proposed system in more complex outdoor scenario.

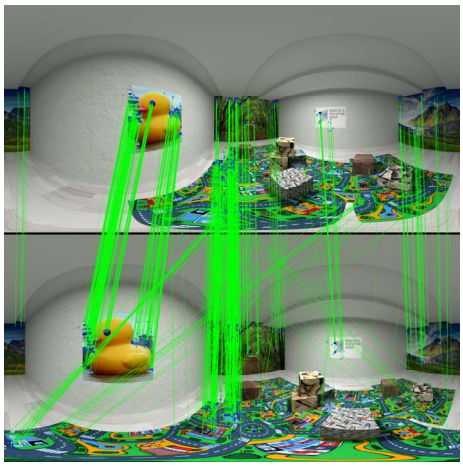


Fig. 5. Feature correspondence with ASIFT after RANSAC: Raw equirectangular images.

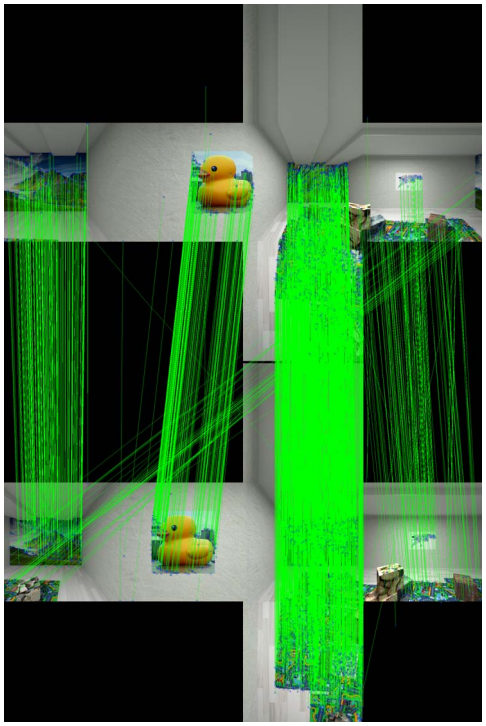


Fig. 6. Feature correspondence with ASIFT after RANSAC: Preprocessed images

TABLE I. FEATURE CORRESPONDENCE IN SYNTHETIC DATA

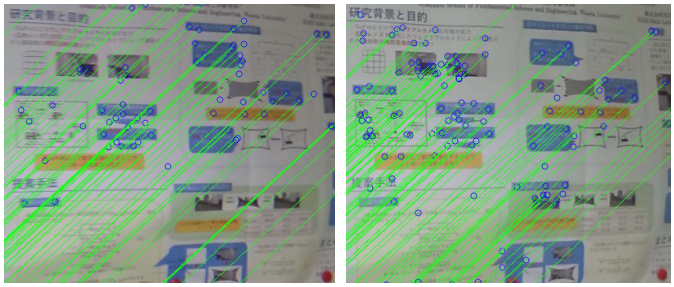
	Number of matched points in synthetic data			
	Raw equirectangular image		***Preprocessed image***	
	Before RANSAC	***After RANSAC***	***Before RANSAC***	***After RANSAC***
SIFT	2296	1174	2561	2133
ASIFT	3658	1225	8041	7394

TABLE II. FEATURE CORRESPONDENCE IN REAL WORLD DATA

	Number of matched points in real world data			
	Without merging		***With merging***	
	Before RANSAC	***After RANSAC***	***Before RANSAC***	***After RANSAC***
SIFT	588	271	665	390
ASIFT	3871	1920	3863	2246

TABLE III. MOTION ESTIMATION IN SYNTHETIC DATA

	Tx	***Ty***	***Tz***	***α [rad]***	***β [rad]***	***γ [rad]***
Frame 1	0.999	6.8 e-4	0.005	0.00	0.002	0.001
Ground truth	1.000	0.000	0.000	0.00	0.000	0.000
Frame 2	1.955	1.3 e-4	0.084	-0.02	0.010	0.003
Ground truth	2.000	0.000	0.000	0.00	0.000	0.000



(a) Before merging (b) After merging

Fig. 7. Feature correspondences with SIFT after RANSAC in merged area of real scene

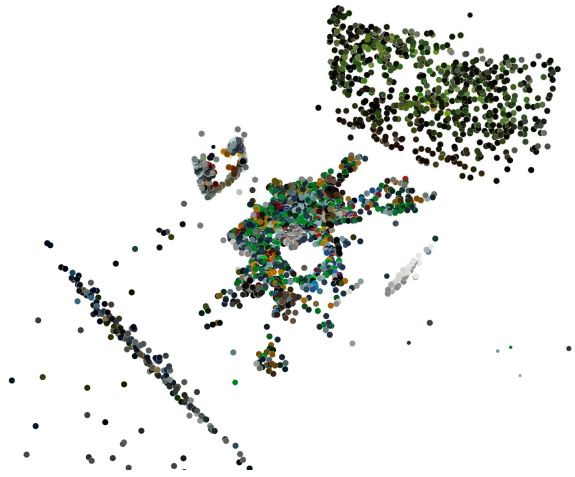


Fig. 8. Sparse 3D reconstruction from synthetic images

REFERENCES

[1] Zhang, Zichao, et al. "Benefit of large field-of-view cameras for visual odometry." *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016.

[2] https://theta360.com.

[3] https://code.facebook.com/posts/1638767863078802/under-the-hood-building-360-video/

[4] Lowe, David G. "Distinctive image features from scale-invariant keypoints." *International journal of computer vision*60.2 (2004): 91-110.

[5] Yu, Guoshen, and Jean-Michel Morel. "ASIFT: An algorithm for fully affine invariant comparison." *Image Processing On Line* 1 (2011): 11-38.

[6] Szeliski, Richard. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[7] Fischler, Martin A., and Robert C. Bolles. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography." *Communications of the ACM* 24.6 (1981): 381-395.

[8] Lourakis, Manolis IA, and Antonis A. Argyros. "SBA: A software package for generic sparse bundle adjustment." *ACM Transactions on Mathematical Software (TOMS)* 36.1 (2009): 2.

[9] https://www.blender.org/

[10] https://opencv.org