Highly Accurate and Fast Two-view Pose Estimation by Fast Reduction of Spherical Image Distortion Effects

Taisei Ando[1*], Junwoon Lee[1], Mitsuru Shinozaki[2], Toshihiro Kitajima[2], Qi An[1] and Atsushi Yamashita[1]

[1]Department of Human and Engineered Environmental Studies, The University of Tokyo, 277-8563, Japan ({ando, leejunwoon, anqi, yamashita}@robot.t.u-tokyo.ac.jp)
[2]Technology Innovation R&D Dept.II, Research & Development Headquarters, KUBOTA Corporation, 590-0908, Japan ({mitsuru.shinozaki, toshihiro.kitajima}@kubota.com) * Corresponding author

Abstract: Spherical images have a wide field of view and are effective for pose estimation, but they have the problem of inherent distortion. Existing methods to reduce the effects of distortion significantly increase computation time and cannot be used for real-time applications. We propose a method that enables accurate and fast feature point-based two-view pose estimation using spherical images by reducing the effect of distortion in equirectangular images. In our approach, one image is generated by rotating an equirectangular image, and feature point detection and descriptor extraction are performed from the two images: the original image and the generated image. The information is then integrated by adopting the least distorted regions of the two images. Our approach works faster than existing distortion reduction methods because of the small number of projection planes. In experimental evaluation, it was shown that our proposed method is faster and equally accurate compared to state-of-the-art methods in pose estimation.

Keywords: Spherical image, Feature point detection, Image projection, Pose estimation

1. INTRODUCTION

Pose estimation is a technique to determine one's position in a 3-D space. Enhancing the accuracy and speed of pose estimation contributes to the automation of movement in applications such as autonomous vehicles and robotic systems. In pose estimation, Global Navigation Satellite System (GNSS), Light Detection and Ranging (LiDAR) and cameras are mainly used. However, GNSS faces challenges in environments where satellite signals are obstructed, such as in indoors or forests [1]. Additionally, LiDAR is vulnerable to strong light conditions and adverse weather [2]. Cameras can provide valuable information even in environments where GNSS and LiDAR are ineffective, making camera-based pose estimation methods a complementary or alternative approach to those based on GNSS and LiDAR. Among camera-based pose estimation methods, feature-based approaches are a powerful option in resource-limited situations due to their lower computational costs.

In feature-based pose estimation, increasing the field of view enhances the accuracy of the estimation [3]. Therefore, using cameras that can capture a wide range can potentially lead to more accurate self-localization. Spherical cameras, with a 360-degree field of view, can capture information from all directions in a single shot. However, when images captured by a spherical camera are projected onto a flat surface to apply conventional image processing techniques, peculiar distortions occur. The difference between a conventional image and an equirectangular image is shown in Fig. 1. While equirectangular images can encapsulate a 360-degree view in a single image, images becomes significantly distorted towards the upper and lower parts [4]. Due to these distortions, using standard image processing techniques for pose estimation can lead to decreased accuracy [5].

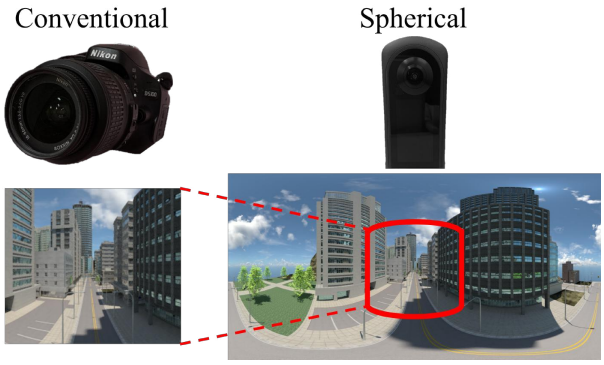

Fig. 1.: Comparison of conventional and spherical cameras

There are some studies on reducing the effects of distortion in spherical images. One of the simplest methods to reduce distortion is the cubemap, which projects a spherical image onto a cube [6]. However, this method has the problem that the image becomes discontinuous at the boundary of each face of the cubemap, making it difficult to acquire appropriate information in the area near the boundary. [4] has proposed tangent plane projection, which effectively reduces distortion in spherical images more than cubemap. However, this method also increases the computational time for feature point detection and descriptor extraction, compromising real-time performance. [7] has proposed a method that involves rotating equirectangular images by 0 degrees, 60 degrees, and 120 degrees to facilitate feature point detection and extraction in areas with minimal distortion. However, this approach disrupts the continuity at the equirectangular image's left and right edges.

An alternative method for addressing distortion involves directly processing spherical images [8–10]. This

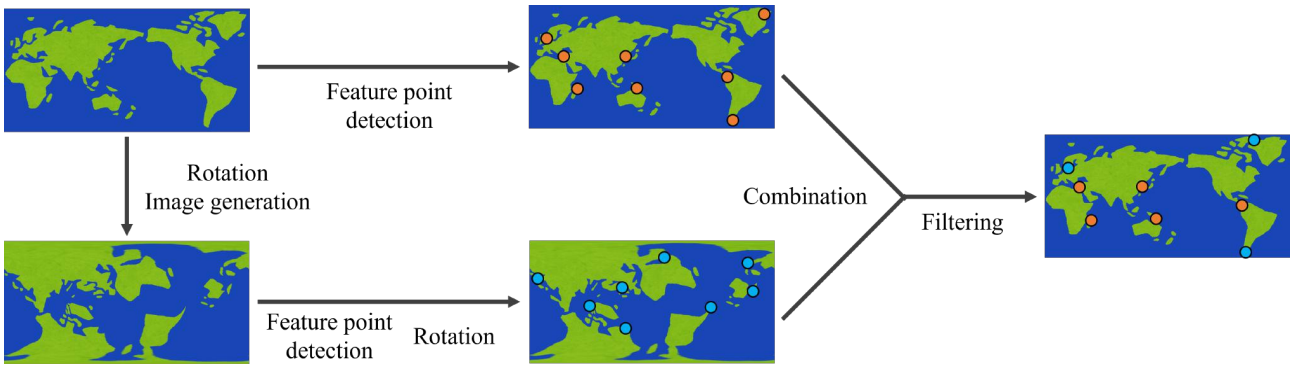

Fig. 2.: Overview of proposed method

strategy is beneficial as it obviates the need to convert images into equirectangular or other formats. However, considering that learning-based methods have performed well in matching tasks with perspective projection images [11–13], there is room for improvement in accuracy over these heuristic approaches.

Therefore, we newly propose a method that enables accurate and fast pose estimation using spherical images by reducing the effect of distortion in equirectangular images.

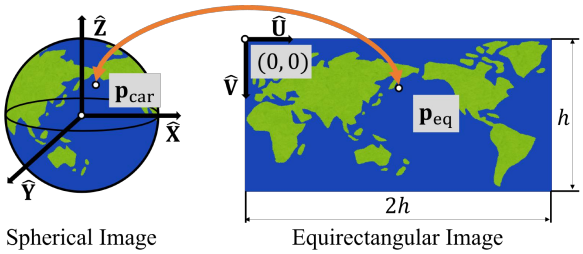

Fig. 3.: Correspondence between 3-D Cartesian coordinates and equirectangular image coordinates

2. PROPOSED METHOD

Where a spherical image is represented by a rectangular image, the image has a peculiar distortion. Notably, as one approaches the top and bottom of the image, these distortions become more pronounced, leading to several challenges during feature point detection and extraction:

- Due to distortions in spherical images, points that are not typically detected as feature points in conventional images are detected as feature points.
- The extraction of feature descriptors is adversely affected by these distortions, thereby reducing the accuracy of feature point matching.
- Although projecting spherical images onto the tangent plane of the polyhedron can mitigate these distortions, it significantly increases computational time.

Equirectangular images, while offering a wide field of view, become more distorted towards the top and bottom. Feature point detection and extraction accuracy are degraded in areas of severe distortion. To address these challenges, our method leverages the characteristic of equirectangular images where distortion decreases closer to the equator of the image.

Our proposed method achieves distortion reduction using fewer projection planes compared to existing approaches, thereby accelerating the process.

2.1 Overview of the Proposed Method

The proposed method effectively mitigates the effects of distortion by rotating the equirectangular image in 3-D space and performing a single projection. An overview of the proposed method is illustrated in Fig. 2. Initially, the coordinates on the equirectangular image are converted into spherical coordinates. These spherical coordinates are then rotated and converted back into equirectangular image coordinates. This transformation is applied to the image to generate a rotated image of the original image.

Subsequently, feature point detection and descriptor extraction are performed on both the original and the generated images. The coordinates of feature points detected in the generated image are transformed to the coordinate system of the original image. Finally, feature points from areas of the original image with small distortion are retained, while feature points from areas of high distortion are replaced with those from the generated image. This technique allows for rapid and effective reduction of distortion impacts on feature point detection and descriptor extraction in equirectangular images.

2.2 Generation of a new equirectangular image with rotated viewpoint

The relationship between the coordinate system of equirectangular images and the 3-D Cartesian coordinate system is illustrated in Fig. 3. The coordinates from equirectangular images can be converted to the 3-D Cartesian coordinate system using the following equation:

$$\mathbf{p}_{\text{car}} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \sin\frac{\pi u}{h} \sin\frac{\pi v}{h} \\ \cos\frac{\pi u}{h} \sin\frac{\pi v}{h} \\ \cos\frac{\pi v}{h} \end{bmatrix}. \quad (1)$$

Conversely, the coordinates from the 3-D Cartesian coordinate system can be converted back to the equirectangular image coordinate system using the following equa-

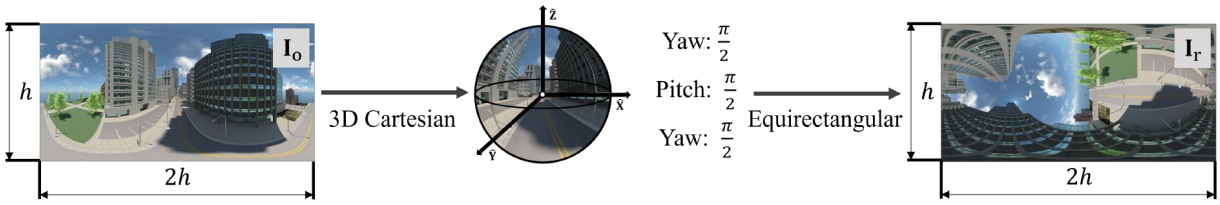

Fig. 4.: Generation of a new equirectangular image with rotated viewpoint

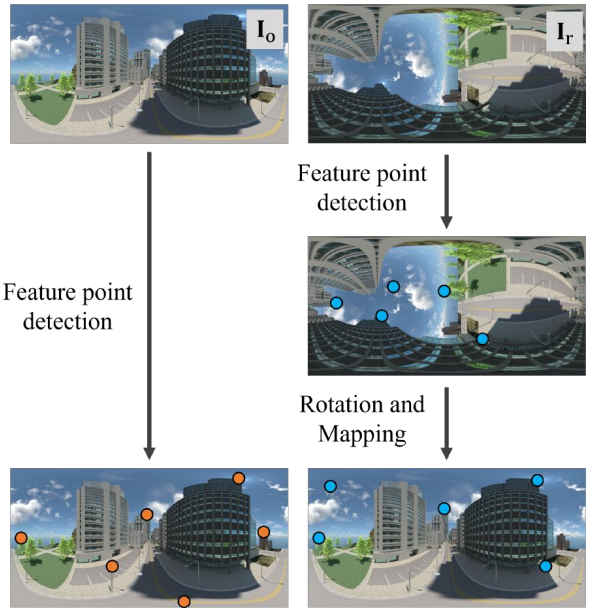

Fig. 5.: Feature point detection and descriptor extraction of two images

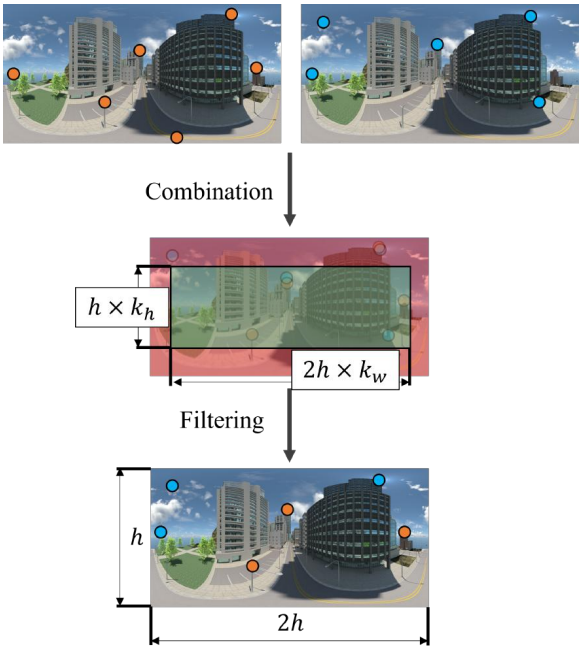

Fig. 6.: Integration of feature points of two images

tion:

$$\mathbf{p}_{\text{eq}} = \begin{bmatrix} u \\ v \end{bmatrix} = \frac{h}{\pi} \begin{bmatrix} \tan^{-1} \frac{x}{y} \\ \cos^{-1} z \end{bmatrix}. \tag{2}$$

Here, a point $\mathbf{p}_{\text{eq}} = (u, v)^{\text{T}}$ on the equirectangular image is mapped to $\mathbf{p}_{\text{car}}$ in the 3-D Cartesian coordinate system. Given that the image's vertical and horizontal resolutions are denoted as h and $2h$ respectively, the ranges of u and v are defined as $0 \leq u \leq 2h - 1$ and $0 \leq v \leq h - 1$.

Then, considering the aforementioned relationships, the transformation of equirectangular images into 3-D Cartesian coordinates is computed. The image is subjected to sequential rotations around the yaw and pitch axes by $\frac{\pi}{2}$ as depicted below:

$$\mathbf{R}_{\text{Y}}(\phi) = \begin{pmatrix} \cos\phi & 0 & \sin\phi \\ 0 & 1 & 0 \\ -\sin\phi & 0 & \cos\phi \end{pmatrix}, \tag{3}$$

$$\mathbf{R}_{\text{Z}}(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{4}$$

The rotated equirectangular image $\mathbf{I}_{\text{r}}$ is then obtained by applying Equation (2). An illustration of the process of rotating a spherical image in 3-D space and outputting an equirectangular image is shown in Fig. 4. This operation allows for the mapping of pixels from the peripheral areas of the original image $\mathbf{I}_{\text{o}}$ to the central regions of the equirectangular image.

2.3 Feature Point Detection and Coordinate Inversion

Feature point detection and descriptor extraction performed on both $\mathbf{I}_{\text{o}}$ and $\mathbf{I}_{\text{r}}$ are illustrated in Fig. 5. In our proposed method, feature point detection and descriptor extraction are carried out on these images using conventional image processing techniques.

Subsequently, the feature point coordinates obtained from $\mathbf{I}_{\text{r}}$ undergo a transformation. Initially, the coordinates on the equirectangular image are converted into 3-D Cartesian coordinates using Equation (1). Then, the coordinates are subjected to sequential rotations of $-\frac{\pi}{2}$ around the yaw, pitch, and yaw axes, as defined by Equations (3) and (4). This procedure allows for the transformation of feature point coordinates back to their positions in $\mathbf{I}_{\text{o}}$.

2.4 Filtering of Feature Points

As discussed in Section 2.3, performing feature point detection on both $\mathbf{I}_{\text{o}}$ and $\mathbf{I}_{\text{r}}$ can result in multiple detections of the same point in 3-D space, which diminishes the accuracy of matching. To address this issue, our

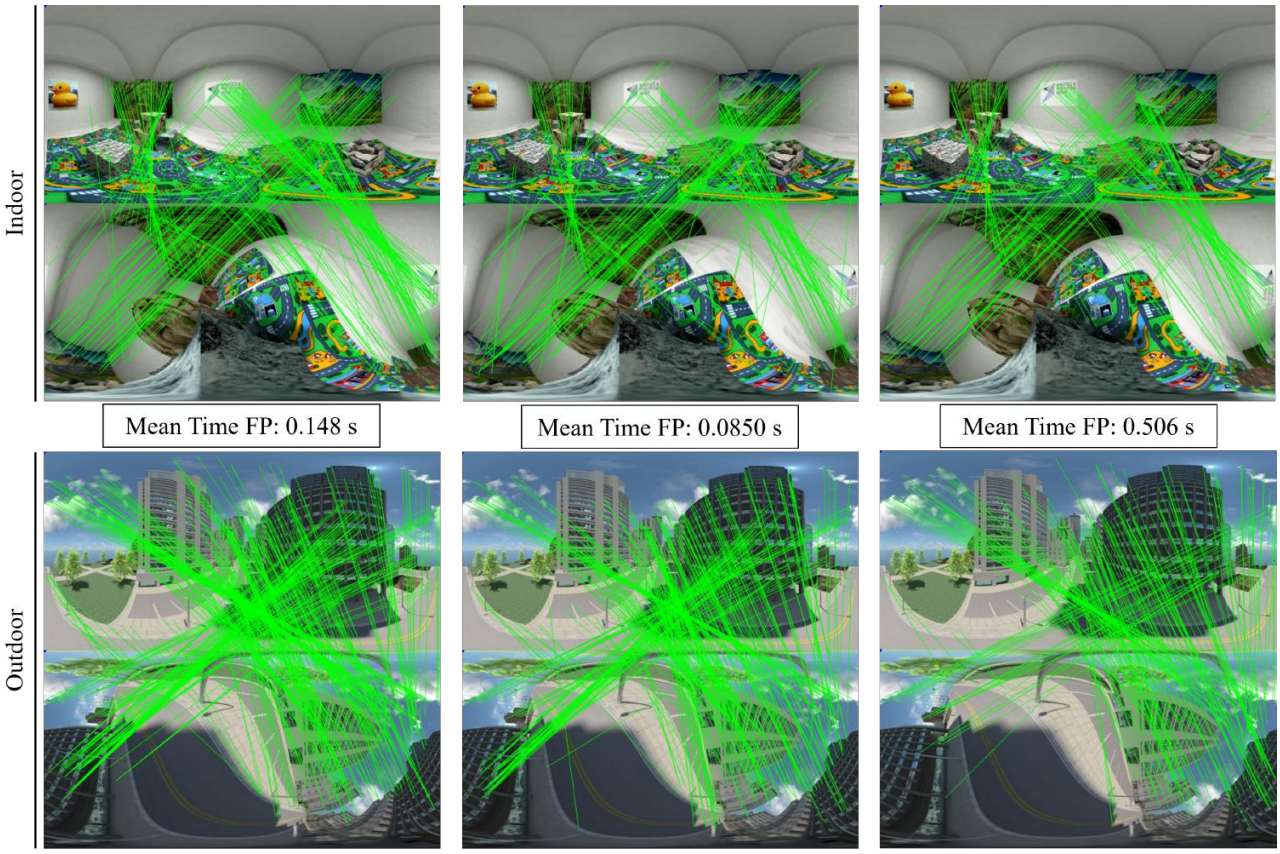

Fig. 7.: Qualitative keypoint matches: The green lines indicate matches that were classified as inliers by RANSAC. The feature detection method employed is SuperPoint. Mean Time FP is mean time to detect feature points.

method applies a filtering process using the equirectangular image coordinates.

Feature point filtering in our proposed method is illustrated in Fig. 6. Here, k_h and k_w are the proportionality constants that determine the size of the cropped area relative to the original image dimensions. Feature points located around the center of $\mathbf{I}_\mathrm{o}$ retain their original features and descriptors, while those outside this central area adopt the features and descriptors from $\mathbf{I}_\mathrm{r}$.

3. EXPERIMENT

3.1 Experimental Setup

To validate the effectiveness of the proposed method, we performed two-view pose estimation using pairs of equirectangular images. The experiment was conducted as follows:

1. Feature detection and extraction: we applied both existing and proposed methods to preprocess equirectangular images. Then, features are detected and extracted from the preprocessed images using existing methods, followed by feature point matching.
2. Pose estimation using the eight-point algorithm: the matched features were used to estimate the pose using the eight-point algorithm [14]. During this process, SK nonlinear optimization [15] and the Random Sample Consensus (RANSAC) algorithm [16] were employed to mitigate the impact of outliers.
3. Accuracy evaluation: the rotation matrices and translation vectors were calculated, and their accuracy was evaluated using the Mean Absolute Error (MAE). Additionally, the calculation speed was assessed based on the feature detection and matching times.

For the pose estimation, the relative pose was calculated as rotation matrices and translation vectors from the feature point matching results. The angular rotation error θ_rot was defined as below:

$$\theta_\mathrm{rot} = \arccos\left(\frac{\mathrm{trace}(\mathbf{R}_1^\mathrm{T}\,\mathbf{R}_2) - 1}{2}\right), \tag{5}$$

where $\mathbf{R}_1$ and $\mathbf{R}_2 \in \mathrm{SO}(3)$ are rotation matrices. Similarly, the angular translation error θ_trans was defined as below:

$$\theta_\mathrm{trans} = \arccos\left(\frac{\mathbf{t}_1 \cdot \mathbf{t}_2}{\|\mathbf{t}_1\|\|\mathbf{t}_2\|}\right), \tag{6}$$

where $\mathbf{t}_1$ and $\mathbf{t}_2 \in \mathbb{R}^3$ are transration vectors.

In the experiment, the proportionality constants k_h and k_w mentioned in Section 2.4 were set to $\frac{2}{3}$ and $\frac{7}{8}$, respectively. We compare the proposed method with a method

Table 1.: Two-view pose estimation: We report the mean time to detect feature points (Time FP), mean time to match feature points (Time MC), mean number of feature points (NUM FP), mean absolute error of rotation (R MAE) and mean absolute error of translation (T MAE). Here, the units for Time FP and Time MC are seconds (s), while the units for R MAE and T MAE are degrees (deg).

	Overall		Indoor			Outdoor		
Method	Time FP	Time MC	Num FP	R MAE	T MAE	Num FP	R MAE	T MAE
SIFT	0.0862	0.0134	1869	9.82	18.9	2704	10.4	23.8
tSIFT	0.567	0.00979	1078	8.13	10.3	1696	5.80	15.7
pSIFT	0.150	0.0107	1739	7.44	11.7	2660	7.12	18.2
ORB	0.0432	0.0558	7314	15.3	27.7	8387	18.6	43.8
tORB	0.223	0.0329	5527	9.24	15.4	6618	9.73	23.3
pORB	0.0736	0.0627	7999	12.7	21.2	9242	10.5	27.5
SuperPoint	0.0850	0.0164	1491	7.77	11.0	2190	9.43	16.2
tSuperPoint	0.506	0.0117	815	6.49	8.17	1221	7.53	14.1
pSuperPoint	0.148	0.0125	1348	5.86	8.11	2054	7.47	11.5

that detects feature points directly from equirectangular images without distortion correction and a method that uses tangent plane projection [4]. Furthermore, the combination of the k-nearest neighbors and Lowe's ratio method [17] are used for feature matching.

In this study, we adopt several feature detection methods, including Scale-Invariant Feature Transform (SIFT) [17], Oriented FAST and Rotated BRIEF (ORB) [18], and SuperPoint [11]. SIFT and ORB are implemented in OpenCV, and SuperPoint is implemented by a open source[1].

The experiments were conducted on Ubuntu 20.04.6 LTS, equipped with a 12th Gen Intel® Core™ i9-12900 processor, 64GB DDR4 memory, and an NVIDIA RTX A4500 GPU.

The dataset [5] used for the experiments consisted of a total of 900 image pairs generated from 9 virtual environments, which were constructed with open source 3-D modeling software such as Blender[2] and UnrealCV[3]. The ground truth for rotation and translation was predetermined during the image generation process. The resolution of the images was 1024×512.

3.2 Results and Discussion

Examples of the matching results for pose estimation between two viewpoints are shown in Fig. 7. These figures demonstrate that our proposed method increases the number of inliers in both indoor and outdoor scenes.

The quantitative evaluation of the pose estimation between two viewpoints is presented in Table 1. The experimental values derived directly from the equirectangular images are denoted as SIFT, those detected using the tangent plane projection method [4] as tSIFT, and those using our proposed method as pSIFT.

In terms of computation time, when using the proposed method, the time required for feature point detection and descriptor extraction was approximately twice and 0.3 times that required when using the tangent plane projection method and methods without distortion reduction, respectively. Moreover, the computational efficiency of the proposed method can be further improved by omitting feature point detection in regions excluded from the filtering process, as outlined in Section 2.4 . Additionally, there is no significant variance in feature matching time between the methods.

When using SIFT/SuperPoint, the number of feature points detected is ordered as follows: the equirectangular image, the proposed method, and tangent plane projection. The higher count of feature points in the equirectangular image is attributed to distortions that cause regions, which are normally not identified as features in conventional images, to be detected as such. Conversely, when using ORB, the sequence is the proposed method, the equirectangular image, and tangent plane projection. The reason why the proposed method detects more feature points than the equirectangular image with ORB is likely related to ORB's correlation between the number of pixels in the image and the number of feature points.

In terms of accuracy, ORB and SIFT generally achieve higher precision than without distortion reduction and lower precision compared to using the tangent plane projection method. On the other hand, the proposed method achieves higher precision when using SuperPoint compared to the tangent plane projection method. The combination of SuperPoint and the proposed method consistently resulted in the highest precision. Furthermore, the accuracy of our proposed method can be enhanced by optimizing the determination of k_h and k_w in Section 2.4.

4. CONCLUSION

In this paper, we propose a method that enables accurate and fast feature point-based two-view pose estimation using spherical images by reducing the effect of distortion in equirectangular images. In our method, spherical image is generated by rotating it in 3-D space, and feature point detection and descriptor extraction are performed on the original and generated images. By employing the information from the center of the two images,

[1] https://github.com/rpautrat/SuperPoint
[2] https://www.blender.org/
[3] https://unrealcv.org/

information is obtained from the area where the effect of distortion is relatively small. The proposed method was evaluated on spherical image pairs dataset of indoor and outdoor scenes. In experimental evaluations, the proposed method achieved comparable pose estimation accuracy as the tangent plane projection method while requiring only about 30% of the computation time. As future work, we aim to achieve further acceleration by refining the filtering component of the proposed method. Additionally, we will conduct more ablation studies to evaluate the effectiveness of the proposed approach.

REFERENCES

[1] C. Gioia, "GNSS Navigation in Difficult Environments: Hybridization and Reliability," *Ricerche di Geomatica 2014*, vol. 1, 2014.

[2] A. Carballo, J. Lambert, A. Monrroy, D. Wong, P. Narksri, Y. Kitsukawa, E. Takeuchi, S. Kato, and K. Takeda, "LIBRE: The Multiple 3D LiDAR Dataset," *Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1094–1101, 2020.

[3] A. Rituerto, L. Puig, and J. Guerrero, "Visual SLAM with an Omnidirectional Camera," *Proceedings of the 20th International Conference on Pattern Recognition (ICPR2010)*, pp. 348–351, 2010.

[4] M. Eder and J.-M. Frahm, "Convolutions on Spherical Images," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR2019W)*, June 2019.

[5] J. Murrugarra-Llerena, T. L. T. Da Silveira, and C. R. Jung, "Pose Estimation for Two-View Panoramas based on Keypoint Matching: a Comparative Study and Critical Analysis," *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR2022W)*, pp. 5198–5207, 2022.

[6] Y. Zhao, C. Wen, Z. Xue, and Y. Gao, "3D Room Layout Estimation from a Cubemap of Panorama Image via Deep Manhattan Hough Transform," *Proceedings of the 17th European Conference on Computer Vision (ECCV2022)*, p. 637–654, 2022.

[7] H. Taira, Y. Inoue, A. Torii, and M. Okutomi, "Robust feature matching for distorted projection by spherical cameras," *IPSJ Transactions on Computer Vision and Applications*, vol. 7, pp. 84–88, 2015.

[8] J. Cruz-Mota, I. Bogdanova, B. Paquier, M. Bierlaire, and J.-P. Thiran, "Scale Invariant Feature Transform on the Sphere: Theory and Applications," *International Journal of Computer Vision*, vol. 98, pp. 217–241, 2012.

[9] Q. Zhao, W. Feng, L. Wan, and J. Zhang, "SPHORB: A Fast and Robust Binary Feature on the Sphere," *International Journal of Computer Vision*, vol. 113, no. 2, pp. 1573–1405, 2015.

[10] H. Guan and W. A. P. Smith, "BRISKS: Binary Features for Spherical Images on a Geodesic Grid," *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR2017)*, pp. 4886–4894, 2017.

[11] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-Supervised Interest Point Detection and Description," *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR2018W)*, pp. 337–349, 2018.

[12] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-Free Local Feature Matching with Transformers," *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR2021)*, pp. 8918–8927, 2021.

[13] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR2020)*, pp. 4937–4946, 2020.

[14] R. Hartley, "In Defense of the Eight-Point Algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 6, pp. 580–593, 1997.

[15] B. Solarte, C.-H. Wu, K.-W. Lu, Y.-H. Tsai, W.-C. Chiu, and M. Sun, "Robust 360-8PA: Redesigning The Normalized 8-point Algorithm for 360-FoV Images," *Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA2021)*, pp. 11032–11038, 2021.

[16] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, vol. 24, no. 6, p. 381–395, 1981.

[17] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 1573–1405, 2004.

[18] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An Efficient Alternative to SIFT or SURF," *Proceedings of the 2011 International Conference on Computer Vision (ICCV2011)*, pp. 2564–2571, 2011.