RELATIVE PANORAMIC CAMERA POSITION ESTIMATION FOR IMAGE-BASED VIRTUAL REALITY NETWORKS IN INDOOR ENVIRONMENTS

M. Nakagawa, K. Akano, T. Kobayashi, Y. Sekiguchi

[*] Dept. of Civil Engineering, Shibaura Institute of Technology, Tokyo, Japan - mnaka@shibaura-it.ac.jp

Commission IV / WG 5

KEY WORDS: Image-based virtual reality, panoramic image, optical flow, camera network, indoor environment

ABSTRACT:

Image-based virtual reality (VR) is a virtual space generated with panoramic images projected onto a primitive model. In image-based VR, realistic VR scenes can be generated with lower rendering cost, and network data can be described as relationships among VR scenes. The camera network data are generated manually or by an automated procedure using camera position and rotation data. When panoramic images are acquired in indoor environments, network data should be generated without Global Navigation Satellite Systems (GNSS) positioning data. Thus, we focused on image-based VR generation using a panoramic camera in indoor environments. We propose a methodology to automate network data generation using panoramic images for an image-based VR space. We verified and evaluated our methodology through five experiments in indoor environments, including a corridor, elevator hall, room, and stairs. We confirmed that our methodology can automatically reconstruct network data using panoramic images for image-based VR in indoor environments without GNSS position data.

1. INTRODUCTION

Virtual reality (VR) is a 3D virtual space in which users can experience a real space with 3D computer graphics using a VR theater or wearable devices. There are many VR applications, such as games, navigation, cultural site recordings, disaster monitoring, and infrastructure management. VR can be categorized a model-based VR and image-based VR.

Model-based VR is a virtual space with a 3D model generated using CAD software products, and the 3D model is generated from point cloud data acquired via laser scanning and structure from motion (SfM) processing (Snavely, 2010). Model-based VR can represent a real space from arbitrary viewpoints. However, its cost is higher, because the 3D model must be prepared. Moreover, when 3D models with many polygons are used, a high rendering processing cost is incurred to reconstruct scenes in the model-based VR.

Image-based VR is a virtual space generated with panoramic images projected onto a primitive model, such as a sphere or a cube. In image-based VR, realistic VR scenes can be generated with lower rendering cost. However, it is not easy to measure geometries of objects in the VR scenes. Panoramic images can be generated through some approaches, such as image integration by post-processing or omnidirectional image registration in real-time inner processing in a panoramic camera. In particular, a panoramic camera can reduce the cost of image acquisition.

Although it is impossible to reconstruct scenes from arbitrary viewpoints with a VR scene in image-based VR, a viewpoint translation can be represented with continuous VR scenes. In image-based VR, camera network data are described as relationships, such as nodes and links, among VR scenes. The network data are generated manually or by an automated procedure using camera position and rotation data. This manual work requires much time to link with each VR scene. On the other hand, when panoramic images are acquired in outdoor environments, network data for image-based VR can be easily prepared using a database of images localized on a map, GPS positioning data (Torii et al. 2010, Agarwal et al. 2015), and azimuth data taken from a magnetic sensor (Yazawa et al. 2009). When panoramic images are acquired in indoor environments, network data should be generated without Global Navigation Satellite Systems (GNSS) positioning data. Moreover, there are many scene changes in images, because the distance between camera and objects is smaller in indoor environments. Therefore, shorter distances between camera positions, such as submeter pitches, are better to achieve smooth translations in image-based VR. When indoor positioning systems can be used, it may be possible to generate network data automatically using the indoor position data. However, common indoor positioning systems are designed with 10 m accuracy for spatial resolution. Thus, camera position data management would be limited to 10 m pitches. Moreover, corridors are difficult environments for image-based VR generation with indoor positioning data, because the multipass problems in radio propagation make the positioning environment in corridors unstable.

Based on these technical issues, a new methodology is required to generate network data to connect the panorama images acquired with submeter steps in indoor environments. Therefore, we attempted to develop an algorithm to assist image-based VR network generation. In this research, we propose a methodology to generate network data with panoramic images. We evaluated the processing performance of our proposed algorithm through five experiments in indoor environments.

2. METHODOLOGY

Our proposed methodology consists mainly of image rectification, feature and corresponding point detection, estimation of camera translation, and network data generation, as shown in Figure 1. The image rectification is a horizontal adjustmentafter image acquisition. Although the feature and corresponding point detection are based on conventional feature matching, geometrical network constraints are applied to them in image combination.

https://doi.org/10.5194/isprs-annals-IV-2-W4-349-2017 | © Authors 2017. CC BY 4.0 License.

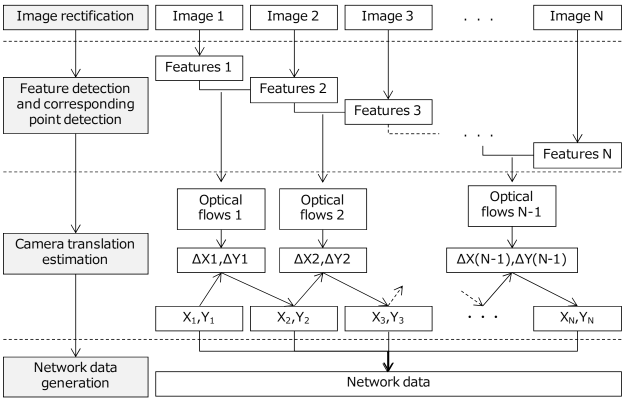


Figure 1. Processing flow

2.1 Image combination in corresponding point detection

Generally, camera translation and rotation parameters are estimated using SfM processing. However, corresponding image detection is a practical bottleneck in camera translation and rotation estimation, because the number of image combinations increases quadratically. When n images are input, the number of image combinations is $n \times (n - 1)/2$. For example, when 10 images are input, there will be 45 image combination patterns. Moreover, when we use panoramic images to reconstruct 3D data with SfM, the panoramic images are converted to centric projection images with different directions to reject outliers after corresponding point detection (Arth et al., 2011). When n panoramic images are converted to centric projection images with eight directions, the number of combination images is $8n \times (8n - 1)/2$ patterns. For example, if 10 images are input, the number of image combination for SfM would be 3160 patterns.

We focus on omitting camera rotation parameter estimation and bundle adjustment in SfM. We also focus on a restriction such as panoramic image acquisition along a straight line or grid lines to reduce the number of image combinations and corresponding point detections. This restriction can improve processing time in image matching. When n images are input, the number of image combination would be $n - 1$ patterns for image matching. For example, when 10 images are input, the number of image combinations would be reduced to nine patterns for image matching.

2.2 Feature matching

Many descriptors for feature matching have been proposed, such as Features from Accelerated Segment Test (FAST) (Rosten et al., 2005), Speeded Up Robust Features (SURF) (Bay et al., 2008), Scale-invariant Feature Transform (SIFT) (Lowe, 2004), Binary Robust Invariant Scalable Keypoints (BRISK) (Leutenegger et al., 2011), Binary Robust Independent Elementary Features (BRIEF) (Calonder et al., 2010), Fast Retina Keypoint (FREAK) (Alahi et al., 2012), Oriented FAST and Rotated BRIEF (ORB) (Rublee et al., 2011), and, MSER (Maximally Stable Extremal Regions) (Obdrzalek et al., 2010). These descriptors have trade-offs among processing speed, processing precision, stability, and robustness. In our experiments, we used SURF to achieve both high-speed processing and high stability, because our algorithm may allow mismatching in feature matching.

2.3 Camera translation estimation

First, optical flows in each panoramic image combination are estimated using corresponding points obtained in feature matching. Next, based on image acquisition at equal intervals along a straight line or grid lines, the optical flows are grouped into four directions in the panoramic image. As shown in Figure 2, there are characteristic optical flows in the panoramic image after camera translation. Thus, a camera translation is estimated with these characteristic optical flows.

Median values of optical flows in each direction are used to estimate the translation to allow mismatching in the feature matching. Although it is possible to estimate absolute camera translation parameters through a backward intersection methodology with known objects, we estimated relative camera translation parameters in our research.

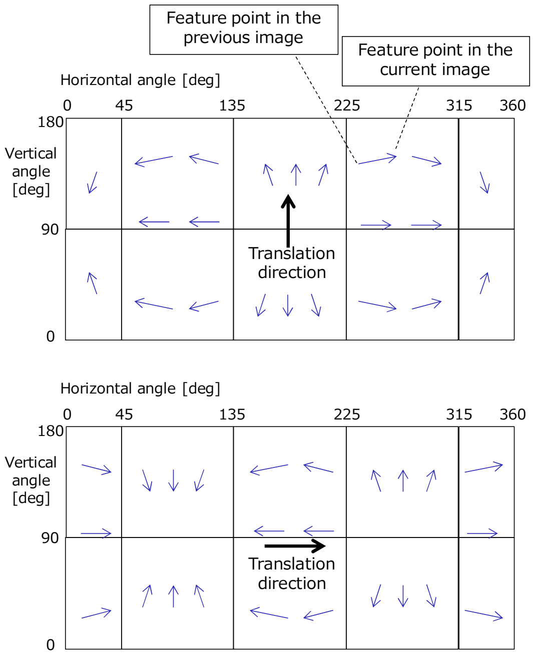


Figure 2. Optical flow estimation for camera translation estimation

2.4 Network data generation

The first camera position was defined as the origin point. Camera positions are defined as nodes and connecting lines adjacent camera positions are defined as links. Each link has a relative translation distance defined as one without a distance unit. Based on these constraints, we estimate a camera trajectory. When the camera trajectory shapes a straight line, each node is simply connected to the nearest node. When the camera trajectory follows a grid, each node is connected to the nearest node under a geometric network constraint, such as four-neighbors (90° angles) or eight-neighbors (45° angles).

3. EXPERIMENT

We used a panoramic camera (THETA S, RICOH) mounted on a tripod, as shown in Figure 3. Panoramic images were captured remotely with a smart phone in corridors, a room, an elevator hall, and stairs, as shown in Figure 4 and Table 1. We kept three constraints in panoramic image acquisition: equal intervals for camera position pitch, fixed camera directions, and fixed camera heights, as shown in Figure 5. We applied inner image processing using a gyro sensor in the camera to generate horizontally rectified panoramic images.

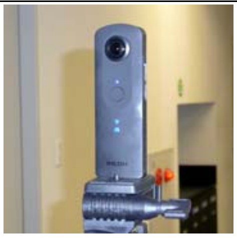
	Image size	14 M pixels (5376×2688)
	Focus distance	10 cm - ∞
	Sensor (s)	1/2.3 CMOS (×2)
	Weight	125 g

Figure 3. Panoramic camera (THETA S, RICOH)

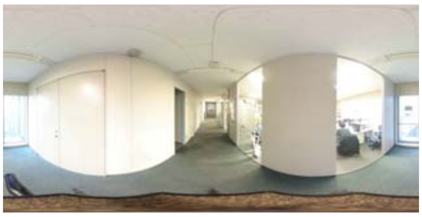
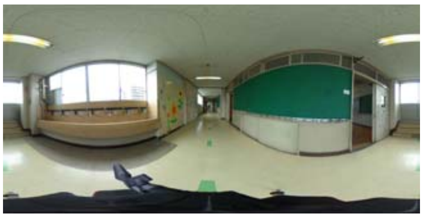
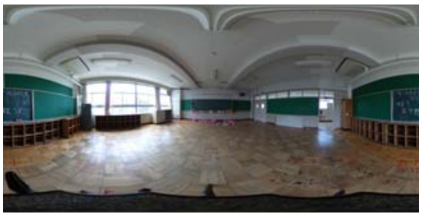
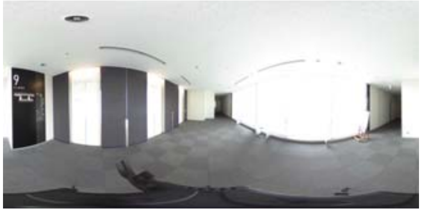
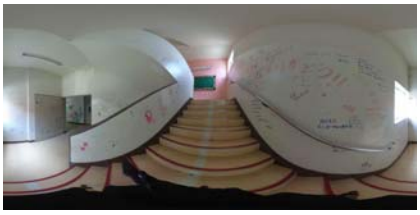
Corridor 1	
Corridor 2	
Room	
Elevator hall	
Stairs	

Figure 4. Experimental environments

Table 1. Data sets

Data set	Camera position interval (average)	The number of acquired images
Corridor 1	50 cm	15
Corridor 2	100 cm	31
Room	60 cm	9
Elevator hall	60 cm	27
Stairs	100 cm	37

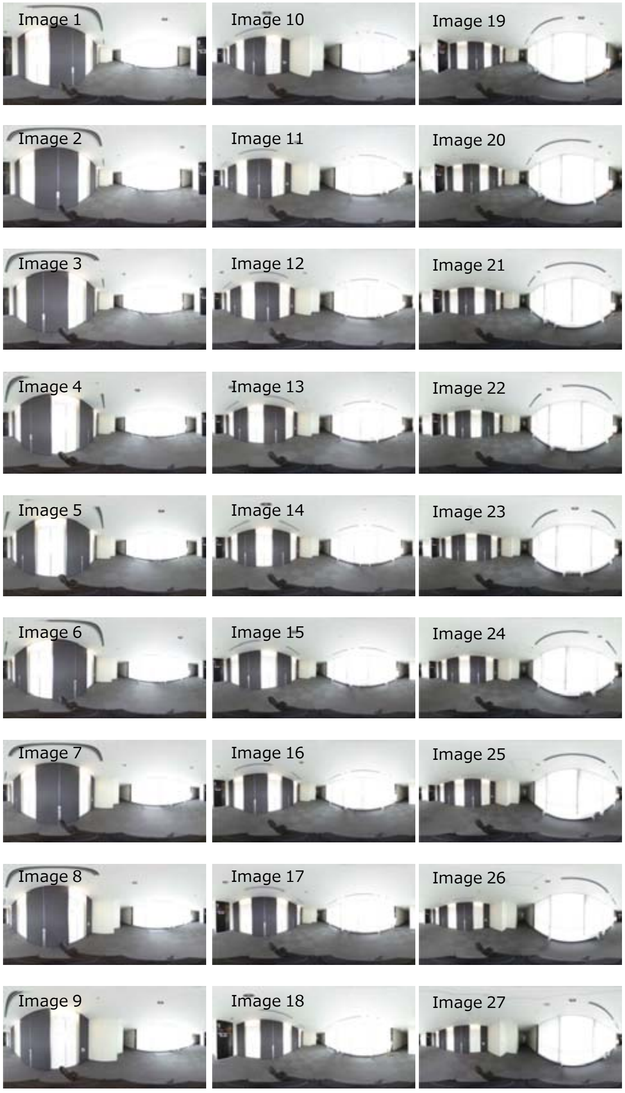


Figure 5. Acquired panoramic images in the Elevator hall

4. RESULTS

4.1 Processing time

Processing for image loading, feature and corresponding point detection (using SURF), and camera translation estimation required several minutes in total for each dataset using an Intel Core i7-U 3.30 GHz processor with MATLAB (single thread), as shown in Table 2.

Table 2. Processing time

Data set	The number of acquired images	Image loading [s]	SURF [s]	Translation estimation [s]
Corridor 1	15	8.8	32.4	4.4
Corridor 2	31	15.3	67.8	8.0
Room	9	4.6	18.0	2.4
Elevator hall	27	15.4	64.5	5.7
Stairs	37	17.5	75.7	9.1

4.2 Camera trajectory estimation

Results for Corridor 1, Room, and Elevator hall were successful for camera translation estimation. However, camera translation estimation failed for Corridor 2 and stairs.

The camera translation in Corridor 1 was correctly estimated, as shown in Figure 6. The X and Y axes indicate the horizontal camera position. The left image shows the actual camera path, and the right image shows our estimated result.

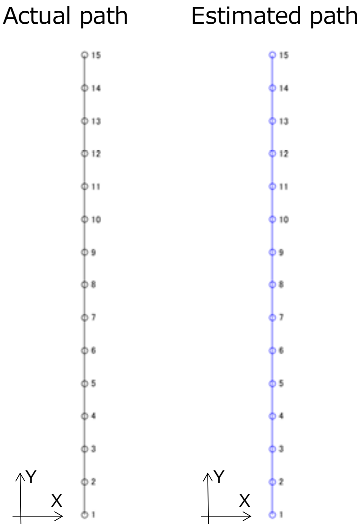


Figure 6. Camera translation estimation result, Corridor 1 (Left image: actual camera path, right image: our estimated result)

The estimated camera translation in Corridor 2 is shown in Figure 7. Horizontal axes indicate estimated camera positions, and vertical axis indicates image identification numbers. Although the actual camera path was a straight line, the estimated path meandered around the straight line.

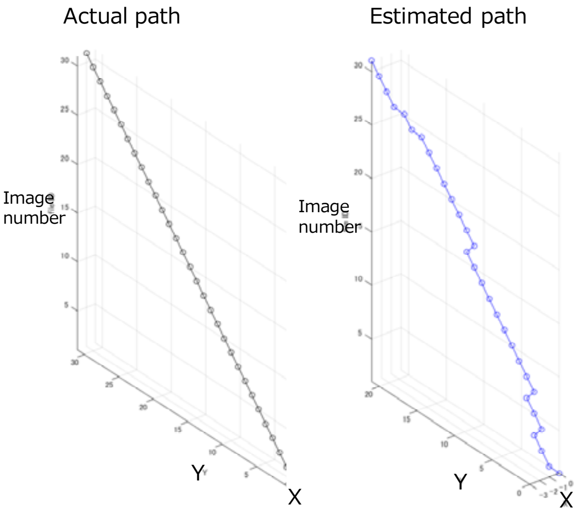


Figure 7. Camera translation estimation result, Corridor 2 (Left image: actual camera path, right image: our estimated result)

Figure 8 shows our estimated camera translation in the room. The X and Y axes indicate horizontal camera position.

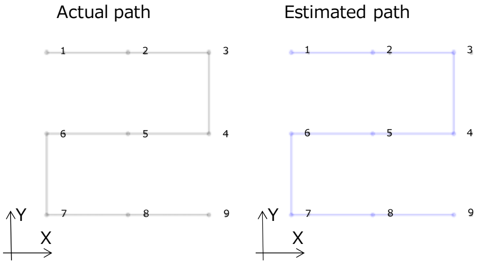


Figure 8. Camera translation estimation result (Room) (Left image: the actual camera path, right image: our estimated result)

Figure 9 shows our estimated camera translation in the elevator hall. The X and Y axes indicate horizontal camera position. Figure 10 shows the camera network data estimated after the camera path estimation.

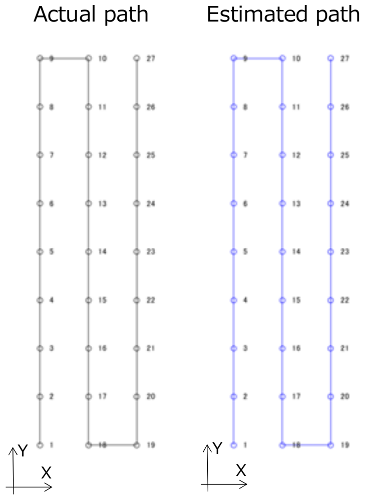


Figure 9. Camera translation estimation result, Elevator hall (Left image: actual camera path, right image: our estimated result)

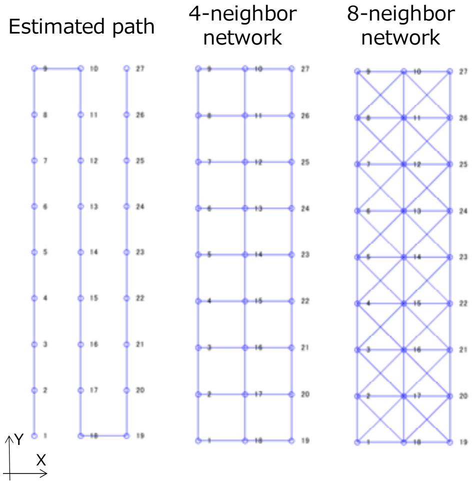


Figure 10. Camera network data, Elevator hall (Left image: estimated camera path, center image: 4-neighbor network, right image: 8-neighbor network)

Figure 11 shows our estimated camera translation for the Stairs. Horizontal axes indicate estimated camera positions, and the vertical axis indicates image identification numbers. Although the actual camera path was a spiral line, the estimated path meandered along the spiral line.

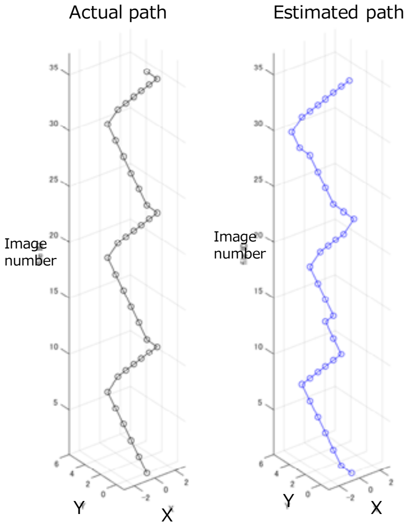


Figure 11. Camera translation estimation result, Stairs (Left image: actual camera path, right image: our estimated result)

4.3 Optical flow estimation

Examples of estimated optical flows in camera translation estimation are shown in Figure 12. We confirmed that the proposed methodology can reconstruct camera translations as camera network data when there is continuity between panoramic images.

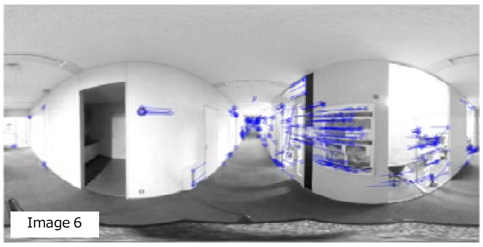
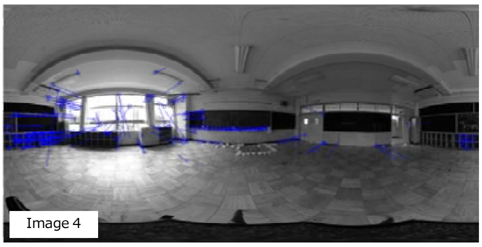
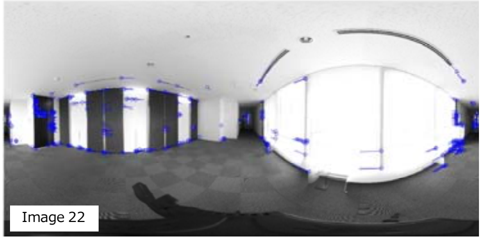
Corridor 1	
Room	
Elevator hall	

Figure 12. Estimated optical flows (Corridor 1, Room, and Elevator hall)

On the other hand, we confirmed that the proposed methodology fails to reconstruct camera translation precisely when there is low continuity between panoramic images, as shown in Figure 13. Continuity and estimation performance depend on camera distances, because our proposed methodology is based on feature tracking processing. The number of extracted features is also significant in estimating camera translation. Appropriate camera distances are required to improve the stability of camera translation estimation. However, it is not easy to estimate suitable camera distances before image acquisition, because the appropriate distances depend on the environment and the objects. Thus, we will investigate a real-time navigation and assistance approach for panoramic camera acquisition in future work.

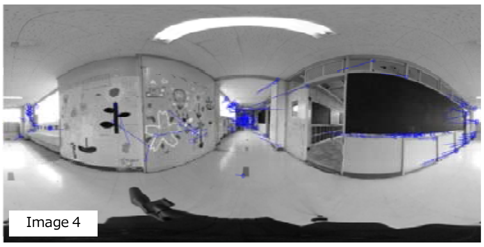
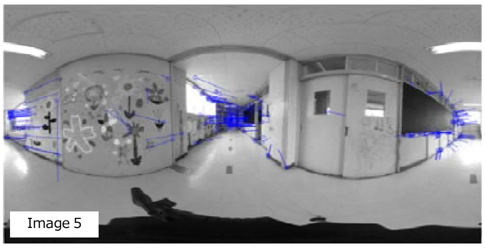
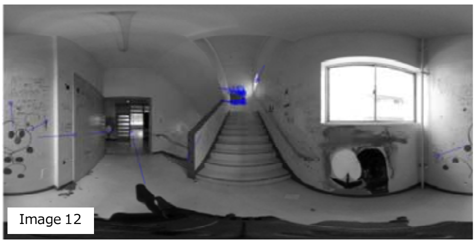
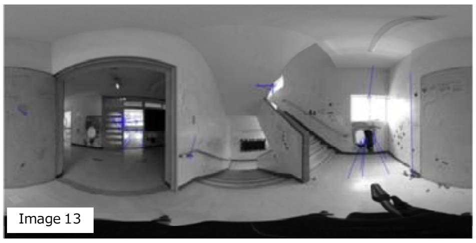
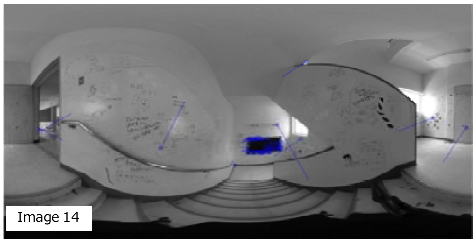
Corridor 2	 
Stairs	  

Figure 13. Estimated optical flows (Corridor 2, and Stairs)

5. SUMMARY

In this paper, we have focused on image-based VR generation using a panoramic camera in indoor environments. We proposed a methodology to automate network data generation using panoramic images for image-based VR spaces. We verified our methodology through five experiments in indoor environments, in corridors, an elevator hall, a room, and on stairs. We also evaluated the processing performance of our proposed algorithm in these experiments. Although the stability of our methodology depends on the camera position intervals and the number of feature points in the images, we confirmed that our methodology can automatically reconstruct network data using panoramic images for image-based VR in indoor environments without GNSS position data.

ACKNOWLEDGEMENT

This work was supported by JACIC Grant Number 2016-08. Moreover, our experiments are supported by Koto city office in Tokyo.

REFERENCES

Snavely, N., 2010, Bundler: Structure from motion (SFM) for unordered image collections.
(from http://www.cs.cornell.edu/~snavely/bundler/).

Torii, A., Sivic, J., Pajdla, T., 2010, Visual localization by linear combination of image descriptors, *IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops*, pp. 102-109.

Agarwal, P., Burgard, W., Spinello, L., 2015, Metric Localization using Google Street View, *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pp. 3111-3118.

Yazawa, N., Uchiyama, H., Saito, H., Servières, M., Guillaume Moreau, 2009, Image Based View Localization System Retrieving from a Panorama Database by SURF, *Proceedings of the 11th IAPR Conference on Machine Vision Applications*, MVA 2009, pp. 118-121.

Arth, C., Klopschitz, M., Reitmayr, G., Schmalstieg, D., 2011, Real-Time Self-Localization from Panoramic Images on Mobile Devices, *Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on*, pp. 37-46.

Rosten, E., Drummond, T., 2005, Fusing Points and Lines for High Performance Tracking, *Proceedings of the IEEE International Conference on Computer Vision*, Vol. 2 (October 2005), pp.1508–1511.

Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., 2008, SURF:Speeded Up Robust Features, *Computer Vision and Image Understanding (CVIU)*, Vol. 110, No. 3, pp.346–359.

Lowe, David G., 2004, Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, Volume 60 Issue 2, pp.91-110.

Leutenegger, S., Chli, M., Siegwart, R., 2011, BRISK: Binary Robust Invariant Scalable Keypoints, *Proceedings of the IEEE International Conference, ICCV*, pp.2548-2555.

Calonder, M., Lepetit, V., Strecha, C., Fua, P., 2010, BRIEF: Binary Robust Independent Elementary Features. *In Proceedings of the 11th European conference on Computer vision*: Part IV, pp.778-792.

Alahi, A., Ortiz, R., Vandergheynst, P., 2012, FREAK: Fast Retina Keypoint, *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.510-517.

Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011, ORB: an efficient alternative to SIFT or SURF, *In 2011 IEEE International Conference on Computer Vision*, pp.2564-2571.

Obdrzalek D., Basovnik, S., Mach, L., Mikulik, A., 2010, Detecting Scene Elements Using Maximally Stable Colour Regions, *Communications in Computer and Information Science*, vol. 82 CCIS, pp 107–115.