Information and Media Technologies 9(3) : 361–365 (2014)
reprinted from: IPSJ Transactions on Computer Vision and Applications 6: 68–72 (2014)

Express Paper

Background Estimation for a Single Omnidirectional Image Sequence Captured with a Moving Camera

Norihiko Kawai[1,a)] Naoya Inoue[1] Tomokazu Sato[1,b)] Fumio Okura[1,c)] Yuta Nakashima[1,d)] Naokazu Yokoya[1,e)]

Received: March 14, 2014, Accepted: April 24, 2014, Released: July 25, 2014

Abstract: This paper proposes a background estimation method from a single omnidirectional image sequence for removing undesired regions such as moving objects, specular regions, and uncaptured regions caused by the camera's blind spot without manual specification. The proposed method aligns multiple frames using a reconstructed 3D model of the environment and generates background images by minimizing an energy function for selecting a frame for each pixel. In the energy function, we introduce patch similarity and camera positions to remove undesired regions more correctly and generate high-resolution images. In experiments, we demonstrate the effectiveness of the proposed method by comparing the result given by the proposed method with those from conventional approaches.

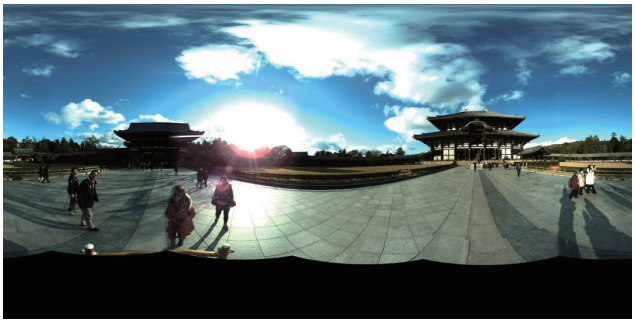
Keywords: background estimation, object removal, omnidirectional image

1. Introduction

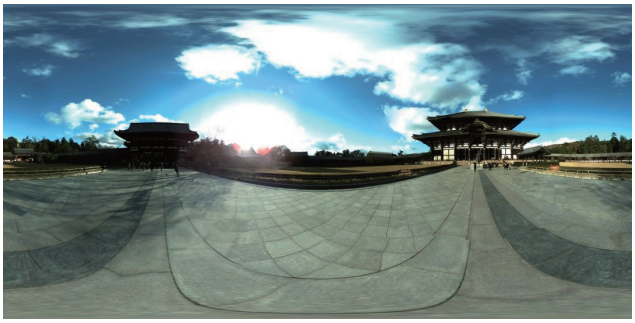
Omnidirectional image sequences captured with a moving camera have a wide variety of applications, such as telepresence and vehicle navigation systems. For example, Google Street View is one of the most well-known telepresence systems that are already available to the public. A vision-based vehicle navigation system [11] is another interesting example of such applications, which uses preliminary captured omnidirectional image sequences to estimate and track the pose of the vehicle using image matching between the sequences and an image from the in-vehicle camera.

Usually, omnidirectional image sequences contain undesired regions that cause serious problems for the applications. Google Street View often suffers from the privacy issue since its images come with the appearances of pedestrians. The vehicle navigation system can fail in pose estimation and tracking when the preliminarily captured omnidirectional image sequences contain a lot of moving objects (e.g., cars and pedestrians), specular surfaces, and the camera's blind spot. Methods for removing such undesired regions by replacing them with an estimated background as shown in **Fig. 1** are strongly required.

To remove such undesired regions and replace them with the actual background, a straightforward approach is (1) finding undesired regions and (2) estimating their background by frame alignment, which makes pixelwise correspondences between background pixels in different frames and pixels in undesired regions in a target frame.



(a) Original



(b) Output

Fig. 1 Example of removal of undesired regions by the proposed method (39th frame).

Flores et al. [4] proposed a method to remove pedestrians from omnidirectional images, in which regions of pedestrians are specified with pedestrian detection algorithm [8]. Our previous paper [7] proposed a method to fill in an uncaptured region caused

[1] Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara 630–0192, Japan
a) norihi-k@is.naist.jp
b) tomoka-s@is.naist.jp
c) fumio-o@is.naist.jp
d) n-yuta@is.naist.jp
e) yokoya@is.naist.jp

Information and Media Technologies 9(3) : 361–365 (2014)
reprinted from: IPSJ Transactions on Computer Vision and Applications 6: 68–72 (2014)

by a blind spot of an omnidirectional camera, in which the region is preliminarily specified with camera calibration. These methods use the assumption that the undesired region is planar for frame alignment. However, the planner assumption limits applicable scenes.

To relax the geometric constraint, some methods have been proposed that take different approaches to align frames in sequences of ordinary perspective images. Matsushita et al. [9] estimate motion in the undesired regions using the optical flows around them. Bhat et al. [1] estimate a dense depth map of the target scene, and Granados et al. [5] use multiple homographies. These methods fill in the manually specified regions using different frames aligned to the target one. However, the methods that require manually specifying undesired regions may not be applied to image sequences with a large number of frames.

Cohen [3] proposed a method to remove moving objects without specifying the target objects, assuming that the alignment among frames is given. This method treats object removal as a labeling problem; a frame index is regarded as a label and an appropriate label that is likely to be the background is assigned to each pixel using energy minimization. The result is generated by using pixel values from different frames. Specifically, the energy function used in this method is defined based on the variance over the intensities of a single pixel in multiple aligned frames. The similar energy functions are also used in Refs. [1], [5], and they are based on the pixel differences. However, using only the pixelwise differences of intensities is sometimes insufficient to remove moving objects and generate good quality of background texture. For omnidirectional images, Uchiyama et al. [12] remove moving objects from an omnidirectional street image without manual specification of them. This method requires multiple image sequences taken along almost the same path because it replaces moving objects in a target frame with background from an image captured at a position very close to that of the target frame. Therefore, this method requires much time to be applied to large environments.

In this paper, we propose a method for removing undesired regions in a single omnidirectional image sequence. Assuming that the background of the undesired regions in some frames is visible in the other frames, we replace the undesired regions with their background without explicitly specifying them in the image labeling framework. In order to relax the assumption of scene geometry, we adopt 3D reconstruction-based frame alignment using the structure-from-motion and multiple-view stereo techniques. In the proposed method, we use patch similarity in our energy function of the image labeling problem, rather than pixelwise differences, which empirically demonstrates better background assignment. In addition, our energy function leverages a camera position of each frame to obtain the background with higher resolution.

2. Background Estimation

2.1 Frame Alignment Based on 3D Reconstruction

In the proposed method, we first estimate a camera pose of each frame and 3D geometry of the scene by applying Structure-from-Motion [13] and Multi-View Stereo (MVS) [6] to a single image sequence as shown in **Fig. 2**. The reasons why we employ 3D reconstruction for frame alignment are that (1) an omnidirectional image can more easily and accurately reconstruct the geometry of the whole scene than an ordinary image because an omnidirectional camera can capture almost the entire field of view, and that (2) the 3D geometry of background, not moving objects, can be obtained by using photo consistency among frames in MVS [6]. We then generate a depth map for each frame from the reconstructed 3D geometry. Usually the reconstructed 3D geometry does not cover an entire frame because objects or the sky that are sufficiently far from the camera cannot be reconstructed. For such regions in the frame, we deem the depth values to be infinity. Next, multiple frames whose camera positions are within a certain distance threshold d from that of a target frame are extracted. The extracted frames are warped to the viewpoint of the target frame using the depth map as shown in **Fig. 3**. The black region on the ground surface in Fig. 3 (b) corresponds to that in the bottom of Fig. 3 (a), which is the omnidirectional camera's blind spot. Among the warped images, the same position's pixels are the corresponding pixels.

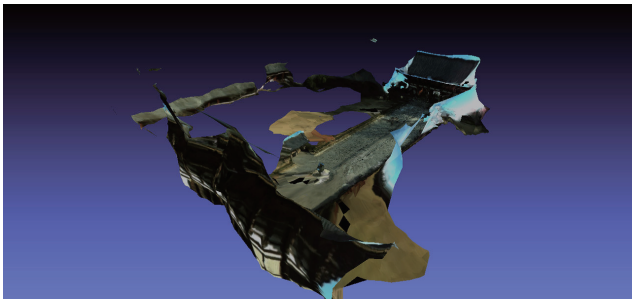
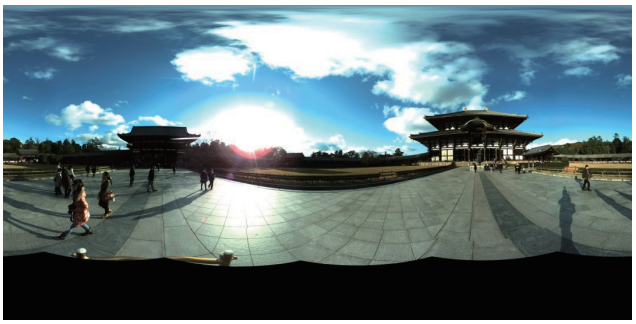
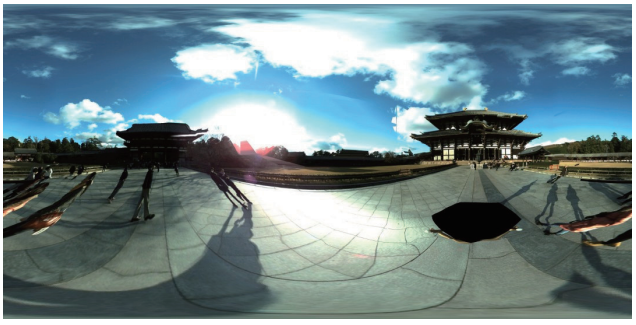


Fig. 2 Reconstructed 3D geometry of target scene.



(a) Original (42nd frame)



(b) Warped image

Fig. 3 Example of image warping. The 42nd frame is warped to the viewpoint of the 39th frame (Fig. 1 (a)).

Information and Media Technologies 9(3) : 361–365 (2014)
reprinted from: IPSJ Transactions on Computer Vision and Applications 6: 68–72 (2014)

2.2 Image Composition Based on Frame Selection

Using warped omnidirectional images of the extracted frames, an appropriate frame, which is likely to be background, for each pixel is selected by minimizing an energy function with graph cut [2]. Given frame t as the target frame, energy function E is defined as:

$$E = \sum_{p \in A} E_1(f_p) + \lambda \sum_{(p,q) \in B} E_2(f_p, f_q), \quad (1)$$

where f_p and f_q are labels for pixel p and q, respectively, representing frame indexes from which the pixel colors are copied. A is a set of all pixels in the omnidirectional image, and B is a set of pairs of adjacent pixels. This pixel adjacency considers that our images are omnidirectional panoramas. λ is a weight to control the contribution of the second term.

E_1 is defined as a linear combination of two terms:

$$E_1(f_p) = L(f_p) + \alpha D(f_p), \quad (2)$$

where α is a weight to balance the two terms.

The first term $L(f_p)$ is based on distance between camera positions of the target frame and frame f_p, and defined as:

$$L(f_p) = 1 - \frac{k}{\|\mathbf{x}_{f_p} - \mathbf{x}_t\|_2^2 + k}, \quad (3)$$

where $\mathbf{x}_{f_p}$ and $\mathbf{x}_t$ are camera positions of frame f_p and the target frame, respectively. The parameter k controls the influence of the distance between camera positions. Among the warped omnidirectional images, the resolution of texture is inversely proportional to the square of distance from each object in the background to the camera. Therefore, a high resolution background image can be obtained if we select frames that are captured from the cameras as close to that of the target frame as possible. To the best of our knowledge, this is the first work that incorporates the spatial relationship among cameras into the energy function.

$D(f_p)$ is a dissimilarity measure of patches centered at pixel p between frame f_p and other warped frames, defined as:

$$D(f_p) = \frac{\sum_{g \in G_{f_p}} SSD(f_p, g)}{N}, \quad (4)$$

where

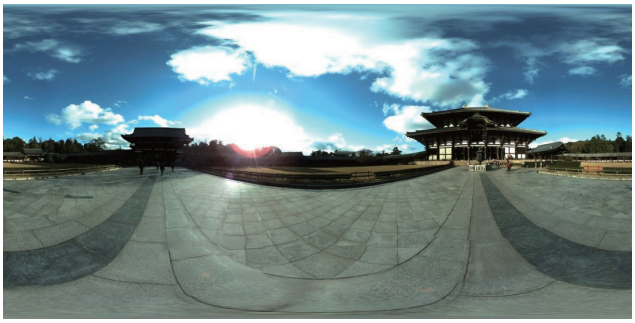
$$SSD(f_p, g) = \sum_{p'} \|\mathbf{I}_{f_p}(p') - \mathbf{I}_g(p')\|_2^2. \quad (5)$$

$\mathbf{I}_{f_p}(p)$ and $\mathbf{I}_g(p)$ are the colors of pixel p in frames f_p and g in the RGB color space. The summation is calculated over pixel p' in the patch of $M \times M$ pixels centered at p. N is a normalization constant so that the value of $D(f_p)$ can be in [0, 1]. G_{f_p} is a set of the frames that given by

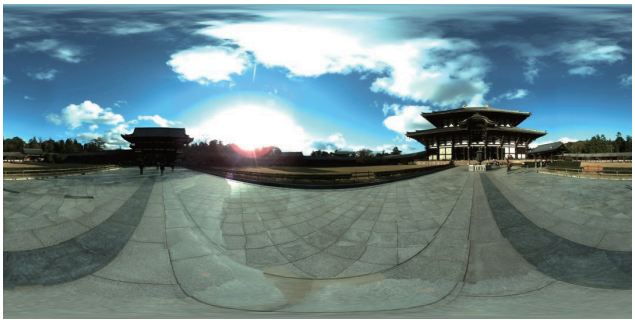
$$G_{f_p} = \{g | SSD(f_p, g) < \underset{i \in F_t}{median}\ SSD(f_p, i)\}, \quad (6)$$

where F_t is a set of the extracted frames for target frame t. Function *median* gives the median value. Unlike pixelwise differences used in Refs. [1], [3], [5], the patch similarity enables us to select frames capturing the background more correctly.

Energy E_2 is a smoothness term and defined as:



(a) With Poisson blending



(b) Without Poisson blending

Fig. 4 Example of generated background image with and without Poisson blending (45th frame).

Table 1 Parameters in experiments.

d	λ	M	α	k
10 [m]	0.025	11 [pixel]	60	10

$$E_2(f_p, f_q) = \mu(f_p, f_q)(\|\mathbf{I}_{f_p}(p) - \mathbf{I}_{f_q}(p)\| + \|\mathbf{I}_{f_p}(q) - \mathbf{I}_{f_q}(q)\|), \quad (7)$$

where $\mu(f_p, f_q)$ gives 0 when $f_p = f_q$, otherwise 1. This term prevents from frequently changing frames between adjacent pixels.

After frame selection, we obtain an omnidirectional background image by copying pixel values of the selected frames with color adjustment by Poisson blending [10] as shown in **Fig. 4**.

3. Experiments

We experimentally demonstrate the effectiveness of the proposed method using an image sequence with 61 frames, comparing the obtained background images by the proposed method with those by conventional approaches, which use pixelwise differences of intensities or do not consider the camera position-based term. For the experiments, we used a PC with Windows7, Core i7-990X 3.47 GHz CPU, and 12 GB of memory. The image sequence was captured with Point Grey's Ladybug3 while moving almost straight. The distance of cameras between each pair of adjacent frames was about 1 meter, and each frame was resized to 1080×540 pixels. Parameters used in the experiments are summarized in **Table 1**. We generated the background images using the following methods:

Method A Proposed method

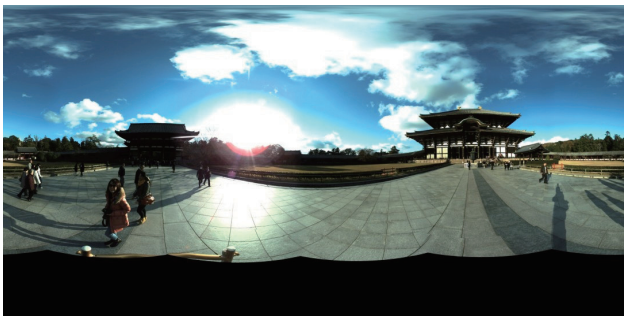
Method B Method using pixelwise differences rather than patch similarity (Eq. (5))

Method C Method without the camera position-based term (Eq. (3))

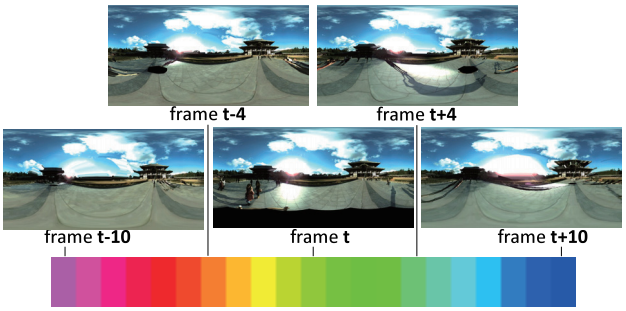
Method D Median of intensities among extracted frames

Figure 5 shows the estimated background images when the

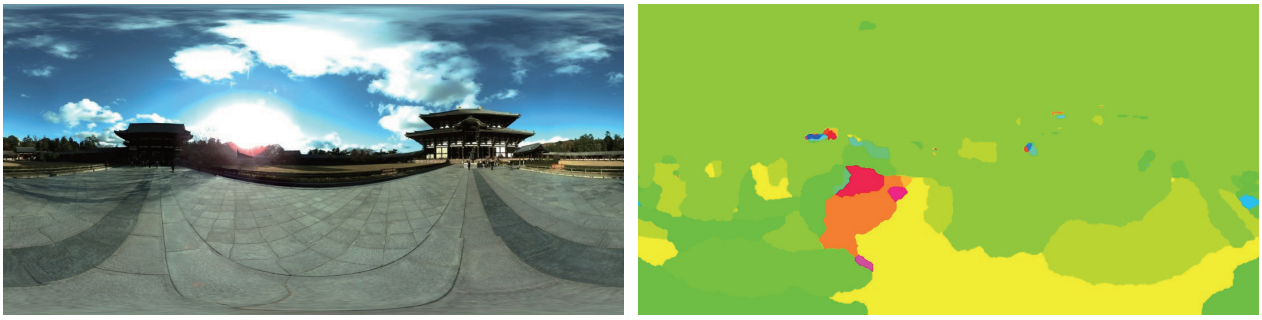
Information and Media Technologies 9(3) : 361–365 (2014)
reprinted from: IPSJ Transactions on Computer Vision and Applications 6: 68–72 (2014)



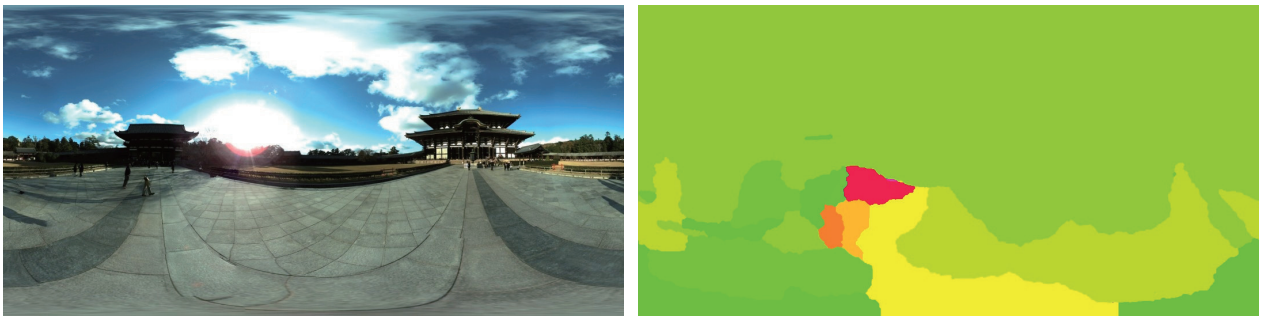
(a) Input (41st frame)



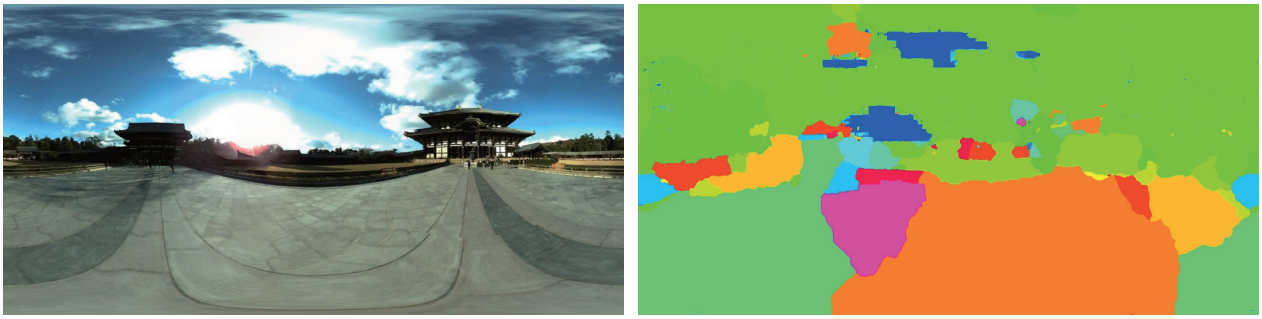
(b) Warped images from different frames and their corresponding colors



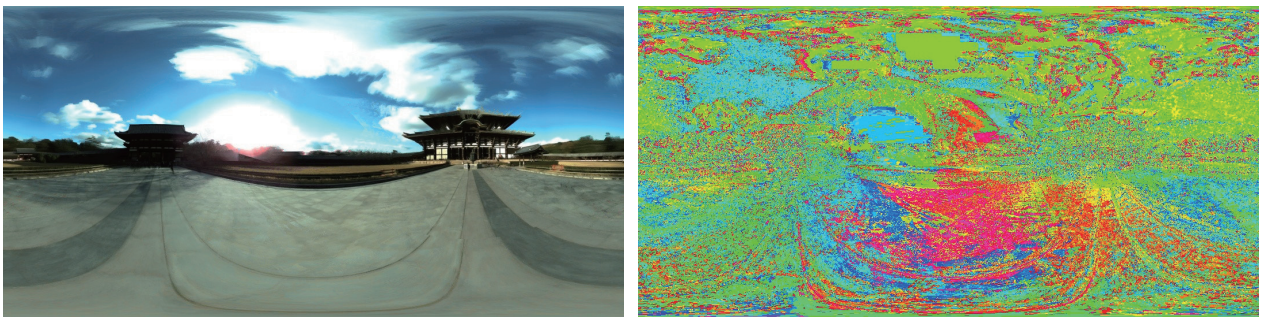
(c) Method A: Proposed method



(d) Method B: Method using pixelwise differences



(e) Method C: Method without considering camera positions



(f) Method D: Median of pixel values among extracted frames

Fig. 5 Experiments for the 41st frame. The left images are composite images and the right images are selected frames in (c), (d), (e), and (f).

Information and Media Technologies 9(3) : 361–365 (2014)
reprinted from: IPSJ Transactions on Computer Vision and Applications 6: 68–72 (2014)

target frame is the 41st frame. Using the distance threshold, 19 frames (from 32nd to 50th frame) were extracted. Figure 5 (a) shows the original 41st frame. Figure 5 (c) to (f) show the generated images and selected frames. In the following, we discuss the experimental results.

As shown in Fig. 5 (c), the proposed method successfully removed undesired regions, i.e., moving objects, specular surfaces, and the blind spot, and the generated background regions gave the detailed texture. On the other hand, the method using pixelwise differences shown in Fig. 5 (d) failed in removing some moving objects and cast shadows. From the comparison between Fig. 5 (c) and (d), we confirmed that the patch similarity was more effective to remove moving objects than pixelwise differences. The method that does not consider camera position-based term gave blurring texture as shown in Fig. 5 (e). Compared to this method, the proposed method generated clearer edges of the tiles on the ground surface. Also as seen in the right image of Fig. 5 (e), a larger number of distant frames were selected than the proposed method. This results in decreasing the resolution of the generated background image. These results indicated that considering the distances among camera positions was crucial for generating background images with high resolution. In the result given by median of pixel values among the extracted frames as shown in Fig. 5 (f), although undesired regions were removed, the resolution of the generated image is the lowest. In the proposed method, it took 157 seconds for the frame selection by graph cut, and 194 seconds for Poisson blending.

4. Conclusion

In this paper, we have proposed a background estimation method for a single omnidirectional image sequence without manually specifying undesired regions. In the proposed method, we used a reconstructed 3D geometry for frame alignment, and we newly introduced patch similarity and camera positions. The experimental results successfully demonstrated that the patch similarity- and camera position-based terms in our energy function were beneficial to correct background frame selection and maintain the resolution of background images. In future work, we need to address the misalignment of multiple frames because the reconstructed 3D model is not always accurate.

Acknowledgments This work was partially supported by JSPS KAKENHI Nos. 23240024 and 25540086.

References

[1] Bhat, P., Zitnick, C.L., Snavely, N., Agarwala, A., Agrawala, N., Cohen, M., Curless, B. and Kang, S.B.: Using Photographs to Enhance Videos of a Static Scene, *Proc. Eurographics Symp. Rendering*, pp.327–338 (2007).

[2] Boykov, Y. and Kolmogorv, V.: An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.26, No.9, pp.1124–1137 (2004).

[3] Cohen, S.: Background Estimation as a Labeling Problem, *Proc. Int. Conf. Computer Vision*, pp.1034–1041 (2005).

[4] Flores, A. and Belongie, S.: Removing Pedestrians from Google Street View Images, *Proc. Int. Workshop on Mobile Vision*, pp.53–58 (2010).

[5] Granados, M., Kim, K.I., Tompkin, J., Kautz, J. and Theobalt, C.: Background Inpainting for Videos with Dynamic Objects and a Free-moving Camera, *Proc. Europian Conf. Computer Vision*, pp.682–695 (2012).

[6] Jancosek, M. and Pajdla, T.: Multi-view Reconstruction Preserving Weakly-supported Surfaces, *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp.3121–3128 (2011).

[7] Kawai, N., Machikita, K., Sato, T. and Yokoya, N.: Video Completion for Generating Omnidirectional Video without Invisible Areas, *IPSJ Trans. Computer Vision and Applications*, Vol.2, pp.200–213 (2010).

[8] Leibe, B., Leonardis, A. and Schiele, B.: Robust Object Detection with Interleaved Categorization and Segmentation, *Int. Journal of Computer Vision*, Vol.77, No.1–3, pp.259–289 (2007).

[9] Matsushita, Y., Ge, W., Tang, X. and Shum, H.-Y.: Full-Frame Video Stabilization with Motion Inpainting, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.28, No.7, pp.1150–1163 (2006).

[10] Pérez, P., Gangnet, M. and Blake, A.: Poisson Image Editing, *Proc. ACM SIGGRAPH*, pp.313–318 (2003).

[11] Uchiyama, H., Deguchi, D., Takahashi, T., Ide, I. and Murase, H.: Ego-localization Using Streetscape Image Sequences from In-vehicle Cmeras, *Proc. IEEE Intelligent Vehicles Symposium*, pp.185–190 (2009).

[12] Uchiyama, H., Deguchi, D., Takahashi, T., Ide, I. and Murase, H.: Removal of Moving Objects from a Street-view Image by Fusing Multiple Image Sequences, *Proc. Int. Conf. Pattern Recognition*, pp.3456–3459 (2010).

[13] Wu, C.: VisualSFM: A Visual Structure from Motion System (2011), available from ⟨http://ccwu.me/vsfm/⟩.

(Communicated by *Mitsuru Ambai*)