# Generation of omnidirectional image without photographer

Ryusei Noda and Norihiko Kawai[0000−0002−7859−8407]

Faculty of Information Science and Technology, Osaka Institute of Technology
1-79-1 Kitayama, Hirakata, Osaka 573-0196, Japan
norihiko.kawai@oit.ac.jp

**Abstract.** In order to create a virtual reality (VR) space using omnidirectional images, it is desirable to use images without the photographer's inclusion. In this study, we propose a method to generate an omnidirectional image without the photographer's inclusion by using multiple images taken by an omnidirectional camera. In the proposed method, the photographer rotates around the omnidirectional camera and takes several images. Next, we perform feature point matching on the omnidirectional images and unify the appearance of all the images by using the amount of translation calculated from the matching. Finally, the images are combined with graph cut and Poisson image editing to produce an omnidirectional panoramic image without the photographer in it.

**Keywords:** Omnidirectional image · Photographer removal · Graph cut.

## 1 Introduction

With the widespread use of omnidirectional cameras, more and more people use them in various situations. The size of the omnidirectional camera is getting smaller, and now it can fit in one hand. Therefore, it is easy to take a panoramic image by holding an omnidirectional camera in one's hand. The omnidirectional panoramic images taken by such an omnidirectional camera are used in the construction of VR spaces for the purpose of virtual experience of remote areas such as Google Street View and real estate previews on the Internet. However, when taking an omnidirectional image while holding the camera in one's hand, the photographer appears in a large part of the image, and the background is partially obscured. For this reason, it is preferable to remove moving objects such as the photographer for the above purpose.

One solution to this problem is to set up an omnidirectional camera on a tripod and capture images remotely. However, this method cannot be used over water or on uneven terrain. In addition, even if a tripod is used, the tripod appear in the image. Although inpainting methods for static images [12, 3, 9, 15] may be able to remove the tripod and the photographer from the image, it requires a lot of effort to specify the target area manually, and the inpainted result may be different from the actual background.

In contrast to this, methods have been proposed to display the actual background in the area of the unneeded objects to be removed by taking multiple images and videos of the actual background and combining them. For example, Cohen [2] proposed a method to generate a background-only image by using a video taken with a fixed camera as input and selecting and copying a frame without moving objects for each pixel. However, since a fixed camera is assumed, alignment between frames is not necessary. Different methods are to use optical flow to align its background and restore the background by copying pixel values from other frames [14, 10]. However, since optical flow is used, they assume that the camera movement between frames is not too large.

On the other hand, object removal for omnidirectional images captured by an omnidirectional camera has also been studied. The method in [5] removes humans by transforming perspective projection images obtained from omnidirectional images taken at different points by homography and combining them. Since this method uses only a homography transformation for alignment between images, misalignment occurs if the background is not flat. The method in [8] performs structure-from-motion [13] and multi-view stereo [7] on omnidirectional images captured while moving, and the estimated camera pose and background shapes are used to align the images and restore the background of moving objects. As described above, the conventional research assumes the input of omnidirectional images captured with different positions.

In this study, we propose a method to generate an omnidirectional image with the photographer in the image removed without using any equipment such as a tripod. The photographer with an omnidirectional camera takes several pictures while rotating around the omnidirectional camera so that the camera position do not change. The method generates an omnidirectional image without a photographer by aligning the images using feature matching and automatically selecting an appropriate input image without a photographer for each pixel.

## 2 Proposed method

### 2.1 Overview

In the proposed method, (1) we first input multiple omnidirectional images taken by the photographer while rotating around an omnidirectional camera. In this study, we assume that the same object is at approximately the same height in each omnidirectional image because the omnidirectional images are generated using the direction of gravity obtained from the accelerometer in the camera. (2) Harris corner detection is performed for each input image. Patches are set up around the feature points detected by Harris corner detection, and the most similar patches are matched by calculating the sum of squared differences (SSD) of pixel values between the patches in the two images. The horizontal translation between the omnidirectional images is calculated by RANSAC and the least-squares method. (3) The appearance of each input image is unified using the calculated amount of horizontal translation. (4) The appropriate image without the photographer is selected for each pixel by graph cut. Finally, (5) the pixel

values are synthesized from the selected images with Poisson image editing to produce an omnidirectional image without the photographer. In the following, steps (2) through (5) are described in detail.

### 2.2 Calculation of translation

Each input image is converted to grayscale and Harris corner detection [6] is performed. In Harris corner detection, pixels where large changes in pixel values are observed are detected, and these are used as feature points. Fig. 1 shows an image in which feature points were extracted by Harris corner detection.

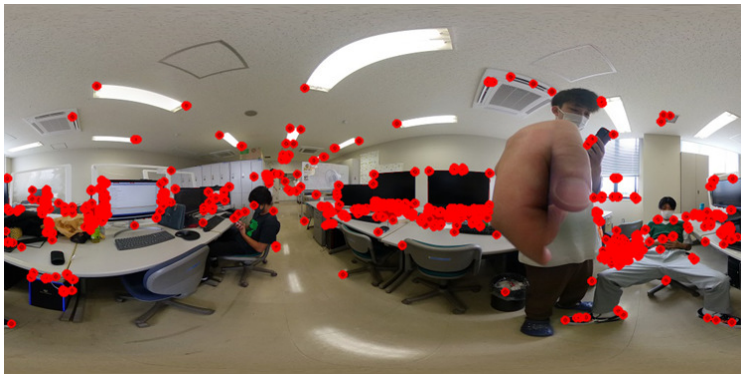


Fig. 1: Example image in which feature points are extracted by Harris corner detection.

Next, the sum of squared differences (SSD) of pixel values is used to correspond the feature points between each input image. Specifically, the method calculates the SSD between the patches centered on the feature points in one of the input images and the patches centered on the feature points in each of the other input images, and finds the feature points with the smallest SSD for each feature point, and corresponds them.

As mentioned above, in each omnidirectional image, this study uses the assumption that the same object is at approximately the same height due to the camera function that outputs omnidirectional images using the direction of gravity of the camera. Therefore, when calculating the SSD between a target feature point and other feature points, the SSD is calculated only for those feature points whose Y-coordinate values of other feature points are within a certain range of the Y-coordinate values of the target feature point. This reduces the computation time and the occurrence of erroneous correspondences.

Next, we calculate the amount of horizontal translation using RANSAC and the least-squares method. RANSAC (Random Sample Consensus) [4] is a robust estimation method that considers the possibility that a given observation

contains outliers and aims to eliminate their effects. In this study, we randomly extract a certain pair of corresponding points from the pairs of corresponding points determined by SSD. We count the number of pairs of corresponding points whose amounts of translations are within a certain range compared to the amount of translation between the selected corresponding points. This process is repeated to determine the pair of corresponding points with the highest number of counts. The corresponding points whose amounts of translations are not within the range compared to the amount of translation of the determined pair are excluded as outliers. After the outliers are removed, the least-squares method is used to find the average of the translations of all the remaining pairs and calculate the amount of translation between the two images. Note that all translations are calculated in the plus direction since the image we are dealing with here is an omnidirectional image, that is, the leftmost pixel is connected to the rightmost pixel.

### 2.3 Unifying the appearance of multiple images

The input images other than the reference one are shifted in the X-axis direction by the amount of translation calculated above. This process makes the X-coordinates of the same object in the background of all the input images equal, and the appearance is unified. Fig. 2(a) shows a reference image and 2(b) shows one of the other images. Fig. 2(c) shows the image after translating Fig. (b) so that its appearance becomes Fig. (a). As can be seen from Figs. (a) and (c), background objects other than the photographer have approximately the same X-coordinates.

### 2.4 Image composition by image selection

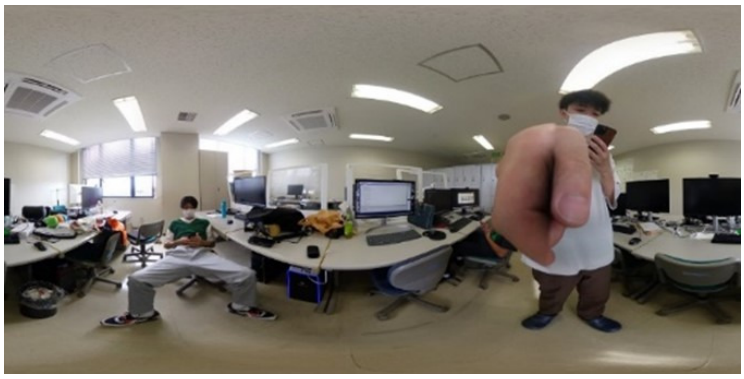
An appropriate image for each pixel is selected from a set of images with unified appearance by energy minimization using graph cut algorithm [1], and the images are combined with Poisson image editing [11] to produce an omnidirectional panoramic image without a photographer. Energy function $E$ is defined as follows:

$$E = \lambda \sum_{u \in A} E_1(f_u) + \frac{\kappa}{2} \sum_{(u,v) \in N} E_2(f_u, f_v), \tag{1}$$

where $f_u$ and $f_v$ are image indices for pixel $u$ and $v$, respectively. The pixel colors of the image indices are modified and synthesized to produce the final result. $A$ is a set of all pixels in the omnidirectional image, and $N$ is a set of pairs of adjacent pixels in the omnidirectional image. In this study, the leftmost and rightmost pixels in an omnidirectional image are also considered as a neighboring pair. $\lambda$ and $\kappa$ are weights to balance the two terms.

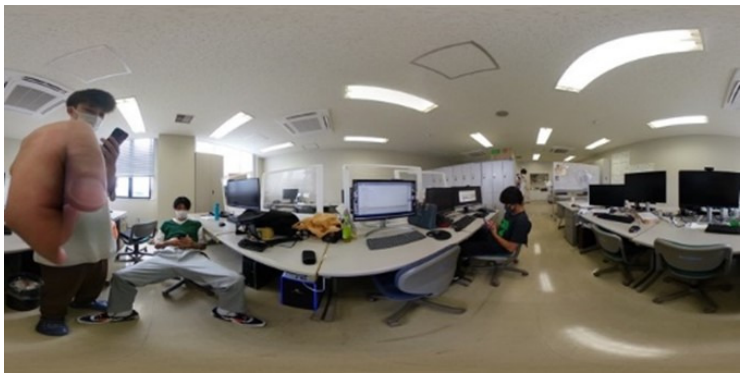
The first energy $E_1$ is a data term, which is based on the plausibility of the background, and defined as follows:

$$E(f_u) = ||\mathbf{I}_{f_u}(u) - \mathbf{M}(u)||, \tag{2}$$



(a) Reference image

(b) Other image



(c) Translated image

Fig. 2: Unifying the appearance by horizontal translation.

where $\mathbf{I}_{f_u}(u)$ is the vector of RGB colors of pixel $u$ in image $f_u$. $\mathbf{M}(u)$ is the vector of RGB colors of pixel $u$ in the smoothed median image that is generated by calculating the median value of all the unified images for each pixel and smoothing the values by a moving average filter. Here, the median is calculated independently for RGB. In the median image, non-photographer colors are selected with high probability. Therefore, by minimizing this energy, the resulting image is similar to the median image, producing an image without a photographer.

The second energy $E_2$ is a smoothness term and defined as follows:

$$E_2(f_u, f_v) = ||\mathbf{I}_{f_u}(u) - \mathbf{I}_{f_v}(u)|| + ||\mathbf{I}_{f_u}(v) - \mathbf{I}_{f_v}(v)|| \tag{3}$$

This term prevents from frequently changing image indices between adjacent pixels. Even when the source image indices are different between neighboring pixels, the indices are switched where the difference in pixel values between the source images is small. This makes the border where the indices are different between adjacent pixels less noticeable.

The overall energy E is minimized by graph cut, but this time the number of image indices is larger than 2. For this reason, we use the $\alpha$-$\beta$ swap algorithm. Specifically, we extract two indices and use graph cut to swap the two indices. This is done for all pairs of indices. This process is done until there are no more index swaps, or until a certain number of swaps are repeated.

Finally, an omnidirectional panoramic image without the photographer is generated by combining the pixel values from the image for each pixel selected by graph cut with Poisson image editing, which makes the color difference less noticeable at the boundary where the image indices switch.

## 3 Experiments

To demonstrate the effectiveness of the proposed method, experiments were conducted in three different scenes. We used an omnidirectional camera RICOH THETA Z1 for input, which outputs an omnidirectional image in which the direction of gravity is the downward direction of the image, although the specific algorithm has not been disclosed. The resolution of the omnidirectional images was resized to 1024x512 pixels for the experiments, and three input images were used in each scene. The patch size for SSD calculation was set to 21x21 pixels. The weights of the graph cuts were changed for each scene in the experiment. We also compared the results with the median images. In the following, we describe the experimental results for each scene in detail.

### 3.1 Scene A: Indoor scene

Fig. 3 shows the input images in scene A. This figure shows that the photographer is at approximately the same position in the image, but the background has shifted to the side. The top left image in Fig. 3 was used as a reference, and the

Fig. 3: Scene A: Three input images.

Fig. 4: Scene A: Translated input images.

other two images were horizontally translated as shown in Fig. 4. As shown in the figure, background objects other than the photographer have approximately the same x-coordinates.

The median image of the three translated input images is shown in Fig. 5. From the figure, we can confirm that the photographer has disappeared, but the image is produced with a blurred texture in the vertical direction. We assume that the original input images were generated by detecting the direction of gravity using an acceleration sensor, but errors in the direction caused by the movement of the hand holding the camera may have resulted in a shift in the vertical texture. As a result, vertical texture blurring occurs when images are merged using the simple median.

Next, we show the results of graph cut in the proposed method. The parameter $\lambda$ in the graph cut was fixed to 100, and we output the results with three different types of $\kappa$. Fig. 6 shows the resulting images by just copying the pixel values that were selected by graph cut with each value of kappa and the corresponding images showing the indices of the source images.

Fig. 5: Scene A: Median image of three translated input images.

When the value of $\kappa$ is small, the resulting image is close to the Median image, and vertical blurring can be often seen. We can also confirm that the indices of the images are frequently changed. When $\kappa$ is 100, the image indices are less interchangeable, and the same indices are clustered in large regions. As a result, blurring of the image in the vertical direction is almost completely eliminated. However, the fluorescent light in the upper right corner and the unnatural skin tone in the middle left corner remain. When $\kappa$ is 500, an omnidirectional image without the photographer is generated with almost no discomfort in any of the regions. However, with a simple copy of the images, we can observe the edge caused by the difference in color at the floor and ceiling.

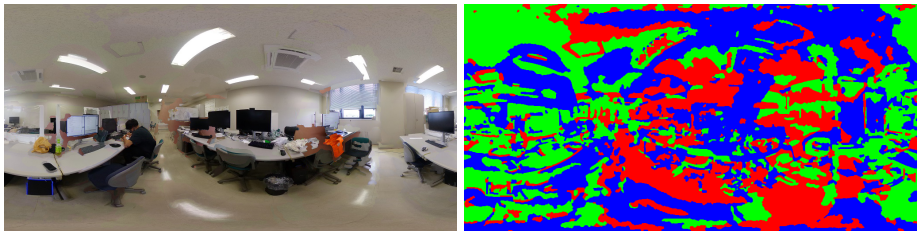
Fig. 7 shows the final result of the proposed method by applying Poisson image editing to Fig. 6(c). As shown in the figure, even at the boundary where the image indices switch, the edges caused by the difference in colors are no longer noticeable, indicating that the image without the photographer has been successfully created.

### 3.2 Scene B: Outdoor scene

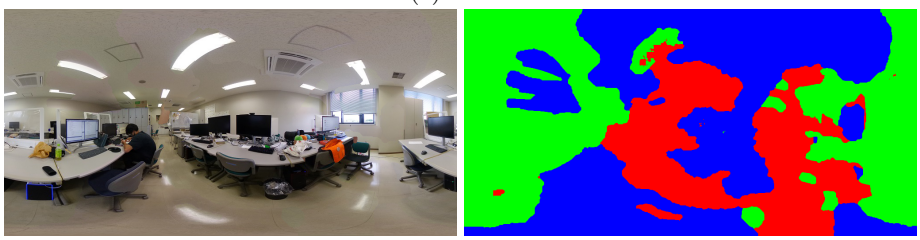
We conducted experiments in an outdoor scene as scene B. Fig. 8 shows the input images in scene B. In this scene, a high building and some trees are captured. These input images were horizontally translated with the left top image in Fig. 8 as a reference.

The median image of the three translated input images is shown in Fig. 9. Also in this scene, the image has blurred textures in the vertical direction.

Fig. 10 shows the result of the proposed method with $\kappa = 500$. In this scene, we were able to generate an omnidirectional image completely without the photographer. However, the shadow of the photographer remained unnaturally on the right side of the composite image. A closer look at this region in the three input images shows that, in the first image it is the photographer's region, in the second image it is this shadow region, and in the third image it is the actual background that is not the shadow of the photographer. Because of the three



(a) $\kappa = 10$



(b) $\kappa = 100$

(c) $\kappa = 500$

Fig. 6: Scene A: Results of graph cut. The left images are resulting ones and the right images indicate the image indices of the source images.

Fig. 7: Scene A: Result of the proposed method.

Fig. 8: Scene B: Three input images.

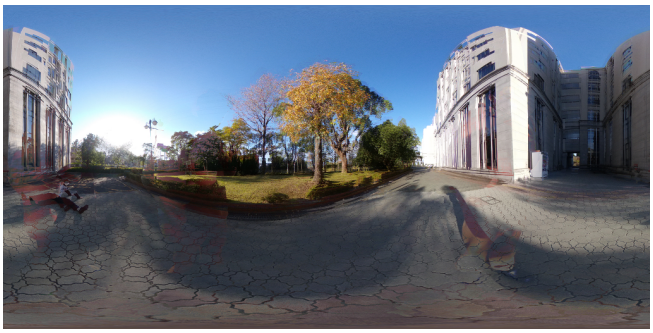


Fig. 9: Scene B: Median image of three translated input images.

different textures, the actual background was not properly selected in this case. We can also see that in this scene an edge is created at the boundary where the index of the image in the sky region switches, due to the different brightness of the sky. Fig. 11 shows the result after applying Poisson image editing to Fig. 10. The boundary edges in the sky regions were made less noticeable.

### 3.3 Scene C: Narrow indoor scene

We conducted experiments in another indoor scene as scene C. Fig. 12 shows the input images in scene C. These images were taken in an elevator, which is a fairly narrow space. There is a mirror in the elevator. These input images were horizontally translated with the left top image in Fig. 12 as a reference.

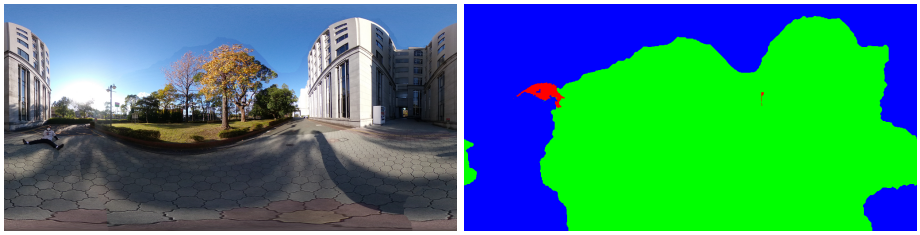


Fig. 10: Scene B: Results of graph cut with $\kappa = 500$. The left image is resulting ones and the right images indicate the image indices of the source images.

Fig. 11: Scene B: Result of the proposed method.

The median image of the three translated input images is shown in Fig. 13. The image also has blurred textures in the vertical direction, and unnatural color tones appear in some regions.

Fig. 14 shows the result of graph cut with $\kappa = 300$. In this scene, we were also able to generate an omnidirectional image completely without the photographer. However, a vertical texture shift can be observed at the boundary where the image indices switch. As mentioned when showing the median image in scene A, there are vertical shifts in the input images. Especially in a narrow space where the distance between the camera and the object is close, a small error in the direction of gravity can lead to a large shift in the image. In addition, the photographer is reflected in the mirror area. Similar to scene B, in the first image, the region is the one where the photographer is in the mirror, in the second image it is the photographer's region, and in the third image it is the actual background. Therefore, the actual background was not properly selected in this case. Fig. 15 shows the final result of the proposed method with Poisson image editing. Although Poisson image editing has made the boundary less noticeable, it is still difficult to compensate for the vertical shift.

Fig. 12: Scene C: Three input images.

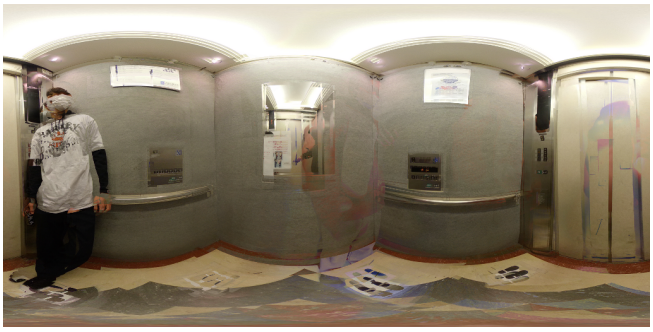


Fig. 13: Scene C: Median image of three translated input images.

### 3.4 Discussion

**Vertical misalignment** Since the input images have vertical misalignment, this misalignment causes a fatal degradation of the quality of the image by the simple median. On the other hand, the proposed method was able to generate an omnidirectional image without the photographer with little noticeable misalignment by selecting an appropriate image index by energy minimization even when there was some vertical misalignment. However, in a narrow space where the object to be photographed is close to the camera, we confirmed that the vertical misalignment still has a significant impact on the quality of the generated image.

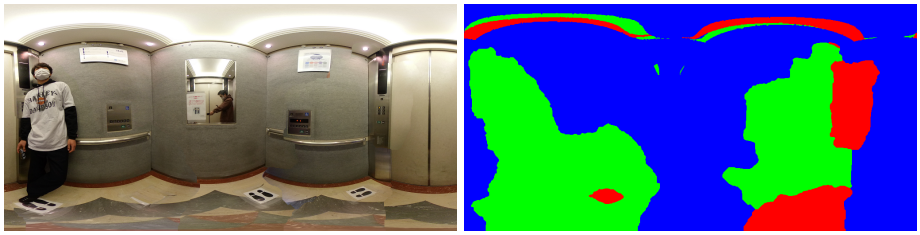


Fig. 14: Scene C: Results of graph cut with $\kappa = 300$. The left image is resulting ones and the right images indicate the image indices of the source images.

Fig. 15: Scene C: Result of the proposed method.

**Number of input images** A prerequisite for the proposed method to work well is that the median image should have approximately the value of the actual background. Although the experiments were conducted with three input images, when there are cast shadows from a strong light source or reflections in a mirror, all three images may have different textures in a certain region. In such cases, an appropriate median image is not created, resulting in an unnatural texture being generated in the resulting image. Therefore, in some scenes, more than three images are needed to obtain successful results.

## 4 Conclusion

In this study, we proposed a method to generate an omnidirectional panoramic image without the photographer in it by using multiple images taken by an omnidirectional camera. In the proposed method, captured images are matched by feature point matching. Then, the amount of horizontal translation calculated from the feature point matching is used to unify the appearance of all the omnidirectional images. Finally, we combine these images by image selection through energy minimization by graph cut algorithm and Poisson image editing to generate an omnidirectional image without the photographer in it.

As future work, it is necessary to correct the vertical misalignment before integrating the images. In addition, although the experiment was conducted with three images, we found that three images were not enough in environments with shadows and reflections. Therefore, it is necessary to conduct experiments using more images.

**Acknowledgements** This work was supported by JSPS KAKENHI Grant Numbers JP18H03273, JP18H04116, JP21H03483.

## References

1. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimizaiton in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(9), 1124–1137 (2004)
2. Cohen, S.: Background estimation as a labeling problem. In: IEEE International Conference on Computer Vision. pp. 1034–1041 (2018)
3. Criminisi, A., Perez, P., Toyama, K.: Region filling and object removal by exemplar based image inpainting. IEEE Transactions on Image Processing **13**(9), 1200–1212 (2004)
4. Fischler, M.A., Bolles, R.C., Bae, S., Yi, J.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
5. Flores, A., Belongie, S.: Re moving pedestrians from google street view images. In: International Workshop on Mobile Vision. pp. 53–58 (2010)
6. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proc. Alvey Vision Conference. pp. 147–151 (1988)
7. Jancosek, M., Pajdla, T.: Multi-view reconstruction preserving weakly-supported surfaces. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3121–3128 (2011)
8. Kawai, N., Inoue, N., an Fumio Okura, T.S., Nakashima, Y., Yokoya, N.: Background estimation for a single omnidirectional image sequence captured with a moving camera. IPSJ Transactions on Computer Vison and Applications **6**, 68–72 (2014)
9. Kawai, N., Yokoya, N.: Image inpainting considering symmetric patterns. In: IAPR International Conference on Pattern Recognition. pp. 2744–2747 (2012)
10. Le, T.T., Almansa, A., Gousseau, Y., Masnou, S.: Object removal from complex videos using a few annotations. Computational Visual Media **22**(5), 267–291 (2019)
11. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. ACM Transactions on Graphics **22**(3), 313–318 (2003)
12. Telea, A.: An image inpainting technique based on the fast marching method. Journal of Graphics Tools **9**(1), 23–24 (2004)
13. Wu, C.: Visualsfm: A visual structure from motion system. `http://ccwu.me/vsfm/`
14. Xu, R., Li, X., Zhou, B., Loy, C.C.: Deep flow-guided video inpainting. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
15. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5505–5514 (2018)