Generation of an Omnidirectional Video without Invisible Areas Using Image Inpainting

Norihiko Kawai, Kotaro Machikita, Tomokazu Sato, and Naokazu Yokoya

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
{norihi-k, tomoka-s, yokoya}@is.naist.jp
http://yokoya.naist.jp/

Abstract. Omnidirectional cameras usually cannot capture the entire direction of view due to a blind side. Thus, such an invisible part decreases realistic sensation in a telepresence system. In this study, an omnidirectional video without invisible areas is generated by filling in the missing region using an image inpainting technique for highly realistic sensation in telepresence. This paper proposes a new method that successfully inpaints a missing region by compensating for the change in appearance of textures caused by the camera motion and determining a searching area for similar textures considering the camera motion and the shape of the scene around the missing region. In experiments, the effectiveness of the proposed method is demonstrated by inpainting missing regions in a real image sequence captured with an omnidirectional camera and generating an omnidirectional video without invisible areas.

1 Introduction

Telepresence systems that enable us to experience a remote site are expected to be used in various fields such as entertainment and education. In these fields, omnidirectional videos captured with a moving omnidirectional camera are sometimes used [1, 2]. However, an ordinary omnidirectional camera cannot capture the entire direction of view due to a blind side as shown in Fig. 1. Thus, such an invisible part decreases realistic sensation in telepresence. In order to achieve telepresence with highly realistic sensation, this research aims at generating an omnidirectional video without invisible areas by inpainting the missing region caused by the blind side. Conventionally, many image inpainting methods for a still image have been proposed [3–5]. Missing regions in not only a still image but also a video can be filled in by applying these methods to each frame in a video. However, textures may discontinuously change between successive frames because the methods use only information in a frame.

On the other hand, methods that fill in missing regions in a video considering temporal continuity have been proposed [6–11]. These methods are classified into two categories. One uses the motion information of a scene in an image sequence

Fig. 1. Omnidirectional panorama image with missing region (black region) caused by the blind side.

[6–9] and the other does not [10, 11]. The former method specifies the appropriate textures for missing regions by calculating the motion of objects in a video or the motion of a camera and fills in the missing regions using the specified texture. The latter method searches whole the video for the spatial-temporal volume similar to that around missing regions and fills in the missing regions using the similar volumes. Both methods can generate a video with temporally continuous change in texture. However, these methods do not consider the change in the appearance of textures caused by the camera motion. Therefore, it is difficult for these methods to successfully inpaint missing regions in an omnidirectional video caused by the blind side of an omnidirectional camera because the appearance of the texture appropriate for a missing region in a frame changes in different frames of a moving omnidirectional camera.

To overcome these problems, this paper proposes a new method that successfully inpaints a missing region compensating for the change in the appearance of textures. Concretely, by assuming that the shape of the blind side of the target scene is planar, the change in the appearance of the texture caused by the camera motion is compensated by projecting omnidirectional images onto the planar surface fitted to the 3-D positions of natural feature points on the ground acquired by structure-from-motion (SFM). In addition, by using the fitted plane and the camera motion, the data region in which appropriate textures for missing regions may exist is determined. Finally, good quality images are obtained by using an image inpainting technique. In this research, we employ an omnidirectional multi-camera system (OMS) that is composed of radially arranged multiple cameras and we assume that the ground exists in the direction of the blind side of a moving OMS.

2 Generation of an omnidirectional video without invisible areas

The flow of the proposed method is as follows. (a) The position and posture of an OMS and 3-D positions of natural feature points are estimated using SFM for

an omnidirectional video. (b) A plane for each frame is fitted to natural feature points near the ground by using the position and the posture of the OMS and the 3-D positions of natural feature points. (c) An image sequence projected on the fitted plane is generated from the omnidirectional video. (d) Data regions in which appropriate textures for missing regions may exist are specified on the projected image plane using the position and posture of the OMS and the fitted planes. (e) A missing region in the projected image plane of each frame is successively inpainted by minimizing an energy function based on the similarity between the texture in the missing region and the specified data region. (f) An omnidirectional video without invisible areas is generated by re-projecting the inpainted image onto the omnidirectional panoramic video with a missing region. In the following sections, each process is described in detail.

2.1 Estimation of extrinsic camera parameters and positions of natural feature points

The position and posture of an OMS and 3-D positions of natural feature points are estimated by SFM [12] for an omnidirectional image. In this method, first, a target scene is captured with a moving OMS. Next, initial extrinsic camera parameters and 3-D positions of feature points are estimated by tracking the natural feature points in a video, which are detected by Harris operator. Finally, the accumulative errors of the camera parameters and the 3-D positions of feature points are minimized by bundle adjustment for whole the video.

2.2 Generation of images projected on planes by estimating shapes around missing regions

In this research, on the assumption that an omnidirectional video is captured while moving on the ground and the shape around a missing region is planar, an image sequence that includes missing regions is generated by projecting the omnidirectional video to the planes in order to compensate for the change in the appearance of textures caused by the camera motion.

Concretely, first, natural feature points for plane fitting are selected from the points obtained by SFM described in Section 2.1. Here, the points that satisfy the following conditions are selected: (i) a point exists in the spherical area whose center is a projection center of a representative camera unit of an OMS and radius is l, and (ii) the height z of a point in the world coordinate system is $(p < z < p+m)$ (p and m are constants) as shown in Fig 2. Next, the expression of the plane that represents the ground in the world coordinate system is set as $z = ax + by + c$, and the parameters (a, b, c) are determined by the least-square method so as to minimize the following cost function L.

$$L = \sum_{i=1}^{n} (ax_i + by_i + c - z_i)^2, \tag{1}$$

where (x_i, y_i, z_i) are the coordinates of a feature point and n is the number of selected feature points. An image sequence is generated by projecting the

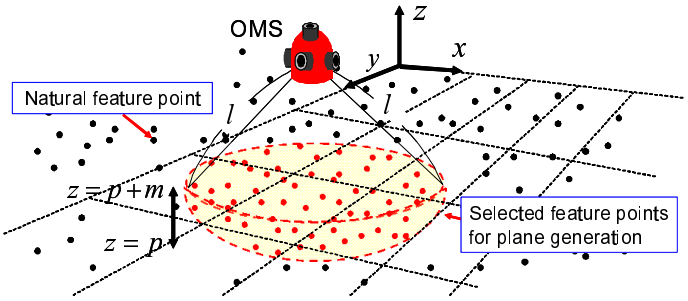


Fig. 2. Selection of feature points around missing region.

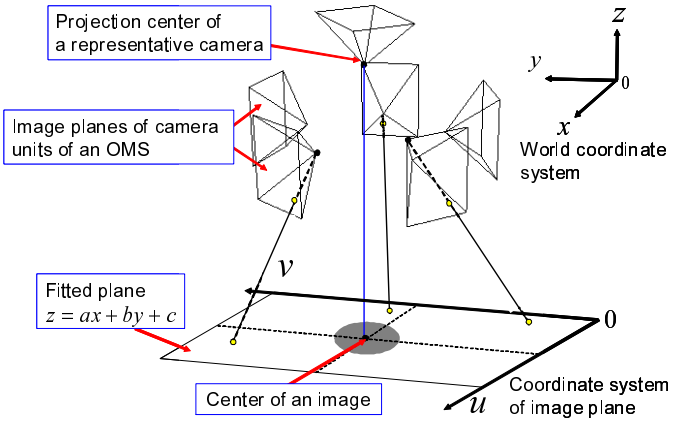


Fig. 3. Generation of an image projected on a plane.

omnidirectional video to the estimated plane for each frame as shown in Fig. 3. Here, in order for a missing region to be the center of the projected image, an intersection point of the plane with the straight line that goes just under an OMS through the projection center of a representative camera of the OMS is set as the center of the image. Additionally, in order to prevent the rotation of the textures in projected image planes, the basis vectors $(\mathbf{u}, \mathbf{v})$ of the image in the world coordinate system are set so as to satisfy the following equation.

$$\mathbf{u} \cdot \mathbf{y} = 0, \tag{2}$$

where $\mathbf{y}$ is one of the basis vectors of the world coordinate system.

2.3 Inpainting a missing region based on energy minimization

A missing region in each frame is successively inpainted by applying an energy minimization method to each image projected on a plane with a missing region generated by the method described in the previous section. In the following,

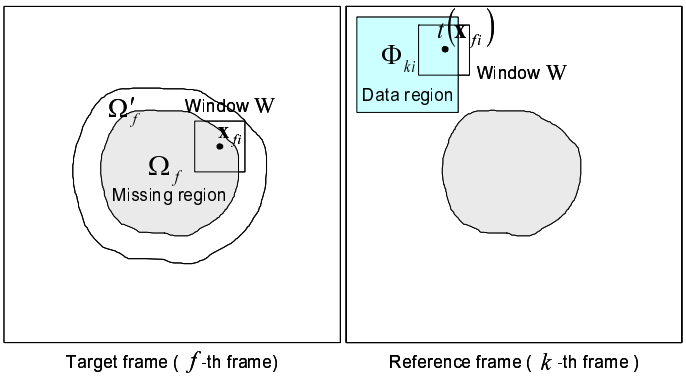


Fig. 4. Missing and data regions in projected images for inpainting process.

the definition of an energy function, a method determining a data region and a method minimizing the energy function are described.

Definition of an energy function As shown in Fig. 4, a missing region in the projected image of the f-th frame (target frame) is inpainted using an energy function based on the similarity of textures between region Ω'_f including missing region Ω_f in the f-th frame and data region Φ_{ki} in the k-th frame (reference frame) ($k \neq f$). Here, Ω'_f is the expanded area of the missing region Ω_f in which there is a central pixel, $\mathbf{x}_{fi}$, of a square window W overlapping region Ω_f and each data region Φ_{ki} corresponding to each pixel $\mathbf{x}_{fi}$ in the f-th frame is individually determined. Energy function E is defined as the weighted sum of SSD (Sum of Squared Differences) between the textures around pixel $\mathbf{x}_{fi}$ in region Ω'_f and $t(\mathbf{x}_{fi})$ in data region Φ_{ki}.

$$E = \sum_{\mathbf{x}_{fi} \in \Omega'_f} w_{\mathbf{x}_{fi}} SSD(\mathbf{x}_{fi}, t(\mathbf{x}_{fi})), \tag{3}$$

where $w_{\mathbf{x}_{fi}}$ is the weight for pixel $\mathbf{x}_{fi}$ and is set as 1 if $\mathbf{x}_{fi}$ is inside of region $\Omega'_f \cap \overline{\Omega_f}$ because pixel values in this region are fixed; otherwise $w_{\mathbf{x}_{fi}} = g^{-d}$ (d is the distance from the boundary of Ω_f and g is a constant) because pixel values around the boundary have higher confidence than those in the center of the missing region.

$SSD(\mathbf{x}_{fi}, t(\mathbf{x}_{fi}))$, which represents the similarity of textures around pixel $\mathbf{x}_{fi}$ and $t(\mathbf{x}_{fi})$, is defined as follows:

$$SSD(\mathbf{x}_{fi}, t(\mathbf{x}_{fi})) = \sum_{\mathbf{q} \in W} \{I(\mathbf{x}_{fi} + \mathbf{q}) - \alpha_{\mathbf{x}_{fi} t(\mathbf{x}_{fi})} I(t(\mathbf{x}_{fi}) + \mathbf{q})\}^2, \tag{4}$$

where $I(\mathbf{x})$ represents the pixel value of pixel $\mathbf{x}$. $\alpha_{\mathbf{x}_{fi} t(\mathbf{x}_{fi})}$ is the intensity modification coefficient. Note that textures around a missing region may change due to the reflection of the light on the ground and the shadow of the camera and operator. Therefore, by using this coefficient, the brightness of textures in data

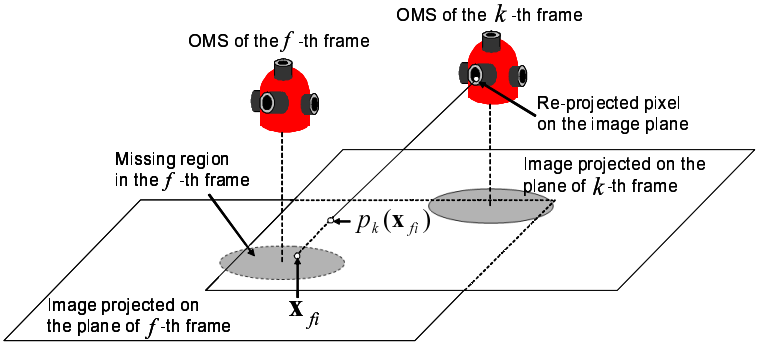


Fig. 5. Projection to the other frame.

regions is adjusted to that in the missing region. In this research, $\alpha_{\mathbf{x}_{fi}t(\mathbf{x}_{fi})}$ is defined as the ratio of average pixel values around pixels $\mathbf{x}_{fi}$ and $t(\mathbf{x}_{fi})$ as follows:

$$\alpha_{\mathbf{x}_{fi}t(\mathbf{x}_{fi})} = \frac{\sqrt{\sum_{\mathbf{q}\in W} I(\mathbf{x}_{fi}+\mathbf{q})^2}}{\sqrt{\sum_{\mathbf{q}\in W} I(t(\mathbf{x}_{fi})+\mathbf{q})^2}}. \tag{5}$$

Determination of a data region A data region in which position $t(\mathbf{x}_{fi})$ of the most similar texture pattern may exist is determined by using the position and posture of a moving OMS estimated in Section 2.1 and the planes generated in Section 2.2. In this research, appropriate textures for the missing region in the f-th frame are expected to be captured in the frames other than the target frame by assuming that the omnidirectional video is captured while moving. Additionally, the parameters of the plane and the position and posture of the OMS in each frame are known. Therefore, regions in which the most similar pattern exists can be determined in the frames other than the target frame by using the geometric relationships of a moving camera system and the ground. Also, an appropriate frame is determined considering the resolution of the similar texture pattern. In the following, we describe the way to determine a region and a frame that are used as a data region for the energy minimization process described in the following section.

First, the 3-D coordinate of pixel $\mathbf{x}_{fi}$ in the target (f-th) projected image is re-projected on the image plane of a camera unit of the OMS in the k-th frame. Then, the pixel coordinate $p_k(\mathbf{x}_{fi})$ of the intersection of the k-th projected image with the straight line that goes through the re-projected pixel on the image plane of the camera unit and pixel $\mathbf{x}_{fi}$ on the f-th projected image is calculated as shown in Fig. 5. In a similar way, pixel coordinate $p_k(\mathbf{x}_{fi})$ in each frame k corresponding to pixel $\mathbf{x}_{fi}$ is calculated. Next, a frame is selected considering the position of $p_k(\mathbf{x}_{fi})$ in a projected image and the difference of frames between the target and the reference frames. In projected images, the resolution of texture becomes lower the farther a pixel is from the center of the image because textures of objects remote from the camera become small in input images of an OMS. In

order to prevent the generation of blurred textures, textures near the center of the image should be used as samples for inpainting. In addition, it is highly possible that temporally close frames have similar brightness of textures. Therefore, the appropriate frame $s(\mathbf{x}_{fi})$ is selected from candidate frames $\mathbf{K} = (k_1, \ldots, k_n)$ by the following equation.

$$s(\mathbf{x}_{fi}) = \underset{k \in \mathbf{K}}{\operatorname{argmin}}(\|p_k(\mathbf{x}_{fi}) - \mathbf{x}_{center}\| + \lambda |k - f|), \tag{6}$$

where candidate frames $\mathbf{K}$ are picked up so that the fixed range of the texture around $p_k(\mathbf{x}_{fi})$ does not include the missing region. $\mathbf{x}_{center}$ is the central pixel in the k-th planar projected image and λ is the weight for the difference of frames. Finally, fixed square area S whose center is pixel $p_{s(\mathbf{x}_{fi})}(\mathbf{x}_{fi})$ is set as a data region $\Phi_{s(\mathbf{x}_{fi})i}$, which is used for the energy minimization process described in the following section. In a similar way, each data region $\Phi_{s(\mathbf{x}_{fi})i}$ corresponding to each pixel $\mathbf{x}_{fi}$ in expanded missing region Ω'_f is individually determined.

Energy minimization Energy function E in Eq. (3) is minimized by using a framework of greedy algorithm in a similar way to [13]. In our definition of energy E, the energy for each pixel can be treated independently if pattern pairs ($\mathbf{x}_{fi}$,$t(\mathbf{x}_{fi})$) can be fixed and the change of coefficient $\alpha_{\mathbf{x}_{fi}t(\mathbf{x}_{fi})}$ in the iterative process of energy minimization is very small. Thus, we repeat the following two processes until the energy converges: (i) search for the most similar pattern keeping pixel values fixed, and (ii) perform a parallel update of all pixel values keeping pattern pairs fixed.

In process (i), data region Φ_{ki} ($k = s(\mathbf{x}_{fi})$) is searched for position $t(\mathbf{x}_{fi})$ of the most similar pattern keeping pixel values $I(\mathbf{x}_{fi})$ fixed. $t(\mathbf{x}_{fi})$ is determined as follows:

$$t(\mathbf{x}_{fi}) = \underset{\mathbf{x} \in \Phi_{ki}}{\operatorname{argmin}}(SSD(\mathbf{x}_{fi}, \mathbf{x})). \tag{7}$$

In process (ii), all pixel values $I(\mathbf{x}_{fi})$ are updated in parallel so as to minimize the energy keeping the similar pattern pairs fixed. In the following, the method for calculating pixel values $I(\mathbf{x}_{fi})$ is described. First, energy E is resolved into element energy $E(\mathbf{x}_{fi})$ for each pixel $\mathbf{x}_{fi}$ in missing region Ω_f. Element energy $E(\mathbf{x}_{fi})$ can be expressed in terms of the pixel values of $\mathbf{x}_{fi}$ and $f(\mathbf{x}_{fi} + \mathbf{q}) - \mathbf{q}$, coefficient α as follows:

$$E(\mathbf{x}_{fi}) = \sum_{\mathbf{q} \in W} w_{(\mathbf{x}_{fi}+\mathbf{q})} \{I(\mathbf{x}_{fi}) - \alpha_{(\mathbf{x}_{fi}+\mathbf{q})t(\mathbf{x}_{fi}+\mathbf{q})} I(t(\mathbf{x}_{fi} + \mathbf{q}) - \mathbf{q})\}^2. \tag{8}$$

The relationship between energy E and element energy $E(\mathbf{x}_{fi})$ for each pixel can be written as follows:

$$E = \sum_{\mathbf{x}_{fi} \in \Omega} E(\mathbf{x}_{fi}) + C. \tag{9}$$

C is the energy of pixels in region $\Omega'_f \cap \overline{\Omega_f}$, and is treated as a constant because pixel values in the region and all pattern pairs are fixed in process (ii). Therefore,

by minimizing element energy $E(\mathbf{x}_{fi})$ respectively, total energy E can be minimized. Here, if it is assumed that the change of $\alpha_{\mathbf{x}_{fi}t(\mathbf{x}_{fi})}$ is much smaller than that of pixel value $I(\mathbf{x}_{fi})$, by differentiating $E(\mathbf{x}_{fi})$ with respect to $I(\mathbf{x}_{fi})$, each pixel value $I(\mathbf{x}_{fi})$ in missing region Ω_f can be calculated in parallel as follows:

$$I(\mathbf{x}_{fi}) = \frac{\sum_{\mathbf{q}\in W} w_{(\mathbf{x}_{fi}+\mathbf{q})}\alpha_{(\mathbf{x}_{fi}+\mathbf{q})t(\mathbf{x}_{fi}+\mathbf{q})}I(t(\mathbf{x}_{fi}+\mathbf{q})-\mathbf{q})}{\sum_{\mathbf{q}\in W} w_{(\mathbf{x}_{fi}+\mathbf{q})}}. \tag{10}$$

In addition, a coarse-to-fine approach is also employed for energy minimization. Concretely, an image pyramid is generated and processes (i) and (ii) are repeated from higher-level to lower-level layers successively. This makes it possible to decrease computational cost and avoid local minima.

2.4 Generation of an omnidirectional video using inpainted images

An omnidirectional video without invisible areas is generated by re-projecting the projected images inpainted in the previous section onto spherical panoramic images with a missing region. Concretely, first, the coordinate of the intersection of the plane with the straight line that goes through the projection center of a camera unit and each pixel in the missing region in the spherical panoramic image is calculated. Next, the pixel value of the calculated coordinate in the projected image is copied to the panoramic image.

3 Experiments

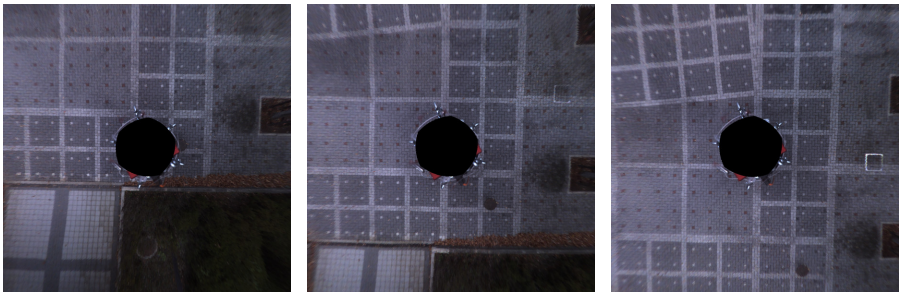
In this section, the effectiveness of the proposed method is demonstrated by inpainting a missing region caused by the blind side of an OMS and generating an omnidirectional video without invisible areas. In the following, the experiment of inpainting for images projected on images is described and a prototype telepresence system using the omnidirectional video without invisible areas is presented.

3.1 Inpainting a missing region in an omnidirectional video

In this experiment, we used Ladybug [14] as an OMS that is composed of 6 camera units and an omnidirectional image sequence (300 frames) is captured. Figure 6 shows the 1st frame of 6 image sequences captured with Ladybug. The position and posture of Ladybug and the positions of natural feature points were obtained by SFM [12] described in Section 2.1. A missing region in each projected image is determined by manually specifying the region in 6 images of the first frame. In addition, a blind region in the projected image is also specified as the missing region.

First, as shown in Fig. 7, images projected on planes were generated by the method described in Section 2.2. The resolution of a projected image was set as 1200 $\times$ 1200 pixels. Round black regions in the images are missing regions

Fig. 6. 1st frame of input image sequence obtained by 6 camera units.



1st frame 51st frame 101st frame

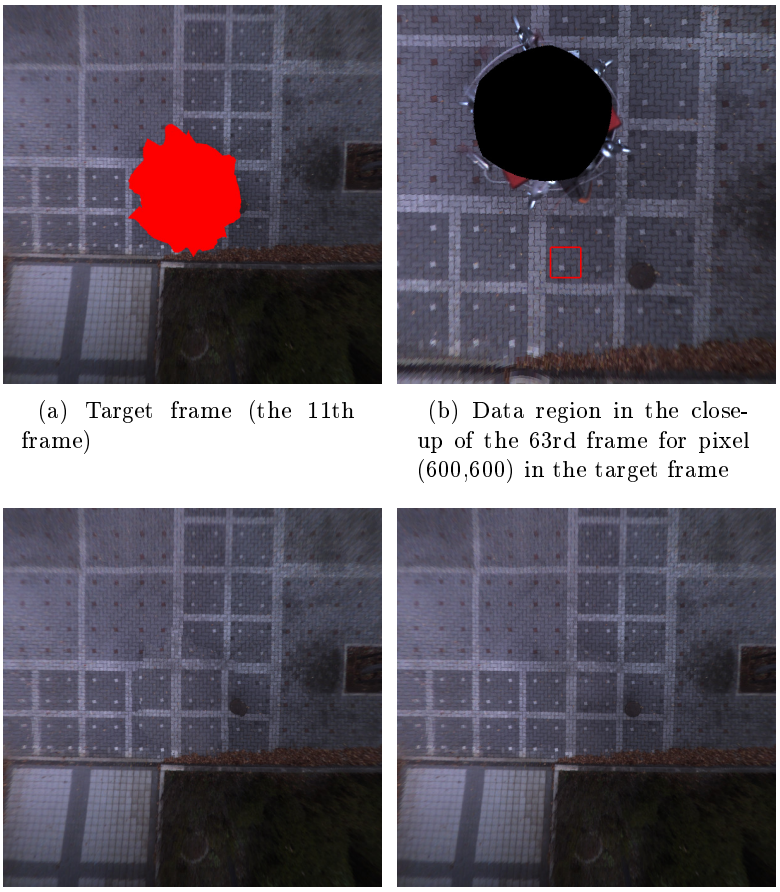
Fig. 7. Images projected on planes.

caused by the blind side of Ladybug. As shown in these figures, textures of tiles on the ground are uniform regardless of the position of pixels and textures of the same objects do not rotate in each frame. As a result, appropriate projected images used for inpainting were generated.

Next, a missing region in each projected image was inpainted. Figure 8 shows the experiment of inpainting for the projected image of the 11th frame. Figure 8(a) shows the target 11th frame in which the missing region is specified and Fig. 8(b) shows the data region in the close-up of the 63rd frame corresponding to pixel (600,600) in the target frame. Figure 8(c) shows the result by projecting pixel values in other frames onto the missing region in the target frame using the position and posture of Ladybug and the generated plane without the inpainting process. From this figure, the geometrical and optical disconnect of textures in the boundary of the missing region appears. We consider this is because of the errors of the estimation of camera parameters by SFM and errors of plane fitting. On the other hand, in the resultant image by the proposed method as shown in Fig. 8(d), textures continuously connect on the boundary and plausible textures are generated in the missing region. Fig. 9 shows the inpainted images corresponding to Fig. 7. In each frame, the missing region is successfully inpainted.

3.2 Omnidirectional telepresence without invisible areas

In this experiment, the effectiveness of the proposed method is demonstrated by making the telepresence system using an omnidirectional video in which missing regions are filled in with inpainted images shown in the previous section. Figure 10 shows the omnidirectional panorama image without invisible areas



(a) Target frame (the 11th frame)

(b) Data region in the close-up of the 63rd frame for pixel (600,600) in the target frame

(c) Result by projection using the plane

(d) Result by the proposed method

Fig. 8. Comparison of results by projection using a plane and proposed method.

generated by projecting the inpainted image (Fig. 9) onto the panoramic image (2048 $\times$1024 pixels). By using the panoramic image as input, we built an omnidirectional telepresence system. Figure 11 shows examples of user's views in the telepresence system. By comparison of the left and right images in Fig. 11, we can confirm that realistic sensation is drastically increased by the proposed method.

4 Conclusion

In this paper, we have proposed a method that generates an omnidirectional video without invisible areas by compensating for the change in the appearance of textures caused by the camera motion and determining a data region

1st frame 51st frame 101st frame

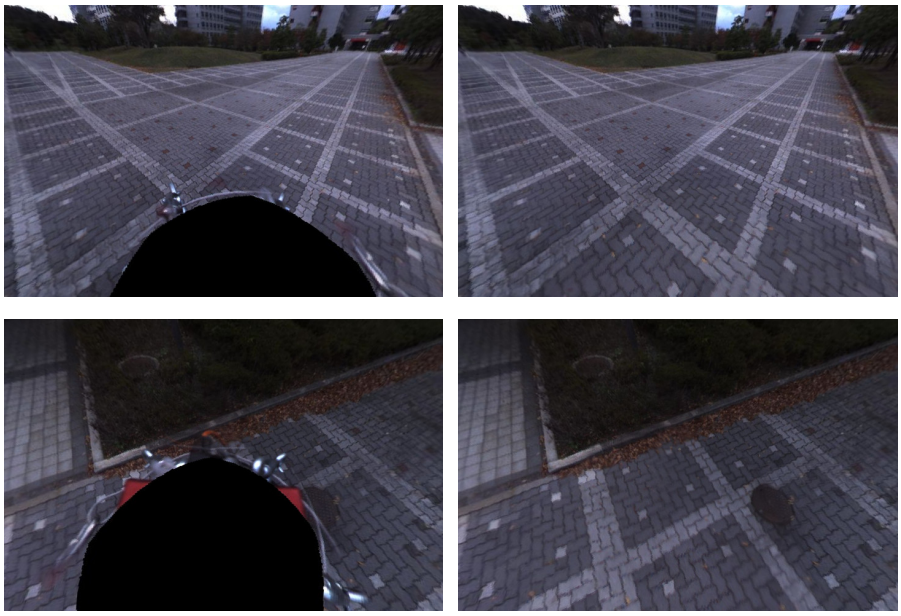
Fig. 9. Inpainted projected images (Corresponding to Fig. 7).

Fig. 10. Filled panoramic image of 1st frame (Corresponding to Fig. 1).

considering the camera motion and the shape of the scene around the missing region. In experiments, missing regions in images projected on planes were successfully inpainted and the omnidirectional telepresence without missing regions was achieved. In future work, we will perform experiments with various scenes. In addition, the proposed method will be evaluated quantitatively by using virtual environments.

References

1. S. Ikeda, T. Sato, N. Yokoya: Immersive Telepresence System with a Locomotion Interface Using High-resolution Omnidirectional Videos. In: Proc. IAPR Conf. on Machine Vision Applications. (2005) 602–605
2. M. Hori, M. Kanbara, N. Yokoya: Novel Stereoscopic View Generation by Image-Based Rendering Coordinated with Depth Information. In: Proc. Scandinavian Conf. on Image Analysis. (2007) 193–202
3. M. Bertalmio, G. Sapiro, V. Caselles, C. Ballester: Image Inpainting. In: Proc. ACM SIGGRAPH2000. (2000) 417–424
4. A. Criminisi, P. Perez, K. Toyama: Region Filling and Object Removal by Exemplar-Based Inpainting. In: IEEE Trans. on Image Processing. Volume 13. (2004) 1200–1212



With a missing region. Without a missing region.

Fig. 11. Looking around using omnidirectional video.

5. N. Komodakis, G. Tziritas: Image Completion Using Global Optimization. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (2006) 442–452
6. Y. Matsushita, E. Ofek, W. Ge, X. Tang, H. Shum: Full-Frame Video Stabilization with Motion Inpainting. IEEE Trans. on Pattern Analysis and Machine Intelligence **28**(7) (2006) 1150–1163
7. J. Jia, Y. Tai, T. Wu, C. Tang: Video Repairing under Variable Illumination Using Cyclic Motions. IEEE Trans. on Pattern Analysis and Machine Intelligence **28**(5) (2006) 832–839
8. Y. Shen, F. Lu, X. Cao, H. Foroosh: Video Completion for Perspective Camera Under Constrained Motion. In: Proc. IEEE Int. Conf. on Pattern Recognition. (2006) 63–66
9. K. Patwardhan, G. Sapiro, M. Bertalmio: Video Inpainting Under Constrained Camera Motion. IEEE Trans. on Image Processing **16** (2007) 545–553
10. Y. Wexler, E. Shechtman, M. Irani: Space-Time Completion of Video. Trans. on Pattern Analysis and Machine Intelligence **29** (2007) 463–476
11. V. Cheung, B. Frey, N. Jojic: Video epitomes. In: Proc. IEEE Conf. on Computer Vision and Patern Recognition. (2005) 141–152
12. T. Sato, S. Ikeda, N. Yokoya: Extrinsic Camera Parameter Recovery from Multiple Image Sequences Captured by an Omni-directional Multi-camera System. In: Proc. European Conf. on Computer Vision. Volume 2. (2004) 326–340
13. N. Kawai, T. Sato, N. Yokoya: Image Inpainting Considering Brightness Change and Spatial Locality of Textures and Its Evaluation. In: Proc. Pacific-Rim Symp. on Image and Video Technology. (2009) 271–282
14. Point Grey Research Inc.: Ladybug. `http://www.ptgrey.com/products/spherical.asp`