Paper

# Free-viewpoint Mobile Robot Teleoperation Interface Using View-dependent Geometry and Texture

Fumio Okura (student member)[†*], Yuko Ueda [†], Tomokazu Sato (member)[†], Naokazu Yokoya (member)[†]

**Abstract** This paper proposes a free-viewpoint interface for mobile-robot teleoperation, which provides viewpoints that are freely configurable by the human operator head pose. The viewpoints are acquired by a head tracker equipped on a head mounted display. A real-time free-viewpoint image generation method based on view-dependent geometry and texture is employed by the interface to synthesize the scene presented to the operator. In addition, a computer graphics model of the robot is superimposed on the free-viewpoint images using an augmented reality technique. We developed a prototype system based on the proposed interface using an omnidirectional camera and depth cameras for experiments. The experiments under both virtual and physical environments demonstrated that the proposed interface can improve the accuracy of the robot operation compared with first- and third-person view interfaces, while the quality of the free-viewpoint images generated by the prototype system was satisfactory for expressing the potential advantages on operational accuracy.

**Key words**: mobile robot, teleoperation, free-viewpoint image generation, view-dependent geometry and texture

## 1. Introduction

Historically, many types of mobile robots have been developed and employed to operate on behalf of humans for various situations[2]. The importance of the teleoperation interface has increased significantly, particularly for working in unknown and/or extreme environments, such as disaster areas with narrow pathways and unforeseen obstacles. Although there has been considerable research devoted to the automatic control of mobile robots[3)4)], most practical robots are still controlled by human operators using video images captured by robot-mounted cameras. These include PackBot[5], which was deployed for surveillance of the Fukushima Daiichi nuclear power plant in Japan after the 2011 earthquake. A human operator should have sufficient skills for controlling a robot to prevent it from colliding with its surroundings so it can safely and effectively complete its assigned tasks. To achieve successful operations, it is important to determine the most effective way to represent to its human operators the field of view surrounding the robot. This objective is even further essential because human vision is the most important sense used by humans to grasp the surrounding environment in teleoperation tasks. To this end, there have been numerous studies on image presentation approaches for teleoperation interfaces of mobile robots[6)–17)].

In terms of image presentation for human operators, existing remote-control interfaces are classified in two categories: those that provide a first-person view from the robot (that is, from the position of the camera on the robot), and those that provide a third-person view (a bird's-eye view from above the robot). While these interfaces are currently used in practical applications, problems with them still remain that decrease robot operation safety. One key problem is the difficulty in grasping distances from the robot to surrounding objects. To address this problem, we propose a mobile-robot teleoperation interface that uses a free-viewpoint image generation technique from computer-vision and -graphics fields. The proposed interface provides the human operator with a novel way to grasp the environment surrounding a robot in a remote site. As demonstrated by our experiments, the interface realizes an intuitive robot operation from viewpoints that are freely configurable by a head-mounted display (HMD) and a head tracker, which provides photorealistic textures of real-world environments, as shown in **Fig. 1**. Our experiments also demonstrate that the proposed interface improves the safety of mobile-robot operations.

This paper is an extended version of the one presented at the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems[1].

Received August 3, 2013; Revised October 18, 2013; Accepted November 20, 2013

† Nara Institute of Science and Technology (NAIST) (Nara, Japan)
∗ JSPS Research Fellow

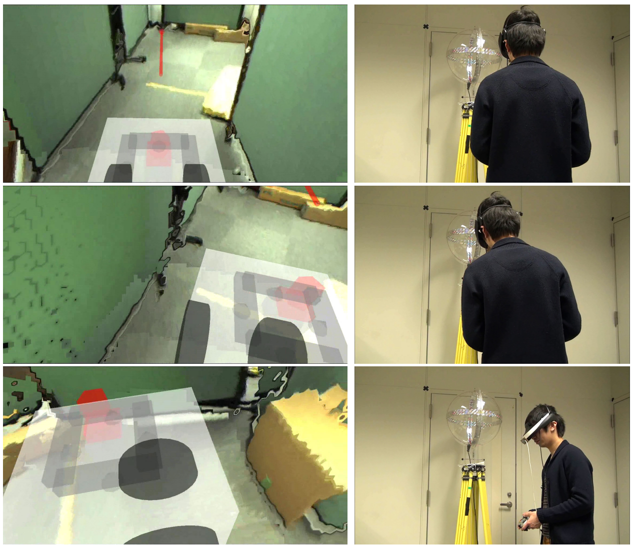


**Fig. 1** Left column: Images from the prototype system presented for the human operator (right column) based on the free-viewpoint teleoperation interface.

## 2. Related work

### 2.1 Mobile-robot teleoperation interfaces

The most common image presentation approach for mobile-robot teleoperation is based on the first-person view[6], which is the scene directly captured by robot-mounted cameras. Most studies and robotic products have employed monocular cameras to provide a first-person view for surveillance in environments such as minefields[7] and sewers[8]. Omnidirectional cameras[18] are often used for first-person-view interfaces that enable operators to configure their view direction to scan the scene. Compared to monocular cameras, omnidirectional cameras have a shorter delay when changing view direction. However, there are two problems with omnidirectional cameras with regard to the operator's understanding of the robot's surroundings. For one, in omnidirectional images, scenes in a downward angle are not visible because the robot occludes them. Second, it is difficult to grasp distances between the robot and surrounding objects.

Interfaces that provide third-person views, which are views from directly above or from above and diagonally behind a robot, have been developed to overcome the above problems. Packbot[5] and Quince[9], mobile-robots employed for the surveillance of the Fukushima Daiichi nuclear power plant, were operated through a third-person-view interface using a pair of robots: one moved forward for surveillance, while the other captured images of the first robot from behind. As another example, a study by Shiroma et al.[10] obtained images of a robot from above by physically mounting a camera on a long arm. Their investigation demonstrated that, in terms of speed and safety, the third-person view is more effective than the first-person view. Although these approaches are expected to facilitate grasping distances to surrounding objects, they do not completely resolve the occlusion problem that occurs in downward scenes.

In most situations, however, it is difficult to directly capture a third-person view from physically mounted cameras; therefore, image processing and/or multi-sensor integration approaches, which combine information captured from the robot, are often employed to generate third-person views. Time Follower's Vision[11], for example, a remote-control vehicle visual-presentation system, provides a viewpoint from behind by displaying images captured several seconds previously as the robot moves forward. To generate more customized viewpoints, three-dimensional (3D) shapes of objects in the surrounding environment acquired from depth sensors mounted on a robot are often used with the images captured by a camera. The interfaces proposed by Saitoh et al.[12], Nielsen et al.[13], and Ferland et al.[14] provide operators with both the first-person view and 3D models reconstructed based on simultaneous localization and mapping (SLAM) approaches[19]. Kelly et al.[15] have realized a photorealistic third-person-view interface by appropriately mapping images to 3D shapes in an outdoor environment. Image-processing-based third-person view interfaces basically combine multiple textures and/or 3D shapes captured from multiple locations to generate the scene at the bottom of the robot, and then artificially superimpose the appearance of the robot. Therefore, these interfaces can improve the visibility of scenes of the bottom or far side of the robot by using a transparent robot appearance, such as the technique used in Time Follower's Vision[11].

In ordinary third-person-view interfaces described above, the operator's viewpoint is fixed or selectable from a few viewpoints that are configured in advance. Although these types of third-person-view interfaces improve the speed and safety of teleoperations, the problems mentioned earlier still remain with the difficulty in grasping the distance to surrounding objects on the far sides of the robot. Because of the limitation of selectable viewpoint variation, the problems are especially apparent in complex environments, such as those containing narrow pathways and obstacles. From this perspective, one of the ultimate forms of third-person-view interfaces is a Virtual Environment Vehicle Interface (VEVI)[16)17)], which provides a freely configurable view; the viewpoint and direction can be freely changed by the operator using an HMD and a head tracker. Although an interface with a freely configurable view is expected to provide intuitive and safe operations, the existing VEVIs[16)17)] have been developed only as virtual-reality interfaces without using real-world textures.

This study realizes an intuitive and freely configurable third-person-view interface using an HMD and a head tracker for providing photorealistic textures of real-world environments. The proposed interface employs state-of-the-art image processing and computer vision methods; therefore, the robot's far side is visible by superimposing a transparent robot appearance. That is, our approach resolves the occlusion problem, while facilitating grasping distances to surrounding objects on the far sides of the robot. We investigate the effectiveness of a prototype free-viewpoint-operation interface for a real environment through evaluations. The approaches used to realize freely configurable views with the textures of a real environment are described in the following section.

### 2. 2 Free-viewpoint image generation

In the fields of computer graphics and computer vision, techniques for generating freely configurable views from multiple images are referred to as free-viewpoint image generation (or arbitrary- or novel-viewpoint image generation). One free-viewpoint image generation approach is known as model-based rendering (MBR). This approach is a traditional computer graphics/vision pipeline that reconstructs 3D shapes of real environments first, and then maps images of the environment over them as textures. At present, 3D shapes can be acquired in real-time from environments ranging from a small desktop space[20)] to a large outdoor environment[21)]. In this approach, the quality of the free-viewpoint images generated by MBR is directly affected by the accuracy of the 3D shapes; that is, unnatural distortions or missing areas in the views are easily exposed. On the other hand, image-based rendering (IBR) generates free-viewpoint images without using explicit 3D shapes. There have been numerous studies on IBR techniques, such as view morphing[22)], light-field rendering[23)24)], and lumigraph rendering[25)26)]. Although IBR reduces missing areas in resulting images, the approach requires images captured at a large number of places and directions. Otherwise, large distortions may appear[27)].

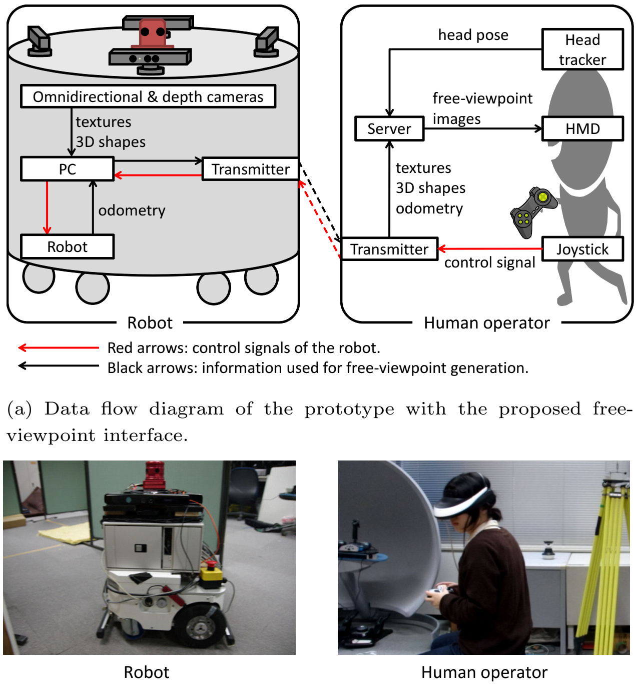
In recent years, the main approach of free-viewpoint image generation is to use hybrid techniques that combine MBR and IBR with the goal of resolving their respective problems. The primary approach of hybrid free-viewpoint generation is view-dependent texture mapping (VDTM), which was originally proposed by Debevec et al.[28)]. This technique selects and blends multiple textures acquired from multiple cameras, and then maps the blended textures onto the 3D models. Although the textures acquired at proper positions are selected in VDTM approaches, large errors in the 3D shapes still cause significant distortion on the resultant free-viewpoint images. State-of-the-art methods of hybrid rendering appropriately transform the 3D shapes depending on the position of the virtual viewpoint, while also selecting and blending the textures[29)30)]. These hybrid-rendering approaches, which are sometimes referred to as view-dependent geometry and texture (VDGT), generate reasonably clear free-viewpoint images, even though the 3D shapes include large errors (e.g., when there are missing regions in 3D shapes).

The proposed interface employs a VDGT hybrid-rendering method similar to the one proposed by Sato et al.[30)]. Because this VDGT method does not achieve real-time processing, we simplified and improved it to realize real-time image generation. In addition, we superimposed a 3D model of the mobile robot on the free-viewpoint images using an augmented reality (AR) technique[31)]. The combined technique of free-viewpoint image generation and AR is referred to herein as augmented free-viewpoint image generation. In the following sections, details of the interface are described with a prototype system example using an omnidirectional camera and four depth cameras. We also discuss the effectiveness of the free-viewpoint interface through evaluations in both virtual and physical environments.

## 3. Augmented free-viewpoint interface

### 3. 1 Overview

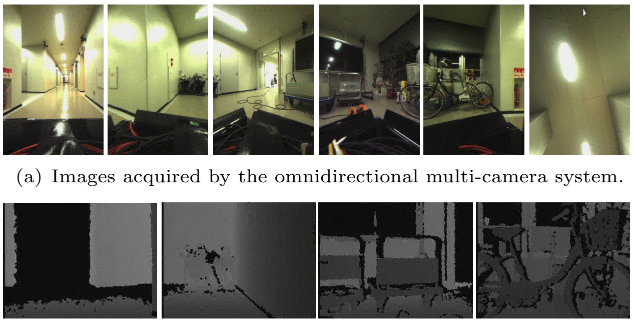
**Fig. 2** provides an overview of the proposed free-viewpoint interface. The robot is equipped with a conventional camera and depth cameras for acquiring environmental information such as textures and 3D shapes of the surroundings. The environmental information and odometry from the wheels of the robot are transmitted to the site of the human operator. A server



(a) Data flow diagram of the prototype with the proposed free-viewpoint interface.

(b) Images of the prototype system.

**Fig. 2** Prototype system based on the proposed interface.



(a) Images acquired by the omnidirectional multi-camera system.

(b) Depth images from four depth cameras.

**Fig. 3** Environmental information acquired from the prototype robot.

receives the information from the robot and generates in real-time augmented free-viewpoint images that are displayed on an HMD. The viewpoint of these generated images synchronously changes with the operator's head pose (position and direction), which is acquired by a head tracker mounted on the HMD. The details of the augmented free-viewpoint image generation method used in the proposed interface are described in Section 3. 4. Although the experiment employed a simple, wheeled robot operating with joystick control signals, including forward/backward movement and rotation, the proposed interface is compatible with other types of mobile robots and control methods.

### 3. 2 Prototype system

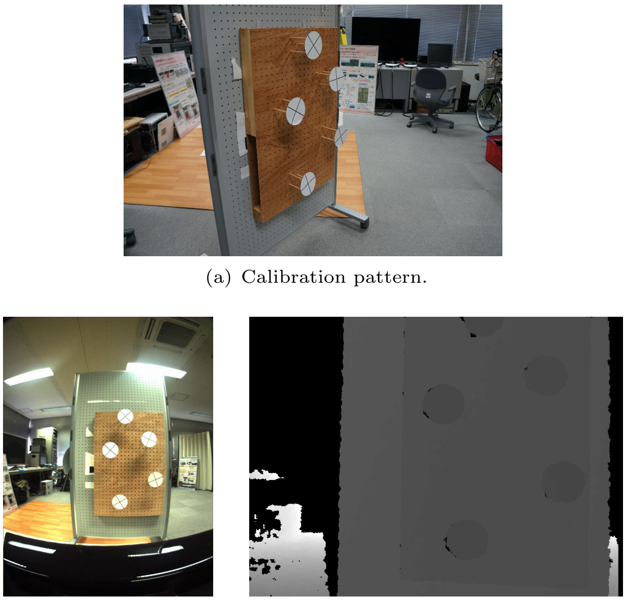
As mentioned earlier, the prototype system specifications are based on the proposed free-viewpoint interface, as shown in Fig. 2(b).

( 1 ) Robot configuration

The prototype robot has an omnidirectional camera and four depth cameras. Examples of the environmental information, including textures and 3D shapes as depth images, captured by the prototype system are shown in **Fig. 3**. The robot (Reference Hardware, Mayekawa Manufacturing Co.) is equipped with four wheels that enable forward/backward movement and rotation. An omnidirectional multi-camera system (Ladybug2, Point Grey Research, Inc.) mounted on top of the robot captures textures of the omnidirectional first-person view. Four depth cameras (Kinect, Microsoft, Inc.) are mounted on four sides of the robot to acquire depth images of the surroundings. The horizontal field-of-view of each depth camera is 57°, which indicates that the four depth cameras cannot simultaneously cover all of the robot's surroundings.

The relative position and orientation between the omnidirectional and depth cameras are calibrated in advance. We prepare a special calibration pattern consisting of a plane and circular constructions, as shown in **Fig. 4**(a). The pattern is captured from numerous positions and directions by omnidirectional and depth cameras mounted on the robot (see Figs. 4(b) and 4(c)). The center point of each marker is manually designated on omnidirectional images. The corresponding marker captured by the depth camera is then selected from the depth image. The 3D centroid of the selected marker is calculated using the depth; it is used as the 3D point corresponding to the 2D point designated on the omnidirectional image. The relative position and orientation are calculated by solving the perspective-n-point (PnP) problem from the 2D-3D correspondences.

Environmental information is newly acquired and transmitted to the server in the prototype system when one of three conditions is met: the moving distance of the robot exceeds a threshold from the latest image acquisition; the rotation angle of the robot exceeds a threshold; or, a certain period of time has elapsed. The omnidirectional and depth images are acquired with rough software-based synchronization. The maximum lag time of the synchronization depends on the maximum frame rate of the sensors (e.g., less than 33 ms in our implementation using 30 fps sensors). In our experiments, we did not find negative effects due to the delay; however, more accurate synchronization might be



(a) Calibration pattern.

(b) Calibration pattern captured by a camera in an omnidirectional multi-camera system.

(c) Calibration pattern captured by a depth camera.

**Fig. 4** Calibration between omnidirectional and depth cameras.

required under operations involving rapid robot movement. The acquired images are combined into one large image and transmitted by a wireless HDMI extender with small delays ($< 1$ ms) for the experimental environment.

(2) Configuration for human operator

The human operator wears an HMD (HMZ-T1, Sony) that displays augmented free-viewpoint images. This HMD is equipped with a receiver for an electromagnetic sensor (Fastrak, Polhemus, Inc.) that operates with a transmitter to measure the six degrees of freedom pose of the operator's head. Augmented free-viewpoint images are generated from environmental information (textures and depth images from the robot), odometry, operator head pose, and the 3D robot model in the server. When only one shot of the depth image set is used to generate the free-viewpoint images, large missing areas appear due to occlusions. In addition, there is lack of depth information from the four depth cameras because they do not cover the entire view from the robot. The proposed interface unifies time-series 3D point clouds to reduce the missing areas. The most recent environmental information received from the robot is combined with older information by 3D point cloud alignment using the odometry information as the initial estimate for each transmission of environment information. The augmented free-viewpoint images are generated in real-time from 3D point clouds and omnidirectional images. The point clouds are combined by the L0-norm-based alignment process, which is a simplified implementation of the method by Hieida et al.[32]. It should be noted that the unification process is implemented as a thread separated from the augmented free-viewpoint image generation process. That is, the calculation cost of the unification process does not affect the frame rate of resultant augmented free-viewpoint image sequences presented to the human operator. The depth-image unification and augmented free-viewpoint image generation processes are described in more detail in the following sections.

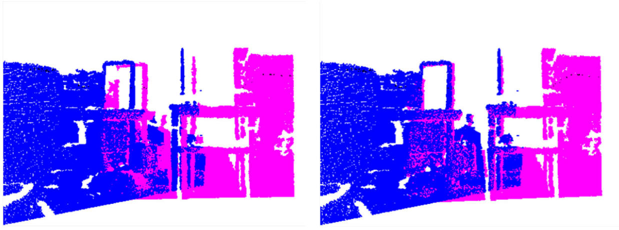
### 3.3 Unification of multiple depth images

The proposed interface accurately estimates the position and orientation of the robot for each image-acquisition operation by aligning the time-series depth images while minimizing the norms among the multiple 3D point clouds. The norms are minimized by using odometry data as the initial estimation. The prototype system employs L0-norm minimization with a two-dimensional exhaustive search algorithm. Using L0-norm is a robust solution for point cloud alignment with large outliers; however, it is difficult to minimize L0-norm using gradient-based minimization algorithms. To achieve a real-time process, the 3D points that exist in a certain interval of height are first projected on a 2D horizontal plane. The minimum value of cost function based on L0-norm is then searched for in a 2D solution space by changing the rotation and translation in tiny intervals around the acquired odometry. Although our implementation searches for a minimum in a direct manner, it would be possible to use pairs of techniques for more efficient searching, such as a simultaneous localization and mapping based on L0-norm minimization[32]. Because the proposed interface does not specify the alignment methods for point clouds, other alignment techniques can be employed, such as modern iterative closest-point algorithms[33], feature-based matching approaches[34], and registration using Gaussian mixture models[35].

When aligning point cloud $\mathbf{p}$ to another point cloud $\mathbf{q}$, L0-norm $|\mathbf{p}_i, \mathbf{q}_j|_0$ is defined as

$$|\mathbf{p}_i, \mathbf{q}_j|_0 = \begin{cases} 0 & (\exists j, |\mathbf{p}_i - \mathbf{q}_j|_2 \leqq \epsilon) \\ 1 & (otherwise) \end{cases} , \qquad (1)$$

where $\epsilon$ denotes a tiny distance that can be regarded as an identical point. The system minimizes the $E(\mathbf{R}, \mathbf{t})$ defined as the sum of L0-norms with changing rotation



(a) Alignment by odometry only. (b) Alignment using L0-norm.

**Fig. 5** Point clouds aligned before and after refinement between the two positions of acquired images, denoted in blue and pink, respectively.

matrix **R** and translation vector **t** from **p** to **q** as

$$E(\mathbf{R},\mathbf{t}) = \sum_i |\mathbf{R}\mathbf{p}_i + \mathbf{t}, \mathbf{q}_j|_0. \qquad (2)$$

The transformation from the robot to the real-world coordinates that is specified by **R** and **t** is also used to augmented free-viewpoint image generation.

Examples of point clouds with and without the L0-norm-based alignment process are shown in **Fig. 5**. Misalignments between two point clouds are reduced by the L0-norm-based alignment process.

### 3.4 Augmented free-viewpoint image generation

Augmented free-viewpoint images are generated from the unified point clouds, operator's viewpoint, and 3D robot model in three steps: 1) view-dependent geometry generation; 2) view-dependent texture mapping; and 3) superimposition of 3D robot model.

A free-viewpoint image generation method with view-dependent geometry and texture is employed for this study[30]. The existing method[30] requires the preliminary reconstruction of 3D shapes using multi-view stereo approaches, and it does not achieve real-time processing. Notably, a nonlinear optimization process in the method[30] that uses a whole sequence to achieve high-quality image generation, requires a considerable calculation cost. We used 3D point clouds acquired with depth cameras and realized real-time processing by not employing the optimization processes in the geometry generation and pixel-wise texture selection. In addition, the 3D model of the robot was superimposed on the free-viewpoint images using a standard AR technique.

(1) View-dependent geometry generation

The view-dependent geometry generation process produces a depth image of the operator's view using point clouds to reduce any missing areas and unnatural distortions in the free-viewpoint images. The depth is estimated from the combined point clouds in the following steps.

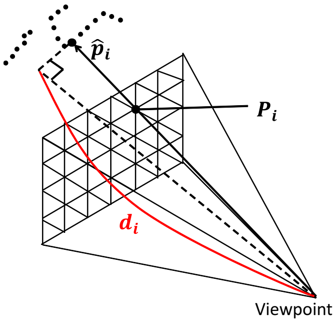


**Fig. 6** View-dependent geometry generation.

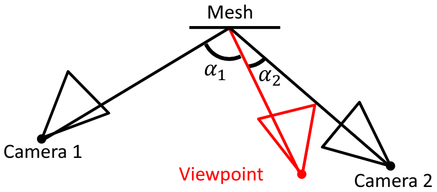


**Fig. 7** View-dependent texture selection. Camera 2 is selected as the mesh texture in this case.

(i) Divide the view plane of the operator into triangular meshes, as shown in **Fig. 6**, and employ $\mathbf{p}_i$ as the vertices of the meshes.

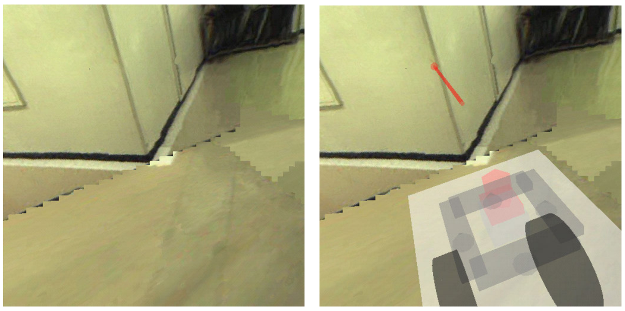
(ii) Project the 3D points onto the view plane. Some 3D points may be far from the appropriate depth, such as those existing over walls.

(iii) Estimate the depth of each vertex $\mathbf{p}_i$, and compare the depths of the projected points neighboring $\mathbf{p}_i$. The depth value $d_i$ of the point $\hat{\mathbf{p}}_i$ that has the smallest depth is employed as the depth of the vertex $\mathbf{p}_i$.

It is possible that there are no $\hat{\mathbf{p}}_i$ corresponding to $\mathbf{p}_i$ because of lack of depth information. This lack is caused by the occurrence of occlusions in a complex environment and by the limited field-of-view of depth cameras. If there are neighboring vertices whose depth values have been successfully estimated, $d_i$ is determined using linear interpolation of the valid depth values. Otherwise, $d_i$ is set as the largest value that can be measured by the depth cameras (4000 mm in our prototype system).

(2) View-dependent texture mapping

For each mesh produced in view-dependent geometry generation, the appropriate texture is selected from time-series omnidirectional textures captured for the given mesh. The generated geometry may include some errors in the 3D shapes. As shown in **Fig. 7**, we define $\alpha$ as the angle between two vectors from the center of the mesh: one is to the camera capturing the texture of



(a) Free-viewpoint image. (b) Augmented free-viewpoint image.

**Fig. 8** Superimposition of 3D robot model.

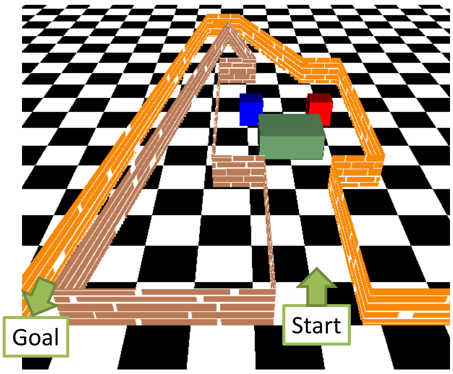


**Fig. 9** Experimental virtual environment.

the mesh, and the other is to the viewpoint to be generated. The pose of the robot estimated in Section 3. 3 is used for the pose of the camera. The proposed method selects the texture that has the smallest $\alpha$ because the distortion of the appearance caused by 3D shape errors is smaller when $\alpha$ is smaller. Finally, the selected texture is projected and mapped onto the mesh. This texture selection strategy is common in some IBR and hybrid rendering approaches[28].

( 3 ) Superimposition of 3D robot model

In the proposed interface, the hand-made 3D robot model is superimposed using the robot pose information estimated by aligning the depth images, as described in Section 3. 3, as well as the odometry information. The generated free-viewpoint images do not include the appearance of the robot itself, as shown in **Fig. 8**(a). The transmission of the depth images, which requires a sufficiently large bandwidth, may cause large delays. Therefore, the robot may not be superimposed in the appropriate position in a free-viewpoint image when only the robot pose estimated in Section 3. 3 is used. To generate augmented free-viewpoint images while considering such delays, changes in the robot pose since the most recent acquisition of depth images are calculated from odometry information; these are used with the depth alignment-based pose information. The prototype system 3D model is transparently superimposed to improve the visibility of the scene on the robot's far side, along with a virtual arrow indicating the robot's travel direction, as shown in Fig. 8(b).

## 4. Virtual environment experiments

When operating a remote robot in a real environment, there are many factors affecting a human operator's experience, such as the quality of the free-viewpoint images generated by the proposed approach. We therefore conducted an evaluation to investigate the interface characteristics using freely configurable views in an ideal (i.e., virtual) environment.

The human operator configuration was the same as the one discussed in Section 3. 2. Participants at a remote site operated a virtual robot in a simulation environment using CG models. The participants performed the following two tasks:

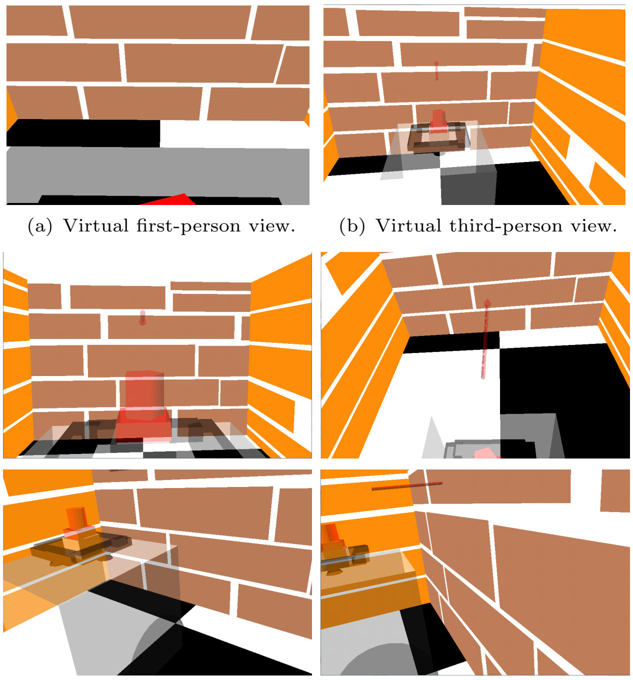
**Task 1:** Run through a path as quickly and safely as possible without colliding with the wall (the path included narrow passages obstacles, as shown in **Fig. 9**).

**Task 2:** Approach the wall as closely as possible, without colliding with it (such behavior is sometimes required in practical situations to accurately operate a robot with arms beside a wall).

For each task, we compared three interfaces: the first-person view, third-person view, and proposed augmented free-viewpoint interface. The third-person view was fixed at a diagonal location behind the robot at approximately 45° above the horizontal location. The images presented to the operators were produced without the proposed free-viewpoint image generation method by rendering the virtual environment from only the configured viewpoint using a traditional graphics library, as shown in **Fig. 10**. The participants were ten people in their twenties or thirties. Most of them were not familiar with the robot operation tasks. Therefore, we first conducted training on joystick robot operation in a virtual environment different from the one used for the experiment. The operations were successfully completed without a wall collision, except for the first-person view interface in Task 1 above (average 1.1 times of collision).

In addition to the investigation using objective factors, we conducted subjective evaluations based on the following three questions to ascertain the operator's impression of each interface:

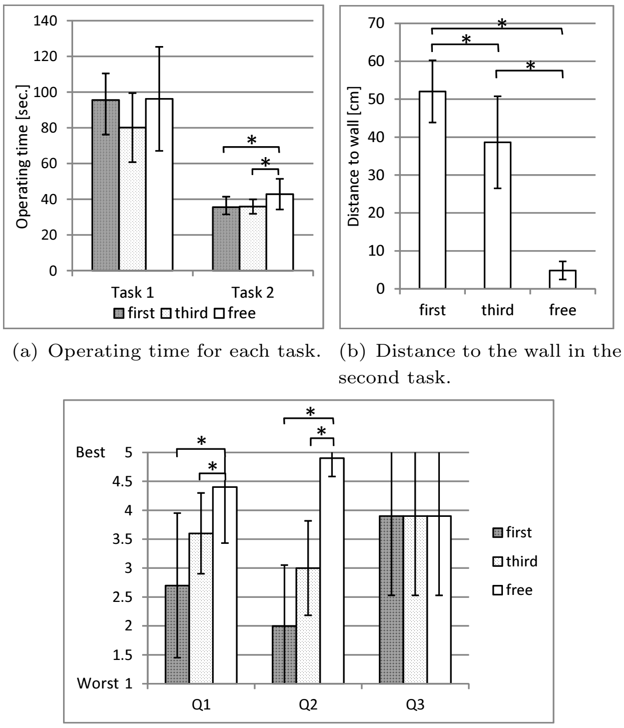
**Q1:** Were the obstacles on the ground easily recog-



(a) Virtual first-person view. (b) Virtual third-person view.

(c) Virtual free-viewpoint.

**Fig. 10** Examples of each interface view in a virtual environment.



(a) Operating time for each task. (b) Distance to the wall in the second task.

(c) Questionnaire results.

**Fig. 11** Results of the virtual environment experiments ('*' indicates a significant difference by a paired multiple comparison test, $p < 0.05$).

nized?

**Q2:** Was it possible to grasp the distance between the robot and the wall?

**Q3:** Could you operate the robot without feeling delays?

The participants answered these questions after each stage of the experiments using a scale from one (worst) to five (best).

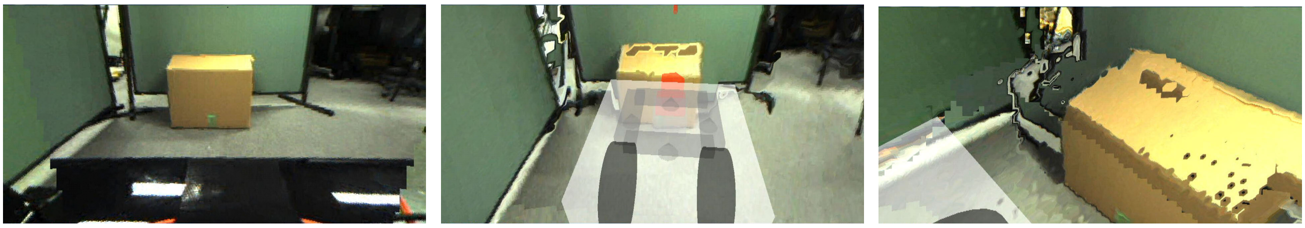
**Fig. 11** shows the results of the experiments in the virtual environment, as well as the pairs that had significant difference $p < 0.05$. We employed a Friedman multiple comparison test with a Bonferroni-adjusted post-hoc Wilcoxon signed-rank test. This comparison scheme is a nonparametric alternative of one-way repeated-measures analysis of variance (ANOVA), with a post-hoc paired t-test with Bonferroni correction. The correction determines the significant level to prevent increasing type-one errors in the multiple comparison test. In the results for Task 1, shown in Fig. 11(a), the operating times to complete the task were not significantly different among the three interfaces. In Task 2, the free-viewpoint interface was significantly more accurate than the other interfaces, as shown in Fig. 11(b); instead a longer operating time was taken with that interface to approach very close to the wall. This implies that the free-viewpoint interface could generate viewpoints that enable the operators to effectively grasp the distance between the robot and the wall, which is shown in the bottom-right of Fig. 10(c).

The results of the questionnaires, presented in Fig. 11(c), were that the free-viewpoint interface had significantly higher ratings for both the first and second questions, which concerned the ability to recognize the surrounding environment. This means that the proposed interface reduces ambiguity in the operator's recognition of the virtual environment surroundings. For the third question, there were no significant differences among three interfaces. They had exactly the same ratings from all participants. The differences among interfaces had fundamentally little impact on perception of the delay, at least in the given environment, which had a tiny delay.

The results of these experiments indicate that the proposed interface offers the advantage of accurate and safe operation over short operating time. The results therefore imply that the proposed interface is potentially valuable for mobile-robot remote navigation tasks, such as surveillance.

## 5. Physical environment experiments

We performed experiments in a physical environment using the prototype system described in Section 3.2.



(a) First-person view.

(b) Third-person view.

(c) Free-viewpoint.

**Fig. 12** Examples of each interface view in a physical environment.

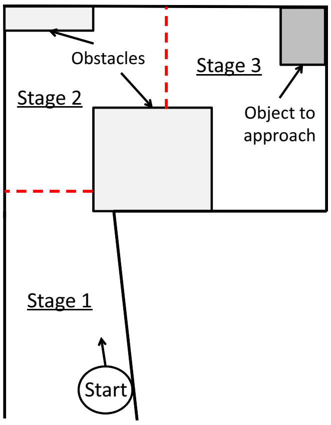


**Fig. 13** Map of experimental physical environment.

The experiment participants again consisted of ten people in their twenties or thirties, who operated the robot using the same three interfaces discussed in the previous section. In this experiment, the first-person view was generated from the most current omnidirectional image, and the participants could freely change their view direction. The third-person viewpoint was fixed at a diagonal position behind the robot, whose images were generated using the same technique used in the proposed free-viewpoint interface. The third-person view also enabled the participants to change their view direction. Examples of the view of each interface are shown in **Fig. 12**. Free-viewpoint images presented for the human operator from a couple of viewpoints are shown in Fig. 1 of Section 1. The runway for the robot used in the experiment was constructed in three stages, as shown in **Fig. 13**. The participants were directed to operate the robot with respect to each stage without colliding with the wall and obstacles. A task was set for each stage as follows:

**Stage 1:** Run through a straight narrow passage as quickly as possible.

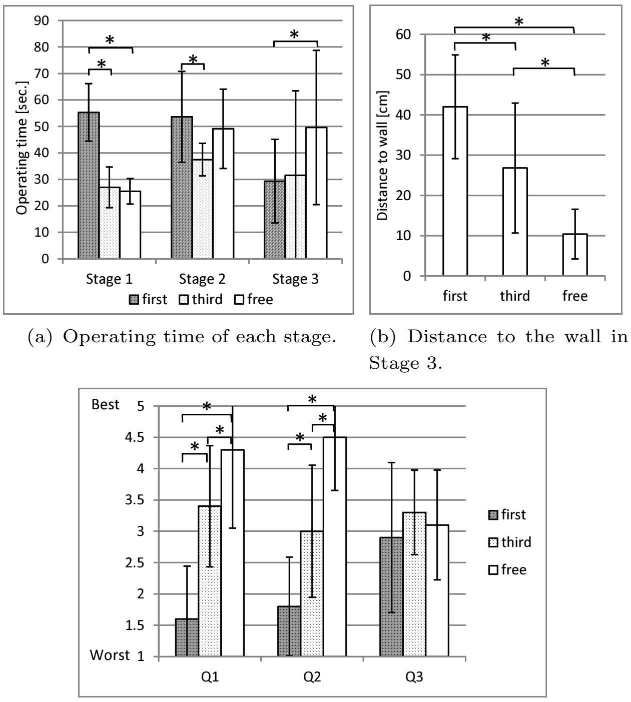
**Stage 2:** Run through a curved passage with obstacles as quickly as possible.

**Stage 3:** Approach the wall as closely as possible.

We evaluated the operating time for each stage, along with the distance to the wall in Stage 3. In Stage 3, a participant collided with the wall with the third-person view and the free-viewpoint interface, respectively. In Stage 2, another participant also collided with the wall with the third-person view interface. The participants were allowed to retry the stage when a collision occurred. In addition to objective factors, we investigated operator subjective issues by posing to participants the same three questions described in Section 4.

In our experiment, the environmental information was newly acquired and unified when one of the following conditions was met: 1) 60 cm movement of the robot, 2) 20° rotation of the robot, or 3) 200 seconds elapsed from the last image capture. Using these conditions, the images were acquired at an average of 13.9 locations through all stages. The unification process of the environmental information was performed in less than 33 ms in the given small environment used in this experiment, where the process does not affect the frame rate of the augmented free-viewpoint video presented to the examinees. The augmented free-viewpoint image was successfully presented on an HMD at 30 fps.

**Fig. 14**(a) shows the operating time for each stage and interface. The figure also shows the pairs that had significant difference $p < 0.05$ calculated by the same manner as outlined in Section 4. In Stage 1, the free-viewpoint interface, together with the third-person view interface, enabled the operators to quickly complete the task. The participants most often fixed their viewpoint when using the proposed interface for running through the straight pathways. In those cases, the free-viewpoint interface exhibited behavior similar to that of the third-person view. When using the proposed interface, the distance to the wall in Stage 3 was significantly smaller than with the others, as shown in Fig. 14(b), but the operating time was longer than in the first-person view. This implies that the operator using the proposed interface takes additional time to closely approach the wall compared to the other interfaces. These results demonstrate the same trend as do the virtual environment experiments: The proposed in-



(a) Operating time of each stage.

(b) Distance to the wall in Stage 3.

(c) Questionnaire results.

**Fig. 14** Physical environment experiment results ('*' indicates a significant difference by a paired multiple comparison test, $p < 0.05$).

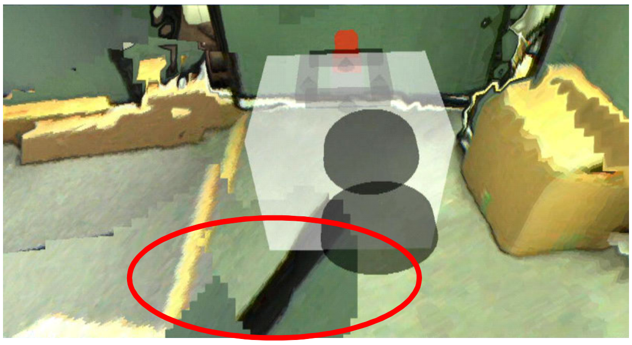


**Fig. 15** Artifact in an augmented free-viewpoint image.

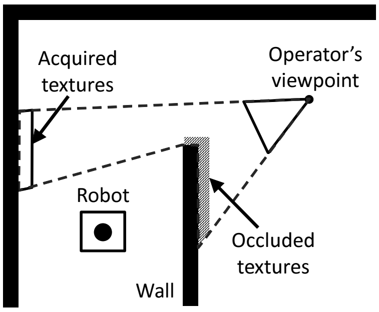


**Fig. 16** Occlusion problem in free-viewpoint image generation. Shaded area denotes where textures are occluded by the wall but required to complete the operator's view.

terface improves operation accuracy. In addition, quick operation was achieved in some cases of the proposed interface used with the third-person view, particularly in simple environments.

The results of the questionnaires are shown in Fig. 14(c). Similar to the evaluation result in the virtual environment, the proposed interface received higher ratings for both the first and second questions, but not the third question. We can find the same trend in ideal (virtual) and physical environments from both objective and subjective aspects. Therefore, the free-viewpoint image generation process in our prototype system successfully expresses the potential advantage of the proposed interface.

## 6. Discussions

### 6.1 Quality of augmented free-viewpoint image

From our experiments in the physical environment, we have confirmed that the quality of images generated by the prototype system is satisfactory for expressing the potential advantage demonstrated in the virtual-environment experiments. Nevertheless, to realize a more effective interface, the image quality should be further improved. In the bottom of **Fig. 15**, for example, artifacts appear in the augmented free-viewpoint image. They could cause false recognition by operators and therefore lack of safe operation. Such artifacts are generated due to the *occlusion problem*, as illustrated in **Fig. 16**. Although environmental information (i.e., 3D shape and textures of the environment) cannot be physically acquired in the region occluded by surrounding objects (e.g., walls), the generation process for the operator's view sometimes requires them. A promising alternative for increasing safety is to display these textures in the free-viewpoint image using another form, such as by filling them with a prominent color to highlight them as occluded textures.

### 6.2 Limitations

Although the proposed free-viewpoint interface does not essentially restrict specific system configuration and implementation, our prototype system based on the proposed interface has some limitations. In the implementation of the depth unification process, we employed the *flat floor* assumption; i.e., our prototype does not accommodate the slope of the floor or tilt of the robot. To overcome this problem, 3D point cloud registration approaches[33][34] can be applied. In addition, the prototype system is specialized for indoor use because of the hardware limitation (e.g., depth cameras cannot be used outdoors). Other robots, sensors, or devices can be employed for our proposed interface by modifying some implementations.

## 7. Conclusions and future work

In this paper, we have proposed a teleoperation interface for mobile robots with a freely configurable viewpoint using photorealistic textures of the physical world. This free-viewpoint enables human operators to intuitively change their viewpoints using an HMD and a head tracker. A free-viewpoint image generation method with view-dependent geometry and texture has been simplified and improved to achieve real-time processing for the proposed interface. In addition, we achieved augmented free-viewpoint image generation in which a 3D model of the robot was superimposed on the free-viewpoint image using AR techniques. Our experiments conducted in both virtual and physical environments have confirmed that the proposed interface has advantages in terms of operational accuracy over time required to complete tasks, while the quality of the generated free-viewpoint images is satisfactory for demonstrating the advantage of our proposed interface.

In future work, we will investigate the effects of delay in the proposed interface in a setting with large delays in environment information transmission, improve the prototype system for more practical situations, and enhance the quality of the free-viewpoint image generation process.

## Acknowledgement

This research was supported by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research, Nos. 24700208 and 23240024, Grant-in-Aid for JSPS Fellows No. 25–7448, and the "Ambient Intelligence" project granted by Ministry of Education, Culture, Sports, Science and Technology (MEXT). Finally, we would like to thank Dr. Kentaro Takemura, former Assistant Professor of Robotics Lab., Nara Institute of Science and Technology, Nara, Japan, for his support in our prototype system implementation.

## References

1) F. Okura, Y. Ueda, T. Sato and N. Yokoya: "Teleoperation of Mobile Robots by Generating Augmented Free-Viewpoint Images", Proc. 2013 IEEE/RSJ Int. Conf. Intelligent Robots & Syst. (IROS'13), pp.665–671 (2013)

2) G. N. DeSouza and A. C. Kak: "Vision for Mobile Robot Navigation: A Survey", IEEE Trans. Pattern Analysis & Machine Intelligence, 24, 2, pp.237–267 (2002)

3) R. Siegwart, I. R. Nourbakhsh and D. Scaramuzza: Introduction to Autonomous Mobile Robots, MIT Press, Cambridge (2011)

4) H. Choset: "Coverage for Robotics–A Survey of Recent Results", Annals of Mathematics & Artificial Intelligence, 31, 1, pp.113–126 (2001)

5) B. M. Yamauchi: "PackBot: A Versatile Platform for Military Robotics", Proc. SPIE, 5422, Unmanned Ground Vehicle Technology VI, pp.228–237 (2004)

6) T. Fong and C. Thorpe: "Vehicle Teleoperation Interfaces", Autonomous Robots, 11, 1, pp.9–18 (2001)

7) D. W. Hainsworth: "Teleoperation User Interfaces for Mining Robotics", Autonomous Robots, 11, 1, pp.19–28 (2001)

8) R. T. Laird, M. H. Bruch, M. B. West, D. A. Ciccimaro and H. R. Everett: "Issues in Vehicle Teleoperation for Tunnel and Sewer Reconnaissance", Proc. 2000 IEEE Workshop Vehicle Teleoperations Interfaces (2000)

9) K. Nagatani, S. Kiribayashi, Y. Okada, S. Tadokoro, T. Nishimura, T. Yoshida, E. Koyanagi and Y. Hada: "Redesign of Rescue Mobile Robot Quince", Proc. 2011 IEEE Int. Sympo. Safety, Security, & Rescue Robotics (SSRR'11), pp.13–18 (2011)

10) N. Shiroma, N. Sato, Y. Chiu and F. Matsuno: "Study on Effective Camera Images for Mobile Robot Teleoperation", Proc. 13th IEEE Int. Workshop Robot & Human Interactive Commun. (ROMAN'04), pp.107–112 (2004)

11) M. Sugimoto, G. Kagotani, H. Nii, N. Shiroma, F. Matsuno and M. Inami: "Time Follower's Vision: A Teleoperation Interface with Past Images", IEEE Comput. Graphics & Appl. Mag., 25, 1, pp.54–63 (2005)

12) K. Saitoh, T. Machida, K. Kiyokawa and H. Takemura: "A 2D-3D Integrated Interface for Mobile Robot Control Using Omnidirectional Images and 3D Geometric Models", Proc. Fifth IEEE/ACM Int. Sympo. Mixed & Augmented Reality (ISMAR'06), pp.173–176 (2006)

13) C. W. Nielsen, M. A. Goodrich and R. W. Ricks: "Ecological Interfaces for Improving Mobile Robot Teleoperation", IEEE Trans. Robotics, 23, 5, pp.927–941 (2007)

14) F. Ferland, F. Pomerleau, C. T. Le Dinh and F. Michaud: "Egocentric and Exocentric Teleoperation Interface Using Real-Time, 3D Video Projection", Proc. Fourth ACM/IEEE Int. Conf. Human-Robot Interaction (HRI'09), pp.37–44 (2009)

15) A. Kelly, N. Chan, H. Herman, D. Huber, R. Meyers, P. Rander, R. Warner, J. Ziglar and E. Capstick: "Real-Time Photorealistic Virtualized Reality Interface for Remote Mobile Robot Control", Int. J. Robotics Research, 30, 3, pp.384–404 (2011)

16) B. Hine, P. Hontalas, T. Fong, L. Piguet, E. Nygren and A. Kline: "VEVI: A Virtual Environment Teleoperations Interface for Planetary Exploration", Proc. 25th SAE Int. Conf. Environmental Syst. (1995)

17) L. A. Nguyen, M. Bualat, L. J. Edwards, L. Flueckiger, C. Neveu, K. Schwehr, M. D. Wagner and E. Zbinden: "Virtual Reality Interfaces for Visualization and Control of Remote Vehicles", Autonomous Robots, 11, 1, pp.59–68 (2001)

18) K. Yamazawa, Y. Yagi and M. Yachida: "Omnidirectional Imaging with Hyperboloidal Projection", Proc. 1993 IEEE/RSJ Int. Conf. Intelligent Robots & Syst. (IROS'93), 2, pp.1029–1034 (1993)

19) M. G. Dissanayake, P. Newman, S. Clark, H. F. Durrant-Whyte and M. Csorba: "A Solution to the Simultaneous Localization and Map Building (SLAM) Problem", IEEE Trans. Robotic & Aut., 17, 3, pp.229–241 (2001)

20) S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison and A. Fitzgibbon: "KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera", Proc. 24th ACM Sympo. User Interface Software & Techn. (UIST'11), pp.559–568 (2011)

21) P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J. M. Frahm, R. Yang, D. Nistér and M. Pollefeys: "Real-Time Visibility-Based Fusion of Depth Maps", Proc. 11th IEEE Int. Conf. Comput. Vision (ICCV'07), pp.1–8 (2007)

22) S. M. Seitz and C. R. Dyer: "View Morphing", Proc. ACM SIGGRAPH'96, pp.21–30 (1996)

23) M. Levoy and P. Hanrahan: "Light Field Rendering", Proc. ACM SIGGRAPH'96, pp.31–42 (1996)

24) T. Naemura, T. Takano, M. Kaneko and H. Harashima: "Ray-Based Creation of Photo-Realistic Virtual World", Proc. Third Int. Conf. Virtual Syst. & Multimedia (VSMM'97), pp.59–68 (1997)

25) S. J. Gortler, R. Grzeszczuk, R. Szeliski and M. F. Cohen: "The Lumigraph", Proc. ACM SIGGRAPH'96, pp.43–54 (1996)

26) C. Buehler, M. Bosse, L. McMillan, S. Gortler and M. Cohen: "Unstructured Lumigraph Rendering", Proc. ACM SIG-

GRAPH'01, pp.425–432 (2001)

27) S. B. Kang, R. Szeliski and P. Anandan: "The Geometry-Image Representation Tradeoff for Rendering", Proc. 2000 IEEE Int. Conf. Image Processing (ICIP'00), 2, pp.13–16 (2000)

28) P. E. Debevec, C. J. Taylor and J. Malik: "Modeling and Rendering Architecture from Photographs: A Hybrid Geometry- and Image-Based Approach", Proc. ACM SIGGRAPH'96, pp.11–20 (1996)

29) M. Irani, T. Hassner and P. Anandan: "What Does the Scene Look Like from a Scene Point?", Proc. Seventh European Conf. Comput. Vision (ECCV'02), pp.883–897 (2002)

30) T. Sato, H. Koshizawa and N. Yokoya: "Omnidirectional Free-Viewpoint Rendering Using a Deformable 3-D Mesh Model", Int. J. Virtual Reality, 9, 1, pp.37–44 (2010)

31) R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier and B. MacIntyre: "Recent Advances in Augmented Reality", IEEE Comput. Graphics & Appl. Mag., 21, 6, pp.34–47 (2001)

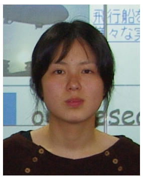
32) Y. Hieida, T. Suenaga, K. Takemura, J. Takamatsu and T. Ogasawara: "Real-Time Scan-Matching Using L0-Norm Minimization Under Dynamic Crowded Environments", Proc. Fourth Workshop Planning, Perception & Navigation for Intelligent Vehicles, pp.257–262 (2012)

33) A. Segal, D. Haehnel and S. Thrun: "Generalized-ICP", Proc. 2009 Robotics: Science & Syst. (RSS'09), 25, pp.26–27 (2009)

34) R. B. Rusu, N. Blodow and M. Beetz: "Fast Point Feature Histograms (FPFH) for 3D Registration", Proc. 2009 IEEE Int. Conf. Robotics & Aut. (ICRA'09), pp.3212–3217 (2009)

35) B. Jian and B. C. Vemuri: "Robust Point Set Registration Using Gaussian Mixture Models", IEEE Trans. Pattern Analysis & Machine Intelligence, 33, 8, pp.1633–1645 (2011)

**Fumio Okura** received his M.E. degree in information science from Nara Institute of Science and Technology in 2011. Since 2011 he has been pursuing his Ph.D. at Nara Institute of Science and Technology. He has been a research fellow of the Japan Society for the Promotion of Science since 2013.



**Yuko Ueda** received her B.S. degree in information and computer science from Nara Women's University in 2011. She received her M.E. degree in information science from Nara Institute of Science and Technology in 2013. She has been working at Sony Corporation since 2013.

**Tomokazu Sato** received his B.E. degree in computer and system science from Osaka Prefecture University in 1999. He received his M.E. and Ph.D. degrees in information science from Nara Institute of Science and Technology in 2001 and 2003, respectively. He was an assistant professor at Nara Institute of Science and Technology from 2003 to 2011, when he became an associate professor.

**Naokazu Yokoya** Naokazu Yokoya received his B.E., M.E., and Ph.D. degrees in information and computer sciences from Osaka University in 1974, 1976, and 1979, respectively. He joined Electrotechnical Laboratory (ETL) in 1979. He was a visiting professor at McGill University in 1986-87 and has been a professor at Nara Institute of Science and Technology since 1992. He has also been a vice president at Nara Institute of Science and Technology since April 2013.