Augmented Visualization: Observing as Desired

Shohei Mori* and Hideo Saito†
* Keio University, Yokohama, Japan
E-mail: s.mori.jp@ieee.org
† Keio University, Yokohama, Japan
E-mail: hs@keio.org

***Abstract*—Over 20 years have passed since a free-viewpoint video technology has been proposed in which a user's viewpoint can be freely set up reconstructed 3D space of a target scene photographed by multi-view cameras. This technology allows us to capture and reproduce the real world as it recorded. Once we capture the world in a digital form, we can modify the world as augmented reality does (i.e., placing virtual objects in the digitized real world). As oppose to this concept, the augmented world also allows us to see through real objects by synthesizing the backgrounds that cannot be observed in our raw perspective directly. The key idea is to generate the background image using multi-view cameras observing the backgrounds at different positions, and seamlessly overlaying the recovered image in our digitized perspective. In this article, our studies regarding such desired view generation techniques are reviewed, and its future directions and open problems are discussed.**

I. Introduction

Once the real world is digitized via multi-view observations, the digitized world can be transferred, modified, and played on a computer from any viewpoint and time points [18]. Such a basic idea is called arbitrary viewpoint image generation and it has already been 20 years since being proposed.

Using this technology, it is no longer necessary to stay in a 2D display with a fixed viewpoint. As typified by head-mounted displays (HMDs), displays are now portable and therefore arbitrary view images are presented in front of the eyes. Combining with head pose estimation, the virtual view is presented, with the head motion in real time [6]. This video technology called augmented reality (AR), can bring the virtual world into real world by overlaying virtual objects onto the captured real space. For example, AR can support users with annotating virtual signs, bringing video game world to the real world by superimposing virtual creatures, etc.

On the other hand, one might notice that only adding information is not enough for truly controlling and reproducing scenes as one desired. Conflicts between old annotations, such as existing signboards and AR annotations, will cause confusion for users, so we want to erase the old ones. Even if an AR creature destroys a wall, the wall still exists there and users cannot see the other side of the collapsed wall. Contrary to AR, video technologies that shrink real information are called diminished reality (DR). Once we digitized the world, naturally, we can choose not to reproduce some parts of the world. Thus, unnecessary objects in the captured images are not reproduced in AR space, but are compensated with other information. For example, when we want to erase an object, we can compensate for its background. While such ideas are not new at all [25], [16], real implementations emerged in the last decades because achieving scene reconstruction and synthesis at the same time requires various, practical challenges.

In this article, the authors introduce the DR technologies necessary for visually modifying the real world as desired from the perspective of free-viewpoint image generation methods.

II. Free-Viewpoint Image Generation: Principle and Advances

Measuring and reconstructing the three-dimensional (3D) shape of target scenes has been studied for nearly 50 years as a basic technology in the field of image processing and measurement. In Computer Graphics (CG) and Virtual Reality (VR) fields, such digitized 3D shapes have been used for taking a picture of scenes from arbitrary viewpoints where input from real world cameras did not exist. In Virtualized Reality [9], Kanade *et al*. demonstrated that 3D shape reconstruction techniques can be applied for dynamic scenes in time space for generating arbitrary view point videos. This technology has been studied as a new type of image presentation technology and started to be used in practical situations such as sports.

The essence of this technology lies in how accurately one can obtain the shape and texture (colors) of the target object. However, the accuracy of the abovementioned 3D shape reconstruction-based approaches is limited and detracts from the final output image quality. On the other hand, the ray space theory [7] -based approaches can generate arbitrary viewpoint images without acquiring the 3D shape of the scene. The ray space theory represents an image as a group of light rays in a space filled with light rays from every direction and has been studied since it was first proposed 20 years ago. This method is implemented by photographing the target scene with a large number of cameras (a camera array) providing a variety of viewpoints. Recently proposed methods using such image data structures can computationally change camera parameters, such as focus and aperture *after* the shooting where before we would change the parameters *before* the shooting when using conventional cameras. Also, methods for photographing pictures in higher resolutions, in terms of space and time beyond the limitation determined by sampling theory, have been proposed. Thus, the future development of computational photography technology is greatly expected.

III. Observing as Desired

The techniques in the previous section are summarized as ones to collect image data photographed and acquired by sensors, such as cameras, or to generate desirable images for observers after the data acquisition. For example, when *free*-viewpoint video technology is used in broadcasting etc., what *free* is usually means the observation position which the photographer or the director of the broadcast had been predetermined. Therefore, free-viewpoint videos allow the audience actually watching the video to move their viewing positions to where they desired. In other words, this is a technology that allows us to change our viewpoint in the digitized 3D world, regardless of the intention of the direct acquisition of the image data. The technique for such desirable images, regardless of the intention or situations at the time of actual image captures, becomes more meaningful as the amount of acquired image data increases. When the camera was invented, "taking a picture" meant to considering every property every time (such as camera position, posture, subject, shutter, etc.) since the cost required for one photograph was large. Therefore, even a single photo conveyed details inexpressible by words to many people.

In modern worlds, plenty of surveillance cameras are installed in societies and images are collected at every moment. Similarly, a lot of images are taken by cameras owned by individuals. However, most of the surveillance and personal camera images are not seen by humans. That is, a lot of images not seen by anyone are stored. Therefore, there is a demand for techniques to re-create and present in desired forms images from these huge image datasets. As an example of the forms, the authors have been conducting research on image generation of "Diminished Reality (DR) [22]." This is a technique to generate and present images of a scene in which undesirable objects are virtually removed from desired viewing positions.

IV. Diminished Reality: Principle and Basic Functions

DR is different from augmented reality (AR) [2], [3], [12] that superimposes virtual objects on the real world to enhance reality. Fig. 1 shows distinctive differences between AR and DR. AR overlays virtual objects to add some positive information to the real world. DR overlays virtual objects as well but the objects are negative information to diminish the reality. Therefore, we could say that DR is a visualization method by diminishing.

Fig. 2 shows schematic figures describing differences of real world, VR, AR, and DR in terms of visible light rays in each reality. In (a), the observer sees real objects in the environment. In VR (b), HMD occludes light rays of the real environment and presents ones from virtual world. In AR (c), real light rays are visible through the HMD and, at the same time, virtual rays are presented. Thus, the virtual object (black star) looks like it is existing in the real environment. In DR (d), real rays are selectively occluded by the HMD. Rays initially occluded by objects are recovered and therefore visible through the HMD.

DR technology is defined as a set of methodologies (*diminishing*, *replacing*, *in-painting*, or making something *see-through*) for altering real objects in a perceived environments in real time to "diminish" reality. Each function can be described as follows; To *diminish* objects, the objects of interest are degraded in colors or textures to get less attention; to make objects *see-through*, backgrounds of the objects are digitized beforehand or in real time in a similar manner to free-viewpoint image generation and overlaid onto the observer's view in accordance with the observer's head motion; to *replace* objects, alternative virtual objects are overlaid onto the real objects to hide them; to *inpaint* objects, plausible background images are generated from the pixels except for region of interest pixels on the fly.

V. Diminished Reality Examples

A. Formulating Invisibility

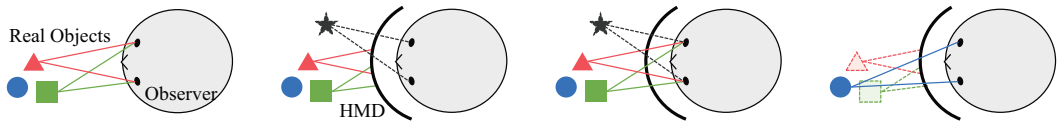
Fig. 3 shows an example of visualizing a work area occluded by a operator's hand and a holding tool in the perspective [20]. The operator might see the work area when he/she moves his/her head, but sometimes has difficulty when the holding tool is large or when a third person watching the first person's view video cannot see the work area since he/she cannot move the viewing position. Using diminished reality technology, the observers can change the transparency of hands and tools in the operator's view. Some arbitrary viewpoint image generation methods can be described as a synthesizing technique of a weighted sum of view-dependent images. Here, multi-viewpoint cameras are surrounded and capturing the environment, and smaller weights of light rays are given to the undesirable objects to make them see-through in the user's perspective.

B. Applying DR to Real Applications

Fig. 4 shows an example of a use of DR technology for cinematography [14]. SFX/VFX makes films popular, but the filmmaking process complicated. Therefore, pre-visualized movies created in the early stage of the filmmaking called PreVis played an important role over the last several decades. The PreVis movies are simple computer graphics movies or movies shot with low-cost cameras to share a creator's vision and to estimate camera movements, necessary personnel, and other potential expenditures. However, vending machines or modern signboards in the environment can ruin the image especially when filmmakers create a historical movie. In Fig. 4, images taken during location hunting (B) are synthesized to the current video stream of a scene (A) with the visually annoying object present in the view of the camera. Since illumination conditions are different from each other, a simple overlay generates conspicuous borders around the region of interest (C), but these borders are reduced by a color compensation technique (D). This kind of PreVis work enables filmmakers not only to share their image, but also to estimate the necessary amount of CG processes in the later stage.

(a) Augmented Reality (a) Diminished Reality

Fig. 1. Conceptual differences in AR and DR
AR overlays arbitrary virtual objects onto the real scene whilst DR creates virtual objects in the background using a free-viewpoint image generation technique.



(a) Real Space (b) Virtual Reality (c) Augmented Reality (d) Diminished Reality

Fig. 2. Differences of real world, VR, AR, and DR
Light rays are selectively presented to the eyes via an HMD.

(a) An input frame set (b) Observer's view and result

Fig. 3. Removing a hand and a tool
Because the cross mark on the board is occluded in the observer's view but visible from the other cameras, it is recovered.

C. Getting a Wider View

Field of view of background observation cameras is limited because they are basically placed ahead of the user's camera, especially in see-through vision applications. To overcome this problem, we built an RGB-D camera combining a range scanner and a fish-eye camera to be used in see-through vision applications (Fig. 5 [23]). This type of semi-transparent image generation method have been studied as AR X-ray Vision [1], [24] and See-through Vision [26]. Using this technology, for example, we can check the cars and people approaching from behind walls and can check if shops behind buildings are open or not. Although we can see behind walls through mirrors for curve, see-through technology does not require us to mentally transform the view in 3D like we usually do. In addition, we can see the backgrounds in a wider field of view. The mechanisms of the see-through representation is quite simple. A straightforward method is to alpha blend an original input image and a diminished reality image. However, the resulting image will have two types of information (i.e., background and foreground) at once. Therefore, effective representation methods of these information has been discussed in AR X-ray Vision area.

Considering DR as a filtering method of information would also involve ethical problems. Therefore, it is a difficult problem to decide what to hide or show, whether to automatically control or to trust the user. Although erasing a vehicle in front will give us a wide field of view, it is easy to imagine that we will collide with the invisible vehicle. The perceived appearance is greatly different depending on when the forward vehicle is displayed semi-transparently, when the focus is on the front vehicle, or when the focus is on the far road. Therefore, besides DR displays, taking into account human sensing technology such as accommodation estimation will become important.

Fig. 4. DR-PreVis
(C) is a simple synthesis of input (A) and pre-fetched (B) images. (D) is color corrected result.

Fig. 5. See-through Vision
RGB-D camera (right) for wide FoV (left).

Fig. 6. DR result under dynamic lighting changes
Illumination changes in ROI is estimated from surrounding pixels.

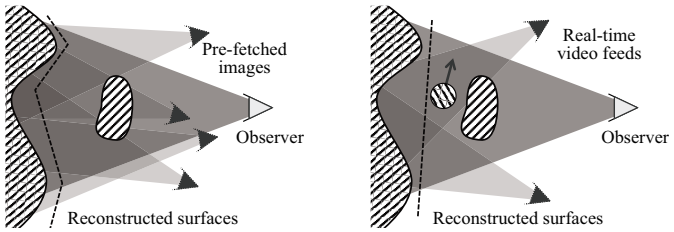


Fig. 7. Pre-observation and real-time observations in DR
Number of cameras and real-time processing is a trade-off in DR.

VI. Image Processing in Diminished Reality

A. Dealing with Viewpoint Changes

In DR, an arbitrary viewpoint image generated from other viewpoints should overlay only regions where the objects of interest exist. Therefore, when an object to be removed is small or far from the observer, 2D alignment may be sufficient even for 3D background scenes [15]. When the object to be removed is large or close to the observer, the background region occluded by the object to be removed increases, and 3D registration is required. Visual simultaneous localization and mapping (vSLAM) [17] seems sufficient for the 3D alignment, but an extension for registration with the background's and vSLAM's coordinates is required. There is also an efficient 3D registration method that utilizes the previously obtained background data [19]. Since the result of synthesis on the screen space is the final output for DR, matching the generated background and the current image should be effective, even if 3D registration is erratic.

B. Handling Lighting Changes

In the case of using image data taken in the past, it is necessary to reduce the photometric inconsistency between the background and the current image because they are imaged at different time. DR methods basically use color tone correction processing on the image plane in order to achieve real-time processing [15], [10]. Even if the light source always moves around within a 3D scene, a simple color tone correction often results in sufficient results in the appearance, if the processing is performed in real time [19].

C. What Needs to be Diminished?

Objects to be removed are determined in advance or selected via a user interface such as click or drag. In the case of determining the objects of interest in advance, such objects are limited to specific types, like walls and people, or object detection is avoided for simplification. When selecting an object to be removed with a user interface, the region must be properly enclosed, and even after that, it must keep tracking along with the movement of the viewpoint. Many methods cover objects with a 3D bounding object or track using image-based tracking methods.

VII. Open Problems in Diminished Reality

A. When to Observe Backgrounds?

Whether background data should be collected beforehand or in real time depends on applications, and therefore is difficult to uniquely determine. Sufficient time and efforts before DR processing allows us to generate background images of high quality that endures observations from any viewing positions. On the other hand, it is undeniable that the object arrangement in the scene has already changed from that in the pre-observation. Even if the placement of objects is unchanged, illumination will change from moment to moment depending

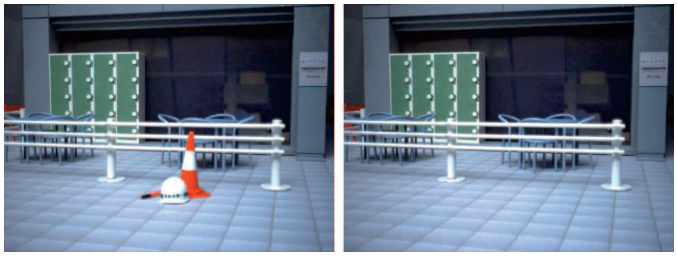


Fig. 8. Input and ground truth images for DR
Geometry and lighting except the target object must be consistent.

on the weather. In order to cope with such changes, processing on the image space is used exclusively, and exaggerated models, such as light source information and reflection models of objects, are not used. The biggest merit of observing the background in real time is that such information changing in real time is included in the background image. With no difference in the optical system of cameras, the above-mentioned optical inconsistency is often eliminated at the stage when the background is synthesized. However, multi-viewpoint observation in real time makes the system cumbersome, and, at the same time, requires computationally expensive processing. In principle, since the background observation viewpoint is located away from the user's viewpoint, the quality of the synthesis result tends to be lower than the method based on the pre-observation.

B. What is the Ground Truth Background?

Making ground truth for DR methods is not an easy task. Similar to free-viewpoint image generation methods, improving the accuracy of the synthesis result is the basic research motivation. However, how much accuracy is enough for DR? We can calculate errors by comparing a DR processed image sequence with the corresponding image sequence without a real object to be removed [21]. However, one must consider when the DR processing result cannot necessarily be defined or it is not necessary to be a true result. For example, in a case of deleting a manhole on a road, the required quality depends on the purpose of the DR process. It is unclear to us whether the pipe system under the manhole should be visualized or asphalt without manholes should be reproduced. Therefore, in such the case, it is necessary to perform a user study to evaluate whether or not the implemented DR method was able to output a visually convincing result.

C. How to Obtain Image Resources?

Particular indications of how many photos should be taken. In many cases, we will need external resources, other than collecting them ourselves, for example, to save time, human resources, and storage. So far, the Internet [15], surveillance cameras [4], and other users' cameras [5] are listed as such resource candidates, but these methods have not yet been implemented in practical applications.

D. How DR Visualizations Effect to Our Perception?

Pseudo haptics [13] is a famous perceptual concept suggesting that our haptic sensations are easily affected by vision in virtual reality. It is known that this perceptual illusion is also present in AR scenarios. In this context, what happens in DR scenarios where a user's hand or holding objects removed in the user's perspective? We obtained some user study results demonstrating that virtually shortened sticks can felt heavier than they actually are. The figure in Fig. 9 shows sticks of the same density, length, and diameter. The three figures show diminished sticks. Note that participants wield the same sticks but felt felt a distinct difference in terms of the weight.

VIII. Conclusions

In this article, the authors discussed features of arbitrary viewpoint image generation and extension of the idea to diminished reality. All examples shown in Section V are based on principles of the two approaches described in Section II (i.e., 3D shape reconstruction and ray space theory). The essence of these approaches are identical in that their purpose is to describe the 3D space regardless of differences in data representation. In DR, the described 3D space is arbitrarily reproduced in accordance with the real space coordinates to see the scene in a desired way. As described in Section VI and VII, DR has special issues to be solved in terms of image processing, such as viewpoint changes with limited image resources under dynamic lighting conditions. While some DR methods achieved high quality object removal results, researchers do not know how the results are perceived. The authors feel that these perceptual issues need further discussion in the near future.

The authors would like to introduce the first international DR survey paper for readers who discovered DR technology through this article [22]. The authors are the starting members of "Technical Committee on Plenoptic Time-Space Technology (PoTS)[1]" founded in Japan in 2015. PoTS started to investigate a way of presenting via spatio-time representation based on the idea of extending ray space theory to the time axis. Through such research activity, the authors hope to further advance the technology.

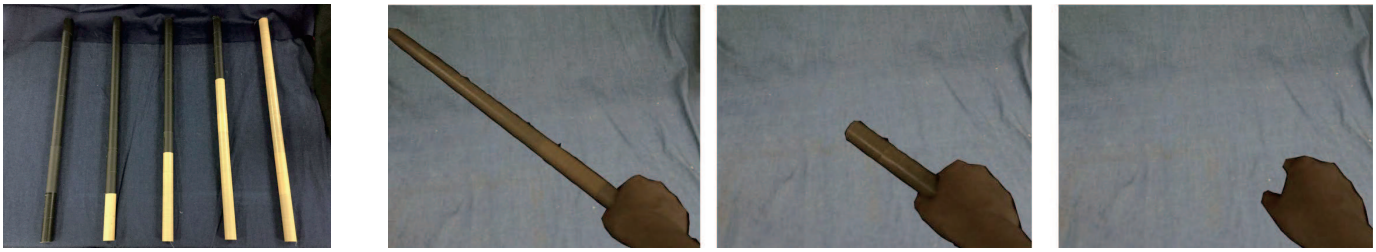
Acknowledgment

The authors are grateful to M. Maezawa, K. Oishi, L. Jinxia, M. Tanaka, and S. Hashiguchi for technical assistance with the experiments. The work presented in this manuscript is supported in part by JSPS KAKENHI Grant Numbers JP17K12729 (Grant-in-Aid for Young Scientists (B)), JP13J09193 (Grant-in-Aid for JSPS Research Fellow (DC-1)), JP16J05114 (Grant-in-Aid for JSPS Research Fellow (PD)), and JP24220004 (Scientific Research (S)).

References

[1] B. Avery, C. Sandor, and B. H. Thomas, "Improving spatial perception for augmented reality x-ray vision," Proc. IEEE VR, pp. 79–82, 2009.

[1] http://www.hvrl.ics.keio.ac.jp/pots/index_eng.html



(a) Sticks to be wielded (b) Diminished sticks

Fig. 9. Wielding a diminished stick
Our user study showed that the shorter the diminished stick to be wielded, the heavier the stick felt.

[2] R. T. Azuma, "A survey of augmented reality." Presence: Teleoperators and Virtual Environments, Vol. 6, No. 4, pp. 355–385. 1997.

[3] R. T. Azuma, "Recent advances in augmented reality," IEEE Computer Graphics and Application, Vol. 22, pp. 34–47, 2001.

[4] P. Barnum, Y. Sheikh, A. Datta, and T. Kanade, "Dynamic seethroughs: Synthesizing hidden views of moving objects," Proc. Int. Symp. on Mixed and Augmented Reality (ISMAR), pp. 111–114, 2009.

[5] A. Enomoto and H. Saito, "Diminished reality using multiple handheld cameras," Proc. Asian Conf. on Computer Vision (ACCV), pp. 130–150, 2007.

[6] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, V. Tankovich, C. Loop, Q. Cai, P. Chou, S. Mannicken, J. Valentin, V. Pradeep, S. Wang, S. B. Kang, P. Kohli, Y. Lutchyn, C. Keskin, and S. Izadi, "Holoportation: Virtual 3D teleportation in real-time," Proc. SIGGRAPH, pp. 741–754, 2016.

[7] T. Fujii and H. Harashima, "Coding of an autostereoscopic 3-D image sequence," SPIE VCIP, Vol. 2308, pp. 930–941, 1994.

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI). Vol. 20, No. 11, pp. 1254–1259., 1998.

[9] T. Kanade, P. J. Narayanan, and P. W. Rander, "Virtualized reality: Concepts and early results," Proc. Workshop on Representation of Visual Scenes, 1995.

[10] N. Kawai, M. Yamasaki, T. Sato, and N. Yokoya, "Diminished reality for AR marker hiding based on image inpainting with reection of luminance changes," ITE Trans. Media Technology and Applications, Vol. 1, No. 4, pp. 343–353, 2013.

[11] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," Matters of Intelligence, Springer, pp. 115–141, 1987.

[12] E. Kruijff, J. E. Swan, and S. Feiner, "Perceptual issues in augmented reality revisited," Proc. Int. Symp. on Mixed and Augmented Reality (ISMAR), pp. 3–12, 2010.

[13] A. Lecuyer, S. Coquillart, A. Kheddar, P. Richard, and P. Coiffet, "Pseudo-haptic feedback: can isometric input devices simulate force feedback?," Proc. IEEE Virtual Reality, pp. 83–90, 2000.

[14] J. Li, J. Saito, S. Mori, S. Ikeda, F. Shibata, A. Kimura, and H. Tamura, "Re-design and implementation of MR-based filmmaking system by adding diminished reality functions," Trans. on Virtual Society of Japan, vol. 21, no. 3, pp. 451–462, 2016. (in Japanese)

[15] Z. Li, Y. Wang, J. Guo, L.-F. Cheong,S. Z. Zhou, "Diminished reality using appearance and 3D geometry of Internet photo collections," Proc. Int. Symp. on Mixed and Augmented Reality (ISMAR), pp. 11–19, 2013.

[16] S. Mann, "Mediated reality with implementations for everyday life," Presence Connect, 2002.

[17] E. Marchand, H. Uchiyama, F. Spindler, "Pose estimation for augmented reality: A hands-on survey," IEEE Trans. on Visualization and Computer Graphics, Vol. 22, Issue 12, pp. 2633–2651, 2015.

[18] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan, "Image-based visual hulls," Proc. SIGGRAPH, pp. 369–374, 2000.

[19] S. Mori, F. Shibata, A. Kimura, and H. Tamura, "Efficient use of textured 3D model for pre-observation-based diminished reality," Proc. Int. Workshop on Diminished Reality as Challenging Issue in Mixed and Augmented Reality (IWDR), pp. 32–39, 2015.

[20] S. Mori, M. Maezawa, N. Ienaga, and H. Saito, "Detour light field rendering for diminished reality using unstructured multiple views," ditto, pp. 292–293, 2016.

[21] S. Mori, Y. Eguchi, S. Ikeda, F. Shibata, A. Kimura, and H. Tamura, "Design and construction of data acquisition facilities for diminished reality research," ITE Trans. on Media Technology and Application (MTA), Vol. 4, No. 3, p. 259–268, 2016.

[22] S. Mori, S. Ikeda, and H. Saito, "A survey of diminished reality: Techniques for visually concealing, eliminating, and seeing through real objects," IPSJ Trans. on Computer Vision and Applications (CVA), 2017, DOI: 10.1186/s41074-017-0028-1.

[23] K. Oishi, S. Mori, and H. Saito, "Diminished reality using by 3D-LIDAR and omnidirectional camera," Proc. National Convention of IPSJ, 4N-08, pp. 2-319–2-320, 2017. (in Japanese)

[24] M. Santos, I. Souza, G. Yamamoto, T. Taketomi, C. Sandor, and H. Kato, "Exploring legibility of augmented reality x-ray," Multimedia Tools and Applications, Col. 75, Issue 16, pp. 9563–9585, 2015.

[25] G. M. Stratton, "Some preliminary experiments on vision without inversion of the retinal image," Psychological Review. Vol. 3, No. 6, pp. 611–617, 1896.

[26] T. Tsuda, H. Yamamoto, Y. Kameda, and Y. Ohta, "Visualization methods for outdoor see-through vision," IEICE Trans. Information and Systems, pp. 1781–1789, 2006.