2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and
2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)

Xth Person View Video for Observation from Diverse Perspectives

Naoki Shimura*, Hidehiko Shishido†, Yoshinari Kameda‡, Kenji Suzuki§, Itaru Kitahara‡
Systems and Information Engineering*, Center for Computational Sciences†‡, Faculty of Engineering§
University of Tsukuba
Tsukuba, Japan
E-mail: s1620791@u.tsukuba.ac.jp*, shishido@ccs.tsukuba.ac.jp†, {kameda, kitahara}@iit.tsukuba.ac.jp‡, kenji@ieee.org§

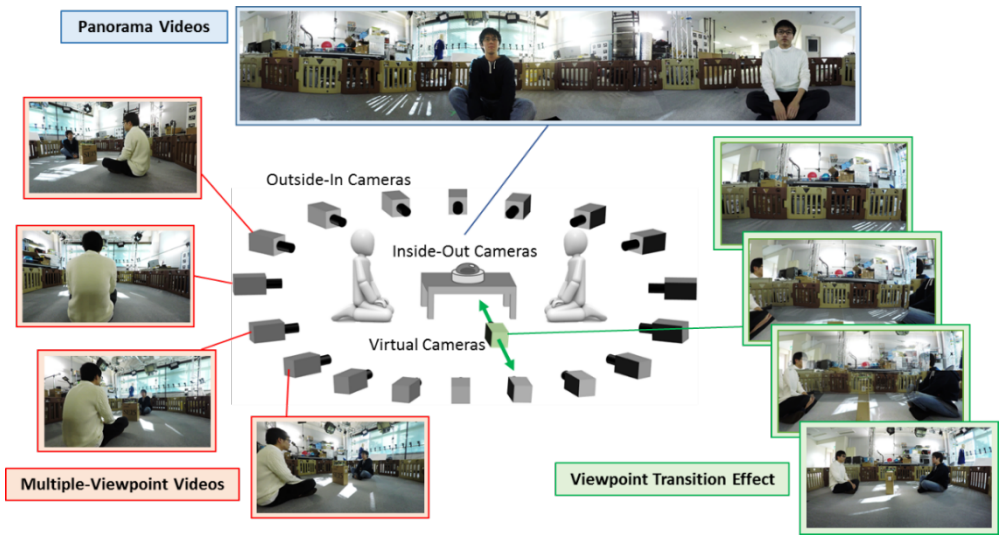

Fig 1. System overview of Xth Person View Video

***Abstract*— This paper proposes "Xth Person View Video" that enables observation from the viewpoints of diverse people and objects using multiple-view video technique. Observation from various perspectives is achieved by properly switching "Outside-In images" captured from the outside to the inside of the target scene and "Inside-Out images" captured from the inside to the outside. To generate Outside-In video, we use the Bullet-Time video technique, which presents images by switching multiple-viewpoint videos based on the capturing camera alignment. "Enhanced Bullet-Time" is used to make appropriate observations on target objects moving in the space. Inside-Out video is generated by extracting the view of an appropriate direction from a panoramic video. To make observers intuitively understand the viewpoint movement between Outside-In video and Inside-Out video, which normally looks different, we employ "Pseudo Viewpoint Transition", which expresses viewpoint movement between two viewpoints via simple 3D CG models. The effectiveness of the proposed method is evaluated through both objective and subjective evaluations.**

Keywords-component; Multi-view images; Omnidirectional image; Viewpoint transition images; Inside-Out/Outside-In shooting;

I. INTRODUCTION

In collaborative work and teaching tasks with multiple people, it is important for everyone to understand each other's intentions. Since being able to see things from your partner's point of view has a great influence on communication, sometimes the teaching side and the taught side are arranged side by side and work under a common view. However, in the case of dynamic work such as sports instruction and children's education, it is difficult to create such a common viewpoint. Our research aims to improve the efficiency of collaborative work and teaching tasks by reproducing viewpoints from various perspectives by utilizing multi-view video media called "Xth Person View Video".

Multi-view images with multiple cameras record a given scene from various directions, so it is suitable for detailed observation of subjects [1]. Bullet-Time [1] [2] is a visual effect presenting multi-view images taken with cameras surrounding a subject while switching according to the camera arrangement. Since a feeling of movement of the viewpoint can be reproduced, Bullet-Time has been utilized for movies, promotional videos and sports [3]. As mentioned above, it is superior to multifaceted observation; however, when viewing images with cameras placed around the subject (Outside-In shot), it is difficult to

978-1-5386-5163-6/18/$31.00 ©2018 IEEE

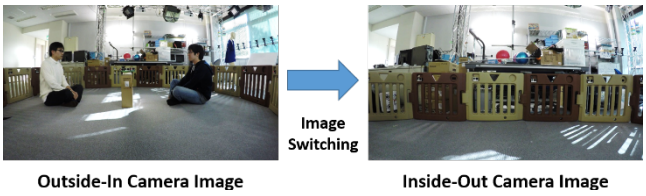

Figure 2. Switching between images shot by two cameras with different shooting distances.

reproduce the appearance as if the user is immersed in the shooting space.

On the other hand, if you shoot the scene with an omnidirectional camera placed inside the space (Inside-Out shooting), you can observe how it looks from the immersive viewpoint. In addition, since the subject is shot from a closer distance, it is possible to observe that subject in detail. However, because its shot is taken from a single viewpoint, it is difficult to observe the subject from multiple directions.

In this paper, we propose an image presentation method to achieve "Xth Person View Video"by combining the advantages of Outside-In and Inside-Out images as shown in Fig. 1, while appropriately switching between the images.

II. Related Works

In a telepresence system, Rekimoto et al. [4] proposed JackIn Space, which uses not only a specific first-person image but also third-person view images taken with another camera. In the conventional telepresence system using only the first-person viewpoint, there is a problem in the fact that it is difficult to understand the situation in a remote place. However, by using the third-person view image, it is possible to see the surroundings of the remote place. In addition, when switching to images from another viewpoint, viewpoint transition images are generated to understand the spatial relationship among the cameras. These viewpoint transition images are generated by using a wide-angle camera attached to a person, and a plurality of depth sensors attached to the remote place. In this method, in addition to Outside-In cameras placed surrounding the subject to understand the situation of the shooting space, we use Inside-Out cameras that are omnidirectional and placed inside the shooting space. Thus, it is possible to observe subjects in detail.

On the other hand, the shooting directions of the Outside-In cameras and the Inside-Out cameras are much different. Thus, when simply switching from the Outside-In camera to the Inside-Out camera causes some drastic changes in the appearance, as shown in Fig. 2, it is difficult for the observer to understand the spatial relationship among the cameras. If the 3D shape data of the shooting space is known, it is possible to reproduce the appearance between the Outside-In imaging and the Inside-Out imaging by using free viewpoint technology [5] [6]. However, reconstructing 3D information of a dynamic scene in real time requires calculation cost.

Tsuda et at. [7] proposed an image-generation method that smoothly transitions between multi-view images captured at different points of view by using 3D models of static objects in a previously generated shooting space. In this method, by generating viewpoint transition images using 3D models of static objects such as a wall or a floor surface, it is possible to smoothly interpolate the appearance of the Outside-In image and the Inside-Out image. As these viewpoint transition images are like transitions from one viewpoint to another, the observer can easily understand the spatial relationship among the cameras.

III. Processing for Generating Outside-In Video

Outside-In video is generated using the Bullet-Time technique. By switching the images taken with the Outside-In cameras placed surrounding the subject according to the camera arrangement, the visual effect as ifof moving around the subject is reproduced. When presenting the observer multi--view images with the optical axis of each camera not intersecting at a point while switching it, the observation position (attention point) of the target object on the screen fluctuates. SoThus, it is difficult to give to the observer a smooth viewpoint movement perception. In this method, we reset the optical axis direction of the virtual camera (applying homography transformation to the images) based on the position of the point of attention, and we convert the appearance of Outside-In images so that the point of attention is observed at the same position on the screen [2].

Structure from Motion (SfM) is applied to the multi--view image [8] [9], to estimate internal and external parameters of the multi--view camera, required for conversion processing. The position vector of the Outside-In camera is M_outand the position vector of the point of attention is X. The direction vector D of the optical axis of the vertical camera is shown as the following equations:

$$\boldsymbol{D'} = \boldsymbol{X} - \boldsymbol{M_{out}}\ , \tag{1}$$

$$\boldsymbol{D} = \frac{\boldsymbol{D'}}{|\boldsymbol{D'}|}. \tag{2}$$

By recalculating equations (1) and (2) when switching to another Outside-In camera image, the optical axis of the virtual camera always intersects the point of attention. Also, for continuous observation before and after switching, it is necessary to maintain the apparent size of the target object. Therefore, in this system, the angle of view of the virtual camera is reset before and after switching. The angle of view of the virtual camera is α, and the distance between the virtual camera and the point of attention is d. The view angle of the virtual camera after switching is shown as the following equation. Fig. 3 shows the result of applying these processes.

$$\alpha_{n+1} = \alpha_n \frac{d_n}{d_{n+1}}\ . \tag{3}$$

IV. Processing for Generating Inside-Out Images

Inside-Out video is achieved by utilizing panorama images. Using an omnidirectional camera makes it possible to capture images in all directions with only one camera; however, the spatial resolution decreases. In this system, in order to observe the subject in detail, the Inside-Out image is taken using eight cameras fixed to a rig as shown in Fig. 4. The rig is fixed inside the shooting space to be almost the same height as the mounting position of the Outside-In cameras, and images viewed from inside the shooting space are taken. We apply stitching

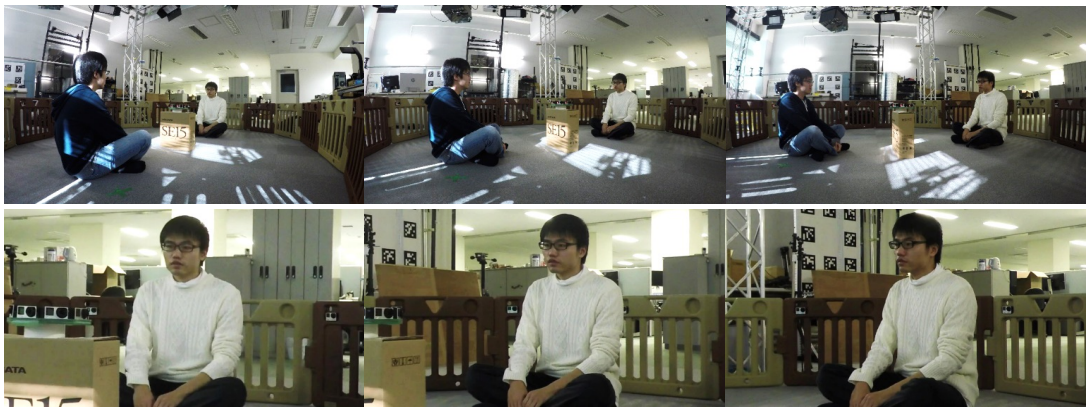

Figure 3. Attention point setting processing of Outside-In cameras.
(Top: Original images, Bottom: Processed images)

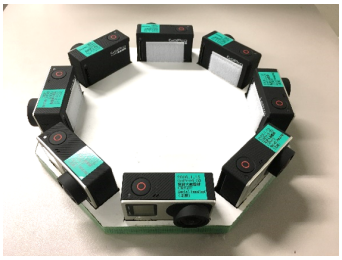

Figure 4. Camera rig used for Inside-Out

processing to eight Inside-Out camera images to generate one panorama image. Fig. 5 shows an example of the synthesized panorama image.

V. Processing for Generating Outside-In Images

This section describes the generation process of the viewpoint transition image between the Outside-In camera and the Inside-Out camera, meaning the moment of switching the image from the Outside-In camera image to the Inside-Out camera image.

A. Generating 3D Models

To generate the viewpoint transition image, 3D models of static objects are necessary. An example of the shooting space is shown in Fig. 6, and now we describe the generation processing of the 3D models. In this case, the Outside-In camera is fixed at a predetermined position on the panel. We generate 3D models of the wall, the floor and the panels using camera parameters estimated by SfM and the outline of the imaging space (sparse 3D Point Cloud). The model of the floor is planar, and it is set downward by the specified value from the height of the Outside-In cameras. The 3D model of the wall is a cylinder that stands upright on the floor, and the model is centered at the point where the optical axes of the Outside-In cameras intersect. The model of a panel is a cuboid of predefined size and is set based on the position and orientation of the Outside-In camera. Multi-view images are pasted onto the estimated 3D models using projection texture mapping [11].

B. Generating Viewpoint Transition Images

The viewpoint transition images are generated by moving the virtual camera between the Outside-In camera and the Inside-Out camera. The position vector $\boldsymbol{M}$ of the virtual camera is expressed by equation (4) using the position vector $\boldsymbol{M_{in}}$ of the Inside-Out camera and the position vector $\boldsymbol{M_{out}}$ of the Outside-In camera.

$$\boldsymbol{M} = k\boldsymbol{M_{in}} + (1-k)\boldsymbol{M_{out}} \quad 0 \le k \le 1 \tag{4}$$

The direction vector $\boldsymbol{D}$ of the optical axis of the virtual camera is expressed by equations (5) and (6) using the position vector $\boldsymbol{X}$ of the target object and the position vector $\boldsymbol{M}$ of the virtual camera:

$$\boldsymbol{D'} = \boldsymbol{X} - \boldsymbol{M}\ , \tag{5}$$

$$\boldsymbol{D} = \frac{\boldsymbol{D'}}{|\boldsymbol{D'}|}. \tag{6}$$

By always directing the direction vector $\boldsymbol{D}$ of the optical axis of the virtual camera to the object of attention, as shown in Fig. 7, the viewpoint is switched from the Outside-In camera image

Figure 5. Example of panoramic image.

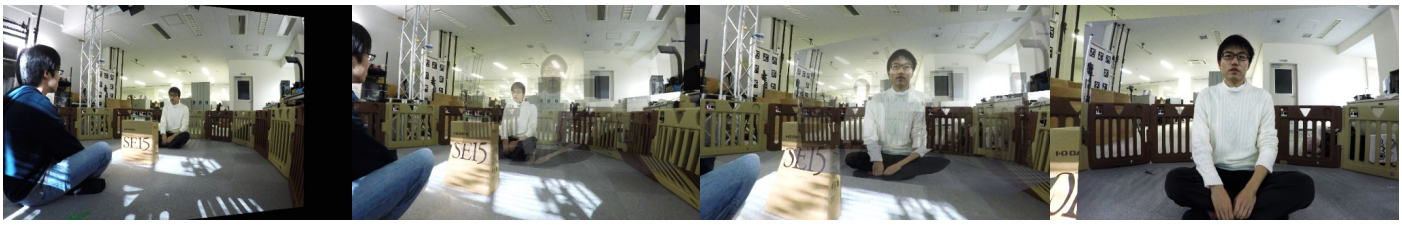

Figure 8. Viewpoint transition image from Outside-In camera to Inside-Out camera.

Figure 6. Shooting space.

to the Inside-Out camera image while capturing the target object. An example of the processing result is shown in Fig. 8.

C. *Selecting Appropriate Camera Process*

When switching from the Outside-In camera image to the Inside-Out camera image, it is necessary to change the viewpoint while always capturing the target object. Therefore, we switch to the Inside-Out camera image in which the target object appears. First, a direction vector $\boldsymbol{D_x}$ is obtained from the center position of the Inside-Out cameras to the target object. The vector $\boldsymbol{D_x}$ is expressed by equation (7) using the position vector $\boldsymbol{M_c}$ of the center of the Inside-Out camera group:

$$\boldsymbol{D_X = X - M_C} \ . \qquad (7)$$

We find the inner product of $\boldsymbol{D_x}$ and the direction vectors of the optical axis of all Inside-Out cameras, and select the Inside-Out camera with the largest inner product.

VI. Shooting And Image Generation

We describe the execution environment of the shooting/image generation. The size of the shooting space is 710 [cm] × 760 [cm]. We used a desktop PC equipped with CPU: Intel Core i7 3.40 [GHz], GPU: NVIDIA GeForce GTX 550 Ti, RAM: 8.00 [GB]. For Outside-In cameras and Inside-Out cameras, we use GoPro HERO4 Black edition (Outside-In cameras: 20 units, Inside-Out cameras: 8 units) from GoPro, Inc., which are capable of wireless synchronous photographing. The resolution is 3,840 [pix] × 2,160 [pix], the frame rate is 30 [fps], the horizontal field angle is 122.6°, and the vertical field angle is 94.4°. Outside-In cameras are placed surrounding the photographic scene, fixed at a predetermined position on panels of the baby fence defining the shooting area. Inside-Out cameras are fixed to a rig inside the shooting space at the same height as the Outside-In cameras. We use Autopano Giga from Kolor [10] to generate a panoramic image combining the captured multiple viewpoint images from inside of the space. VisualSFM [12] is

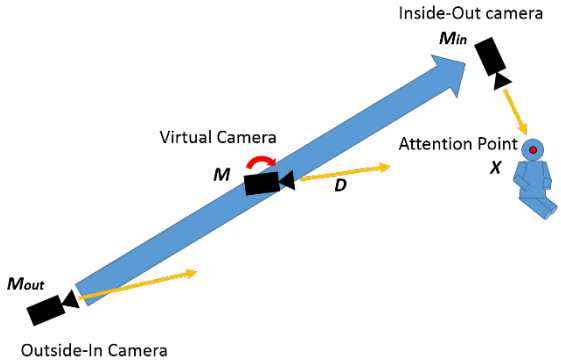

Figure 7. Virtual camera motion.

used for SfM camera calibration. The processing takes 82 seconds for camera calibration and initial setup takes about 8 seconds, but the processing after starting to browse can be executed in real time (video-rate). The time required for the processing is 82 seconds for the camera calibration and the initial setting time takes about 8 seconds. The processing after the start of the sequential browsing can be executed at the video rate (30 frames per second).

VII. Experiment

This section introduces quantitative evaluation to verify the spatial resolution of viewing videos with this system, and a qualitative evaluation using NASA-TLX to evaluate psychological load at the observation task. In this experiment, we calculate the viewing area and the spatial resolution of the images of two people sitting facing each other. In addition, the observation task of the shooting space using this system is carried out, and evaluated by NASA-TLX.

A. *Quantitative Evaluation*

One of the features of the proposed method is to enable detailed observation of the subject. The effectiveness of the proposed method is verified using the map, visualizing the spatial resolution of the subject observed in the image by this system.

1) *Evaluation Method*

As shown in Fig. 9, we take a picture of the situation where two persons are sitting face to face with the system introduced in Section 6. We apply SfM to the captured multi-view images and estimate the camera parameters. Based on the estimated parameters, rough 3D models of the floor, panels and persons in the imaging space are generated. At this time, the generated models are divided into fine patches. For each patch, we calculate the spatial resolution observed in the displayed video,

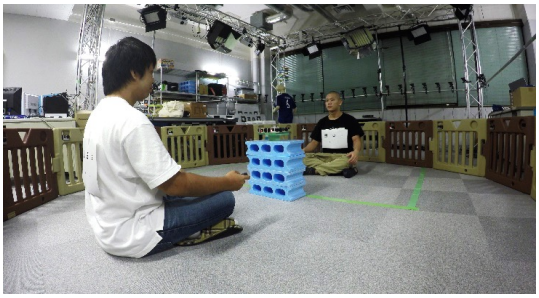

Figure 9. Shooting environment.

and make it the intensity value of the patch. In this study, the spatial resolution is defined as the actual size of the 3D area captured by one pixel in the image. When the shooting distance from the camera to the model is d, the angle of view of the camera is a, and the resolution of the image is p. The spatial resolution r is expressed by the following equations:

$$r = d\tan\theta, \tag{8}$$

$$\theta = \frac{a}{p}. \tag{9}$$

For this experiment, the GoPro HERO4 Black edition is used. The resolution is 3,840 × 2,160 [pix], the horizontal angle of view is 118.2°, and the vertical field angle is 69.5°. The spatial resolution normalized the smallest value (highest resolution) among the spatial resolutions of all the cameras. By comparing this process with the case of using the Inside-Out cameras, and the case without (not using) the Inside-Out cameras, change in the spatial resolution is verified. By setting the Inside-Out cameras at the center of the shooting space, the shooting distance between the camera and the person is shortened, so it is expected that the spatial resolution mainly in the frontal area of the person will be improved.

2) Result

We consider the results regarding the visualization of spatial resolution. The upper part of Fig. 10 shows the result when the Inside-Out cameras are set, and the lower part of Fig. 10 shows the result when the Inside-Out camera is not set. When the Inside-Out cameras are set, the white area with high spatial resolution increases compared to when the Inside-Out cameras are not set. The effect appears in the frontal area of the person being observed; details such as facial expression and gaze are remarkably clearer.

B. Qualitative Evaluation

In the observation task using this system, we perform qualitative evaluation using NASA-TLX, which is widely used in ergonomics to conduct usability tests, including the mental load of an observer.

1) NASA-TLX

NASA-TLX [13] is an evaluation method of asking a subject to perform a certain task and asking a question about this work to obtain a score of the six mental loads in the work of the subject. There are six measures of mental load: MD (Mental Demand), PD (Physical Demand), TD (Temporal Demand), EF (Effort), FR (Frustration), and OP (Own Performance). Mean Weighted

Figure 11. E chart.

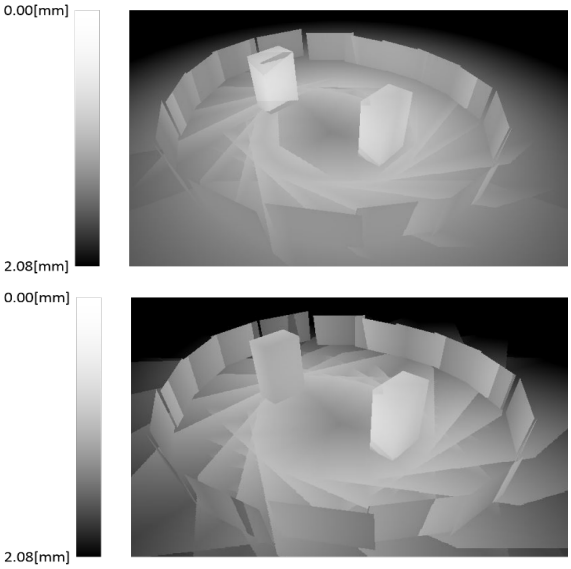

Figure 10. Spatial resolution maps. (Top: With Inside-Out cameras, Bottom: Without Inside-Out cameras)

Workload Score (WWL) is the average score of all the scales weighted by paired comparison for each measure. In this experiment, we ask the subjects to use this system to observe the images in which two persons are sitting facing each other, as in Fig. 9, and read the charts taped to their fronts and backs. The observation task is done twice; first with the Inside-Out cameras set, second with them not set. After completion of each observation task, a NASA-TLtest is performed to verify the mental burden that this system exerts on the observer.

2) Evaluation Method

In this experiment, the observation task of images shot in advance is performed using the proposed system. The display used for image presentation is EIZO EV2736W with a resolution of 2,560 × 1,440 [pix] and a pixel pitch of 0.233 × 0.233 [mm]. Also, the presented images are displayed with 1,728 × 972 [pix]. In the shooting space, as shown in Fig. 9, two persons are sitting facing each other. On the front and back of each person, a chart of five E symbols shown in Fig. 11 is affixed. The sizes of the five figures are 4, 3, 2, 1, and 0.5 [cm] respectively. We ask the subjects to perform the task of telling the direction of the figures on these charts. Subjects were 11 people aged 22 to 28 (average: 23.9, standard deviation: 1.87). Each subject, after practicing once, performed the observation twice with the Inside-Out images and without the Inside-Out images respectively. The presentation order of the two conditions was determined randomly for each subject. Every time the observation task was completed, a questionnaire by NASA-TLX was carried out.

3) Result

We describe the correct answer rate of the character reading task using this system and the result of mental burden estimation by NASA-TLX. Fig. 12 shows the correct answer rate of the character reading task. The accuracy of correct

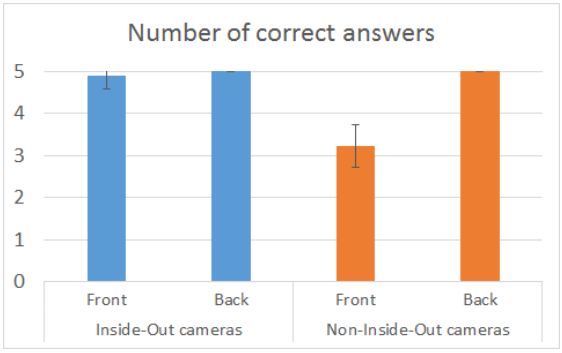

Figure 12. Number of correct answers in assignment. (mean±SD)

reading of figures in front of a person is higher when the Inside-Out cameras are set. This seems to be because the spatial resolution in the human frontal area improved as described in 7.1.

When the distance from the observer's eyes to the display is 600 [mm] and the size of the figure on the display is s, the viewing angle v is expressed by the following equation:

$$v = \arctan\frac{s}{600}. \quad (10)$$

Using equation (10), the viewing angle is 0.156° for the 2-cm figure and 0.067° for the 1-cm figure. On the display, the 2-cm figure is 7 [pix] (1.631 [mm]) and the 1-cm figure is 3 [pix] (0.699 [mm]). Therefore, even with the image resolution, there is a limit in the case of only using the Outside-In cameras. On the other hand, when the Inside-Out cameras are set, the correct answer rate of character reading for a person's frontal area is close to a perfect score. Setting the Inside-Out cameras makes the shooting distance short. Thus, it is possible to observe details that could not be seen using only the Outside-In cameras. There is not much difference in the character reading task on the rear of a person because, even if Inside-Out cameras are set, the observer performs the character reading task using the Outside-In camera images. Fig. 13 shows the result of NASA-TLX. When the Inside-Out cameras are set, it results in better WWL, which is a comprehensive evaluation index. In addition, there is a significant difference in EF, FR, and especially in OP (Analysis of variance, $p< 0.01$). In the case with Inside-Out cameras, EF is reduced because it is not necessary to switch to other camera images to be able to read the characters. In addition, FR is reduced. This is thought to be because subjects feel more stress when trying to read with only the Outside-In cameras. Also, when using Inside-Out cameras, OP improves. This is due to the observer being able to read almost all of the characters as shown in Fig. 12, causing them to feel achievement in the intended task. Therefore, it is confirmed that the mental burden is reduced by performing observation tasks using this system.

VIII. Conclusion

This paper proposed an image-browsing system that combined the advantages of two perspectives by using Inside-Out camera images in addition to Outside-In camera images. We enabled more detailed observation by shooting images from the interior of the shooting space, which cannot be observed with only conventional Outside-In images. To verify the effectiveness of the proposed system, we performed quantitative

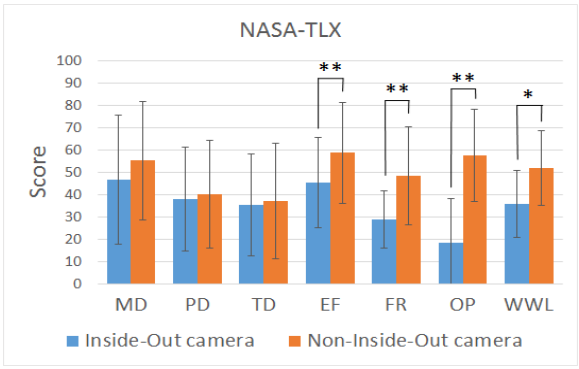

Figure 13. Result of NASA-TLX. (* $p< 0.05$, ** $p<0.01$, mean±SD)

evaluation based on the spatial resolution map. Qualitative evaluation of this system was also conducted using NASA-TLX. One of the future works is to display Xth Person View "movie". We can separately generate panoramic-view video and bullet-time video in real-time. So, our next step is combining them with generating viewpoint transition image in real-time. If the number of capturing cameras increase, it is possible to realize more smooth viewpoint switching. However, as the result, the amount of the video data becomes huge. It is necessary to develop multimedia streaming method using image compression technology.

This work was supported by JSPS KAKENHI Grant Numbers 17H01772 and JST CREST Grant Number JPMJCR14E2, Japan.

References

[1] K. Ikeya, K. Hisamoto, M. Katayama, T. Mishina, and Y. Iwadate, "Bullet time using multi-viewpoint robotic camera system" , In Proceedings of the 11th European Conference on Visual Media Production Article No. 1, 2014.

[2] N. Akechi, I. Kitahara, R. Sakamoto, and Y. Ohta., "Multi-resolution bullet-time effect", In SIGGRAPH Asia 2014 Posters, Article No. 30, 2014.

[3] Timeslice Films, "Timeslice Films - Pioneers Of Multi-Cam", http://timeslicefilms.com/, 2017.

[4] R. Komiyama, T. Miyaki, and J. Rekimoto, "JackIn Space: Designing a seamless transition between first and third person view for effective telepresence collaborations", In Proceedings of the 8th Augmented Human International Conference, Article No. 14, 2017.

[5] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese,

[6] H. Hoppe, A. Kirk, S. Sullivan, "High-quality streamable free-viewpoint video", In ACM SIGGRAPH 2015, Vol. 34, No. 69, 2015.

[7] T. Tsuda, I. Kitahara, Y. Kameda, and Y. Ohta, "Smooth video hopping for surveillance cameras", In ACM SIGGRAPH 2006 Sketches, pages 129, 2006.

[8] C. Wu, "Towards linear-time incremental structure from motion", In Proceedings of the 2013 International Conference on 3D Vision, pages 127-134, 2013.

[9] C. Wu, S. Agarwal, B. Curless, and S.M. Seitz, "Multicore bundle adjustment", In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, pages 3057-3064, 2011.

[10] A. Jenny, "June. Kolor | Autopano - panorama software", http://www.kolor.com/autopano/, 2017.

[11] M. Segal, C. Korobkin, R. Van Widenfelt, J. Foran, and P. Haeberli, "Fast shadows and lighting effects using texture mapping", SIGGRAPH '92 Proceedings of the 19th annual conference on Computer graphics and interactive techniques, 1992.

[12] C. Wu, "VisualSFM: A Visual Structure from Motion System", http://ccwu.me/vsfm/, 2017.