PAPER

Applicability of camera works to free viewpoint videos with annotation and planning

Ryuuki SAKAMOTO[†], *Nonmember*, **Itaru KITAHARA**[†,††], *Member*, **Megumu TSUCHIKAWA**[†††], **Kaoru TANAKA**[††††], *Nonmembers*, **Tomoji TORIYAMA**[†], *and* **Kiyoshi KOGURE**[†], *Members*

SUMMARY This paper shows the effectiveness of a cinematographic camera for controlling 3D video by measuring its effects on viewers with several typical camera works. 3D free-viewpoint video allows us to set its virtual camera on arbitrary positions and postures in 3D space. However, there have been neither investigations on adaptability nor on dependencies between the camera parameters of the virtual camera (i.e., positions, postures, and transitions) nor the impressions of viewers. Although camera works on 3D video based on expertise seems important for making intuitively understandable video, it has not yet been considered. When applying camera works to 3D video using the planning techniques proposed in previous research, generating ideal output video is difficult because it may include defects due to image resolution limitation, calculation errors, or occlusions as well as others caused by positioning errors of the virtual camera in the planning process. Therefore, we conducted an experiment with 29 subjects with camera-worked 3D videos created using simple annotation and planning techniques to determine the virtual camera parameters. The first point of the experiment examines the effects of defects on viewer impressions. To measure such impressions, we conducted a semantic differential (SD) test. Comparisons between ground truth and 3D videos with planned camera works show that the present defects of camera work do not significantly affect viewers. The experiment's second point examines whether the cameras controlled by planning and annotations affected the subjects with intentional direction. For this purpose, we conducted a factor analysis for the SD test answers whose results indicate that the proposed virtual camera control, which exploits annotation and planning techniques, allows us to realize camera working direction on 3D video.

key words: *free-viewpoint video, 3D video, camera work, cinematographic, semantic differential*

1. Intoroduction

Such filming techniques as switching or moving cameras based on expertise are intentionally applied when movies or TV shows are filmed. Some treatises explicitly explain the expertise and claim that camera moving and switching makes it possible to attractively or precisely explain and present the situation [1]–[3]. We hereafter use the term "camera work" to refer to controlling cameras based on expertise. The importance of camera work is also discussed in Computer Graphics (CG). Automatic camera control methods for "camera work" have been investigated that exploit planning techniques on three-dimensional CG spaces [4]–[7].

Technology called 3D or free-viewpoint video has generated views at arbitrary viewpoints in a 3D space from multiple video streams with a Virtualized Reality technique [8], and several practical methods that exploit it have also been proposed. The goal of these practical studies includes support for watching sports [9]–[11], sending telepresence [12], watching traditional performing arts [13], [14], and attending a teleconference [15], [16].

3D video allows viewers to set the virtual camera at arbitrary positions and postures in 3D space. Since a virtual camera can freely move in 3D space, we assume that in theory many camera works are adaptable on 3D video. Precedent practical researches exploiting 3D video allowed users to manipulate virtual cameras; however, they did not consider the camera working of the virtual camera. Research has also failed to investigate the effects of outcome video, which depends on positions, postures, or transitional parameters of virtual cameras, on viewer impressions. However, when applying camera works to 3D video using planning techniques proposed in precedent research, generating ideal output video is difficult because it may include defects due to image resolution limitations, calculation errors, or occlusions as well as others caused by the positioning errors of the virtual camera in the planning process.

This paper clarifies the applicability of cinematographic camera control in 3D video by measuring the psychological effects on subjects who viewed camera-worked 3D videos containing such defects. Hereafter, we call such video, which is made as 3D video with camera work, *Cinematographic 3D video*. First, we measured the effects of defects on viewer impressions by conducting a semantic differential (SD) test. Second, we examined whether cinematographic camera working affected the subjects with intentional direction.

The cinematographic 3D videos used in the experiment were made by general techniques, and thus the concept and techniques can be used for diverse applications when adaptation of the camera work to 3D video is worthwhile. For example, cinematographic 3D video

Manuscript received February 5, 2007.
Manuscript revised April 27, 2007.
Final manuscript received October 1, 2007.

[†]The author is with the ATR Knowledge Science Laboratories
[†††]The author is with NTT
[††]The author is with the Univ. of Tsukuba
[††††]The author is with the Japan Advanced Institute of Science and Technology

may be used in educational applications for filming production because it can re-create camera work whenever required. In current film production, actors have to repeat the same scene shot by camerapersons to apply different camera works on identical scenes. Alternatively, cinematographic 3D video may also be used for Previsualization (also known as Previs or Animatics), which aims to create preproduction video utilizing 3D CG to check shots before actual shooting [17] because cinematographic 3D video allows directors to check shots on more sophisticated video than video using 3D CG.

The rest of this paper is organized as follows. We give an overview of related works in Section 2 and show how to generate cinematographic 3D videos for the experiment in Section 3. The experiment's procedure and results are shown in Section 4. We conclude with limitations in Section 5.

2. Related works

The investigations on computers that control camera work can be divided into two categories: automatically switching cameras set in the real world [18], [19] and controlling virtual cameras and CG characters in the 3D CG world [4]–[7].

Note the investigation by Inoue et. al [18] categorized in the former group. Based on camera work rules called imaginary lines and explained in the grammar of film language, they tried to switch real cameras set in appropriate positions. Based on the research, the assertiveness of situational explanations improves when switching is done along the rules. This research, however, deals with 3D video, and so its target field differs from this paper.

Studies of the latter group are developing applications in the field of Artificial Intelligence by focusing on virtual camera control in 3D CG animation. In these studies, there are two kinds of worlds: where the camera is independent from CG characters' action [5]–[7] and where it is dependent [4], [18].

For example, CamPlan [4] realized various camera works on independent 3D CG animation from moving cameras. This system decided the camera parameters of camera works with distance, rotation, and height adjustments. On the contrary, TVML [5] focuses on controlling 3D characters with a script as well as a virtual camera and characters. In 3D video, since we focus on objects and virtual cameras independent of each other, our world is close to CamPlan's world. However, these studies will not also work on 3D video.

Several applications exploiting 3D video have been proposed [9]–[12], [15], [20]. MR-PreVis project represents one of the method to support the previsualization with mixed-reality techniques, which allows us to combine a virtual model with a real scene [20]. No research, however, has investigated the automatic control of virtual cameras and their psychological effects.

3. Making cinematographic 3D video for experiment

This study focuses on the general applicability of camera works on 3D video. For that purpose, we prototyped a system that generates 3D video to which we applied camera works, because no appropriate 3D video system has been reported that is applicable to camera works and no method applies camera works to 3D videos. The system consists of three sub-softwares: capturing/generating 3D videos, annotating objects, and planning virtual camera positions based on camera works.

In the system each sub-software holds a generality because it is used for well-known or simple methods of making 3D video and planning camera works. Although annotation, mentioned below, that notes object areas is not proposed in any present study, it is simple to use in other systems.

We captured two scenes as 3D video and generated three cinematographic 3D Videos using this system. This section briefly describes the system's procedure and cinematographic 3D Videos used on the psychological experiments mentioned below. Detailed procedures are found in references presented previously [21].

3.1 Captured scenes

Camera work treatises use a certain number of pages to explain camera works which target less than two or three people[1]–[3]. The technique discussed here focuses on ordinary scenes with a few people; for example, people having trivial daily conversations are directed and explained by camera work. .

Therefore, as filming subjects we chose two scenes: a dialogue scene with two people and a walking scene with one person. Fig. 1 shows the actor movements in these scenes. Hereafter, the scene showing the upper side of Fig. 1 is referred to as the "walking scene" and the lower side as the "dialogue scene." Such a situation as a walking scene can often be seen in movie introductions, and dialogue scenes can be seen throughout the whole story. An actress performed the walking scene for 20 sec, and two actors performed the dialogue scene for 8 sec.

We captured these scenes with eight (walking scene) or five (dialogue scene) PCs and calibrated USB2.0 color cameras that can capture VGA images with 20 fps. The sizes of the captured spaces were approximately $5.5 \times 5.5 \times 2.5$ m (walking scene) and $5.5 \times 2.5 \times 2.5$ m (dialogue scene), and environmental cameras were set on the wall to surround the space shown as Fig. 1. Pentium IV 2.8-GHz PCs as environmental cameras captured a video stream from each camera and segmented the objects.

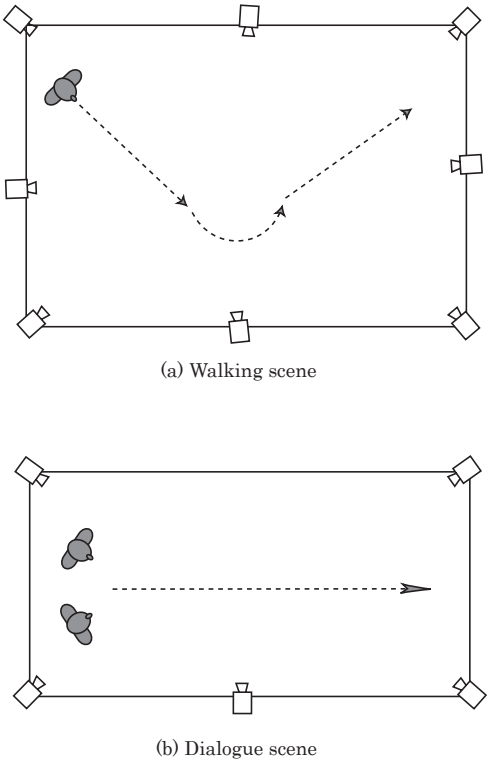


Fig. 1 Diagrams of action(s) of the target person/people and positions of real cameras: (a) walking and (b) dialogue scene. In the walking scene, a person walked around a room with eight environmental cameras. In the dialogue scene two people talked with each other and walked through a room with five environmental cameras.

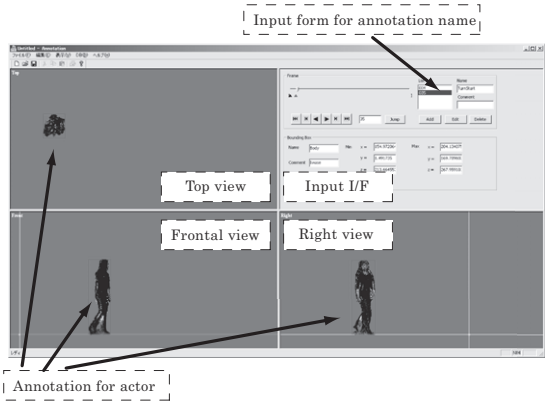


Fig. 2 Screen shot of software for annotating target. Screen of interface is divided into four regions: top, right, and front view of the 3D model, and the input form area. Users can depict the spatial region and direction of target at each frame by mouse.

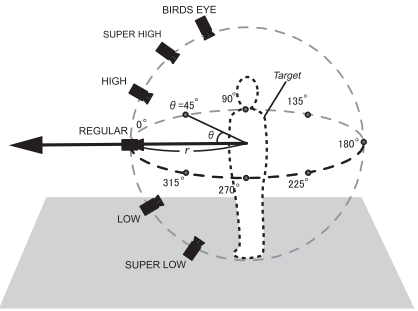


Fig. 3 Angles for deciding initial planning positions. Horizontal and vertical angles for target direction are set as the initial parameters.

3.2 Composing 3D video

To generate 3D video, several approaches have been proposed, including (1) blending two or more video images pixel by pixel [22]–[24], (2) using a fine controllable 3D model prepared manually in advance with texture from real cameras [25], and (3) reconstructing 3D models from environmental cameras [26], [27]. We adopted the third approach because it is the most standard scheme for capturing human activity indoors. We used the following processes to generate 3D video in this test.

step1 Segmenting the foreground regions, which include the object, by an intensity-based background subtraction method [28]

step2 Reconstructing 3D models of the target objects as a voxel volume with the "shape-from-silhouette" method. The 3D space including the object was modeled at a resolution of 300 * 300 * 300 on a 1 * 1 * 1 cm voxel grid.

step3 Rendering the 3D shape using a microfacet billboarding technique and combining this and the room model as an image of one frame

step4 Making a video stream by processing the above steps

3.3 Annotating targets for camera work

On one hand, since the above camera planning studies on 3D animation premise a world completely reined by the system, all parameters of objects are known by the camera planners. On the other hand, in the world of 3D video, some information needs to be prepared separately from modeling because it is unknown by the planner from their position: posture and direction parameters of models created with Shape-from-Silhouette. Therefore, we assume the following set of information called annotation:

(A1) Frame number
(A2) Parameters about existing space of the partial or whole object
(A3) Direction of object

.

In the experiment, annotations are manually given by the software shown in Fig. 2. This software, however, does not deal with the exact data of the 3D model

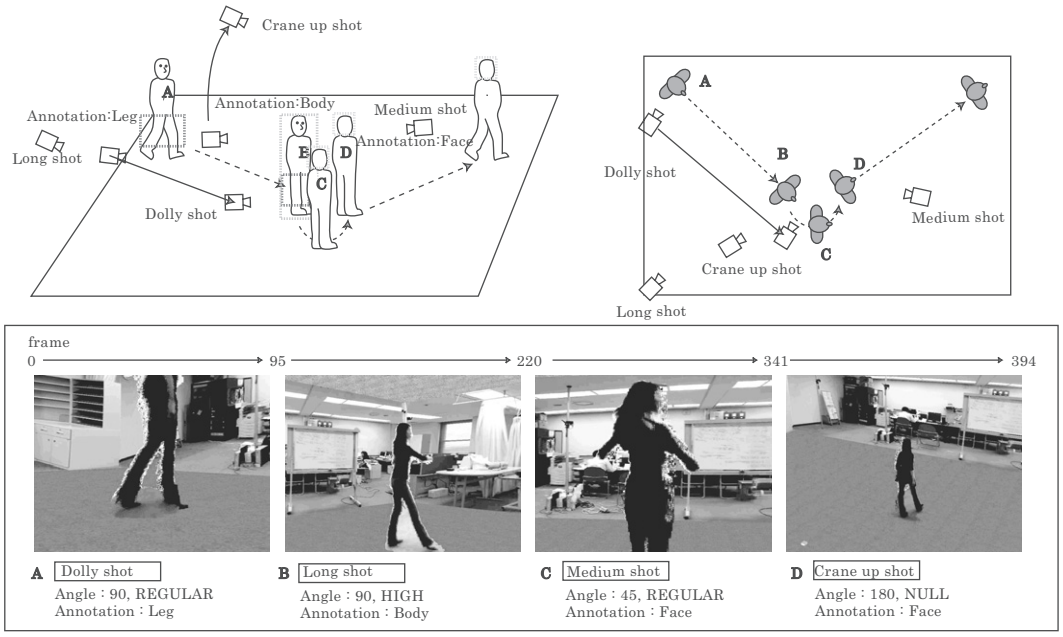


Fig. 4 Overview of camera positions and transition of camera work with annotations for walking scene (SUSPENSE_W). Upper-left portion of figure shows movement of actress with spatial annotations and camera positions of each shot from bird's eye view. Upper-right shows them from aerial view. Lower shows timing chart of shots with shot name, arguments for the shot, and screenshots. In this camera work, dolly, long, medium, and crane up shots were started at 0, 95, 220, and 341 frames, respectively.

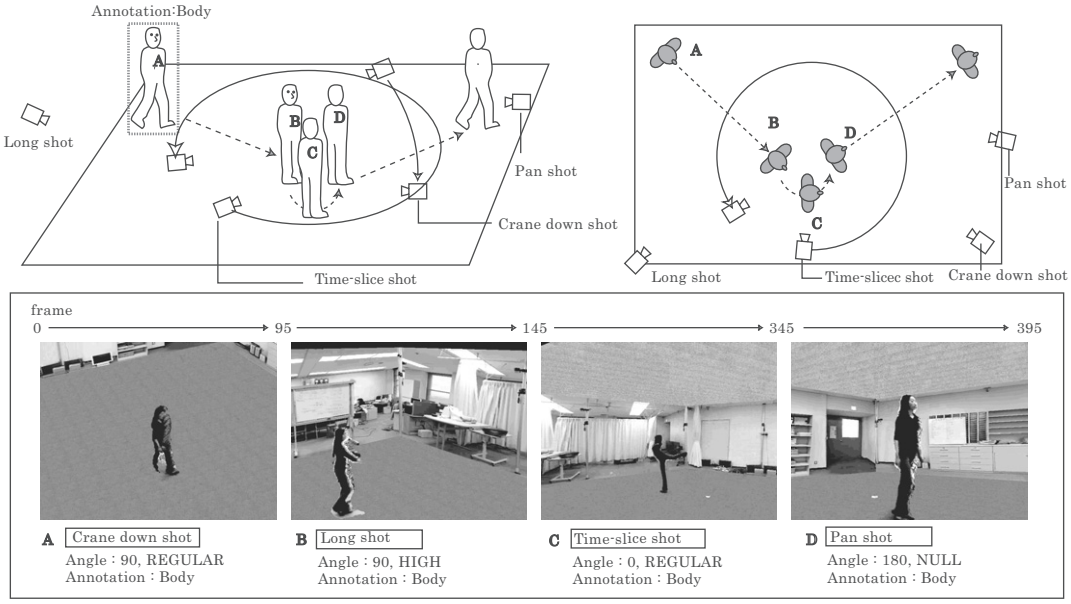


Fig. 5 Overview of camera positions and transition of camera work with annotations for walking scene (DRAMATIC_W)

as (A2) information, but the vertexes of a cuboid covering the target 3D model as approximated information. Because if exact model data are used, the number of vertexes that must be saved becomes huge. (A3) functions as a vector from the center of the cuboid. In this case, when an object for N frames is annotated, N cuboids and vectors are required. Thus, to reduce annotation efforts, the system has an automatically linear interpolation function through early frames to later ones, when annotations are specified.

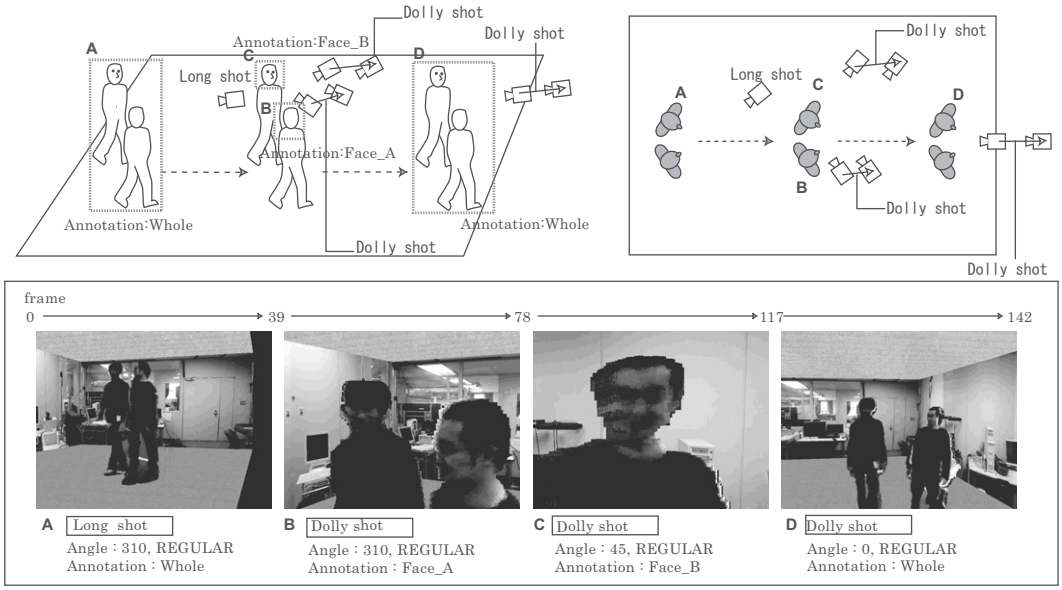


Fig. 6 Overview of camera positions and transition of camera work with annotations for dialogue scene (REVERSE_D)

Table 1 Shots supported by planning software

Action	Name of shot
Fixed	Fix shot, Bust shot Medium shot, Long shot
Moving independently	Crane up shot, Crane down shot Raise up shot, Spin around shot Time-slice shot
Moving with the target	Pan shot, Dolly shot
Zooming	Zoom out shot, Zoom in shot Whip zoom shot

3.4 Planning for cinematographic virtual camera control

Treatises have explained camera works with minimum elements called "shots." A shot is defined as a continuous strip of frames filmed with a single camera that moves consecutively or rests quietly for variable durations. In the research, no definite solution describes which shot is suitable for which scene, even though shot effectiveness and combination restrictions are discussed.

For this reason, we consulted a film expert concerning appropriate for the walking and dialogue scenes. She suggested general camera works for each scene with six kinds of shots: long, medium, dolly, crane up, crane down, pan, and time-slice (bullet-time).

Here, we consider how to apply these shots to the scenes captured as 3D video. TVML is a computer language for camera planning in 3D CG space with 3D objects. Because all related objects move simultaneously with the movement of virtual cameras in TVML's world, we cannot use TVML in this experiment. Hence, we prototyped camera planning software by Java that plans the positions and postures of virtual cameras frame by frame by referring to the annotation. The software can automatically calculate camera parameters on each frame when given an annotation, a shot name, and initial angles (see Fig. 3) as arguments [29]. The shot names listed in Table 1 are the shots supported by the software.

3.5 Generated outcome videos with camera work

Following the suggestion, we created three cinematographic 3D videos from raw videos capturing the walking and dialogue scenes with the pilot software. First, a camera work was applied to the walking scene utilizing dolly, long, medium, and crane up shots. The upper part of Fig. 4 represents the camera positions and transitions of the virtual camera. The figure also provides annotations for the actress as "Annotation:***". The lower part of the figure shows the sequence of the shots, arguments (angle information and annotation), and snapshots. Since this camera work does not shoot the bottom half of the actress' face and leaves an uncertain impression, we call this outcome SUSPENSE_W.

Second, we shot more camera working footage of the same walking scene with different shots. The schematic depiction of annotation and camera movement is shown in Fig. 5. Since this camera work adopts a dynamic shot often called time-slice, we call this output DRAMATIC_W.

Finally, camera working for a dialogue scene was conducted. It appears in Fig. 6. When shooting dialogue scenes with two people, cameras generally face each person and are positioned on opposite sides of each other and are then switched alternately. We adopted

this technique and call this video REVERSE_D.

In these outcome videos, defects were caused by errors of generating 3D videos. For example, arms are cut in the third shot of Fig. 4, and the man's face is collapsed in the third shot of Fig. 6. These errors, which are mainly caused by segmentation problems [30], [31] and the limitations of Shape-from-Silhouette [32], are difficult to completely overcome.

4. Evaluations

4.1 Camera parameter errors caused by planning

To examine the accuracy of camera parameters generated by planning software, we compared camera parameters SUSPENSE_W and REVERSE_D with the ground truth data. The ground truth data were made by the film expert we consulted about camera work for the walking and dialogue scenes. To input the ground truth data, she used the original software that enables us to place and face the virtual camera in 3D space by mouse operation. The software reads the 3D video data of an arbitrary frame and displays it. In the following, videos made from each ground truth datum are called SUSPENSE_W_T and REVERSE_D_T.

Figure 7 represents the differences of camera parameters between SUSPENSE_W and SUSPENSE_W_T. (a) in Fig. 7 provides a plot chart of the virtual camera positions on SUSPENSE_W and SUSPENSE_W_T. Although both trajectories are relatively close, an arc on the left side indicating SUSPENSE_W_T is widely discoursed from SUSPENSE_W. This part is on the track of the crane up shot, and the difference is caused by the camera positions of SUSPENSE_W that are affected by their target positions. On the contrary, the camera of SUSPENSE_W_T independently rose toward the ceiling. (b) is the line plot of the distances between the camera positions of SUSPENSE_W and SUSPENSE_W_T. As shown in the chart, errors remained constant from frames 95 to 220 and 225 to 340; however, they increased during the crane up shot (after the 341th frame). (c) shows the size ratio between screen height and target annotation. The height of the annotations is calculated as the maximum size of the vertical aspect viewed from the virtual camera. (d) shows *cos* value of the angle formed by three points: the camera position of SUSPENSE_W, the center point of the target annotation, and the camera position of SUSPENSE_W_T. Although there are three substantial spikes in (b) and (d) at approximately frames 95, 220, and 340, this is simply a consequence of the timing differences of changing the shots. Fig. 8 represents the differences of camera parameters between REVERSE_D and REVERSE_D_T. Errors are more constant than SUSPENSE_W because REVERSE_D was composed with undynamic shots.

These results indicate that some errors are unavoidable when planning camera positions. Error increases, especially when using transition shots.

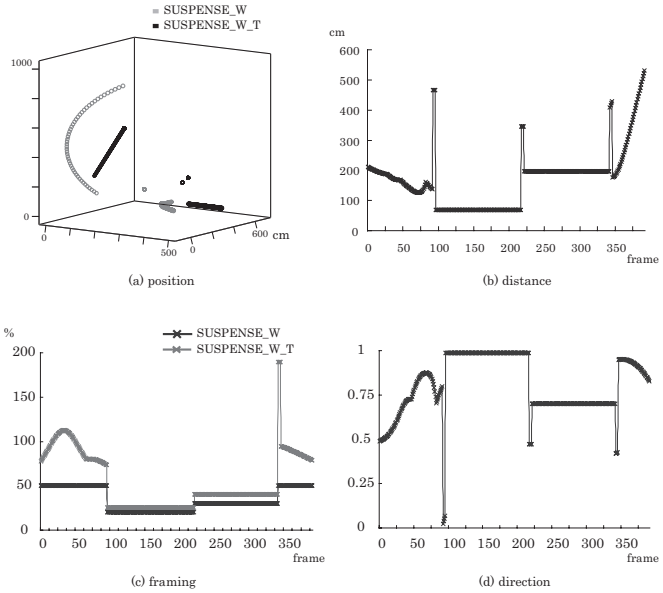


Fig. 7 Comparison of camera parameters between SUSPENSE_W and SUSPENSE_W_T. (a) is a plot chart showing trajectories of each virtual camera; (b) is a line plot of distance between them; (c) shows size ratio between screen height and target appearance in each video; (d) shows *cos* value of angle formed by SUSPENSE_W, the target, and SUSPENSE_W_T.

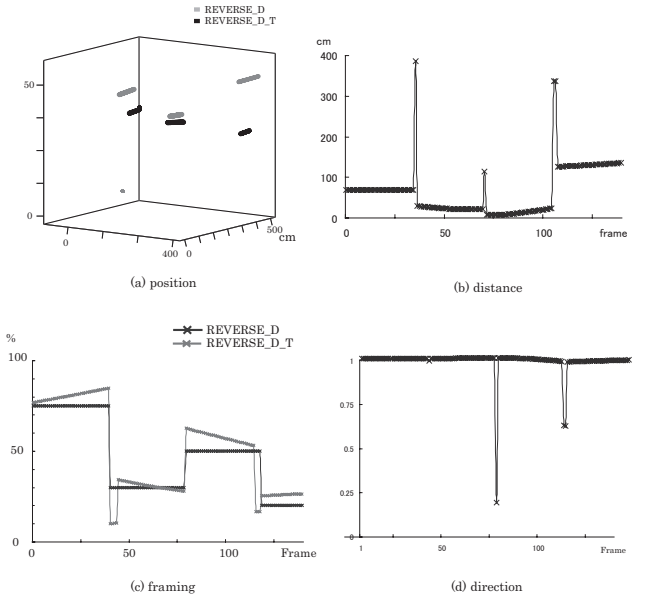


Fig. 8 Comparison of camera parameters between REVERSE_D and REVERSE_D_T

4.2 Impression analysis

In this section, we verify whether the errors indicated in the previous section are significantly noticeable by viewers with a psychological evaluation called the Se-

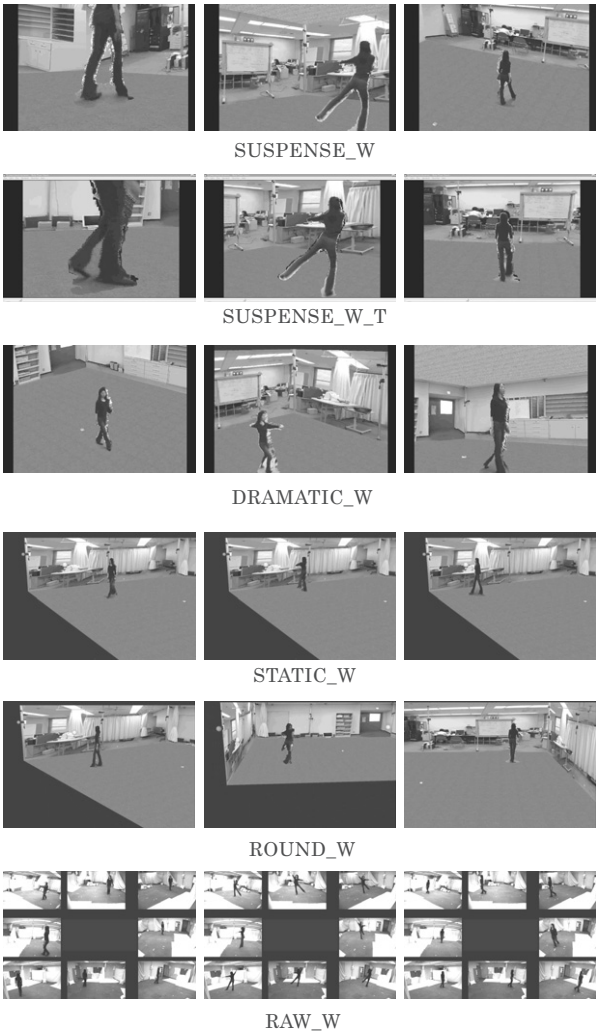


Fig. 9 Screen shots of each video in walking scene

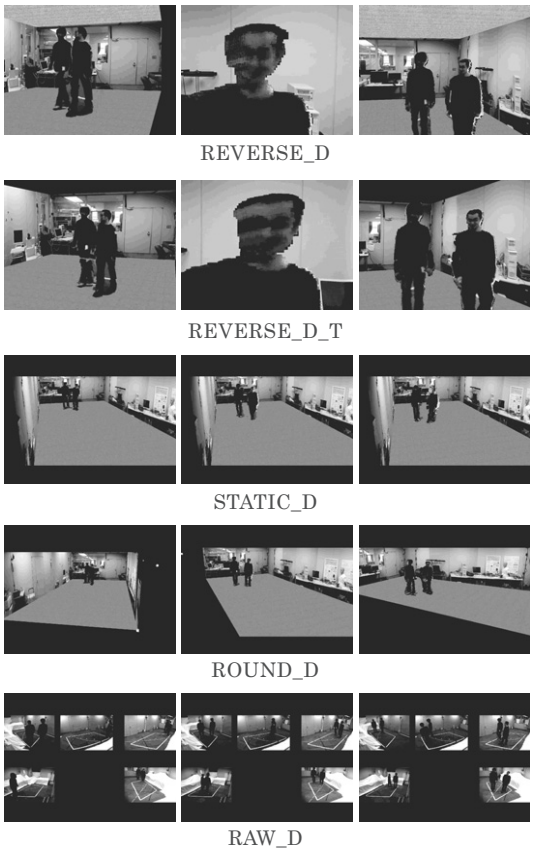


Fig. 10 Screen shots of each video in dialogue scene

mantic Differential (SD) test. This test is widely used for measuring people's impressions of certain objects with bipolar adjective pairs [33]. If the errors are not critical, impressions between cinematographic 3D and ground truth videos will not be measured as different.

4.2.1 Experimental procedure

Aside from SUSPENSE_W, SUSPENSE_E, DRAMATIC_W, REVERSE_D, and REVERSE_E, we prepared the following videos to compare, and these were also presented to the subjects. All of these videos are shown in Figs. 9 and 10. Each picture of the figures shows the screen shot of the videos, and the left is the beginning and the right is the end.

· **Raw video**
Here are the raw video streams captured by six to seven real environmental cameras. Each stream is synchronized and displayed in tiled order into one video. Hereafter, we call the walking scene video RAW_W and the dialogue scene RAW_D.

· **View from a virtual camera without moving**
These are the video streams from the viewpoint of the virtual camera. The virtual camera positioned at the side wall and directed at the actors does not move. The camera height was set about head level. Hereafter, we call the video of the walking scene STATIC_W and the dialogue scene STATIC_D.

· **View from a virtual camera controlled by human**
These are the video streams from the viewpoint of the virtual camera. Its position was directed at the actors and the trajectory of the position arcs around the object(s) without being smoothly controlled by the author. The camera height was set about head level. This is assumed to be a normal user's view of present 3D video systems. Hereafter, we call the walking scene video ROUND_W and the dialogue scene ROUND_D.

The bipolar adjective pairs for the SD test were determined from preliminary hearing surveys given to people who viewed all videos, except the subjects. We prepared sheets describing the pairs on the right column of Table 2, and a 7 scale was set between each pair. Some items were negatively worded and then reversed to avoid social desirability effects. To simplify discussion, each pair is named on the left column of Ta-

Table 2 Adjective pairs for semantic differential test. In the experiment, only Japanese adjective words were showed to subjects. The words in column "Name" are indexes to explain this paper's results.

Name	Adjective pair (Japanese)	
A_CLUMS	smooth (滑らかな)	clumsy (ぎこちない)
A_LIGHT	serious (重厚な)	light (軽快な)
A_RELAX	speedy (スピーディな)	relaxed (ゆったりとした)
A_BRIGH	dim (暗い)	bright (明るい)
A_WEEK	strong (力強い)	weak (弱弱しい)
A_RELIE	tense (緊迫した)	relieved (安心した)
A_QUIET	dynamic (ダイナミックな)	quiet (おとなしい)
A_INFOR	formal (堅苦しい)	informal (うちとけた)
A_FAST	slow (遅い)	fast (早い)
A_DRAMA	monotonous (単調な)	dramatic (ドラマチックな)
A_INTER	boring (退屈な)	interesting (興味深い)
A_SIMPL	complex (複雑な)	simple (単純な)
A_COMPR	incomprehensible (判りにくい)	comprehensible (判りやすい)
A_MILD	drastic (激しい)	mild (穏やかな)
A_HUMAN	mechanical (機械的な)	humane (人間的な)
A_CHEER	gloomy (陰気な)	cheerful (陽気な)
A_SUBST	empty (空虚な)	substantial (充実した)
A_UNCLE	clear (明確な)	unclear (不明瞭な)
A_USUAL	impressive (衝撃的な)	usual (平凡な)
A_CALM	rough (荒々しい)	calm (落ち着いた)
A_SLUGG	swift (すばやい)	sluggish (ゆっくりとした)

ble 2. The subjects included 29 bachelor's or master's students. Before the experiments, they were informed about the SD test sheet with an instruction form. After that, we showed each video to subjects one by one. The order of the videos was randomly changed for each subject to counterbalance.

After the experiment, we compared all videos on each adjective pair by the Bonferroni t-test of a one-way ANOVA.

4.2.2 Results for planning error

Here, we will discuss the relation between the impressions and the precisions of camera parameters. Figs. 11 and 12 provide a mean value of SUSPENSE_W and SUSPENSE_W_T, and REVERSE_D and REVERSE_D_T, respectively. In these figures, obviously all adjective pairs obtained have approximately the same scores on both scenes; there was no significant difference among them by multiple comparisons. Therefore camera parameter errors do not affect viewer impressions. This result indicates that the errors caused by planning and annotation were not critical problems of impression disparities.

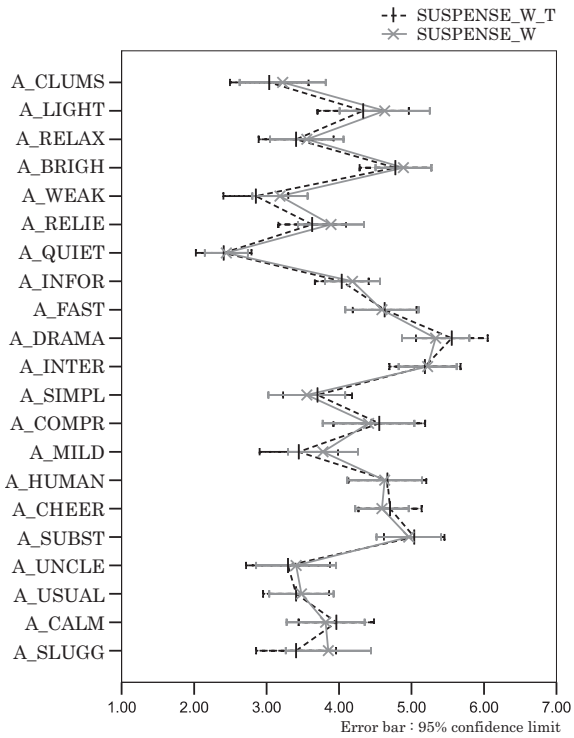


Fig. 11 Mean values of each adjective on SUSPENSE_W and SUSPENSE_W_T

4.3 Impressions

In this part, we will discuss whether the cinematographic 3D videos affected the subjects with intentional direction.

As indicated by the scores plotted in Fig. 13, such negative evaluations as monotonous (at A_DRAMA) and boring (at A_INTER) were gauged. The time necessary to bore viewers seems quite short since both walking and dialogue scenes are less than 20 seconds and the actors are in motion; nevertheless, sustaining viewer attention without any virtual camera movement is difficult.

Focusing on ROUND_W and ROUND_D, for some adjectives there are significant differences between them and the other videos. For example, for adjective A_CLUMS, significant differences to all cinematographic 3D videos were found as well as for A_HUMAN, SUSPENSE_W, and REVERSE_D. These adjectives are related to naturalness of camera control,

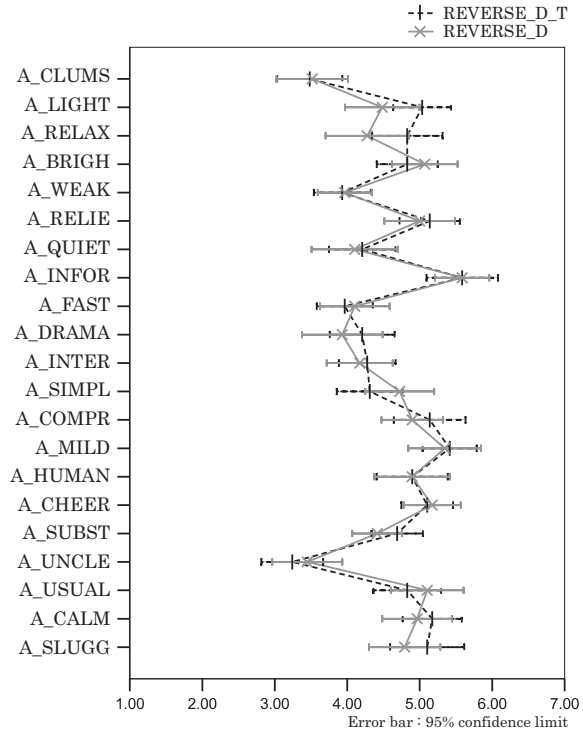


Fig. 12 Mean values of each adjective on REVERSE_D and REVERSE_D_T

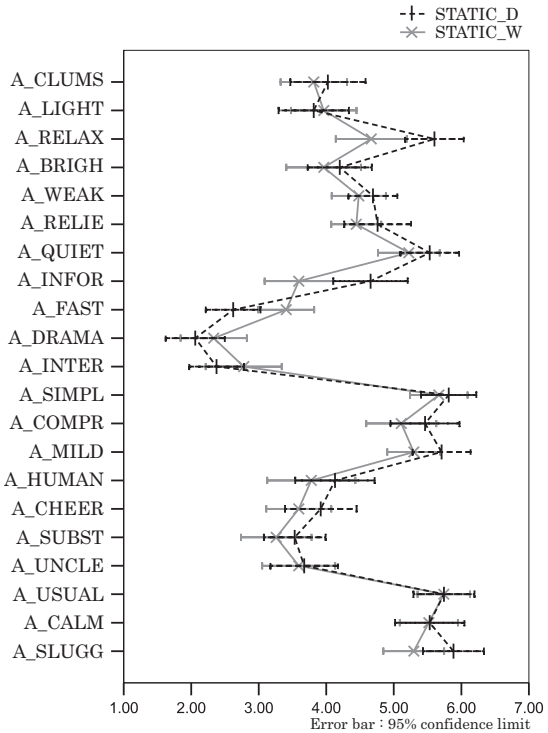


Fig. 13 Mean values of each adjective pair on STATIC_D and STATIC_W

so clumsy control of the virtual camera on ROUND_W and ROUND_D differed from the steady camera control usually seen on TV or movies; ROUND_W and ROUND_D produced uncomfortable feelings in viewers.

Comparing the mean value of the walking and dialogue scenes in Figs. 11 and 12, they obviously form different shapes, and so each camera work created different impressions in viewers. However, we cannot deny the possibility that such differences may be attributed to different impression of the contents themselves.

The results of multiple comparisons between SUSPENSE_W and DRAMATIC_W show significant differences of A_RELAX, A_WEAK, A_QUIET, A_MILD, A_USUAL, and A_CALM(p=.012, .023, .028, .006, .012 and .012), which are associated with characteristics of shots. Since SUSPENSE_W and DRAMATIC_W shot the same scene, these differences simply reflect the shots of camera work. These results indicate that it is possible to satisfactorily direct camera works on 3D video and also camera work with planning and annotation works.

Factor analysis of the SD test answers for each scene was extracted from four factors. Each factor loading is shown in Tables 3 and 4.

For the walking scene (Table 3), the proposed 4-factors solution explains 64.2% of the total variance and received an acceptable value in the Kaiser-Meyer-Olkin measure of sampling adequacy (.893). The data were factor analyzed using the principal factor method and a promax (oblique) rotation for 4 factors. The proposed factors are (1) calm, (2) vigor, (3) comprehension, and (4) speed.

Figure 14 shows the mean values of the factors that include "calm," "vigor," and "comprehension." SUSPENSE_W, SUSPENSE_W_T, and DRAMATIC_W, which are the cinematographic 3D videos, are clustered close, and their "comprehension" factor scores are especially high. This result indicates that camera works extends the understandability of 3D video.

As mentioned above, the actress was shot by first half of scene in SUSPENSE_W to make the video shady and suspenseful. Since DRAMATIC_W scores higher than SUSPENSE_W and SUSPENSE_W_T in the "comprehension" aspect, this intention was achieved.

In the dialogue scene, the proposed 4-factors solution explains 64.0% of the total variance and received an acceptable value from the Kaiser-Meyer-Olkin measure of sampling adequacy (.874). The data were factor analyzed using the principal factor method and a promax (oblique) rotation for 4 factors that included (1) calm, (2) vigor, (3) attraction, and (4) comprehension.

Figure 15 shows the mean values of "calm," "vigor," and "comprehension" factors. Although REVERSE_D and REVERSE_D_T are also clustered and their "comprehension" scores are especially high, the STATIC_D and ROUND_D scores are also relatively high compared with the walking scene because we

Table 3 Factor loading matrix for walking scene

	Factors			
	1	2	3	4
A_USUAL	0.829	-0.423	-0.225	0.190
A_QUIET	0.826	-0.408	-0.277	0.383
A_DRAMA	-0.807	0.664	0.332	-0.136
A_MILD	0.786	-0.163	-0.150	0.530
A_WEAK	0.741	-0.459	-0.528	0.286
A_CALM	0.703	-0.063	0.018	0.411
A_SIMPL	0.696	-0.195	0.149	0.323
A_SLUGG	0.695	-0.332	-0.239	0.664
A_SUBST	-0.693	0.692	0.555	-0.113
A_RELIE	0.596	0.147	-0.132	0.254
A_BRIGH	-0.437	0.790	0.532	-0.166
A_CHEER	-0.542	0.780	0.543	-0.131
A_INTER	-0.738	0.746	0.492	-0.068
A_INFOR	-0.155	0.732	0.246	-0.065
A_HUMAN	-0.219	0.663	0.423	0.032
A_CLUMS	0.268	-0.618	-0.396	0.081
A_LIGHT	-0.118	0.597	0.228	-0.321
A_UNCLE	0.283	-0.521	-0.778	0.056
A_COMPR	-0.035	0.388	0.765	0.050
A_RELAX	0.555	-0.275	-0.201	0.746
A_FAST	-0.591	0.414	0.227	-0.646
Cumulative percent of variance explained	40.0%	55.5%	60.1%	64.2%

Table 4 Factor loading matrix for dialogue scene

	Factors			
	1	2	3	4
A_QUIET	0.855	0.293	-0.718	0.333
A_MILD	0.842	0.605	-0.484	0.504
A_SLUGG	0.837	0.290	-0.478	0.357
A_USUAL	0.754	0.476	-0.712	0.491
A_CALM	0.722	0.572	-0.489	0.536
A_RELAX	0.714	0.132	-0.459	0.067
A_FAST	-0.709	-0.069	0.486	-0.027
A_SIMPL	0.642	0.370	-0.596	0.576
A_WEAK	0.522	-0.022	-0.484	-0.103
A_INFOR	0.394	0.808	-0.039	0.600
A_RELIE	0.555	0.803	-0.281	0.493
A_HUMAN	0.308	0.787	0.118	0.640
A_CHEER	0.081	0.782	0.193	0.533
A_BRIGH	0.097	0.745	0.251	0.469
A_LIGHT	0.085	0.696	0.064	0.350
A_CLUMS	-0.199	-0.515	0.031	-0.275
A_INTER	-0.567	0.056	0.898	-0.089
A_DRAMA	-0.589	-0.037	0.854	-0.030
A_SUBST	-0.223	0.298	0.588	0.171
A_COMPR	0.292	0.538	-0.057	0.874
A_UNCLE	0.025	-0.398	-0.112	-0.677
Cumulative percent of variance explained	33.9%	55.5%	60.5%	64.0%

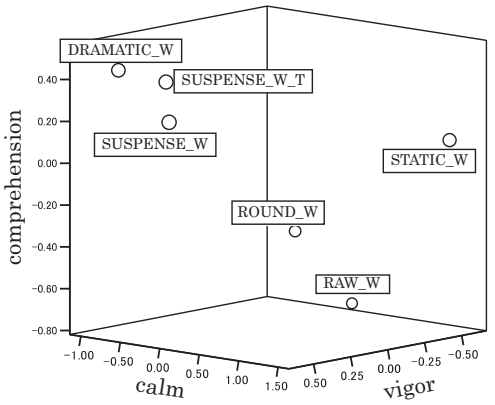


Fig. 14 Plot of mean values of factor scores in walking scene

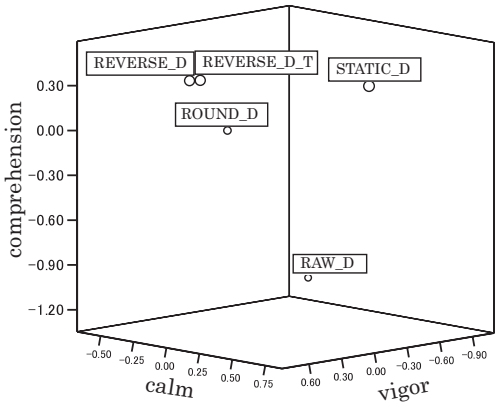


Fig. 15 Plot of mean values of factor scores in dialogue scene

assumed that dynamic shots were not used in REVERSE_D and they did not make a major difference among REVERSE_D, STATIC_D, and ROUND_D.

Finally, let us confirm that the above results were not mainly caused by attractivity differences of the contents of each scene but camera working differences. Fig. 13 provides the mean values of STATIC_W and STATIC_D whose camera work was identical but capturing different scenes. As shown in the figure, each line chart almost formed the same shape, and there are just three significant differences by paired t-test: A_RELAX(p=.018), A_INFOR(p=.006), and A_FAST(p=.017). Thus, the attractivity of each scene hardly affected impressions.

These results indicate that exploiting the proposed virtual camera control, annotation, and planning technique allowed us to realize camera working direction on 3D video.

5. Conclusion and limitations

This paper described the effectiveness of camera working in 3D video with a psychological test. For the test, we prepared three outcome videos that applied camera works to 3D videos and exploited simple techniques for making 3D video and planning camera parameters, and then we presented these videos and several competitive videos to subjects. From the test results, we concluded that camera work that exploits a simple annotation and planning technique is applicable to 3D video with enough quality for viewers, and the impression of cinematographic 3D video reflects the intention of the director composing the camera works.

It is unclear whether the camera works mentioned in this paper are applicable to scenes on which previous practical studies with 3D video focused, such as sports, because ordinary camera works explained in the treatises are mainly designed as applications to such normal scenes as daily life. Precedent studies did not mention the camera works for sports domains. Therefore, when applying camera work to these scenes, we must consider special camera works suitable for each sport. In this paper, we manually solved changing shots using the software, but the difficulty of using this method for sports scenes remains because the length of capturing video in sports scenes requires too many changing shots. In future work, we will investigate these issues by autonomous annotation using image recognition and planning techniques to select suitable shots.

Acknowledgments

We would like to thank Mika Satomi for her help with the making basic camera works. This research was supported by the National Institute of Information and Communications Technology.

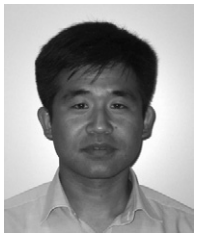
References

[1] D. Arijon, Grammar of the Film Language, Silman-James Press, 1991.
[2] S.D. Katz, Film Directing Shot by Shot: Visualizing from Concept to Screen, Michael Wiese Film Productions, 1991.
[3] S.D. Katz, Cinematic Motion: Film Directing : A Workshop for Staging Scenes, Michael Wiese Film Productions, 1992.
[4] N. Halper and P. Olivier, "Camplan: A camera planning agent," In Smart Graphics. Papers from the 2000 AAAI Spring Symposium, pp.92–100, 2000.
[5] M.Douke, M.Hayashi, and E.Makino, "A study of automatic program production using tvml," Short Papers and Demos, Eurographics '99, pp.42–45, 1999.
[6] B. Tomlinson, B. Blumberg, and D. Nain, "Expressive autonomous cinematography for interactive virtual environments," AGENTS '00: Proceedings of the fourth international conference on Autonomous agents, pp.317–324, ACM Press, 2000.
[7] W.H. Bares and J.C. Lester, "Cinematographic user models for automated realtime camera control in dynamic 3D environments," the Sixth International Conference on User Modeling, pp.215–226, 1997.
[8] T. Kanade, P. Rander, and P.J. Narayanan, "Virtualized reality: Constructing virtual worlds from real scenes," IEEE MultiMedia, vol.4, no.1, pp.34–47, 1997.
[9] I. Kitahara and Y. Ohta, "Scalable 3d representation for 3d video in a large-scale space," PRESENCE, vol.13, pp.164–177, 2004.
[10] N. Inamoto and H. Saito, "Free viewpoint video synthesis and presentation from multiple sporting videos," IEEE International Conference on Multimedia & Expo (ICME2005), 2005.
[11] K. Kimura and H. Saito, "Video synthesis at tennis player viewpoint from multiple view videos," IEEE VR2005, pp.281–282, 2005.
[12] S. Prince, A.D. Cheok, F. Farbiz, T. Williamson, N. Johnson, M. Billinghurst, and H. Kato, "3-d live: real time interaction for mixed reality," Proceedings of the 2002 ACM conference on Computer supported cooperative work, pp.16–20, ACM, 2002.
[13] K. Tomiyama, M. Katayama, Y. Orihara, and Y. Iwadate, "Arbitrary viewpoint images for performances of japanese traditional art," 2nd Conference on Visual Media Production CVMP 2005, pp.68–75, 2005.
[14] S. Nobuhara and T. Matsuyama, "Evolution of 3d video technology," 4th International Symposium on Computing and Multimedia Studies, pp.72–79, 2006.
[15] P. Kauff and O. Schreer, "An immersive 3d video-conferencing system using shared virtual team user environments," CVE '02: Proceedings of the 4th international conference on Collaborative virtual environments, pp.105–112, ACM Press, 2002.
[16] P. Eisert, "Virtual conferencing using 3d model-assisted image-based rendering," 2nd Conference on Visual Media Production CVMP 2005, pp.183–191, 2005.
[17] J.M. Gauthier, Building Interactive Worlds in 3D, Elsevier, 2005.
[18] A. Inoue, H. Shigeno, K. Okada, and Y. Matsushita, "Introducing grammar of the film language into automatic shooting for face-to-face meetings," Proceedings of the IEEE/IPSJ Symposium on Applications and the Internet (SAINT2004), pp.277–280, 2004.
[19] P. Doubek, I. Geys, T. Svoboda, and L.V. Gool, "Cinematographic rules applied to a camera network," Omnivis2004, The fifth Workshop on Omnidirectional Vision, Camera Networks and Non-Classical Cameras, pp.17–29, May 2004.
[20] R. Tenmoku, R. Ichikari, F. Shibata, A. Kimura, and H. Tamura, "Design and prototype implementation of mr pre-visualization workflow," Int. Workshop on Mixed Reality Technology for Filmmaking, 2006.
[21] I. Kitahara, R. Sakamoto, M. Satomi, K. Tanaka, and K. Kogure, "Cinematized reality: Cinematographic camera controlling 3d free-viewpoint video," 2nd IEE European Conference on Visual Media Production (CVMP2005), pp.154–161, 2005.
[22] L. McMillan and G. Bishop, "Plenoptic modeling: An image-based rendering system," Computer Graphics, vol.29, pp.39–46, 1995.
[23] S.J. Gortler, R. Grzeszczuk, R. Szeliski, and M.F. Cohen, "The lumigraph," Computer Graphics, vol.30, pp.43–54, 1996.
[24] M. Levoy and P. Hanrahan, "Light field rendering," Computer Graphics, vol.30, pp.31–42, 1996.
[25] J. Carranza, C. Theobalt, M.A. Magnor, and H.P. Seidel, "Free-viewpoint video of human actors," ACM Trans. Graph., vol.22, no.3, pp.569–577, 2003.
[26] A. Laurentini, "The visual hull concept for silhouette-based image understanding," IEEE Trans. Pattern Analisys ans Machines Intelligence (PAMI), vol.16, no.2, pp.150–162, February 1994.
[27] W. Matusik, C. Buehler, R. Raskar, S.J. Gortler, and L. McMillan, "Image-based visual hulls," SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques, pp.369–374, ACM Press/Addison-Wesley Publishing Co., 2000.
[28] H. Kim, I. Kitahara, K. Kogure, N. Hagita, and K. Sohn, "Personal satellite virtual camera," PCM 2004, pp.87–94, 2004.
[29] H. Kim, R. Sakamoto, I. Kitahara, and K. Kogure, "Cinematized reality: Cinematographic 3d video system for daily life using multiple outer/inner cameras," IEEE Workshop on Three-Dimensional Cinematography (3DCINE'06), 2006.
[30] A.M. Elgammal, D. Harwood, and L.S. Davis, "Non-

parametric model for background subtraction," ECCV '00: Proceedings of the 6th European Conference on Computer Vision-Part II, pp.751–767, Springer-Verlag, 2000.
[31] E.D. Gelasca, T. Ebrahimi, M. Karaman, and T. Sikora, "A framework for evaluating video object segmentation algorithms," CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop, p.198, IEEE Computer Society, 2006.
[32] C. Buehler, W. Matusik, L. McMillan, and S. Gortler, Creating and Rendering Image-Based Visual Hulls, Massachusetts Institute of Technology, 1999.
[33] C.E. Osgood, G. Suci, and P. Tannenbaum, The Measurement of Meaning, University of Illinois Press, 1967.

Ryuuki Sakamoto received MS and PhD degrees in Knowledge Science at the Japan Advanced Institute of Science and Technology (JAIST). He is currently a researcher in the Knowledge Science Lab (KSL) at the Advanced Telecommunications Research Institute International (ATR), Japan. His research interests include information visualization and man-machine interface and CSCW. He is a member of ACM, the Information Processing Society of Japan, and the Japan Creativity Society.

Itaru Kitahara received M.E. degrees in Science Engineering from the University of Tsukuba, Japan in 1996. In 1996 he joined the Sharp Corporation. From 2000 to 2003, he was a research associate of the Center for Tsukuba Advanced Research Alliance, University of Tsukuba. He received a PhD degree in Systems and Information Engineering from the University of Tsukuba in 2003. Since 2003, he has been a researcher at ATR. From 2005, he has been an assistant professor at University of Tsukuba. He was awarded the IEEE VR (Conference on Virtual Reality) Honorable Mention Award in 2003. His research interests include computer vision, mixed reality and intelligent video media. He is a member of IEEE.



Megumu Tsuchikawa received a B.E. degree in Mechanical Engineering from Waseda University, Japan. He is a manager of the Intellectual Property Center at NTT, Japan. His research interests include image processing and image understanding.

Kaoru Tanaka received a MS degree in Knowledge Science at the Japan Advanced Institute of Science and Technology (JAIST). Since 2005, he has been a doctoral student at the same institute. His research interests include knowledge sharing systems and human-computer interaction. He is a member of the Information Processing Society of Japan (IPSJ).

Tomoji Toriyama received a PhD from Toyama Prefectural University, Japan. He is currently a group leader in the Knowledge Science Laboratories at ATR, Japan. His research interests include Large-Scale Integration (LSI) circuit architecture design methodology, human interfaces, and image processing.

Kiyoshi Kogure is the director of the ATR Knowledge Science Laboratories. His research interests include intelligent environments, intelligent agents, and natural language processing. He received his PhD in engineering from Keio University. He is a member of the Information Processing Society of Japan, the Japanese Society for Artificial Intelligence, the Association for Natural Language Processing, the Japanese Cognitive Science Society, and the Acoustical Society of Japan.