Proc. of IEEE Conference on Computer Vision and Pattern Recognition, vol.2, II-131–II-137, Dec. 2001.

Production of Video Images by Computer Controlled Cameras and Its Application to TV Conference System

Masaki Onishi Takehiko Kagebayashi Kunio Fukunaga
Graduate School of Engineering, Osaka Prefecture University
1-1, Gakuen-cho, Sakai, Osaka, 599-8531 Japan

Abstract

TV conference systems have been widely used recently. A participant of each site proceeds with a TV conference using video image on a screen and voice of a partner site. In this case, a fixed video camera shoots a scene of a site in general. The video image taken by a fixed camera, however, is lacking in changes. Also the fixed camera does not take shots efficiency that the participants of a partner site pay attention. As one of the candidates to avoid these defects, there is a method that the computer-controlled cameras shoots the scene. In this paper, we propose an algorithm of shooting the best shot by computer-controlled cameras. The shooting algorithm is mainly decided by estimating an area of the image with high degree of attention, which is given by not only visual information but also auditory information. By using an experimental system, we confirm the effectiveness of our approach by examining a questionnaire from the participants of TV conferences.

1. Introduction

TV conference systems have been widely used as communication network develops and a high speed and large capacity communication is put to practical use. In a TV conference, a participant of each site advances the conference the image on a screen and voice of a partner site. The video image taken by a fixed camera, however, is lacking in changes. Also the fixed camera does not always take shots that the participants of a partner site pay attention.

As one of the candidates to avoid these defects, there is a method of producing a video image taken by a cameraman. This approach requires much technical manpower, while the image is taken so as to be natural for the participants. As another approach, there is a method that shoots an object scene by a computer-controlled video camera. In this approach, it is important to design camerawork how to take an image of the participants in a site.

A representative method composes a TV conference system by switching a shot of the scene based on the knowledge of camerawork (knowledge of an expert cameraman) for a TV discussion program, and the image taken by this type of the camerawork is appraised high comparing with the image taken by a fixed camera.

Many articles have been reported on taking a scene of a TV conference [1] and a distance learning system [2, 3] by computer-controlled cameras. One approach is a system using auditory information, and the other is a system using visual information. An example of the visual approaches is to examine the distribution of spatiotemporal information over the video image, and shoot and focus an area of producing large spatiotemporal information [2]. This type of camerawork assumes that the area to which human being pays attention is the range with large spatiotemporal information, and has succeeded in the camerawork on the case of a distant learning system. There is, however, no system using both the auditory and visual information.

In this paper, we propose an approach of camerawork that takes not only visual attention of spatiotemporal information but auditory attention of spatial distribution of human voice into consideration, and confirm the effectiveness of our method based on questionnaire concerning to the video image from the participants of the TV conference.

2. Camera control for video image production

In this section, we discuss a system of producing a TV conference image taken by fully computer-controlled cameras (shooting cameras). In the first, the system calculates the distribution of spatiotemporal information for every image taken by the monitor camera to estimate a range where the participants move actively. On the other hand, the system examines the distribution of location of sound source of human voice obtained by the sound localization at every unit time. In the next, the system decides a noticeable area (area to which human pays attention) of the scene on the basis of the distribution considering both of the auditory information (distribution of sound source) and visual informa-

tion (distribution of spatiotemporal information). Based on the noticeable area, the system selects the noticeable participant(s). The system tries to produce the best TV conference video image keeping the suitable shot using the computer-controlled video camera.

3. Extraction of auditory information by sound localization

3.1. 3-Dimensional sound localization

The sound localization is to estimate the position of the sound source by observing and analyzing arrived sound wave. There are three representative methods. One is a beam former method [4], the second is a method based on time delay between the microphone signals and the last one is a method of eigen space of the signal [5]. These three approaches mainly focus on estimation of the direction of the sound source, and do not take the estimation of the 3-dimensional position of sound source into consideration. In this section, we explain how to estimate the 3-dimensional position in our approach.

The accuracy of sound localization depends on the distance between the microphones. The accuracy is getting high as the distance between the microphones increase. While the longer distance placement of the microphone set expands the measurable range of the sound localization, the wave frequency of the (voice) sound is getting lower when we estimate uniquely the time delay between the arrived sounds. In this system, we place the microphone set as shown in Figure 1 considering the physical scale of the system and ease of mobility. Here the parameter M_d denotes the distance between the microphones and is set to 45 cm in our system. The sampling frequency is set to 25 kHz, and the sampling time becomes 40 μs. The system samples four microphone inputs simultaneously, and represents these four signals by g_1, g_2, g_3, and g_4 respectively.

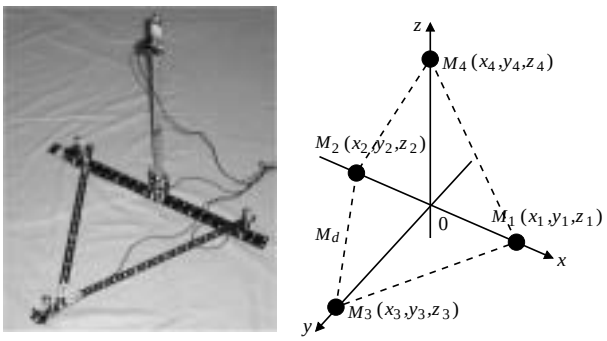


Figure 1. Microphone layout.

In the next, the system examines whether a voice signal appears or not in the input signals. In the case when the voice signal appears in the input signals, the system carries out the sound localization every 20 ms based on the time delay between the four microphone signals. The system always prepares a histogram of the positions of the sound source obtained by sound localization in the last several seconds.

3.2. Sound localization based on the time delay of microphone signals

There is a strong correlation among the input signals, when the voice signal is received by the microphones placed as shown in Figure 1. Mutual correlation between two signals $g_i(t)$ and $g_j(t)$ is defined by the following equation,

$$\phi(t_m) = \sum_{t=1}^{N_\phi} g_i(t) g_j(t + t_m). \quad (1)$$

The system examines the maximum value of $\phi(t_m)$ while varying the parameter t_m from 0 to $t_{\max}$. The value t_m which maximizes $\phi(t_m)$ is the delay time between the signals $g_i(t)$ and $g_j(t)$.

As we suppose the placement of four microphones is given by Figure 1, we can easily obtain three independent time delays from the microphone set. It is possible to estimate the 3-dimensional position of the sound source by using these three independent time delays. In order to calculate the position of the sound sources in a real time, the system firstly divides the 3-dimensional space into voxels, then calculates the delay time among the four microphone signals for each of all voxels, and composes a table of delay time for each voxel in advance. Immediately after the system receives the microphone signals and calculates the delay time among the four signals, the system estimates a voxel of the sound source by comparing the set of delay time with the set of delay time for each voxel by use of the table of the delay time.

4. Extraction of visual information using spatiotemporal information [2]

4.1. Probability density function of color parameter

We define an appearance probability of a color parameter of a given pixel $\boldsymbol{P}_i = [\, x_i \; y_i \,]^T$ in the form of a probability density function based on distribution of color parameters of the neighborhood pixels, and suppose that the probability function of color parameter $\boldsymbol{X}_i = [\, R_i \, G_i \, B_i \,]^T$ of the pixel

$\boldsymbol{P}_i$ follows a 3-dimensional normal distribution function.

$$p(\boldsymbol{X}_i) = \frac{1}{(2\pi)^{\frac{3}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\boldsymbol{X}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu})\}. \tag{2}$$

Here, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ denote a mean value and a covariance matrix of the color parameter $\boldsymbol{X}_i$ at the neighborhood pixels and are given by following equations.

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{X}_k, \tag{3}$$

$$\boldsymbol{\Sigma} = \frac{1}{n}\sum_{k=1}^{n} (\boldsymbol{X}_k - \boldsymbol{\mu})(\boldsymbol{X}_k - \boldsymbol{\mu})^T. \tag{4}$$

As the probability $p(\boldsymbol{X}_i)$ is defined by the distribution of the color parameter of the pixel at the neighbor region, this value can be interpreted to be a degree of appearance of the color parameter $\boldsymbol{X}_i$ at the pixel $\boldsymbol{P}_i$.

4.2. Spatial information and temporal information

Here let us define generating information of the pixel of an image. The information $H(\boldsymbol{X}_i)$ of pixel $\boldsymbol{X}_i$ is given by the following equation (2). This definition of the information of the pixel is based on the definition of information, which is given by the Shannon's communication theory.

$$H(\boldsymbol{X}_i) = -\log_2 p(\boldsymbol{X}_i). \tag{5}$$

This information $H(\boldsymbol{X}_i)$ is getting large as the value of $p(\boldsymbol{X}_i)$ decreases. We can examine the distribution of the generating information of an image by calculating the information of each pixel of the image.

In this paper, we discuss two kind of information, one is originated in a spatial variation of the color parameter of the pixels of an image and the other is temporal variation of the color parameter at a fixed pixel. The spatial information is given by the variation of color parameters on an image at the time t, and is getting large in general in the case when the pattern of an image becomes complicated. On the other hand, the temporal information is defined by the variation of the color parameters along the time axis at the fixed pixel $\boldsymbol{X}_i$.

To show the relationship of the spatial and temporal information, let us examine Figure 2. The figure shows $m \times n$ pixels along the spatial direction on one frame image and l pixels along the temporal direction, namely last l frame images. Taking the $m \times n$ neighbor pixels on a frame image into account, the information $H_s(\boldsymbol{X}_i)$ represents the spatial information. In the case when the l pixels along the time axis at $\boldsymbol{X}_i$, the temporal information is given by $H_t(\boldsymbol{X}_i)$.

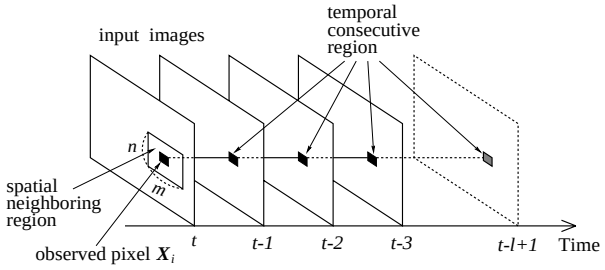


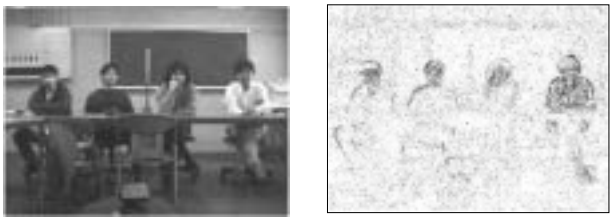
Figure 2. Neighboring and consecutive region.

4.3. Mutual information between spatial and temporal information

It is necessary that the video image of the conference includes the spot, which includes the participants at the virtual conference room as much as possible when taking a video image. The necessary spot of the scene for the persons at the virtual conference room is a region composed by pixels with both spatial and temporal information, such as the region of a speaking participant(s). By applying a way of thinking of mutual information between spatial and temporal information, we define mutual information $I(\boldsymbol{X}_i)$ of the region to which the participants pay attention by the following equation.

$$\begin{aligned} I(\boldsymbol{X}_i) &= \log_2 \frac{p_{st}(\boldsymbol{X}_i)}{p_s(\boldsymbol{X}_i) \cdot p_t(\boldsymbol{X}_i)}, \\ &= H_s(\boldsymbol{X}_i) + H_t(\boldsymbol{X}_i) - H_{st}(\boldsymbol{X}_i). \end{aligned} \tag{6}$$

Here, $H_{st}(\boldsymbol{X}_i)$ denotes the spatiotemporal information of the pixel $\boldsymbol{X}_i$ on the area of the neighbor $l \times m \times n$ pixels where $m \times n$ pixels to the spatial direction and l frame images to the time direction. Figure 3 (a) shows a input image and (b) shows the mutual information. The dark pixels have large information.



(a) Input video image. (b) Distribution of generated information.

Figure 3. Distribution of visual information.

5. Tracking of image area of the noticeable participant

5.1. Selection of the noticeable participant using the integrated information of visual and auditory information

To examine an image area where the human being pays attention, we extract 2 types of information, visual and auditory information. As the two kinds of information are essentially different from each other, it is necessary to define an integrated degree of attention from the visual and auditory information.

In order to obtain the image area to be paid attention, this system projects the point of sound source obtained by sound localization to the x-axis (the horizontal axis) on the 2-dimensional image plane of the monitor camera. And the system counts the number of projected points of the sound source in the slot on the x-axis every t_s seconds and represents a degree of attention by the number of points in the slot. In this way, the system can extract plural sound sources in the case that plural participants speak to each other.

In the next, as visual information, the system calculates the generating information (spatiotemporal information) of each pixel on the image taken by the monitor camera, and then obtains the value contained in each of the slots on the horizontal axis of the image plane. Based on the value in the slot, the system obtains the histogram for each of slots on the horizontal axis every t_v seconds. We consider the value of the histogram to be a degree of attention of the slot on the horizontal axis of the image. In general, it is supposed that the visual stimulus does not disappear immediately after generation but continues in a short time. We suppose the duration time to be t_d second.

After obtaining both the auditory and visual information, the system estimates the integrated degree of attention by unifying the degree of attention given by the auditory information and visual information. On the basis of the experience of the experiments, the integrated degree of attention is decided mainly by the auditory information, while the visual information is considered to be the supplementary information. Under this assumption, we define the integrated degree of attention by the sum of the degree of attention given by auditory information and the degree of attention given by visual information only in the case when the degree of attention of auditory information is greater than a threshold. In the case when the degree of attention of auditory information is smaller than a threshold, the integrated degree of attention is set to zero.

In the next, the system selects the noticeable participant(s) (namely, the participant(s) to whom human being pays attention) using the integrated degree of attention. The selection procedure of the noticeable participant(s) is given by the followings.

(1) Obtain the integrated degree of attention within the range of each participant in the image.

(2) Calculate the average value of the integrated degree of attention for all the participants.

(3) Calculate the difference between the integrated degree of attention for each participant and the average value of the integrated degree of attention.

(4) Decide the noticeable participant(s) if the difference is greater than a threshold.

The shooting algorithm of camera is based on the selection of the noticeable participant(s).

5.2. Control algorithm of a shooting camera

The several reports of the camerawork on a discussion program have been reported. The representative one [1] divides the participants into a group of speaking participant(s) and the group of non-speaking participant(s), and makes the video camera to shoot the area including the group of the speaking participant(s). This approach supposes that an operator always indicates speaking participant(s) during the program. Based on this indication, the system controls the shooting camera(s) using the knowledge of the camerawork concerning to the TV discussion program.

In this paper, we propose an algorithm that is able to select the noticeable participant(s) (speaking participant(s)) by itself on the basis of the integrated degree of attention as defined above and the system makes the camera to shoot the noticeable participant(s) using the selection of speaking participants. To decide the shooting range on the basis of the role of the participant(s), we classify the shots like the Table 1. This classification is prepared for avoiding a monotone video image by changing the number of participants appeared on an image.

In the next, let us define a timing to change the shots. Here, we define the following two conditions,

(1) The time when a noticeable participant(s) are replaced by another participant(s).

Table 1. Classification of shot.

Shot	Definition
s_1	Shot of a noticeable participant
s_2	Shot of a noticeable participant and neighbor
s_3	Shot of noticeable particiants
s_4	Shot of a non-speaking participant
s_5	Shot of non-speaking participants
s_6	Shot of all participants

Table 2. Transition probability matrix on noticeable participant(s) exchange.

		to					
		s_1	s_2	s_3	s_4	s_5	s_6
f	s_1	45.1%	18.5%	3.4%	4.1%	13.9%	15.0%
r	s_2	40.1%	30.4%	11.7%	1.2%	2.4%	13.4%
o	s_3	30.8%	25.9%	34.7%	1.3%	2.3%	5.1%
m	s_4	44.3%	31.5%	11.9%	2.8%	8.5%	1.0%
	s_5	36.7%	31.5%	8.3%	8.9%	4.1%	11.5%
	s_6	32.2%	23.3%	30.1%	0.0%	4.4%	10.0%

Table 3. Transition probability matrix on overpass duration.

		to					
		s_1	s_2	s_3	s_4	s_5	s_6
f	s_1	14.8%	21.7%	25.6%	21.2%	6.7%	11.0%
r	s_2	30.4%	17.2%	22.5%	10.7%	7.2%	12.0%
o	s_3	34.4%	18.0%	25.4%	10.7%	2.1%	9.4%
m	s_4	33.2%	22.9%	26.0%	4.0%	2.9%	12.0%
	s_5	42.2%	19.0%	24.3%	7.2%	2.3%	5.0%
	s_6	43.1%	24.1%	31.4%	1.4%	0.0%	0.0%

(2) The time when the duration of the same shot exceeds a threshold time.

The item (1) intends to change the shot in the case when the degree of importance to shoot the previous noticeable participant(s) is getting low as the new noticeable participant(s) appear in the conference. The item (2) intends to maintain the interest of the participant(s) in the image in order to avoid a monotone image in the case when the same shot continues a long time. Here, the system sets the duration time to 10 seconds. As human being, however, gets tired from too much switching over the shot to another in a short time, the system does not change the shot until a same shot continues more than a threshold time. The system sets this switching time to 2 seconds.

In the next, we define a transition probability matrix so as to evaluate the most suitable shot as the next shot when the system switches the shot. The transition probability matrix is given by calculating all of the rates (r_{ij}) when the system switches from a kind of shot (s_i) to another kind of shot (s_j) on the discussion programs, and the matrix is defined as shown in Table 2 and Table 3 for each of the condition of switching the shot.

6. Experiments and discussions

In the first, we explain some experiments on sound localization to confirm the estimation of the position of the sound source, then explain the experiments on a video image production under the condition of an actual conference in order to confirm the effectiveness of our method. The evaluation of the conference image is done by the results of a questionnaire survey.

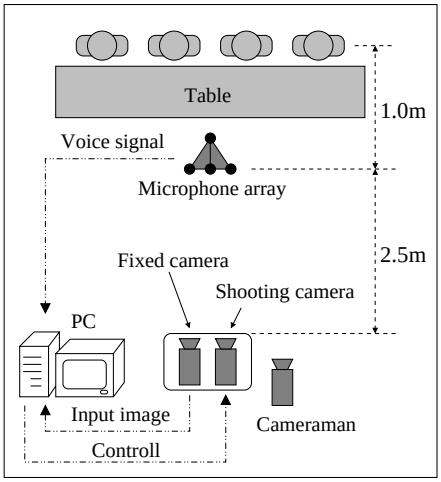


Figure 4. Microphone-set and camera layout.

6.1. Experiments on sound localization

In this experiment, we estimated the 3-dimensional position of a sound source that was a male voice (10 seconds duration). The result of the experiments is shown in Table 4. The position (x, y, z) of the sound source denotes the position where a participant uttered actually and the result of sound localizations shows the horizontal angle, the vertical angle and the distance from the center of the microphone set. According to the result of the sound localization, the range capable of estimating the position was limited within 1.5 m from the center of the microphone set.

6.2. Experiments on production of conference image

In the next, we set up the conference, which was supposed to be a TV-conference, and made experiments on taking a video image of a TV-conference. Figure 4 shows the placement of the microphone set and the video camera in the conference room. And, the computer-controlled camera system took the conference image by controlling the camera in a real time using our proposed method.

The system received the sound signals from the microphone set through an AD converting board, and also received the video signal from the monitor camera at the same time. Computing the sound localization and the spatiotemporal information of the image from the monitor camera,

Table 4. Result of sound source localization.

Position of sound source [m]			Mumber of data	Result of sound localization								
				Horizontal angle [deg]			Vertical angle [deg]			Distance [m]		
x	y	z		correct	mean	variance	correct	mean	variance	correct	mean	variance
0	1.0	0.3	421	0.00	0.98	2.65	16.70	17.15	0.33	1.04	1.42	0.14
0	1.0	0.7	335	0.00	-0.53	16.58	34.99	29.17	2.01	1.22	1.44	0.12
0	1.5	0.3	357	0.00	1.12	0.63	11.31	9.96	0.53	1.53	1.15	0.14
0	1.5	0.7	258	0.00	-2.23	4.13	25.02	29.34	1.36	1.66	1.07	0.13
0	2.0	0.3	127	0.00	0.34	16.65	8.53	18.29	3.91	2.02	1.25	0.19
0	2.0	0.7	305	0.00	4.68	7.77	19.29	24.88	2.69	2.12	1.26	0.16
0.5	1.0	0.3	340	26.57	19.49	2.92	15.02	15.90	1.30	1.16	1.25	0.14
0.5	1.0	0.7	269	26.57	22.90	1.21	32.05	32.71	0.14	1.32	1.43	0.10
0.5	1.5	0.3	158	18.43	3.33	11.91	10.74	15.99	2.57	1.61	1.42	0.15
0.5	1.5	0.7	197	18.43	14.51	2.03	23.88	26.44	0.31	1.73	1.11	0.09
1.0	1.0	0.3	118	45.00	26.99	9.05	11.98	20.27	1.38	1.45	1.27	0.22
1.0	1.0	0.7	92	45.00	26.88	8.76	26.33	25.48	2.05	1.58	1.26	0.24
1.0	1.5	0.3	112	33.69	9.70	15.78	9.45	20.29	6.74	1.83	1.25	0.14
1.0	1.5	0.7	127	33.69	9.67	6.72	21.22	34.09	2.71	1.93	1.30	0.16
1.5	1.0	0.3	110	56.31	18.10	16.63	9.45	24.31	4.46	1.83	1.23	0.25
1.5	1.0	0.7	111	56.31	8.50	20.26	21.22	23.49	2.29	1.93	1.03	0.17
1.5	1.5	0.3	158	45.00	6.18	21.63	8.05	23.43	5.89	2.14	1.27	0.18
1.5	1.5	0.7	111	45.00	10.66	20.94	18.26	25.84	4.59	2.23	1.15	0.18

$$\text{Horizontal angle} = \tan^{-1}\frac{x}{y}, \quad \text{Vertical angle} = \tan^{-1}\frac{z}{\sqrt{x^2+y^2}}, \quad \text{Distance} = \sqrt{x^2+y^2+z^2}.$$

the system controlled the shooting camera. The resolution of the image from the monitor camera was set to 320×240 pixels. The number of participants in the conference was supposed to be 4 participants, and the monitor camera was arranged so as to catch all the participants (4 participants). The time rate needs to calculate the sound localization was 0.6 second out of every 1 second, the time needs to calculate image processing was 0.4 second, which corresponds to the 2.5 frames processing per second. The parameters (l, m, n) denote the size of spatiotemporal space when calculating the spatiotemporal information. In our experiments, we set (l, m, n) to (3, 3, 3) taking the real time processing into consideration.

Figure 5 shows examples of the video image taken by our shooting camera. The images in the left column are examples taken by a fixed camera, while the right column shows the images taken by our shooting camera at the same condition. These images show the change from a wide shot to a zoom-up-shot to shoot a noticeable participant(s).

6.3. The questionnaire for image evaluation

Fourteen students evaluate the video image that is taken by our shooting camera in order to confirm the effectiveness of our method by the questionnaires. The object images are the following 3 kinds of images.

(1) Images taken by a fixed camera.
(2) Images taken by the shooting camera using our method.
(3) Images taken by a cameraman.

The 10 items of the questionnaire to evaluate the conference image are selected referring to the related works [1, 2]. The results of the questionnaire is shown in Table 5.

The value of the evaluation for each item is the range from 5 (high) to 1 (low), where the middle value 3 corresponds to the evaluation of the image taken by a fixed camera (a non-controlled camera). The evaluated values in the method taken by a cameraman and our proposed method denote the average value of all respondents. There is no evaluation concerning to Q.6 because of no switching to the shot of a third party in the case of the manual operation of camera.

6.4. Discussions

In this section, we discuss the characteristics of our method based on the results of the questionnaire. The evaluation over the video image using our method is higher than the image taken by a cameraman on the several items of the questionnaire, especially high on the items of Q.7 and Q.8. These results show that the approach of avoiding a monotone and boring image is effective. The results of the items Q.1 and Q.2 express the facts of specifying the noticeable participants and of easy to represent the atmosphere of the conference by switching to the shot of the third party sometimes. On the other hand, the image taken by a cameraman is beyond our approach concerning to the evaluation on the items Q.9 and Q.10. Also the evaluation of the image taken by a cameraman is higher than the image taken by the fixed

Figure 5. Example of created video images.

camera. The evaluation of the image using our method is lower than the one taken by the fixed camera. The reason seems to be too many switches from a shot to another by satisfying the condition of switching to focus another noticeable participants under our algorithm. Also the evaluation on the item Q.2 is higher because of switching to a shot of the third party and to the shot of shooting the whole scene, while the item on Q.3 is same as the method with a fixed camera. According to the results of the experiment and the questionnaire, our approach using the knowledge of switching rule under the discussion program is able to avoid a monotone and a boring video image, and produces a vivid video image maintaining the attention of the participants.

7. Conclusions

In this paper, we have defined auditory information and visual information that represent a degree of producing information at a point in the space of the object scene. Then

Table 5. Result of questionnaires.

Question item		Man	Com
Q.1	Easy to look at the participant(s) expression	3.57	**4.36**
Q.2	Atmosphere of the conference	2.71	**3.43**
Q.3	Easy to recognize the speaking participant(s)	**3.79**	3.14
Q.4	Easy to understand the participant(s) gestures	**3.79**	3.71
Q.5	Effectiveness switching	3.50	**3.79**
Q.6	Effectiveness switching to the non-speaking participant(s)	–	3.57
Q.7	A tedious video image	2.79	**1.57**
Q.8	A monotone video image	3.00	**1.57**
Q.9	Easy to look at video image	**3.43**	2.64
Q.10	Stress of a video image	**2.93**	3.29

In this table, "Man" shows cameraman operation video image, and "Com" shows computer operation video image. The level 3 estimation (middle level) was set to be the case of the video image by the fixed observation camera. The value of boldface type shows excellent evaluation.

we have proposed a method to seek the noticeable participant (a participant to whom human being pays attention) by integrating these two kinds of information among all participants in the case of producing TV conference image. In order to produce the vivid video image of TV conference, we have examined an algorithm of controlling the video camera using the noticeable participant and the knowledge of the camerawork of the TV discussion program. In the last, the effectiveness of our method has been confirmed by evaluating the TV conference video image taken by the computer-controlled camera using the rules based on the knowledge of the camera control.

References

[1] T. Inoue, K. Okada and Y. Matsushita, "Videoconferencing System Based on TV Programs," *Information Processing Society of Japan Journal*, vol. 37, no. 11, pp.2095–2104, Nov. 1996. (in Japanese)

[2] M. Onishi, M. Izumi and K. Fukunaga, "Production of Video Images by Computer Controlled Camera Operation Based on Distribution of Spatiotemporal Mutual Information," *Proc. 15th International Conference on Pattern Recognition*, vol. 4, pp.102–105, Sep. 2000.

[3] M. Minoh and Y. Kameda, "Image A 3D Lecture Room by Interpreting Its Dynamic Situation," *Proc. of The 4th Int. Workshop on Cooperative Distributed Vision*, pp.371–412, Mar. 2001.

[4] Y. Bresler and A. Macovski, "Exact maximum likelihood parameter estimation of superimposed exponential," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol.34, no.5, pp.1081–1089, Oct. 1986.

[5] R. O. Schmid, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Trans. on Antennas and Propagation*, vol.34, no. 3, pp.276–280, Mar. 1986.