# EFFECTS OF AUTOMATIC VIDEO EDITING SYSTEM USING STEREO-BASED HEAD TRACKING FOR ARCHIVING MEETINGS

*Yoshinao Takemae, Kazuhiro Otsuka and Junji Yamato*
NTT Communication Science Laboratories, NTT Corporation
3-1 Morinosato-Wakamiya, Atsugi-shi, Kanagawa, 243-0198, Japan
E-mail: {takemae,otsuka,yamato}@eye.brl.ntt.co.jp

## ABSTRACT

This paper presents an automatic video editing system based on head tracking for archiving meetings. Systems that archive meetings are attracting considerable interest. Conventional systems use a fixed-viewpoint camera and simple camera selection based on participants' utterances. However, conventional systems fail to adequately convey who is talking to whom and nonverbal information about participants etc. We focus on the participants' head orientation since this information is useful in detecting the speaker and who the speaker is talking to. In order to automatically estimate each participant's head orientation, our system combines several modules to realize stereo-based head tracking. The system selects the shot of the participant that most participants are looking at, based on majority decision. Experiments on presenting videos to viewers confirm the effectiveness of our system in several 3-participant conversations.

## 1. INTRODUCTION

Meetings are one of the most important activities in many workgroups. Often, due to scheduling conflicts or travel constraints, some cannot attend their scheduled meetings. We could overcome these problems by archiving the meetings and teleconferences if we had a system that was really effective. This research also impacts the field of Computer Supported Cooperative Work (CSCW).

Our purpose is to develop an automatic video editing system that can clearly convey the contents of multiparty conversations to viewers afterward. To this end, we focus on two fundamental components: *1) conversation direction*, which shows who is talking to whom, and *2) addressee's response* to speakers, including changes in facial expression and gaze. These components are extremely crucial pre-conditions in determining the contents of conversation. The reason for this is that conversation is constructed from a series of pairs of speaker's utterances and addressee's responses.

In this paper, we propose an automatic video editing system based on stereo-based head tracking for conveying the contents of multiparty conversations to the viewers. Our system can automatically detect participants' head 3D position and orientation during a conversation. Based on the detection results, our system selects the shot of the participant that most participants' heads are facing. We conduct experiments to evaluate the effectiveness of our system. The following sections summarize related work and our approach, and introduce the proposed system. We present details of our experiments, and discuss the results.

## 2. RELATED WORK

While this study focuses on archiving meetings and watching them afterwards, a considerable overlap exists between this domain and teleconferencing. Most conventional systems use a fixed-viewpoint camera. In large multiparty situations, participant face size is small. Hence these systems cannot convey sufficient nonverbal information such as changes in facial expressions and gaze. These visual cues greatly contribute to the viewers' understanding of the participants' intentions and emotions. Other conventional systems use visual representations that arrange multiple participants' shots captured by multiple cameras on one display. However these systems impose heavy cognitive loads on the viewer who must select video windows, and so they hinder the understanding of the conversation.

The solution to this problem is automatic camera selection in which multiple video streams of the multiple participants are appropriately ordered before being distributed. Cluster et al. developed a system called "Distributed Meetings" [4]. The system employs camera selection based on participants' utterances in addition to a panorama view shot. However, this approach cannot adequately convey whom the speaker is talking to and addressees' responses such as a rigid face with silence, since only the speaker is shown. Inoue et al. proposed a camera selection scheme based on a probability model obtained by analyzing the duration and the transition of shots in debate programs on TV [5]. This method provides viewers with video sequences that show speakers' shots only or other participants' shots. However, this approach fails to convey the flow of actual conversations because it uses a probability model, which has no relation to the actual conversations.

For conveying the contents of conversation in TV programs and films to the viewers, a number of cutting techniques are often used [6]. Cutting is equivalent to camera selection. According to "A Theory of Montage" [7], cutting techniques allow discontinuous shots to be formed into

*0-7803-9332-5/05/$20.00 ©2005 IEEE*

a montage that hopefully reconstructs the scenes of the conversation. By controlling the viewers' attention, they allow viewers to actively interpret and discern the relations between shots. For this reason cutting techniques such as ”L Cutting” and ”Shot/Reverse Shot” are used to handle conversations. Such cutting techniques reflect the experience of professional video directors and editors, and it is difficult for computers to completely reproduce their acquired knowledge.

## 3. OUR APPROACH

In a previous work we proposed a video editing rule based on the majority decision of participants' gaze in multiparty conversation [1]. This novel approach exploits participants' gaze behavior to select the most effective shots of participants. This is based on the following assumptions:

1) A person gazes at another when that person is of interest: participants try to acquire visual cues such as facial expression and the gaze direction of the other participants, in order to interpret the others' intention and emotion. This gaze behavior is called ”the monitoring function of gaze” [8]. 2) A person who receives the gaze of more participants has more important information than the others with regard to the conversation.

Experiments indicated that the videos produced by the proposed method can more accurately and clearly convey the conversation direction and addressee's response than conventional visual representations such as camera selection based on the participants' utterance and multiple view shots in 3- to 5-participant conversations. However, we did not implement this method as an automatic video editing system, since participants' gaze direction was extracted manually from captured videos.

Recently, many researchers have developed vision-based gaze tracking systems [2]. However, it is difficult for current vision-based approaches to robustly estimate participants' gaze direction in a multiparty conversation without heavy restrictions on user behavior. In this paper, we focus on the participants' head position and orientation instead of participants' gaze direction for the following reasons: 1) Head orientation is closely related to gaze direction. For example, from data captured in multiparty conversations, in 87% of frames the participants' head and eye gaze pointed in the same direction [3]. 2) Vision-based approaches can robustly estimate the participants' head pose without hindering the conversation.

In this study, to automatically detect whom each participant is looking at, we focused on stereo-based head tracking since it imposes no loads on the participants during the conversation [9]. We propose an automatic video editing system based on the results of detecting head orientation.

## 4. PROPOSED SYSTEM

### 4.1. System Overview

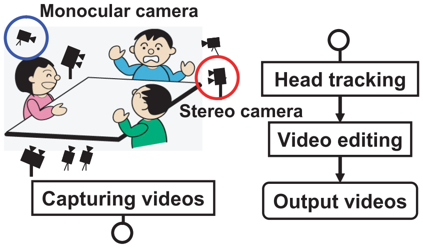

**Fig. 1**. Overview of the system

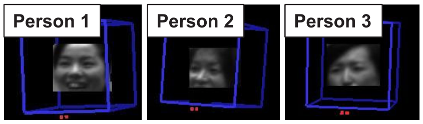

**Fig. 2**. Head tracking results from a 3-participant conversation.

Figure 1 overviews our automatic video editing system which uses stereo-based head tracking. First, stereo images of each participant's head and bust shots of each participant and whole view shot (monocular images) for presentation to the viewers are captured and recorded. These stereo images can also be substituted as the shots for presentation to the viewers. Second, 3D head orientations of the participants are automatically estimated with sequences obtained from each stereo camera. Based on 3D head orientation of each participant, whom each participant is looking at is extracted. Finally, our system automatically selects the shots of the participant that most participants are looking at.

### 4.2. Head Tracking

In order to obtain the 3D head position and orientation of each participant, we allotted a stereo-based head tracker to each participant [9]. Figure 2 shows the results of head tracking in a 3-participant conversation. This tracker uses adaptive view-based appearance models created from a two-frame registration algorithm which combines the robustness of ICP (Iterative Closest Point) and the precision of the normal flow constraint. This technique takes advantage of the depth information available from a stereo camera, which makes it less sensitive to lighting variations. This technique has a rotational RMS error smaller than 3°. By thresholding each participant's 3D head orientation, whom each participant is looking at or averting her/his face from is extracted per frame.

### 4.3. Video Editing

The system selects the bust shot of the participant that most participants are looking at, based on a majority decision, so as to better reflect the progress of the conversation. Figure 3 shows the time transition of participant's head orientation in one part of a 3-participant conversation and an example of the video sequences (P) produced by our system. The participants were debating whether we should legally recognize the death penalty in Japan. In this part, Person 3 was talking to (addressing) Person 2. In Figure 3 (d), the arrows indicate each participant's head orientation. The person that most participants' are looking at is shown in gray. The focus of participants' head orientation alternated between the

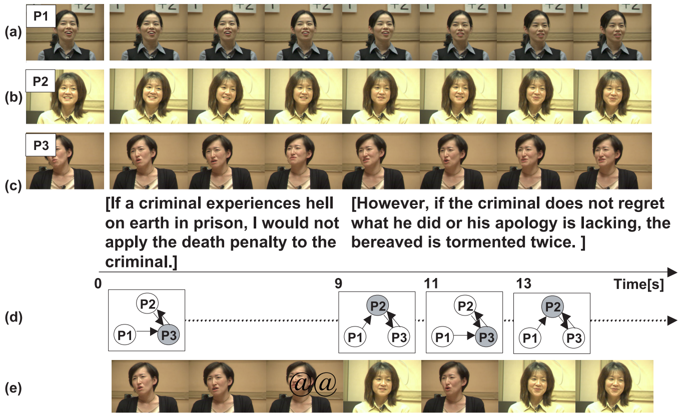
(a)
P1
(b)
P2
(c)
P3
[If a criminal experiences hell on earth in prison, I would not apply the death penalty to the criminal.]
[However, if the criminal does not regret what he did or his apology is lacking, the bereaved is tormented twice. ]
0
9
11
13
Time[s]
(d)
P2
P1
P3
(e)
@@

**Fig. 3**. Participants' head orientation and a video sequence produced by our system for one part of a 3-person conversation.

(a), (b), and (c) show the behavior of Person 1, Person 2, and Person 3, respectively. (d) shows transitions of participants' head orientation. Arrows indicate each participant's head orientation. The person that most participants are looking at is shown in gray. (e) shows the video sequence produced by our system.

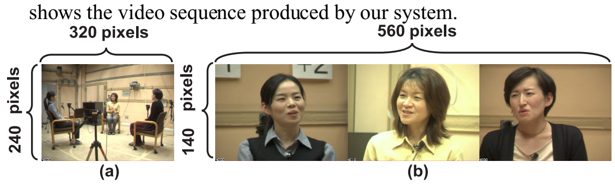

**Fig. 4**. Examples of whole view shot (a) and multiple view shot (b).

speaker and the addressee with the progress of the conversation as shown by the behavior displayed in Figure 3 (d). Consequently our system alternated between camera selection for the speaker and for the addressee (See Figure 3 (e)).

## 5. EXPERIMENTS

We conducted experiments to verify the effectiveness of our system using 3-participant conversations. Subjects, who did not participate in the debates, viewed the resulting videos, and evaluated them.

### 5.1. Collecting Conversation Data

We focused on face-to-face 3-participant debates. Two groups participated in the debates. The participants in each group were Japanese females (average age was 34.3). Bust shots of each participant, a whole view shot, and three stereo camera sequences were recorded. Frame size was 320 × 240 pixels. Pin microphones recorded the utterances. Each group debated about topics such as whether we should legally recognize the death penalty in Japan. The two debates took about 6 minutes and 8 minutes.

### 5.2. Visual Representations Compared

Videos (P) produced by our system were compared to the following four visual representations. Three are currently used for archiving meetings and teleconferences. Face view resolution was the same, 320 × 240 pixels, for all methods.

***1) Whole view shot (W).*** All participants are captured in one shot as shown in Figure 4 (a). This cannot adequately convey the changes in facial expressions and gaze to the viewers because participant face size is too small. The resolution of this video is 320 × 240 pixels.

| Question No. | Questions |
|---|---|
| Q.1 | Did you clearly see who the speaker was ? |
| Q.2 | Did you clearly see the changes in the speaker's facial expressions and gaze? |
| Q.3 | Did you clearly see whom the speaker was talking to (addressee) ? |
| Q.4 | Did you clearly see the changes in addressee's facial expressions and gaze? |
| Q.5 | Did you clearly the relation with regard to position among the participants? |

**Table 1**. Questionnaire.

***2) Multiple view shot (M).*** This places bust shots in one row in order to express the spatial relations between participants (see Figure 4 (b)). This does not completely preserve the geometric arrangement of participants and makes it difficult for viewers to recognize whom the speaker is gazing at. The resolution of this video is 560 × 140 pixels.

***3) Speaker shot (S).*** The moment a participant starts an utterance, a bust shot of the speaker is shown. This has the effect of clearly conveying who the speaker is to the viewers. Utterance intervals of participants were extracted based on thresholding power information of recorded voice. The resolution of this video is 320 × 240 pixels.

### 5.3. Method

***Subjects.*** The paid subjects, who did not participate in the debates, were 59 Japanese people (27 males and 32 females, average age was 28.9). Subjects were divided into four groups. The first group, the second group, the third group and the fourth group viewed each of the visual representations produced using approaches (P), (W), (M) and (S), respectively. The number of subjects in the groups was 15, 15, 14, and 15, respectively.

***Materials.*** The two above mentioned debates were used. Each debate was edited using the four different visual representations. Each editing results was presented to the subjects only once.

***Questionnaire.*** To evaluate the effectiveness of our system, we asked the subjects to complete a questionnaire (See Table 1). Q.1 determines the clarity of recognizing the speaker. Q.2 determines the clarity of recognizing the speaker's nonverbal information. Q.3 determines the clarity of recognizing whom the speaker was talking to: the addressee. Q.4 determines the clarity of recognizing the addressee's nonverbal information. Q.5 determines the clarity of recognizing the relation in position among the participants. This is an important element in clearly conveying to the viewers whom the speaker is talking. In each item, subjects selected one statement from a 7-point scale: -3 (strongly disagree) to 3 (strongly agree).

### 5.4. Results and discussions

The effectiveness of visual representation style was analyzed using the data from the questionnaires. We used one-

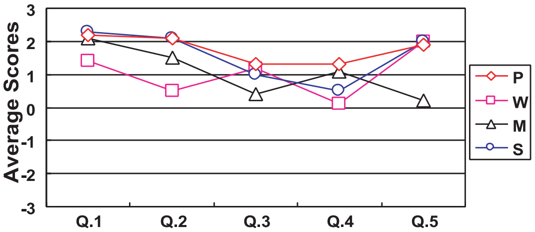

**Fig. 5**. Survey results in each question.

(P), (W), (M) and (S) present videos produced by our system, whole view shot, multiple view shot, and speaker's shot respectively.

factor ANOVA with visual representation type as the independent variable. If a significant difference was found, Tukey's multiple comparison was applied.

Figure 5 shows average scores of each visual representation in each question. For Q.1 and Q.2, (P), (S) and (M) were evaluated more highly than (W) ($p < .01$, $p < .01$). For Q.3, (P), (S) and (W) were evaluated more highly than (M) ($p < .03$). For Q.4, (P) and (M) were evaluated more highly than (W) and (S) ($p < .01$). For Q.5, (P), (S) and (W) were evaluated more highly than (M) ($p < .03$).

We discuss the characteristics of the four visual representations based on the results of the questionnaire below.

***1) Whole view shot (W).*** From the results of Q.1, Q.2 and Q.4, It was difficult for the subjects to recognize the speaker's nonverbal information such as mouth movements and face expression in (W) because face size was too small.

***2) Multiple view shot (M).*** The results of Q.3 and Q.5 indicate that (M) failed to adequately convey whom the speaker was talking to. This reason is as follows. (M) provided insufficient geometric coordination between participant shots, it was difficult for the subjects to recognize the position relation among participants. This characteristic made it difficult for the subjects to recognize whom the speaker's gaze direction and head orientation were directed at.

***3) Speaker's shot (S).*** From the results of Q.1 and Q.2, (S) can convey more clearly convey who is the speaker and speaker's nonverbal information. This is because the shot of the speaker was selected. Contrary to our predictions, (S) had high scores in Q.3. This reason is as follows. The utterances are related to the subsequent utterances in the conversation. For example, there are pairs of suggestion-acceptance[rejection], question-answer etc. Hence, many subjects may recognize the current addressee based on utterance context although a shot of the addressee was not shown.

***4) Our system (P).*** From the results of Q.1 and Q.2, (P) conveys as clearly who is the speaker and the speaker's nonverbal information as (S). Considering the results of both Q.3 and Q.4, (P) can more clearly convey who is the addressee and the addressee' nonverbal information than the others. The reason is as follows. The focus of participant head orientation alternated between the speaker and other participants including addressee. This characteristic caused alternate selection of the shot of speaker and that of the other participants including the addressee according to the progress of the conversation.

Considering all the results, we conclude that the proposed system is more effective than the current alternative view styles of whole view shot, multiple view shot, and speaker's shot.

## 6. CONCLUSION

This paper proposed an automatic video editing system using stereo-based head tracking in face-to-face conversations for conveying conversation contents to the viewers. Our system can automatically track participants' head orientation and use the results to select the shot of the person that most participants are looking at. We conducted experiments to evaluate the effectiveness of our system. The results show that our system is more effective than existing automatic visual representation schemes for archiving meetings and teleconferences. This work offers a realistic framework for automatic video editing systems based on the cue of participant head orientation. In future work, we will evaluate the effectiveness of our system with more than three participants using our video editing rule, the usefulness of which has been proven for more than three participants.

## 7. ACKNOWLEDGMENTS

We would like to thank Prof. Trevor Darrell and Dr. Louis-Phlippe Morency of MIT Computer Sciences and Artificial Intelligence Laboratory for their helpful comments on head tracking and letting us use their useful head tracking software.

## 8. REFERENCES

[1] Takemae, Y., Otsuka, K., and Mukawa, N., ”Impact of Video Editing Rule based on Participants' Gaze in Multiparty Conversation”, Ext. Abstracts CHI '04, pp.1333–1336, 2004.

[2] Matsumoto, Y., Ogasawara, T., and Zelinskey, A., ”Behavior Recognition Based on Head Pose and Gaze Direction Measurement”, Proc. IEEE International Conference on Intelligent Robots and Systems, pp.262–267, 2000.

[3] Stiefelhagen, R., Zhu, J., ”Head Orientation and Gaze Direction in Meetings”, Ext. Abstracts CHI '02, pp.858–859, 2002.

[4] Cutler, R., et al., ”Distributed Meetings: A Meeting Capture and Broadcasting System”, Proc. of ACM Multimedia '02, pp.503-512, 2002.

[5] Inoue, T., Okada, K., and Matsushita, Y., Learning from TV Programs: Application of TV Presentation to a Videoconferencing System, Proc. of ACM UIST '95, pp.147-154, 1995.

[6] Arijion, D., ”Grammar of the Film Language”, Silman-James Press, Los Angeles, 1976.

[7] Glenny, M., Tayler, R. (eds). S. M., ”Eisenstein Selected Works Volume 2, Towards a Theory of Montage”, British Film Institute, 1991.

[8] Kendon, A., ”Some Function of Gaze-Direction in Social Interaction”, Act. Psychologica, Vol.26, pp.22–63, 1967.

[9] Morency, L.-P., Rahimi, A., and Darrell, T., ”Adaptive View-based Appearance Model”, Proc. IEEE conference on Computer Vision and Pattern Recognition, pp.803-810, 2003.