VIDEO QUALITY ANALYSIS FOR AN AUTOMATED VIDEO CAPTURING AND EDITING SYSTEM FOR CONVERSATION SCENES

T. Nishizaki, R. Ogata, Y. Kameda, and Y. Ohta

Graduate School of SIE
University of Tsukuba
Tsukuba, 305-8573,Japan

Y. Nakamura

ACCMS
Kyoto University
Kyoto, 606-8501, Japan

ABSTRACT

This paper introduces video quality analysis for automated video capture and editing. Previously, we proposed an automated video capture and editing system for conversation scenes. In the capture phase, our system not only produces concurrent video streams with multiple pan-tilt-zoom cameras but also recognizes "conversation states" *i.e.*, who is speaking, when someone is nodding, etc. As it is necessary to know the conversation states for the automated editing phase, it is important to clarify how the recognition rate of the conversation attributes affects our editing system with regard to the quality of the resultant videos. In the present study, we analyzed the relationship between the recognition rate of conversation states and the quality of resultant videos through subjective evaluation experiments. The quality scores of the resultant videos were almost the same as the best case in which recognition was done manually, and the recognition rate of our capture system was therefore sufficient.

1. INTRODUCTION

There is a great deal of demand for automated capturing and editing of conversation scenes, which are useful to review events for people who could not attend. There have been a number of related studies regarding the recording of meetings, lectures, etc.[1][2], and editing of recorded videos [5][6]. Our targets are ordinary conversation scenes where two or three people are talking around a table.

Previously, we proposed an automated video capture system [3] and an automated video editing system [4] for conversation scenes. Our proposed capture system not only produces concurrent video streams with multiple pan-tilt-zoom cameras but also recognizes conversation states, such as utterances of demonstrative pronouns and conjunctions, occurrences of utterances, nodding, locations of participants, etc. Although recognition of these states is essential four our proposed automated editing system, it is not plausible to assume perfect recognition. Therefore, it is necessary to clarify how the recognition accuracy affects the editing system with regard to the quality of the resultant videos. In the present study, we analyzed the relationship by determining the results of subjective evaluation of video editing. Our results indicated that the recognition accuracy of our system is sufficient to achieve quality close to that of videos edited manually assuming perfect recognition.

2. CONVERSATION ATTRIBUTES

We first discuss our video capture and editing system to describe how the conversation states are recognized in the capture phase.

Fig. 1. Concurrently recorded videos

In the capture phase, the system produces concurrent video streams keeping appropriate picture compositions by controlling multiple pan-tilt-zoom cameras [3]. Several snapshots of the video clips produced by the automated capturing system are shown in Figure 1. In the figure, two people walked up to a table, sat down, had a conversation, and left the table.

At the same time, conversation states, such as nodding, utterances of demonstrative pronouns and conjunctions, and occurrences of utterances were recognized. Utterance-related information was recognized by IBM ViaVoice, and nodding was counted by our method based on a computer vision technique.

0-7803-9332-5/05/$20.00 ©2005 IEEE

In the editing phase, using the conversation states, the system assembles short video clips in the concurrent video streams recorded in various picture compositions, and produces a final edited video based on editing preferences given by the user. Our editing method is based on optimization with constraints satisfaction [4]. Due to space constraints, we will skip the details of capturing and editing phases; please refer to [3][4] for details.

In this paper, we discuss the influence of the recognition rate of conversation states. To analyze the relationship between the recognition rate of conversation states and the quality of the resultant videos, we produced various edited videos based on various recognition rates of conversation states, and conducted subjective evaluation experiments in which subjects watched and evaluated the various edited videos. Note that the experiment used the optimal editing preferences given in [4].

3. PERFORMANCE OF OUR CAPTURE SYSTEM

To evaluate camera controls in the video capture system, we conducted a subjective evaluation experiment. In the experiment, 30 subjects watched short video clips recorded by the capture system, and scored the following four factors with values from 1 to 5.

- A.1 "How was the rotation speed of the camera?" "Slow", 1; and "fast", 5.
- A.2 "How was the frequency of the camera control?" "Low frequency", 1; "high frequency", 5.
- A.3 "Did the short video clips have good picture composition ?" "Poor", 1; "good", 5.
- A.4 "Was the camera control adequate for the situation?" "Not adequate", 1; "adequate", 5.

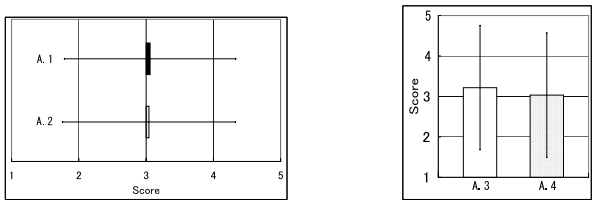

Fig. 2. Evaluation of automated capture

Note that these factors did not require the subjects to discuss the effects of editing of the produced video clips. Figure 2 shows the average scores for each factor. In A.1 and A.2, the best score was "3", while in A.3 and A.4, the best score was "5". In the figure, vertical/horizontal lines on the bars indicate deviation. As shown in A.1 and A.2 in Figure 2, our method had an overall score of 3, indicating that our method realized almost the best control of speed and frequency of moving the pan-tilt-zoom cameras. In contrast, there was room for improvement of picture composition and adequate camera control, as shown in A.3 and A.4 in Figure 2. These scores of A.3 and A.4 were also supported by comments returned by the subjects; some noted that the accuracy of picture composition requires improvement. It is necessary to improve this factor in our future studies.

Table 2. Occurences of utterances

# of video clips	Average precision	Average recall
50	88%	81%

4. INFLUENCE OF RECOGNITION RATE ON RESULTANT VIDEO QUALITY

As the recognition rate of conversation states in the capture system has an influence on automated editing, we first report the recognition rate of our capture system and then discuss the relationship between the recognition rate and the quality of video editing.

4.1. Recognition rate of Conversation States

In this paper, we discuss four conversation states: nodding, utterances of demonstrative pronouns, utterances of conjunctions, and occurrences of utterances. We conducted four experiments to evaluate the recognition rate of our capture system. The first three used 15 video clips of about 15 minutes in total length to evaluate nodding, 47 video clips of about 38 minutes in total length to evaluate utterances of demonstrative pronouns, and 47 video clips of about 35 minutes in total length to evaluate utterances of conjunctions. The results are shown in Table 1 with precision and recall rates; the precision and recall rates were mostly ¿90utterances, another experiment was conducted on 50 recorded videos, each of about 120 seconds in length and containing an average of about 100 seconds of utterances (Table 2).

4.2. Evaluation

To analyze the influence of recognition rate on the resultant video quality of automated editing, we conducted a subjective evaluation experiment. We created five types of automatically edited videos of the same scene by changing recognition rates in five ways, and compared the resultant videos. Figure 3 shows examples of the five edit types that subjects watched, and Table 3 shows the recognition rates of the edit types. Type 1 was regarded as an ideal situation, which assumed that all the conversation states were recognized perfectly. In Type 1, there was no error recognition and no recognition miss. Type 2 corresponded to our

Table 1. Nodding and utterances of keywords

	Clips	Total time [min]	Number	Detected	Error	Failure	Precision	Recall
Nodding	15	15	80	62	6	18	90.3%	77.5%
Demonstrative pronouns	47	38	54	50	5	4	90.0%	92.6%
Conjunctions	47	35	52	48	0	4	100.0%	92.3%

Table 3. Five types of different recongnition rates [%]

	Nodding		Demonstrative pronouns		Conjunctions		Occurrences of utterances	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Type 1	100	100	100	100	100	100	100	100
Type 2	90	78	90	93	100	92	88	81
Type 3	0	0	100	100	100	100	100	100
Type 4	100	100	0	0	0	0	0	0
Type 5	0	0	0	0	0	0	0	0

Fig. 3. Snapshots of five edit types

capture system. Type 3 assumed that utterance states were recognized completely, while nodding was not recognized. On the other hand, Type 4 assumed recognition of no utterance state, while nodding was recognized completely. Finally, in Type 5 no conversation state was recognized. In the experiment, we applied these five editing types to four conversation scenes, and created 20 resultant videos. Subjects watched the videos, and scored their impressions of the following six factors with values from 1 to 5.

- B.1 “Did you understand the statuses of the speakers?”
  “No”, 1; “Yes”, 5.
- B.2 “Did you understand the statuses of the listeners?”
  “No”, 1; “Yes”, 5.
- B.3 “Did you recognized the locations of all persons?”
  “No”, 1; “Yes”, 5.
- B.4 “Did you feel the atmosphere of the conversation?”
  “No”, 1; “Yes”, 5.
- B.5 “Was view switching good?” “No”, 1; “Yes”, 5.
- B.6 “How did you feel about the frequency of view switching?” “Boring”, 1; “busy”, 5.

In these experiments, comparison between Type 1 and Type 2 was important to evaluate the performance of our system.

The results are shown in Figures 4 and 5. Figure 4 shows the averages of the impressions of B.1, B.2, ... B.5, where the best score was “5”, while Figure 5 shows the average impressions of B.6, where the best score was “3”. In the figures, vertical/horizontal lines on the bars indicate deviation.

4.3. Discussion

According to the scores of Type 1 and Type 2 shown in Figures 4 and 5, there were no differences between Type 1 and Type 2 in any of the factors, B.1, B.2, ... B6. The results indicated that the recognition rate of the conversation states recognized by the capturing system was sufficient, and some recognition errors of the system did not have a severe influence on video quality.

According to the editing preferences given to the system in the experiment, a long shot picture composition is sometimes used to capture all the people within a frame, and is inserted in more or less all of the videos of all five types. In addition, if no conversation state is recognized, a long shot is inserted frequently [3][4].

In B.1, evaluations for Type 1, Type 2, and Type 3 were high. This was because the editing system tended to use

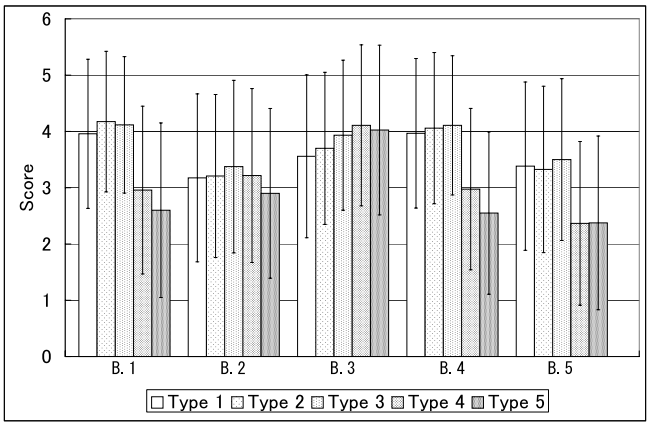

Fig. 4. Evaluations of B.1 to B.5

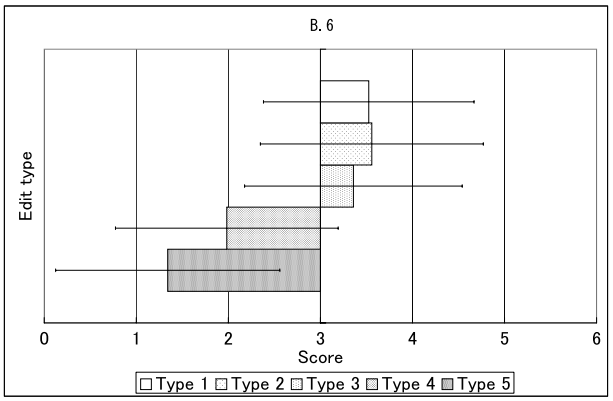

Fig. 5. Evaluations of B.6

speaker shots according to the recognition rates given by Type 1, Type 2, and Type 3.

In B.2, all types had almost the same score. This indicated that subjects could understand aspects of listeners if only long shots were inserted. Thus, the insertion of listener shots did not improve the impressions of B.2.

In B.3, all types again showed almost the same score. We had envisioned that evaluations for Type 4 and Type 5 would be much higher than the other types because we initially felt that a long shot would allow subjects to recognize the locations of all participants, and long shots were inserted frequently for these types that had a few valid states. However, the results indicated that almost half of the subjects could recognize the locations even if a long shot was inserted only a few times.

In B.4, evaluations of Type 1, Type 2, and Type 3 that used speaker shots were high. Therefore, occasional insertion of speaker-related video clips improved understanding of the atmosphere of the conversation.

In B.5, all types were given only low scores. Especially, Type 4 and Type 5 were marked very low because there were no states that were useful for appropriate switching. Thus, it is necessary to improve the timing of switching in the editing system.

In B.6, we found that the frequency of view switching was appropriate because evaluations of Type 1, Type 2, and Type 3, which had many states, had the best scores.

5. CONCLUSIONS

We discussed video quality analysis with our automated video capturing and editing system for conversation scenes. In the capture phase, our capture system not only produced concurrent video streams with multiple pan-tilt-zoom cameras but also recognized conversation states. These states were essential for the automated editing system, and we showed how the recognition rate affects the quality of the resultant videos. We discussed the relationship between the recognition rate of conversation states and the quality of resultant videos based on the results of subjective evaluation experiments.

The results indicated that the quality of the resultant videos was scored as almost same as the best case in which perfect recognition was done manually, and the recognition rate of our capturing system was thus sufficient.

However, there is room for future improvement; it is necessary to improve the accuracy of composition in the capture system and timing of view switching in the editing system. It is also necessary to explore and evaluate recognition of other conversation states.

6. REFERENCES

[1] M. Ozeki, Y. Nakamura, and Y. Ohta, “Human behavior recognition for an intelligent video production system”, IEEE Proc. PCM, pp.1153–1160, 2002.

[2] M. Murakami, S. Nishiguchi, Y. Kameda, M. Minoh, “Effect on Lecturer and Students by Multimedia Lecture Archive System”, 4th ITHET (ITHET2003), pp.377–380, 2003.

[3] T. Nishizaki, R. Ogata, Y. Nakamura, Y. Ohta, “Video Contents Acquisition and Editing for Conversation Scenes”, KES2004, 2004.

[4] R. Ogata, Y. Nakamura, Y. Ohta, “Computational Video Editing Model based on Optimization with Constraint-Satisfaction”, Proc. 4th PCM, pp.CD-ROM, 2A1-2 (6 pages), 2003.

[5] Y. Atarashi, Y. Kameda, M. Mukunoki, K. Kakusho, M. Minoh, K. Ikeda, “Controlling a Camera with Minimized Camera Motion Changes under the Constraint of a Planned Camera-work” Workshop on Pattern Recognition and Understanding for Visual Information Media, in Cooperation with ACCV, 2002, pp.9–14, 2002.

[6] M.Onishi, T.Kagebayashi, K.Fukunaga, “Production of Video Images by Computer Controlled Cameras and Its Application to TV Conference System”, Proc. of IEEE Conference on CVPR, Vol.2, pp.131–137, 2001.