EVALUATION OF SELF-EDITING BASED ON BEHAVIORS-FOR-ATTENTION FOR DESKTOP MANIPULATION VIDEOS

Motoyuki OZEKI and Yuichi NAKAMURA

ACCMS, Kyoto University, Yoshida Honmachi, Kyoto 606-8501 Japan
{ozeki, yuichi}@media.kyoto-u.ac.jp

ABSTRACT

In this paper, we discuss a user interface issue with regard to automatic video editing based on the speaker's intentions. In our experiments, the subjects used previously developed video capturing system to employ 4 types of editing methods by making 3 types of presentations. Subjective evaluation revealed that the editing method that used behaviors-for-attention obtained a good score for a presentation in which the subject was provided with specific instructions regarding the tasks to be performed. In the case of a presentation without a scenario, an editing method using a footswitch and a posture obtained a higher score. It can be concluded that a combination of both behavior-based and footswitch-based editing would provide a good environment for content acquisition.

1. INTRODUCTION

With the wide use of video-based media such as e-learning, video conferencing, and video instructions, automatic video content production has become one of the key technologies. For this purpose, we developed a multicamera system for recording instructions on desktop manipulations such as cooking, do-it-yourself (DIY), or scientific experiments; this system has been presented in Figure 1. We have proposed a camera control method that can adjust the trade-off between the following two requirements; 1) tracking a target and keeping it at the center of the screen and 2) fixing a camera angle and view field in order to suppress shaky and irritating view changes[1]. We also proposed an online editing method (camera switching method) that selects the most appropriate shot based on particular human behaviors that draw the attentions of viewers. Hereafter, we refer to these types of behaviors as "*behaviors-for-attention*." We had demonstrated that the quality of the generated video is satisfactory and not inferior to TV programs with regard to several criteria[2].

A number of methods that support video editing in an online or an offline process have been proposed[3][4][5]. Our editing method features a speaker (a lecturer) assigning editing triggers to the system by his/her behavior. We refer to this scheme of editing as "*self-editing*." This scheme allows a speaker to convey his/her editorial intentions to the system, and to control viewers' attention on the basis of the editorial effects. However, self-editing also introduces an additional editing burden on the speaker. Thus, it is necessary to investigate the usability of the system from the viewpoint of a user interface. In this paper, we demonstrate the following: 1) the most preferable editing interface in a self-editing scheme, and 2) how the evaluation results vary depending on the forms of presentation. Our philosophy features in previous works that evaluate only generated video by video capturing systems[6][7].

In the following sections, we first briefly explain our idea of a self-editing scheme based on behaviors-for-attention. We then describe its purpose and the evaluation method that has been examined in this paper, demonstrate the experimental results, and discuss them.

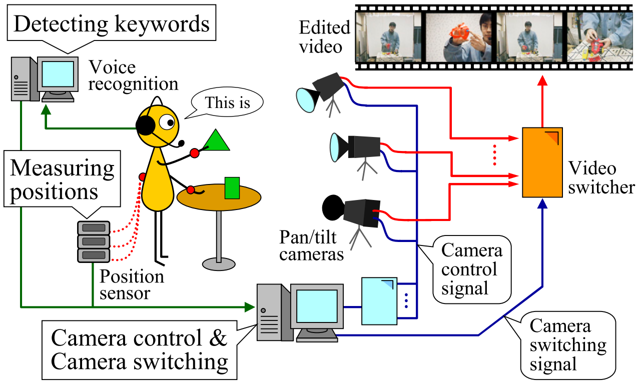


Fig. 1. Outline of our video production system.

2. EDITING BASED ON BEHAVIORS-FOR-ATTENTION

In our scheme for editing desktop manipulation videos, switching between a medium shot of a human + a workspace and a close-up shot of the hand(s) or an object can be considered as the most essential editing schemes. A medium shot provides an overview of the manipulation for viewers. A close-up shot draws the viewers' attention to the manipulation or the object. Thus, our videos are comprised of switching among 1 medium shot and 3 close-up shots, each of which captures the right hand, the left hand, and both the hands, respectively. In our system, these 4 types of shots are captured by 4 cameras assigned to each target.

In order to automate this editing scheme, we need to determine the time at which the system requires to switch to particular shots, and the triggers that can be assigned for switching. We use behaviors-for-attention that significantly draw viewers' attention as one of the triggers. At the appearance of behaviors-for-attention, the system switches from a medium shot to a close-up shot that is assigned by the behavior. It then switches back to a medium shot when the manipulation following the behaviors-for-attention ends. Figure 2 shows our editing rule and presents examples of behaviors-for-attention that frequently appear in desktop manipulations, and that direct viewers' attention to important sections.

In order to detect behaviors-for-attention, we proposed a simple method that utilizes the cooccurrence of motion cues and speech cues. As shown in Figure 2, we used "direct request" and "deictic ut-

1-4244-0367-7/06/$20.00 ©2006 IEEE

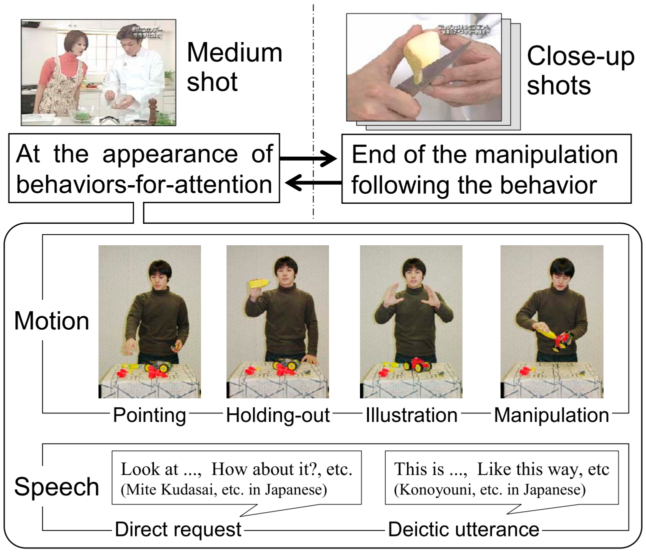

Medium shot
Close-up shots
At the appearance of behaviors-for-attention
End of the manipulation following the behavior
Motion
Pointing
Holding-out
Illustration
Manipulation
Speech
Look at ..., How about it?, etc. (Mite Kudasai, etc. in Japanese)
This is ..., Like this way, etc (Konoyouni, etc. in Japanese)
Direct request
Deictic utterance

Fig. 2. Editing rule and examples of behaviors-for-attention.

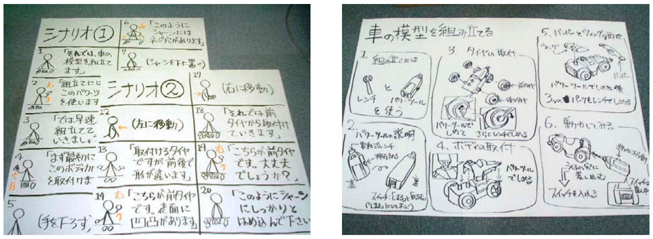

Detailed-scenario Outline-scenario

Fig. 3. Scenarios used in the experiment.

Table 1. Questionnaires for speakers.

(a) Can you edit to suit your requirements perfectly?
(b) Are you comfortable with this interface?
(c) Can you use this interface for more than 30 minutes?
(d) Do you want to use this interface if you get accustomed to it?

terance" as speech cues. Motion cues are simply detected by measuring an arm stretch beyond a certain threshold[1]. The most appropriate close-up shot is determined in the following manner: When both the hands are stretched and when they are close to each other, the system selects a close-up shot of both the hands. If both the hands are held apart from each other, the system selects a close-up shot of the hand that is at a higher position than the other. For more details of this method, please refer to [8]. The end of the manipulation following the behaviors-for-attention is detected when either or both the hands move out of the camera screen or when a speaker lets his/her hand(s) down.

3. EVALUATION OF USER INTERFACE

In this evaluation, we intended to investigate 1) what the most preferable interface in a self-editing scheme is and 2) how the evaluation results vary depending on the form of presentation. The presentations pertain to assembling a toy car (approximately 4 minutes).

We focus on the performance comparison of editing methods; the method using natural behaviors and methods using specific behaviors. We are also interested in what types of specific behavior is better. The editing interfaces investigated by us are as follows:

(A) Behaviors-for-attention:
This interface uses behaviors-for-attention, which has already been explained in section 2.

(B) Oral-keyword:
This interface uses the keywords that are spoken during the presentation. Since each camera is assigned to a specific target, that is, right hand, left hand, both the hands, or the entire scene (for a medium shot), these terms are used as keywords for switching to the camera.

(C) Footswitch-and-posture:
This interface uses a foot-switch instead of speech cues in the method of behaviors-for-attention. A shot change occurs when a speaker pushes a footswitch. An appropriate camera is selected by the posture, which is the same manner as the interface for behaviors-for-attention. Afoot-switch allows a speaker to use both the hands.

(D) Manual-editing:
This is prepared for reference. One of the authors operates a video switcher while viewing a presentation on a monitor.

For the form of presentations, we consider the following 3 situations; 1) recording video instructions, 2) unidirectional realtime streaming, and 3) distance learning with question and answers.

Detailed-scenario:
A detailed scenario is provided to a subject in the form of a cartoon strip. This scenario specifies what to speak and how to perform; the cartoon strip used in the experiment is shown in the left-hand side of Figure 3. This type of presentation is suitable for recording a video such as a manual whose contents are almost fixed.

Outline-scenario:
The outline of a presentation is provided to a subject through an assembling instruction, as shown in the right-hand side of Figure 3. This type of presentation is suitable for lectures or ordinary presentations.

Interaction-form :
The subject is requested to answer the questions asked. This type of presentation is essential to the question-answer process in distance learning.

We chose 6 subjects who had never used our system. Each subject was asked to make a presentation on the assembly of a toy car using each of abovementioned types of presentations; after this presentation, the subjects were asked to rate each interface (1: bad/no - 3: neutral - 5: good/yes) based on 4 criteria shown in Table 1. In order to familiarize the the subjects with our system, we conducted experiments in the following order: detailed-scenario > outline-scenario > interaction-form. When conducting experiments in the interaction-form, we gathered 6 students as interrogators, and provided 2 sites that are connected only by video transmission. In order to reduce the effect of the differences among subjects, we specified the questions before the actual presentation (7 questions for each interface). Each

[1] In our system, magnetic sensors are attached to the speaker's hands and waist for controlling the cameras and measuring an arm stretch.

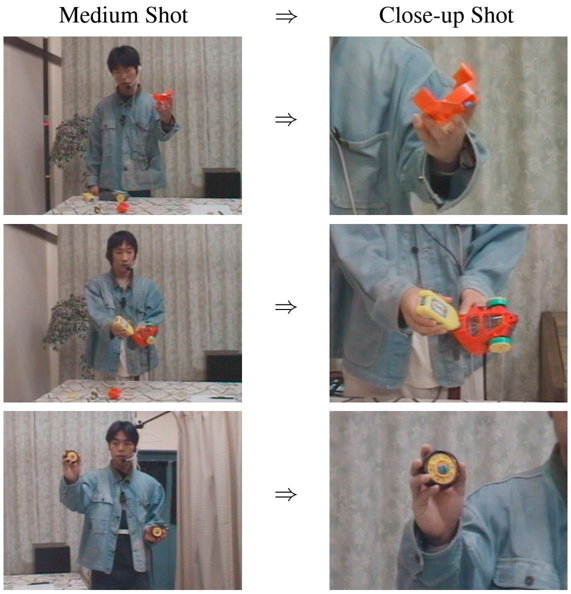


Fig. 4. Examples of captured videos.

subject was required to switch shots more than once when answering each question. Figure4 shows examples of captured videos.

4. RESULTS AND DISCUSSION

The success rates of editing—that is, the rates at which the shot changes according to the subject's requirements—are shown in Table 2. In this experiment, we did not have any false alarms. Thus, the precision is 100%, and each rate in the table means "recall" rate. The overall success rate of each category is around 80%–90%, although that of the "oral-keyword" category is slightly low. The failures of both "behaviors-for-attention" and "oral-keyword" are mainly caused by voice recognition errors; further, it is observed that "footswitch-and-posture" has an advantage in this regard. The fact that the success rate of "manual-editing" is not 100% indicates that editing undertaken by humans is not perfect; this is because online editing requires considerable skills and concentration.

The evaluation results for the criteria presented in Table 1 are shown in Figure 5. Although there is no significant difference, the general tendency can be observed in the graph. In the automated interfaces, "behaviors-for-attention" obtained the highest scores, and "footswitch-and-posture" rank second highest with regard to scoring. With regard to criterion (d)—"Do you want to use this interface if you get accustomed to it?"—the score of "behaviors-for-attention" is approximately equal to that of "manual-editing." This implies that behaviors-for-attention are also good triggers for detecting speakers' intentions for editing. In the interaction-form, we encounter a problem; with regard to criterion (a)—"Can you edit to suit your requirements perfectly?"—the score of "behaviors-for-attention" is observed to be the lowest. This is because the specified context of the questions and answers does not always provide the speaker with a good opportunity to use behaviors-for-attention, and occasionally, the speaker fails to perform one of the behaviors-for-attention that can be recognized by the system. Based on these results, we can conclude that behaviors-for-attention are good triggers for editing by a speaker, and adding a footswitch is of assistance as a supplementary interface, particularly with regard to interactive communication.

Table 2. Success rate of editing. "Switch" means the success rate of switching from a medium shot to a close-up shot, and "Switch&Back" means the success rate of besides from the close-up shot to the medium shot.

Detailed-scenario				
	(A)	(B)	(C)	(D)
Switch	94.8 %	78.1 %	86.4 %	100.0 %
Switch&Back	92.7 %	75.0 %	85.4 %	93.8 %
Outline-scenario				
	(A)	(B)	(C)	(D)
Switch	96.6 %	88.2 %	84.5 %	95.9 %
Switch&Back	94.3 %	83.9 %	81.6 %	89.8 %
Interaction				
	(A)	(B)	(C)	(D)
Switch	83.7 %	82.4 %	81.6 %	96.0 %
Switch&Back	81.6 %	78.4 %	73.5 %	90.2 %

(A) behaviors-for-attention (B) oral-keyword
(C) footswitch-and-posture (D) manual-editing

In a free format questionnaire, many subjects wrote that they received delayed responses after they provided a trigger to the system, particularly with regard to "behaviors-for-attention" and "oral-keyword." This is caused by the delay in speech recognition. In order to solve this problem, we are planning to use an open-source speech recognition software developed at our center, and make appropriate improvements to this software for detecting triggers. On the contrary, one subject wrote that he preferred the interfaces for self-editing to those for "manual-editing" because he was not comfortable with unwanted camera selections by the human editor. This is one of the interesting opinions that demonstrate the advantages of self-editing.

Evaluation of Edited Videos

As mentioned above, the score of "behaviors-for-attention" was the lowest for criterion (a) in the interaction-form. In order to examine the effect of this result on viewers, we asked the interrogators to score each of the edited videos transmitted from the speaker's site. The subjects were not told which interface was used at each trial. The questionnaire are listed in Table 3. The result of the evaluation is shown in graph on the left-hand side of Figure 6. "Behaviors-for-attention" obtained a low score for criterion (f). Based on this, we conjecture that beginners find it relatively difficult to estimate the timing of switching or delay using the behaviors-for-attention, and this leads to unorganized presentations. This implies that if subjects have sufficient practice, their presentations and videos appear smooth and natural.

In order to confirm this, one of the authors who got accustomed to all of our interfaces performed presentations, and examined the scores. We gathered 12 other subjects, and asked them to score the videos edited using the 4 methods. The results are shown in graph on the right-hand side of Figure 6. The scores of "behaviors-for-attention" increased and were satisfactory.

Through these experiments, we conclude that the obtained videos may appear relatively unnatural if a beginner uses the interface by employing the behaviors-for-attention. However, this can be overcome by training. It is also interesting to note that although the score of the video is relatively unnatural, the individual making the presentation has made a good impression, as shown in Figure 5.

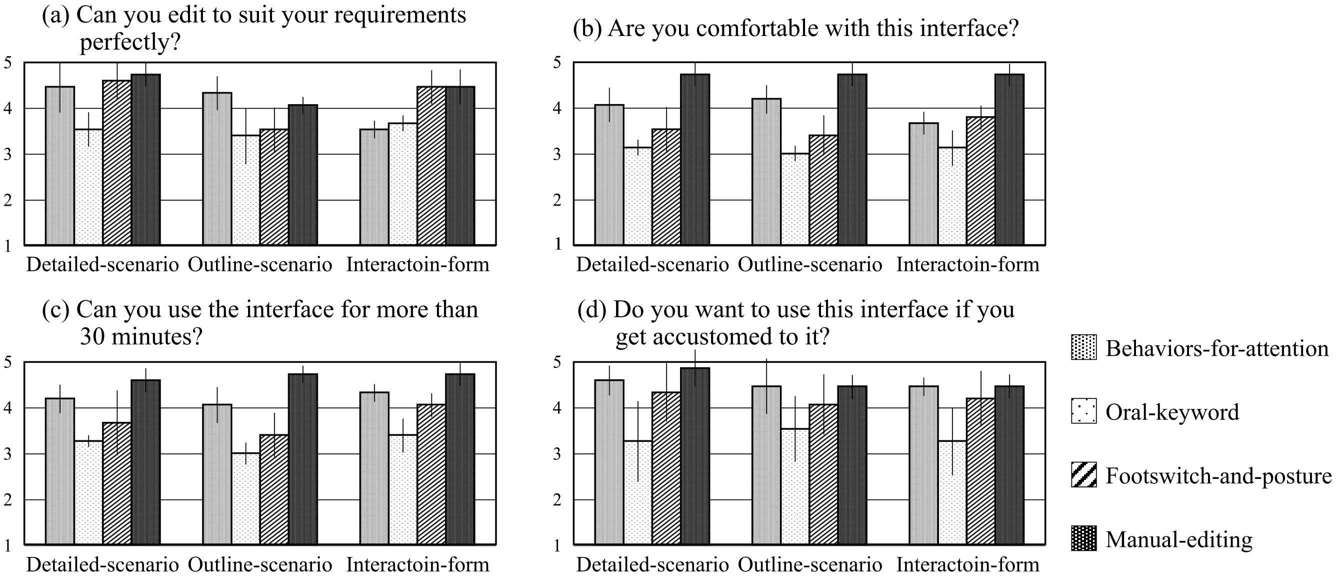
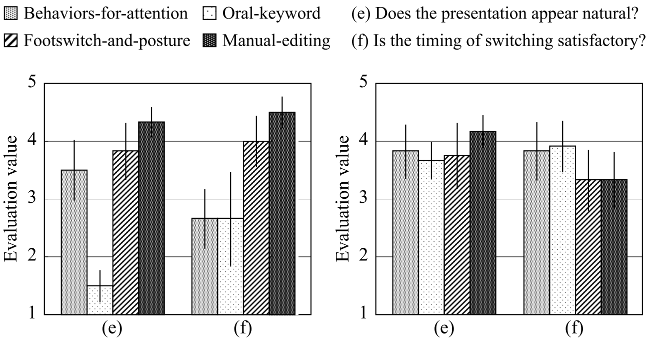


Fig. 5. Results of questionnaire listed in Table 1. A larger value indicates better scores. Each error bar means standard deviation.

Table 3. Questionnaires for viewers.

(e) Does the presentation appear natural? (f) Is the timing of switching satisfactory?



The videos edited by the subjects unaccustomed to the interfaces.

The videos edited by the author accustomed to all interfaces.

Fig. 6. Results of questionnaire listed in Table 3. Ehe error bar means standard deviation.

5. CONCLUSION

We investigated several types of user interfaces for automatic video editing based on the speaker's intentions. We compared 4 types of editing methods using 3 types of presentations. The results revealed that "behaviors-for-attention" obtained a high score when a scenario was provided, and "footswitch-and-posture" obtained a high score without a scenario. Consequently, using both "behaviors-for-attention" and "footswitch-and-posture" is good way for ensuring variety in a presentation. The experiment reported in this paper constitutes the first step toward self-editing. A considerable number of further evaluations are required, e.g., examining the differences based on the contents, types, and length of presentations, experiences on the system, and etc.

The results obtained here can be applicable to other kinds of video materials and lectures. For example, in a lecture with a blackboard, a lecturer often uses behaviors-for-attention expecting that the pointed portion is well paid attention also in remote sites. We can use our system in combination with a manual editing, e.g., making suggestions of where to make the edits while allowing the human editor to make the final decision.

6. REFERENCES

[1] M. Ozeki, Y. Nakamura, and Y. Ohta, "Automated camerawork for capturing desktop presentations," *IEE Proc. Vision, Image & Signal Processing*, 2005.

[2] M. Ozeki, Y. Nakamura, and Y. Ohta, "Video editing based on behaviors-for-attention – approach to professional editing by a simple scheme –," *Proc. ICME*, pp. TP9–4(CDROM), 2004.

[3] A. Girgensohn et al., "A semi-automatic approach to home video editing," *Proc. UIST '00*, ACM Press, pp. 81–89, 2000.

[4] M. Gleicher and J. Masanz, "Towards virtual videography," *Proc. ACM Multimedia*, pp. 375–378, 2000.

[5] N. Babaguchi et al., "Personalized abstraction of broadcasted american football video by highlight selection," *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 575–586, 2004.

[6] Q. Liu, Y. Rui, A. Gupta, and JJ. Cadiz, "Automating camera management for lecture room environments," *Proc. ACM CHI*, pp. 442–449, 2001.

[7] Y. Rui, A. Gupta, and J. Grudin, "Videography for telepresentations," *Proc. ACM SIGCHI*, vol. 5, no. 1, pp. 457–464, 2003.

[8] M. Ozeki, Y. Nakamura, and Y. Ohta, "Human behavior recognition for an intelligent video production system," *Proc. Pacific-Rim Conf. on Multimedia*, pp. 1153–1160, 2002.