Human Behavior Recognition for an Intelligent Video Production System

Motoyuki Ozeki, Yuichi Nakamura, and Yuichi Ohta

IEMS, University of Tsukuba, 305-8573, Japan
{ozeki, yuichi, ohta}@image.esys.tsukuba.ac.jp

Abstract. We propose a novel framework for automated video capturing and production for desktop manipulations. We focus on the system's ability to select relevant views by recognizing types of human behavior. Using this function, the obtained videos direct the audience's attention to the relevant portions of the video and enable more effective communication. We first discuss significant types of human behavior that are commonly expressed in presentations, and propose a simple and highly precise method for recognizing them. We then demonstrate the efficacy of our system experimentally by recording presentations in a desktop manipulation.

1 Introduction

There is now a great demand for audiovisual or multimedia contents in various fields. Content production is, however, a difficult task, which requires both considerable cost and skills. For example, a number of assistants and considerable time for recording and editing are often needed for audiovisual education. Thus, it is widely recognized that automated video production is one of the key technologies on multimedia.

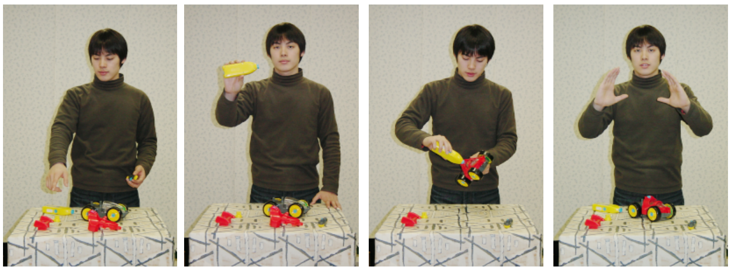
For this purpose, we are investigating a framework for effectively capturing presentations and producing comprehensible videos for teaching/operating /instruction manuals. We have so far constructed the framework's camera system that allows the appropriate video capturing of the targets[1][2]. As the next step, we need a mechanism for emphasizing *the focus of attention* which the members of the audience are expected to recognize.

In this paper, we will first consider the relation between the focus of attention and human behaviors regarding desktop manipulations, and then propose a multimodal method for detecting the focus. We will then present some experiments which demonstrate the performance of our system.

2 Communicating the Focus of Attention

2.1 Research Objective

For desktop manipulation, we assume the following situation as shown in Figure 1:



(a) pointing (b) holding-out (c) manipulation (d) illustration

Fig. 1. Typical types of behavior in presentations

- One person is speaking and presenting desktop manipulation.
- There are virtual audiences, but presenter does not receive questions from them in realtime.

This situation is common to a variety of video presentation contexts, *e.g.,* video manuals or cooking shows. The objective of our research is to realize, in such situations, a *virtual cameraman* who shoots and captures important portions, and to realize a *virtual editor* who selects important shots and edits the videos.

This paper discusses the latter problem: how we can make the most use of a multi-view (multi-angled) video. The key points of this topic regard tagging and editing. Tagging is the problem of describing what information is included in a video at a certain time. Editing is the problem of selecting relevant portions in the sense of time and view, and of making a video suitable for a given purpose.

As one important approach to this topic, we investigated the detection and utilization of a speaker's typical behaviors: detecting a speaker's behaviors which are intended to draw the viewers' attention, tagging in terms of the recognition of those behaviors, and editing by selecting the most appropriate view.

Several related works deal with lectures using automated video capturing or archive systems[3]–[8]. For our purpose, however, different approaches are required:

- The targets that should be captured in desktop manipulations are different from those in lecture scenes. Some targets, such as hands, move fast and in complicated patterns. The combination of simple tracking and simple view switching may result in a shaky and unpleasant video.
- Typical types of behavior that appear in desktop manipulations are different from those of lecture scenes. As we will discuss below, we have to focus on important behaviors that are not considered in the above-mentioned studies.

2.2 System Overview

Figure 2 shows the basic system derived from our research. For realtime measurement of the speaker's position and the object's positions, we currently use

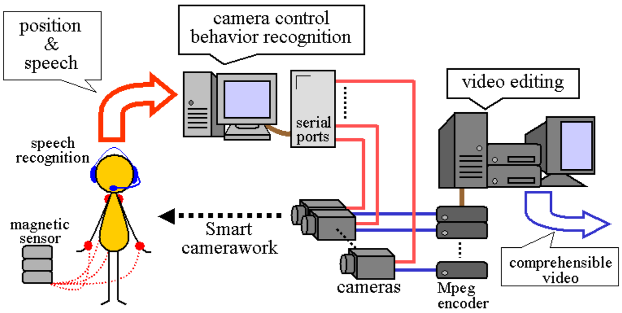


Fig. 2. Overview of the system

magnetic sensors. Pan/tilt cameras are controlled by using the measured positions, and videos taken by those cameras are transmitted, switched, and recorded in MPEG format. For behavior recognition, the system uses a speaker's movements and speech recognition output.

The result of behavior recognition are not only used for tagging the captured videos, but are used for switching to the most appropriate output view. This enables a speaker to give a presentation while checking the status of the system. The switched view can be directly presented to viewers, or the system can edit the recorded video afterward based on the obtained tags.

3 Important Types of Behavior in Desktop Manipulation

The following types of behavior frequently appear in presentations, aimed at drawing the viewers' attention:

pointing: Pointing with one's hand forces the audience to look at the directed area, as shown in Figure 1(a). This corresponds to *deictic movement* in Ekman's classification [9]. The focus is on the indicated object, location, or direction.

holding-out: Holding out, or presenting, an object toward the audience, usually at a position higher than the waist and lower than the eyes (Figure 1(b)). The focus is on the held object.

manipulation: Demonstrating important operations is a typical behavior, as shown in Figure 1(c). It can be a virtual manipulation belonging to *illustrators* in Ekman's classification. The focus is on the manipulation.

illustration: Illustrating a shape, size, or motion by moving hands draws the viewers' attention to it, as shown in Figure 1(d). This also corresponds to *illustrators* in Ekman's classification. The focus is on the locus or the motion of the hands.

Since discrimination between manipulation and illustration is sometimes difficult in actual presentations and their functions are similar, hereafter we classify them together in this paper. In regard to pointing, we currently deal only with pointing at an object within the presenter's reach[1]. Since this diminishes the difference between pointing and holding-out, we also classify these two behaviors together.

4 Behavior Recognition

We have to deal with the above two important types of behavior, pointing/holding-out and manipulation/illustration. For this purpose, we propose simple and fast methods utilizing using motion and speech clues. If the system detects both speech clues and motion clues within a certain period, the system accepts them as a corresponding behavior. We previously investigated the occurrence between motion clues and speech clues[10], and the statistics showed that they cooccur within 2 seconds in around 90% cases. Since the speech recognition sometimes has a delay longer than 2 seconds, we set the tolerance of the delay to 3 seconds.

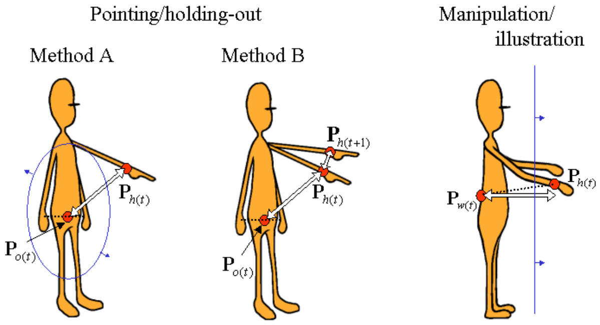


Fig. 3. Behavior recognition

4.1 Motion Clues for Pointing/Holding-out

One of the most distinct features of pointing/holding-out behavior is the arm position: an arm is stretched in front of a speaker. Two other features concern arm motions: a quick stretch of an arm and a sudden stop. We compared two methods, method A and method B, which use arm position, and both arm position and hand motion, respectively.

Using Arm Position (Method A): The system simply detects moments when the arm stretches beyond the threshold. When both hands are stretched and they are close to each other, the system regards the movement as a pointing/holding-out behavior using both hands. If both hands are apart from each other, the

[1] Pointing at an object beyond the presenter's reach is left for future research, since another research is required for delineating the location of the indicated object.

system regards the movement as a pointing/holding-out behavior with a single hand whose position is higher than the other's.

Arm Stretch (AS) is calculated using the following equation, as shown on the left in Figure 3.

$$AS_{(t)} = |\mathbf{P}_{h(t)} - \mathbf{P}_{o(t)}|$$

where $\mathbf{P}_h$ is the hand position, and $\mathbf{P}_o$ is the position when the hand is put down.

Using Arm Motion (Method B): If the system focuses only hand position, it can mis-detect other movements to which a speaker does not intend to call attention. To cope with this problem, in method B the system also checks a quick arm stretch and a sudden stop. If the above features are detected for both hands simultaneously, the system regards the movement as pointing/holding-out behavior utilizing both hands.

Arm Stretch Change (ASC), *Pseudo Velocity (PV)*, and *Pseudo Velocity Change (PVC)* are calculated using the following equations, as shown at the center of Figure 3.

$$ASC_{(t)} = AS_{(t)} - AS_{(t-1)}$$

$$PV_{(t)} = |\mathbf{P}_{h(t)} - \mathbf{P}_{h(t-1)}| \ , \ PVC_{(t)} = PV_{(t)} - PV_{(t-1)}$$

4.2 Motion Clues for Manipulation/Illustration

Since a manipulation/illustration movement is originally a simulation or a demonstration of movements or shapes, there is no fixed pattern for it. To deal with this behavior, we are currently using hand position, whether the hands are on/above the desk. It is calculated using the following equation, as shown on the right of Figure 3.

$$|\underline{\mathbf{P}}_{h(t)} - \underline{\mathbf{P}}_{w(t)}| > Th_{wh}$$

where $\mathbf{P}_w$ is the position of the speaker's waist, and $\underline{\mathbf{P}}$ means the horizontal component of $\mathbf{P}$.

4.3 Speech Clues

Speech suggests the presence of an important behavior, and some types of speech also specify the focus of attention. For example, phrases that include a demonstrative pronoun/adjective/adverb, such as "this(is a ...)", "this (switch)", or "this long", frequently appear in speech, and they strongly suggest a focus of attention.

Figure 4 shows the speech clues that we are currently using, and shows the foci of attention suggested by the clues. The balloon located above each arrow gives the words and corresponding movements that trigger the switching. And the focus of attention is shown above each image. Since this system is designed for Japanese, the Japanese words are the targets of speech recognition.

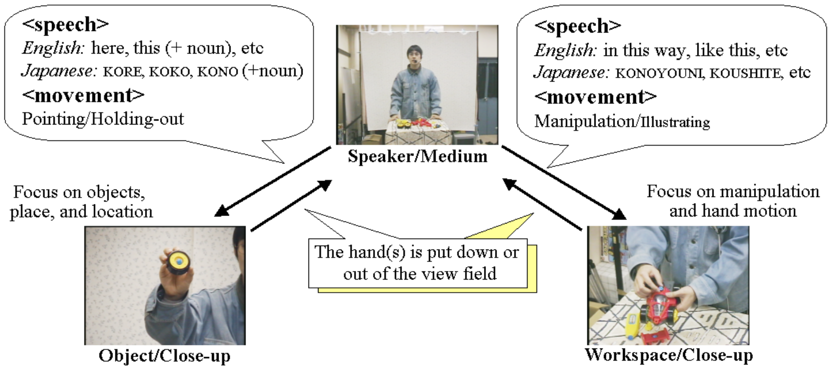

<speech>
English: here, this (+ noun), etc
Japanese: KORE, KOKO, KONO (+noun)
<movement>
Pointing/Holding-out
Speaker/Medium
<speech>
English: in this way, like this, etc
Japanese: KONOYOUNI, KOUSHITE, etc
<movement>
Manipulation/Illustrating
Focus on objects, place, and location
The hand(s) is put down or out of the view field
Focus on manipulation and hand motion
Object/Close-up
Workspace/Close-up

Fig. 4. Typical examples of speech, and video switching condition

5 Experiments

5.1 Evaluation of Behavior Recognition

We examined the performance of our behavior recognition methods by applying our system to real presentations. We gathered 6 students without any professional experience in teaching or giving instructions. Each subject was asked to give a demonstration of the assembly of a toy car. Before the presentation, the subjects were briefed on how the system works, and were asked to express their intentions clearly by means of motion and speech. The subjects were able to see the system's responses by looking at the switched view that indicates the results of focus detection.

Some portions of the video edited according to behavior recognition are shown in Figure 5. In this experiment, three cameras were used, which captured close-up shots of an object held by the speaker, middle shots of the speaker, and close-up shots of the workspace. The switching conditions are illustrated in Figure 4. As we can see in Figure 5, the system properly recognized the focus of attention, producing an effective video.

The left side of Table 1 shows the performance of our system. Without prior training, around 70% of the subjects' behaviors are correctly recognized. All detection failures of method A arose from speech recognition errors. The same can be said for manipulation/illustration.

For the evaluation of our system as a user interface, we asked each subject to fill out a questionnaire after the experiment. Roughly speaking, the subjects had positive impressions of the system. Once we explained the system mechanism, all the subjects were able to quickly adjust to giving presentations on the system. Most of subjects stated that they were not severely constrained by the system and that they were satisfied with the obtained videos. In regard to the detection methods for pointing/holding-out, most of the subjects stated that method A is better than method B because of its better detection rate and fewer constraints.

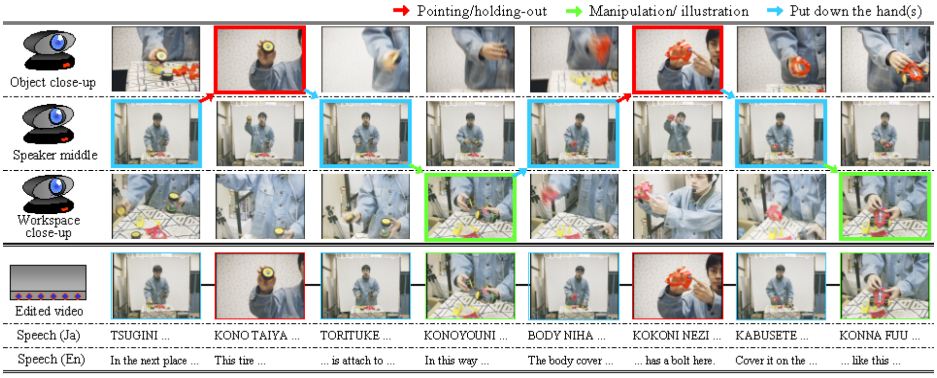


Fig. 5. Example of video switching
The phrases below the images are the transcribed speech. The upper line shows the actual speech in Japanese, and the lower shows the translation into English. The sample movies (mpeg format) can be obtained at
http://www.image.esys.tsukuba.ac.jp/˜ozeki/e_movies.html.

Table 1. Recognition result for real presentations (left side) and comparison of automatic editing (right side)

P/H (MA)		P/H (MB)		M/I	
R(%)	P(%)	R(%)	P(%)	R(%)	P(%)
75	94	64	98	75	98

P/H:Pointing/Holding-out
M/I:Manipulation/Illustration
MA(B):using Method A(B) for P/H
R:Recall, P:Precision

Editing Method	Matched frames#(%)			
	P1	P2	P3	Total
Speech only	52.1	57.4	54.9	55.0
Motion only (MA)	71.7	48.9	54.4	57.4
Motion only (MB)	64.1	61.5	47.0	57.2
Motion and Speech	78.8	87.9	80.2	82.6
Random Editing	50.7	46.0	39.0	44.9

5.2 Evaluation of Automatic Editing

We verified our video editing scheme by subjective evaluation. For this purpose, we captured three kinds of desktop manipulations – P1: assembling a toy car, P2: attaching an I/O adopter to a notebook PC, and P3: cooking a sandwich (emulation), each of which are from 50 to 60 seconds long. Three kinds of edited videos created from these data are compared:

- A video manually edited by one of the authors who can clearly recognize the speaker's behavior.
- The automatically edited video that matches best the manually edited video.
- A randomly edited video in which video switching is periodic. The interval is the average of that of the manually edited one. The probability for each shot is close to[2] that of the manually edited video.

The right side of Table 1 shows the results of the comparison. Each figure in the table shows the number of frames in which the same shot is chosen as

[2] Since the number of the shots are finite, we cannot arbitrarily set this probability.

in the manually edited videos. As we can see here, editing the results obtained by using speech and motion clues shows the best match to the manually edited video. This proves that our multimodal recognition method is more accurate than other methods such as that which relies only on speech.

6 Conclusion

This paper introduced our novel framework for intelligent video capturing and production. We discussed typical types of behavior intended to draw the viewers' attention and proposed multimodal recognition methods. Experimentally, our simple and fast methods demonstrated good performance. The subjects were generally satisfied with our system and with the obtained videos.

As a goal of our future research, we will attempt to achieve a more detailed behavior recognition. For editing videos in wider variety of ways, much more information is required; for example, the system needs to recognize what a speaker is doing during manipulation/illustration movements.

References

1. M. Ozeki, Y. Nakamura, and Y. Ohta, "Camerawork for intelligent video production – capturing desktop manipulations," *Proc. ICME*, pp. 41–44, 2001.
2. M. Ozeki, M. Itoh, Y. Nakamura, and Y. Ohta, "Tracking hands and objects for an intelligent video production system," *Proc. ICPR*, pp. 1011–1014, 2002.
3. L. He et al., "Auto-summarization of audio-video presentations," *Proc.ACM Multimedia*, pp. 489–498, 1999.
4. S. Mukhopadhyay and B. Smith, "Passive capture and structuring of lectures," *Proc.ACM Multimedia*, pp. 477–487, 1999.
5. Y. Kameda, M. Minoh, et al., "A study for distance learning service - tide project -," *Proc. International Conference on Multimedia and Expo*, pp. 1237–1240, 2000.
6. Y. Kameda, M. Mihoh, et al., "A live video imaging method for capturing presentation information in distance learning," *Proc.International Conference on Multimedia and Expo*, 2000.
7. A. Bobick, "Movement, activity, and action," *MIT Media Lab Preceptual Computing Section*, vol. TR-413, 1997.
8. A. Bobick and C. Pinhanez, "Controlling view-based algorithms using approximate world models and action information," *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 955–961, 1997.
9. P. Ekman and W. Friesen, "The repertoire of nonverbal behavior : Categories, origins,usage,and coding," *Semiotica*, vol. 1, pp. 49–98, 1969.
10. Y. Nakamura et al., "MMID: Multimodal multi-view integrated database for human behavior understanding," *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 540–545, 1998.
11. K. Ohkushi, T. Nakayama, and T. Fukuda, *Evaluation Techniques for Image and Tone Quality (in Japanese)*, chapter 2.5, SHOKODO, 1991.