Conversational Contents Acquisition — Sensing and capturing of a variety of human activities anytime, anywhere

Yuichi Nakamura[1] and Yoshinari Kameda[2]

[1] Academic Center for Computation and Media Studies, Kyoto University,
Sakyo, Kyoto, 606-8501, JAPAN
yuichi@media.kyoto-u.ac.jp

[2] Dept. of Systems and Information Engineering, University of Tsukuba,
1-1-1 Tennodai, Tsukuba, 305-8573, JAPAN
kameda@image.esys.tsukuba.ac.jp

Abstract. This paper introduces our works that aim to realize conversational contents. Capturing systems for conversations, presentation, personal experiences have been developed in our laboratory, and they usually show good performance for acquiring good sources of conversational contents. We introduce the basic idea of our framework, actual systems, and our future plan.

1 Introduction

Introducing conversation functions is an attractive and promising approach for realizing advanced multimedia for a variety of applications. As an essential topic for this approach, we have been investigating automatic acquisition of video based multimedia contents, say *conversational contents*, that can be used as e-Learning, training, video manuals, etc. Such data acquisition is one of the most important issues, since those types of contents requires flexible interaction with a variety of presentation forms using a variety of data in response to questions or interaction with the users.

In this paper, we first present our works for capturing and indexing conversation, presentation, and personal experiences. In those works, we placed much emphasis on taking data in various ways from the real world. Those works are approaches for relatively small-scale or controlled situations, and we are planning further research in large-scale or non-restricted situations. The concept of this approach will also be presented in this paper.

2 Capturing Conversation and Presentation

The most straightforward data acquisition for conversational contents is capturing conversation scenes such as meetings. Recorded data, for example, can be directly used for meeting minutes, with which we can review details of a meeting

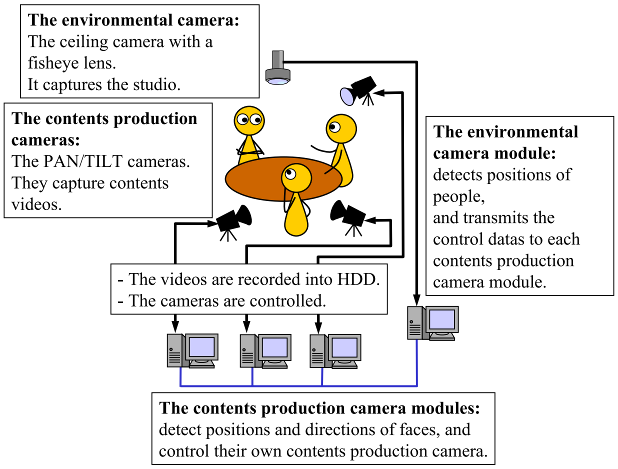

The environmental camera:
The ceiling camera with a fisheye lens.
It captures the studio.
The contents production cameras:
The PAN/TILT cameras.
They capture contents videos.
The environmental camera module:
detects positions of people,
and transmits the control datas to each contents production camera module.
- The videos are recorded into HDD.
- The cameras are controlled.
The contents production camera modules:
detect positions and directions of faces, and control their own contents production camera.

Fig. 1. Conversation capturing system

on our demands. Not a few projects have started for realizing such automatic meeting recorders that enable flexible accesses to recorded data, *e.g.*, "who spoke what", "how they agreed on xxxx", etc.

Moreover, since a conversation is one of the most common ways for human communication, the archived data are useful also as explanations concerning a variety of topics. We are concentrating much on this point, and we use each movie clip of a conversation scenes as a movie clip for a presentation.

For this purpose, we are investigating meeting capturing and editing. One important point for this archiving process is to obtain attractive shots as seen in movies or TV programs without disturbing natural conversations, and another is to obtain useful indices for data retrieval on demand. To meet those demands, we constructed a system as shown in Figure 1 that observes the positions of talking people, and automatically controls cameras for filming them with appropriate picture compositions[1].

Our another research topic is presentation archiving. Presentations are good sources for e-Learning or manuals, and we are constructing an automated system for capturing, editing, and question answering. For this purpose, we need to capture much more details of human behaviors than conversation cases. For example, we need a close-up shot of a specific parts with detailed explanation if we want to learn an assembly of electronic circuits.

We first developed an automated video capturing and indexing system for a desktop presentation as shown in Figure 2[2], and showed automatic editing is possible by recognizing typical behaviors of a presenting person[3]. One key point of this work is that we realized a flexible camerawork that can be adapted to various purposes of capturing by adjusting its parameters. Another point is that we introduced a new editing scheme that uses *behavior-of-attentions*,

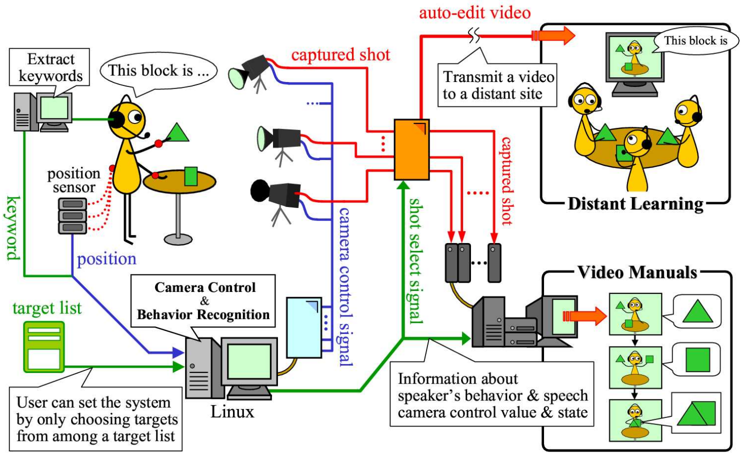
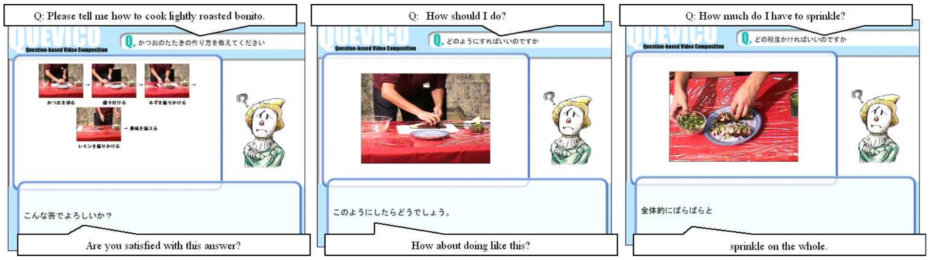

auto-edit video
This block is
Transmit a video to a distant site
Distant Learning
Extract keywords
This block is ...
captured shot
position sensor
keyword
position
camera control signal
shot select signal
captured shot
Video Manuals
target list
Camera Control & Behavior Recognition
User can set the system by only choosing targets from among a target list
Linux
Information about speaker's behavior & speech camera control value & state

Fig. 2. Automated capturing system for desktop presentation


Q: Please tell me how to cook lightly roasted bonito.

Q. かつおのたたきの作り方を教えてください

こんな答でよろしいか？

Are you satisfied with this answer?

Q: How should I do?

Q. どのようにすればいいのですか

このようにしたらどうでしょう。

How about doing like this?

Q: How much do I have to sprinkle?

Q. どの程度かければいいのですか

全体的にぱらぱらと

sprinkle on the whole.

Fig. 3. QUEVICO: question answering scheme

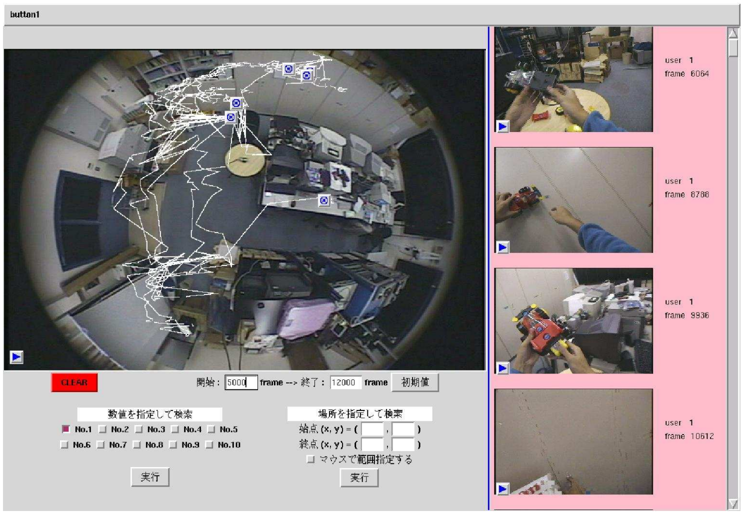
which are the behaviors to draw viewers' attention explicitly or suggestively in presentations[4].

Then, we have some experiences with our question answering scheme, and proved that answering by images or video clips, as shown in Figure 3, is possible for relatively simple tasks[5]. One key point is that the system determines which modality is suitable for answering each question, *i.e.* a video segment, a text, or an audio segment is appropriately chosen for explanation.

Another research is an intelligent classroom[6], which is more practically used to archive lectures that are held as regular courses in Kyoto University. We have developed advanced camera controlling methods[7][8][9] for this purpose.

3 Capturing Personal Experiences

Not only conversations or presentations are the sources for conversational contents. We can easily understand it because we want to tell our experiences to other people, and those frequently become main topics of our conversations.



開始：5000 frame --> 終了：12000 frame

数値を指定して検索

☒ No.1 □ No.2 □ No.3 □ No.4 □ No.5
□ No.6 □ No.7 □ No.8 □ No.9 □ No.10

場所を指定して検索

始点 (x, y) = (,)
終点 (x, y) = (,)
□ マウスで範囲指定する

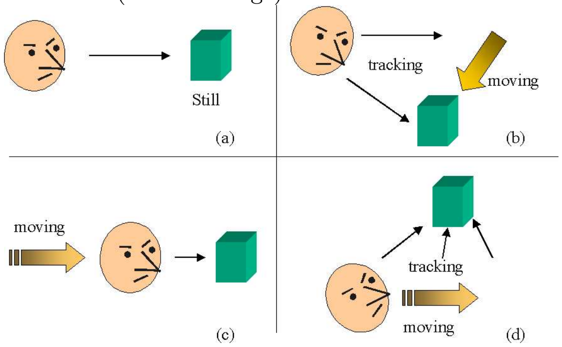
user 1
frame 6064

user 1
frame 8788

user 1
frame 9936

user 1
frame 10612

(a) Personal activity record (the right images) browsers with a surveillance camera view (the left image)



(b) behaviors for paying attention

Fig. 4. Video analysis for archiving personal activities

Capturing and archiving our activities anytime anywhere is a quite attractive issue. For this purpose, we proposed a novel method for analyzing video records captured by a head-mounted camera and an environmental camera as shown in Figure 4(a). This process aims to make retrieval of personal activities easier by detecting important portions from the videos. We showed that the user's behaviors that appear when he/she pays attention to something as shown in Figure 4(b) are useful for this purpose[10, 11]. The links between a wide-angled view from a ceiling camera and a detailed view from a head-mount camera are also effective for recalling experiences such as "who did what and where".

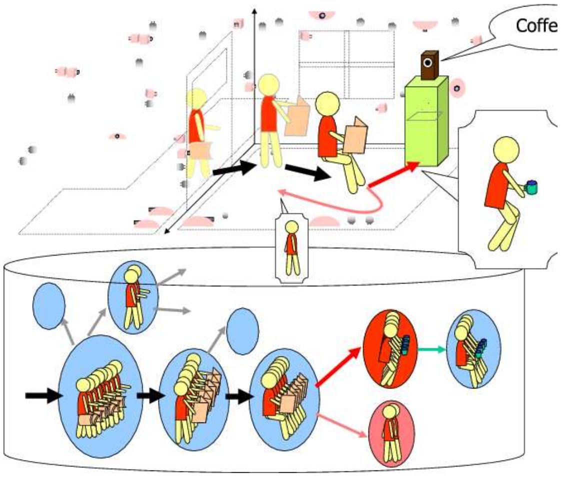


Fig. 5. Massive sensors for capturing human life space: a number of uncalibrated sensors are scattered, and they capture and correlate a large amount of data.

4 Capturing Less Constrained Scenes

We employed, so far, custom-made sensors for each of the above situations. We need further research for sensing and capturing in less constrained environments, since a variety of events would happen at a variety of places.

Our new approach is to develop a novel framework that employs a large number of sensors and massive data. Suppose that image sensors, *e.g.*, CCD cameras, and acoustic sensors, *e.g.*, microphone arrays, are arbitrarily and non-uniformly distributed as shown in Figure 5 and they are initially uncalibrated. Our goal is to establish a new theory by which any events can be detected and recorded by those sensors. The most challenging portion is to retain consistency of observation over sensors, which is achieved by identifying sensors each other, exchanging live data with other sensors, and cooperating each other to assure event detection. Thus, we expect that massive and scattered sensors, their automatic calibration, and their learning capability can be a breakthrough for capturing arbitrary events in a large-scale space.

5 Summary

We briefly introduced our approaches for acquiring conversational contents. Some are concerning capturing videos and indexing for conversation and presentation situations, another is for capturing personal experience by a head-mount camera and environmental cameras. We are also investigating further mechanism for more general situation in a large-scale space.

References

1. T.Nishizaki, et al., “Video Contents Acquisition and Editing for Conversation Scene”, Proc. Eighth Int’l Conference on Knowledge-Based Intelligent Information & Engineering Systems (to appear), 2004
2. M. Ozeki, Y. Nakamura, and Y. Ohta. ”Capturing Desktop Presentations — Camerawork design and evaluation in virtual and real scenes”, Proc. 1st European Conference on Visual Media Production, pp.211-220, 2004
3. R.Ogata, et al., ”Computational Video Editing Model based on Optimization with Constraint-Satisfaction”, Proc. Pacific-Rim Conference on Multimedia, 2003
4. M. Ozeki, Y. Nakamura, and Y. Ohta. ”Video Editing based on Behaviors-for-Attention – Approach to professional editing by a simple scheme –”, Proc. IEEE Int’l Conf. on Multimedia and Expo, TP9-4(cdrom), 2004.
5. H.Izuno, et al., “QUEVICO QA Model for Video-based Interactive Media”, Proc. International Workshop on Content-Based Multimedia Indexing, 2003
6. Y.Kameda et al., “CARMUL: Concurrent Automatic Recording for Multimedia Lecture”, IEEE Int’l Conf. on Multimedia and Expo, Vol.1, pp.129-132, 2003
7. Y.Kameda, et al., “Reduction of Camera Motion Adjustments under a Planned Video Composition with Pan-Tilt Camera”, Proc. Asian Conference on Computer Vision 2004, Vol.1, pp.216-221, 2004
8. S.Nishiguchi, et al., “A Sensor-fusion Method of Detecting A Speaking Student”, IEEE Int’l Conf. on Multimedia and Expo, Vol.1, pp.677-680, 2003
9. T. Rutkowski, et al., “Identification and Tracking of Active Speaker’s Position in Noisy Environments”, Proc. Int’l Workshop on Acoustic Echo and Noise Control, pp.283-286, 2003
10. Y.Nakamura, J. Ohde, Y.Ohta, “Structuring Personal Activity Records based on Attention — Analyzing Videos from Head-mounted Camera”, Proc. 15th International Conference on Pattern Recognition, Track4, pp.220–223
11. S.Kubota, Y.Nakamura, Y.Ohta, “Detecting Scenes of Attention from Personal View Records” Proc. Workshop on Machine Vision and Applications, 2002