CAMERAWORK FOR INTELLIGENT VIDEO PRODUCTION – CAPTURING DESKTOP MANIPULATIONS

Motoyuki Ozeki *Yuichi Nakamura* *Yuichi Ohta*

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

ABSTRACT

In this paper, we introduce an intelligent system for video recording. First, we categorized targets and purposes of shooting, and discuss the cameraworks appropriate for them. Then, we propose camera control algorithms to realize such cameraworks. Based on this idea, we built a prototype pan-tilt camera control system, in which multiple cameras with different purposes automatically track and shoot the targets. We evaluated our system through recording of some presentations on desktop manipulation. The effectiveness of our algorithm was verified through some experiments.

1. INTRODUCTION

With the recent progress of multimedia technology, multimedia contents are widely recognized as useful teaching materials or instruction/operating manuals. Contents production is, however, a difficult task, which requires both considerable costs and skills. For producing videos, we need cameramen who shoot at targets with appropriate cameraworks and directors who intelligently select the best shots and arrange them. These costs of employing cameramen and directors are not usually affordable for small scale purpose, *e.g.* for preparing teaching materials.

To cope with this problem, we are investigating an intelligent system for automated video production. We first examined the cameraworks by considering "target to capture" and "aspect-of-target to capture" in the context of presentations such as instructions on desktop manipulations. Then, we built a prototype system with multiple pan-tilt cameras controlled based on the cameraworks. We applied the system to typical presentations on desktop manipulations, and verified the performance of the system.

2. CAPTURING A SCENE OF DESKTOP MANIPULATION

For automated video production, we have to tackle with the following problems:

camera control: We need *virtual cameramen* by an automated camera system for shooting at the right target with appropriate cameraworks. In most of the current systems with fixed cameras, an important portion is often out of the field or too small to be paid attention.

event recognition and video editing: We need *virtual directors* which intelligently chooses the best views and emphasize important portions. To automate this process, it is also essential to recognize the events occurring in a scene and to tag the captured videos.

We are developing a system which realizes the above functions. Fig.1 shows an overview. In this system, the 3D position

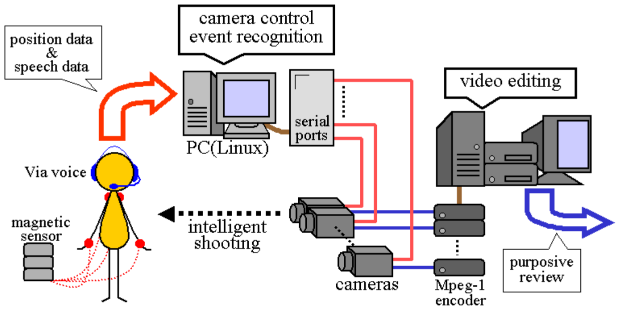


Figure 1: Overview of the system[1]

of a speaker is measured by magnetic sensors, and multiple pan-tilt cameras are controlled according to the speaker's position and movements. Videos taken by those cameras are transmitted and recorded in MPEG-1 format. This framework also includes event recognition process which outputs are used for video editing. According to the event recognition results, the system switches the views or chooses the relevant portions for an explanation. Thus the system gives views that the speaker wants to show or that viewers want to see.

In this paper, we will describe the above camera control portion: required cameraworks, camera control algorithms, and some experiments. As related researches, there are recent works on lecture archiving [1][4][5]. Although cameras are automatically manipulated in some of those systems, we usually need much more sophisticated cameraworks for effectively capturing presentations on desktop manipulations. Specifically, we have to deal with the following problems:

- Close-up shots are necessary to show the details of important objects or important manipulations. In taking such shots, a hand or an important object moves fast and sometimes goes out of view field frame quickly. We need intelligent camera control for fast and robust tracking.
- Moreover, we have various targets to be paid attention, for example, speaker's face, behavior and hands, objects and etc., and the camerawork required for each target is different from others.

In this sense, automated camera control is one of the most important and interesting topics for computerized video production.

3. CAMERAWORK FOR MANIPULATION ON DESKTOP

The purpose of camera control is basically to capture a target with appropriate size and at a good position in an image. The problem of tracking a target, however, is not so simple. For example, when we are shooting at a hand manipulating an object,

[1] Because of the lack of space, we have to attach reduced figures. Please take a look at http://www.image.esys.tsukuba.ac.jp/∼ozeki/ICME2001/index.html for the original ones.

suitable camerawork will be different between the following two purposes: (a) emphasize the details of the object; (b) emphasize the manipulation. For purpose (a), we prefer an extremely close-up shot in which the object is tracked and always kept at the center of the screen. For purpose (b), on the other hand, it is better to fix a camera angle so that we can easily understand the movements of the hand or the object. Consequently, we have to consider camera controls according to the subject and the purpose of a shot.

3.1. Target and Aspect

We consider cameraworks from two points of view: what *target* we want to shoot, and what *aspect-of-target* we want to capture. Basically, the above target is an object to be tracked by a camera, and the above aspect-of-target determines how to track it.

target: In presentations, there are several important targets which we have to pay attention. Currently, we consider four types of target.

<**speaker**>: a speaker, a lecturer, or an instructor.

<**workspace**>: a *dynamic* space where manipulation such as assembling or cooking is proceeding.

<**object**>: an important object to be paid attention.

<**place**>: an important *static* place to be paid attention.

For each target, we prepare three types of shots: long shot, medium shot, and close-up shot.

aspect-of-target: We categorized aspect-of-target as follows:

<**circumstance**>: Target's circumstance which includes position, trajectory, or spatial relationship to other objects. This is effective for giving the overview of a presentation or manipulation with a wide-angled view field.

<**movement**>: Movements of a target with frequent small motions such as hand motions in manipulations.

<**appearance**>: Target's appearance to be stared. As for presentations, a speaker often holds an important object toward the viewers in order to show the details of the object. In this case, it is necessary to capture the target at the center when the object motion stops.

In case of <circumstance>, it is required to fix a camera as long as possible so that viewers could easily observe target's position in a scene. In case of <movement>, it is required to track the target with suppressing small camera movements and get a stable view. In case of <appearance>, it is required to track as smoothly and quickly as possible with keeping the target at center of view field.

3.2. Camera Control Algorithms

Considering the above problems, we propose (1) camera motion smoothing by the Kalman filter and (2) camera motion suppression by *virtual-frame control*. We can adapt a camerawork for various purposes by tuning the parameters for the above methods.

Smoothing by the Kalman filter By smoothing, we expect that sensor noise and small irritating motions such as trembles are eliminated. For that purpose, we use the Kalman filter with the rigid body motion model as system dynamics. A state variable $\mathbf{x}_k$ and a state transition matrix $\mathbf{F}$ are as follows.

$$\mathbf{x}_k = \begin{pmatrix} x \\ \dot{x} \\ \ddot{x} \end{pmatrix} \qquad \mathbf{F} = \begin{pmatrix} 1 & \Delta & \frac{1}{2}\Delta^2 \\ 0 & 1 & \Delta \\ 0 & 0 & 1 \end{pmatrix}$$

where, Δ is a sampling interval of a measurement. $\mathbf{x}_k$ is a state vector containing the current values of position, velocity, and acceleration.

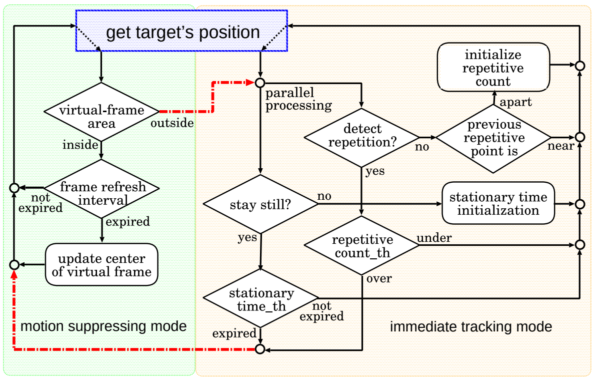


Figure 2: Flow of the virtual-frame control algorithm

The behavior of the Kalman filter we used depends on the ratio of the process noise variance to the measurement noise variance. We consider this ratio (hereafter abbreviated as *noise_variance_ratio*) as one of the camera control parameters which govern the smoothness of tracking. If the ratio is small, the camera tracks more smoothly. On the contrary, the camera tracks more precisely, if the ratio is large.

Virtual-frame control The virtual-frame control algorithm switches the tracking mode to *immediate tracking mode* when the target goes outside the virtual-frame, and switches back to *motion suppressing mode* when the target stays still or a repetitive motion is observed. In immediate tracking mode, a camera quickly and exactly track the target. In motion suppressing mode, camera motion is suppressed while a target stays in a virtual-frame assumed on an image.

This virtual-frame is a rectangle placed at the center of an image, and its size is specified by the ratio(*virtual_frame_size*) to the image size. A camera is moved so that the center of a virtual-frame is located at the target's average position during the last few seconds (*frame_refresh_interval*). The followings are the triggers for switching the tracking mode to motion suppressing mode.

stationary target position: A target is within a small region (*stationary_range_th*) for over a certain duration (*stationary_time_th*).

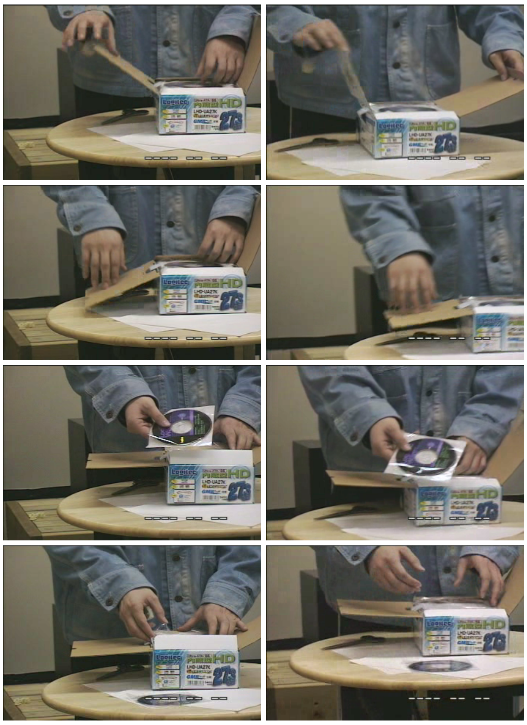
repetitive target motion: A target is moved repeatedly over a certain count (*repetition_count_th*), for example, an object is shaken by a hand. This repetition is detected by checking the sign changes of the target's motion vector.

Fig.2 shows the flow of the algorithm. If virtual_frame_size or frame_refresh_interval is large, a camera angle tends to be fixed. This causes inexact tracking and fixed views. Similarly, if we make any of stationary_time_th, stationary_range_th, and repetition_count_th small, we also get more stable views.

Fig.3 shows the result of shooting at a desktop manipulation in which a person opens a box. The left column shows the sequence of images captured without the virtual-frame control, and the right column shows those with the virtual-frame control. As we can see here, most of uncomfortable view field movements are eliminated in (b).

3.3. Setting of The Camera Control Parameters

The relations between aspect-of-target and the camera control parameters are shown in Fig.4. In case of <circumstance>, we use large virtual_frame_size, long stationary_time_th, and small noise_variance_ratio so that the viewers can easily grasp the circumstances in which a target is moving. Since the tracking be-



(a)With the virtual-frame control (b)Without the virtual-frame control

Figure 3: Manipulation on opening a box (the right hand is tracked)

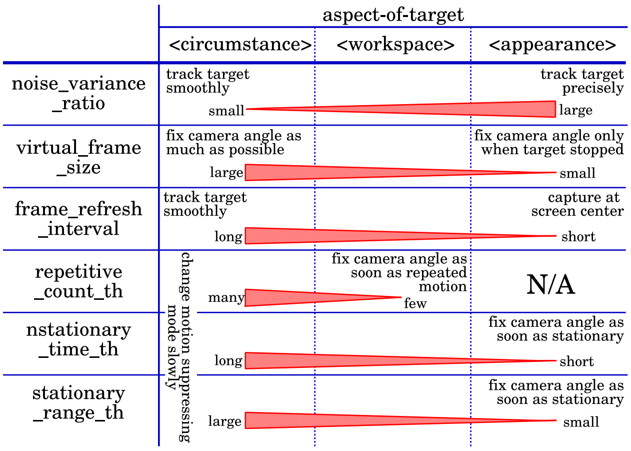


Figure 4: Correspondence between the aspect-of-target and the camera control parameters

comes inexact with large virtual_frame_size, we need to make the stationary_time_th, the stationary_range_th, and the repetition_count_th large so that the tracking mode cannot easily be switched to motion suppressing mode. In case of <movement>, the repetition_count_th is set small in order to quickly detect repetitions. Virtual_frame_size and frame_refresh_interval are also set small expecting that the target is captured at the center of the image. For <appearance>, the virtual_frame_size and frame_refresh_interval are set small in order to capture the target at the center of the image as long as possible. Additionally, to quickly stop the camera motion when a target stops, the stationary_time_th and the stationary_range_th are set small, and the noise_variance_ratio of the Kalman filter is set large.

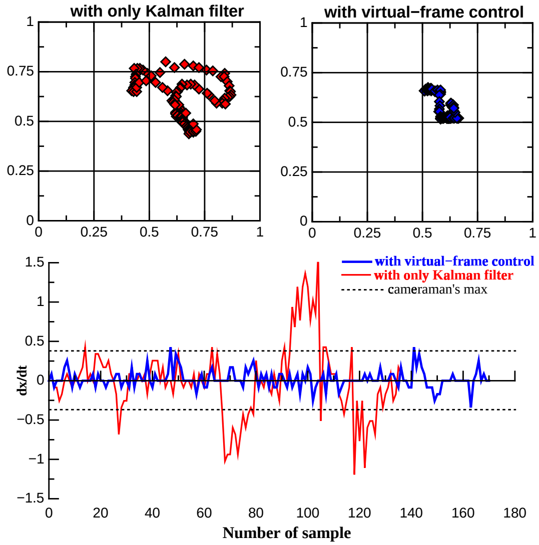


Figure 5: Target's normalized position and velocity with/without frame-control (upper: normalized position, lower: normalized horizontal velocity)

Table 1: The Camera Setting Using Examples of Shoot

Camera	Target Label	Control Method	Basing point
Camera1	speaker/M	<circumstance>	a point on the waist
Camera2	object/C	<appearance>	a point on the right hand
Camera3	workspace/M	<movements>	a middle point of both hands

M: Medium, C: Close-up

4. EXPERIMENTS

As already shown in Fig.1, our system has multiple pan-tilt cameras[2] controlled by a host computer. 3D position of a target is measured by a magnetic positional sensor[3] with the frequency of 30Hz. Although we did not mention the event recognition in this paper, the system detects deictic, pointing, or illustrating movements by integrating speech recognition[4] and movement recognition.

Evaluation of camera control algorithms First we evaluated the camera motions. Since some papers [2][3] reported cameramen's characteristics in tracking objects, we compared the characteristics of our system with real cameramen's.

Fig.5 shows an evaluation of capturing the scene which is already shown in Fig.3. This shows the effect of the virtual-frame control algorithm. In this figure, we plotted the apparent(image) position where the stationary point in a scene is located. With the virtual-frame control, the apparent velocity[5] of a target almost always stays less than the maximum value by professional cameramen. On the other hand, the apparent velocity can easily exceed the maximum without the virtual-frame control. This causes shaky and irritating view field motions. Thus we verified that our camera control algorithms are effective for recording ordinary presentations.

Examples of shooting presentations Here we show an experiment for an actual presentation, in which a person explained how

[2] EVI-D30(Sony)

[3] Flock of Birds(Ascension Technology Corporation)

[4] IBM ViaVoice

[5] The apparent position is plotted with normalized in terms of the screen size. The apparent velocity is the ratio of the difference in normalized position between video two consecutive frames.

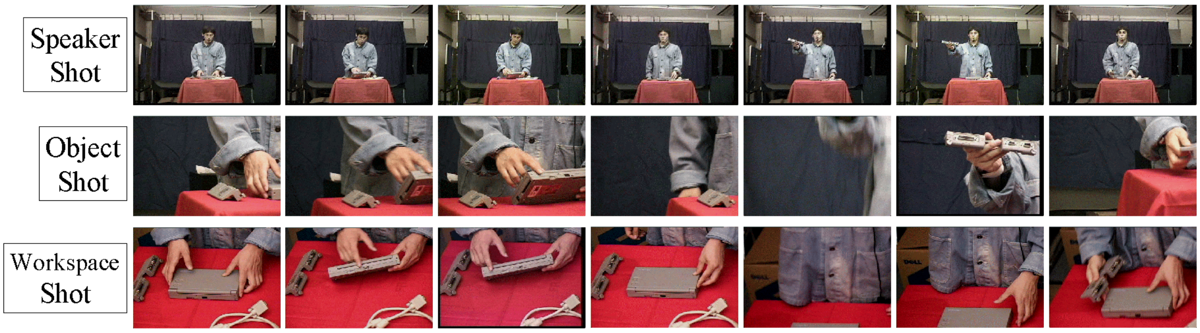


Figure 6: Videos from three cameras

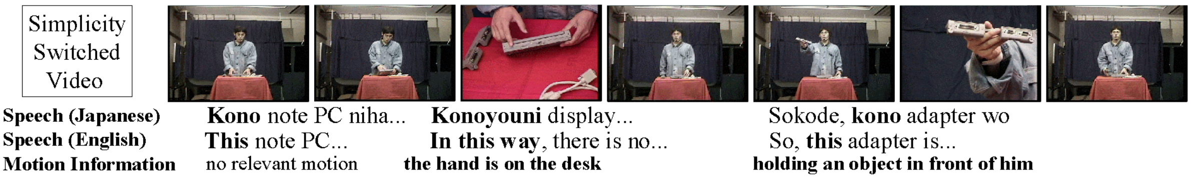


Figure 7: Result of camera switching

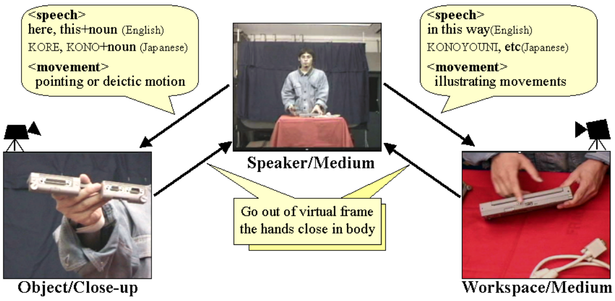


Figure 8: Condition of camera switching

to attach a display cable to a small notebook PC.

Three cameras are used, and the setting for each camera is shown in Table.1. Camera1 captures the speaker's behavior, and tracks with the camera control parameters for <circumstance>. Camera2 shoots at the referent object which the speaker holds and draws the viewers' attention. Camera3 shoots at the workspace from a high position (around 2m high) so that it can captures better view of the desktop which is sometimes difficult to see from the position of camera1 or camera2.

By capturing the presentation through those three cameras, we obtained three streams of videos as shown in Fig.6. As we can see in this result, the speaker's behavior, some important objects, and desktop manipulations were effectively captured, and the captured views were good stuff for communicating the presentation.

Example of video editing As one application of our system, we can realize automated video editing by the combination of video capturing and event recognition. By selecting the most relevant view according to the events, we can obtain a comprehensible video as shown in Fig.7. This selection is fully automated by using an electronic switcher controlled by a host computer.

The switching condition is briefly shown in Fig.8. If a speaker wants to draw the viewers' attention to his/her in manipulation on a desk, this intention appears in speech as some short phrases such as "in this way", "by doing this", and so on [6]. At the same time, the speaker moves his/her hands on the desk. By detecting this behavior by speech recognition and motion detection, the system selects the view through camera3 (shooting at <workspace>). Similarly, if a speaker wants to draw attention to the object he/she is holding in front of him/her, the intention appears in speech as "this", "here", and so on [7]. The speaker also moves his/her hand to the viewers so that they can easily notice it. By recognizing this behavior, the system selects the views by camera2.

Comparing the video in Fig.6 with the video in Fig.7, we can easily understand that the system selects appropriate views and the result is quite satisfactory.

5. CONCLUSION

We proposed a novel camera control framework for intelligent video production. For this purpose, we focused on cameraworks required for capturing presentations on desktop manipulations, then proposed camera control algorithms for tracking various targets in various ways. Our system works well for ordinary presentations as we can easily see that the results are better than the videos taken by a single or a few fixed camera(s).

Detailed evaluation is, however, left for future work. We need to tackle with comparison with other methods or systematic subjective evaluation. Also there is much room for discussion on categorizing targets and cameraworks.

6. REFERENCES

[1] Yoshinari Kameda, Michihiko Mihoh, et al. A live video imaging method for capturing presentation information in distance learning. *ICME*, 2000.

[2] Daiichiro Kato, et al. Analysis of the camera work of television cameramen while tracking subjects. *ITE*, 1996.

[3] Daiichiro Katou, et al. Automatic control of a robot camera for broadcasting based on cameramen's techniques and subjective evaluation and analysis of reproduced images. *JSPA*, 2000.

[4] S. Mukhopadhyay and B. Smith. Passive capture and structuring of lectures. *Proc.ACM Multimedia*, 1999.

[5] Naoki Ohno and Katsuo Ikeda. Video stream selection according to lecture context on remote lecture. *5th Symposium on Intelligent Information Media*, 1999.

[6] The only Japanese speech recognition is implemented. Some demonstrative pronouns, *e.g.* *"KONOYOUNI" or "KOUYATTE"*, are the keywords.

[7] demonstrative pronouns such as "KONO" or "KORE" in Japanese.