Studies of Automatic Video Generation from Real World

Yoshinari KAMEDA† Michihiko MINOH† Katsuo IKEDA‡
Center for Information and Multimedia Studies, Kyoto University†
Graduate School of Informatics, Kyoto University‡
Yoshidahonmachi, Sakyoku, Kyoto, Japan, 606-8501, Japan
E-mail: {*kameda, minoh*}*@media.kyoto-u.ac.jp*†, *ikeda@i.kyoto-u.ac.jp*‡
URL: *http://www.imel1.kuis.kyoto-u.ac.jp*

Abstract

We propose two methods that enables us to make videos automatically from daily scenes. One is for shooting human activities held in a fixed place like lectures, plays, and so on, and the other is for reviewing video taken at a place where people go by.

In the former method, multiple active cameras are needed to shoot multiple objects in the scene. Therefore, a new cinematography is required that can cope with many active cameras simultaneously and can select the best video according to the situation of the scene. We have been developping automatic camera system which estimates the situation of the scene based on image understanding technique and controls the cameras so as to obtain the best shots of the activity.

With the latter method, videos presented to them are to be skimmed so that video viewers can save their time. We have been working on this issue and have done experiments for years and discuss the way to use stored video clips.

1 Introduction

As WWW technology becomes popular today, one of the main issues on serving WWW is how to create WWW content and what kind of contents we should put on WWW. Our research and development division of the Center for Information and Multimedia Studies (CIMS) in Kyoto university works on this issue for several years and have been developping new methods that can make multimedia contents from real world automatically.

We mainly take up the scenes that are related with human activities and can be commonly seen in universities. For example, we have daily classes and student activities anywhere in campus in our university.

To let people view such activities on internet, various kinds of presentation techniques have to be used. Video and audio media are typical and essential media among them because we usually understand the human activities by our eyes and ears.

Our research group studys visual medium first, and then plans to utilize audio medium for generating new type of WWW contents.

Because video medium is usually obtained by video cameras, we consider two types of camera environment for taking videos of human activities.

- Multiple active cameras (dense arrangement)
- Multiple static cameras (sparse arrangement)

The first type is suitable to capture the events which arises at a certain fixed place. We can apply this type of camera environment to lectures, plays, and some kinds of sports, which are commonly seen in our campus. As these activities consist of more than one person, we have to understand where the people are in the scene and who is shot and which camera is used for the shot. Until today, these issues are handled by arranging many cameramen and directors and so it is difficult to film such activities many times because it is usually hard to prepare those staffs. Therefore, we are asked to design a new method through which multiple active cameras are controlled and an appropriate camera is selected to shoot the activity. This method cuts out subspace within which an interesting behavour is observed and determine the best camera to shoot it according to spatial analysis of the scene.

Concerning the second type, a good point is wide application because it is easy to set each camera in a certain scene. On applying this environment, it is important how to extract attractive visual information and how to show it in video style because cameras are set in the scenes where events will happen but no one knows when they arise. In other words, a method that is applied in this type of camera environment should cut out interesting subsequences in a long sequence of video.

Hence we describe two issues of our research in this paper which are related with the two types of camera environment shown above. The two issues are :

- Video camera control and selection of multiple active video cameras
- Video skimming for people in rooms for long period

These two issues are especially focused on visual information taken by video cameras.

Currently, the first issue aims to support distance learning in universities. As classrooms are generally too large to shoot by only one video camera and there many persons to shoot, we have to prepare several cameras and control them according to a situation in the classroom. We will mention our approach in Section 2.

The second issue is to support browsing video which is taken by one video camera for long period at a place people go in and out. This is a new approach of using cameras that are connected to LAN. As it is one of the hardest works to review the video for long time, we have developped video skimming method that makes video length shorter. Details are described in Section 3.

2 Control and Selection for Multiple Cameras

2.1 Discussion for Video Presentation

When multiple cameras are used to capture an activity in a fixed place, how many video images should be displayed to people who want to view it ?

There is an approach which displays all the video images captured by multi-camera system to viewers. This idea is derived from the viewpoint that it is good for the viewers to receive as much information as possible. However, as the purpose for viewing the videos is not to watch the place closely but to see the activity there, some of the videos are insignificant and useless for them.

Therefore, we first consider behavour of people in visual sense when they come accross an activity in real world. They usually focus on a part of the activity because they can see only one subspace at a time. To do that, they move their heads and eyes so that they pay attention to their surrounding environment.

Therefore, it is desirable to generate only one video sequence to the viewers although multiple active cameras are used behind. The video is to be generated by controlling directions of the active cameras and selecting the best camera according to a situation measured by the system.

In our approach described from the next section, we use active cameras which can pan, tilt, and zoom so that our method can realize the functions like what real cameramen do.

2.2 Automatic Camera Control Method

One of the features of our approach is to introduce observation cameras. There are two kinds of cameras in a classroom in our approach. One is observation camera, and the other is shoot camera.

The observation camera is like an human cameraman's naked eye. Its role is to follow the target to shoot. The cameras for this function usually zoom down at their limit so as not to miss movement of the target in the classroom.

On the other hand, the shoot camera is just like a video camera which human cameraman uses. The shoot cameras are forced to turn right and left, up and down, and zoom in and out to obtain good video image against the target.

Although the observation camera is like cameraman's eye and the shoot camera is like the video camera, we have to consider which shoot camera should be used when an object is observed at certain observation camera. We call this description "camera control relation." A representation of the camera control relation has two fields. One field is observation camera, and the other is a list of controlled shoot cameras. We draw this relation in Figure 1.

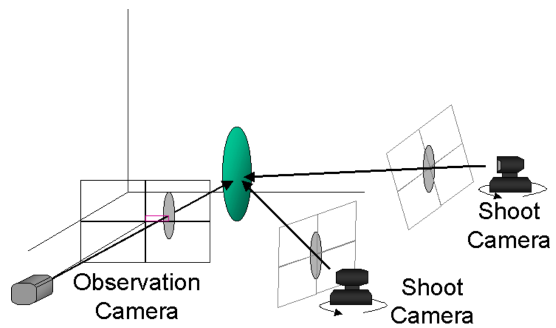

Figure 1: Camera Control Relation

When the observation camera detects movement of an object, the shoot cameras under control of that observation camera changes their direction so as to place the object in the center of their video images. Note that one observation camera can have multiple shoot cameras, and observation cameras can share the same shoot cameras in their camera control relation.

By sharing the shoot cameras, each observation camera can generate wide variety of video images because they are taken from different camera locations. As a drawback of camera sharing, there may be a situation that one shoot camera is forced to shoot two (and more) objects at the same time if different objects are observed simultaneously at the two observation cameras and they share the shoot camera. This kind of situation is inevitable, but we can reduce the possibilty of this situation to practically low level by assigning the shoot cameras under the rule below.

- Do not share shoot cameras among observation cameras observing different objects.

For example, it is acceptable to share shoot cameras among the observation cameras if they are to observe a lecturer, and there is only one lecturer in the room.

There is also a desirable rule on configuring the cameras in the classroom. That is :

- Cover the space of the classroom as large as possible by the viewing volumes of the observation cameras

A viewing volume is a subspace inside which an observation camera can watch objects with its viewing angle. Since normal video cameras can be modeled by perspective projection, a shape of the viewing volume is a pyramid shape. (See Figure 2.)

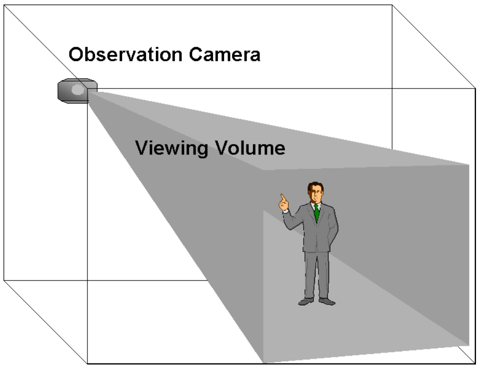

Figure 2: Viewing Volume

Object detection is done by inter-frame subtraction. Therefore, the observation cameras cannot change their directions during the detection. And, pan and tilt parameters for the shoot cameras are determined acccording to the location of the detected movement on the image planes of the observation cameras[1].

2.3 Automatic Camera Selection Rule

There are generally several criteria to define video clip (cut). From our experience, we think people want to see moving objects rather than static objects. And, as more important insight, we think people prefer to see an object which just starts moving rather than an object which keeps moving.

So, in our approach, each observation camera sends a request to video selector to select one of the shoot cameras it is controlling when it detects movement of an object. Once after the movement is detected and send the request, it will not send a selection request while the object keeps moving. To detect the beginning of the movement, we introduce time parameter τ_i for an observation camera i. Suppose a movement is detected on the camera i. In this case, the camera i does not recognize the beginning of new movement unless there is a no-motion period longer than τ_i seconds. An example of video selection function is shown in Figure 3. In this example, there are 4 observation cameras. The horizontal axes mean time flow, and the vertical axes indicate motion detection level at each observation camera. A grey arrow indicates a selection request. If the motion detection level comes higher than certain threshold value, the request is generated. In the middle of the sequence in Figure 3, the camera No.4 cannot generate a request after the video of the camera No.3 is selected because the beginning of the second movement is too close (smaller than τ_4) to the end of the previous movement.

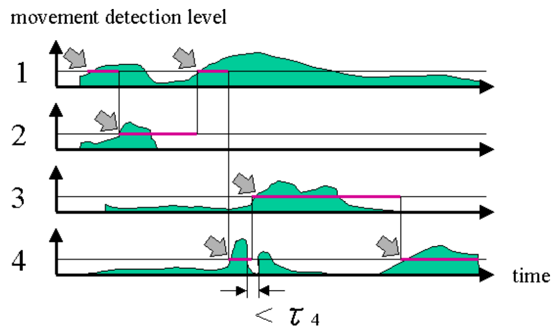

Figure 3: Video Selection

2.4 Experiment

2.4.1 Target Activity

Since there are many kinds of activities held in a fixed place, approaches to control multi-camera system depend on the activities to be captured.

From this viewpoint, it is important for us to choose an activity carefully which we take up as a target. The activity should be done or played usually and commonly so as to prove applicability of our method.

Hence we take up lectures in a classroom as our target application of our method.

2.4.2 System Design

We have a protoype system for serving distance learning lectures. Figure 4 shows overview of the classroom where the system is installed. There are 8 cameras inside the classroom. Figure 5 is a snapshot of the side.

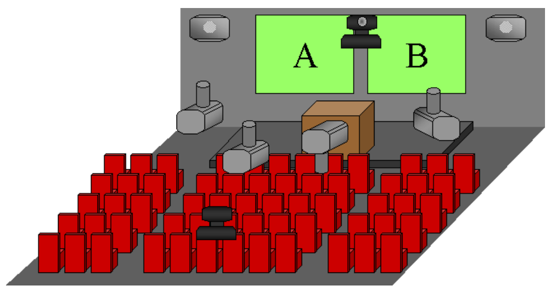

Figure 4: Overview of a Site

Figure 5: Snapshot of the Site

Figure 6 is top-view of the classroom. In the Figure 6, cameras expressed by dark grey arrow are SONY EVI-D30, and cameras expressed by light grey arrow are SONY EVI-G20. We assume that the lecturer walks mainly around the upper rectangle near the screens. The lower big rectangle indicates a region of student seats in the classroom. Except for camera No.4, all the cameras are hanged from the ceiling.

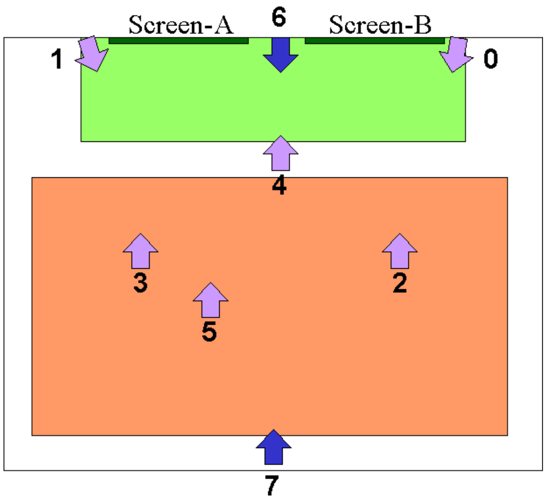

Figure 6: Camera Layout

Table 1: Camera Function

No.	Task	Function
0	Students	Observation
1	Students	Observation
2	Lecturer	Observation
3	Lecturer	Observation
4	Lecturer	Shoot/Observation
5	Lecturer	Shoot
6	Students	Shoot
7	Lecturer	Shoot

Table 2: Camera Control Relation

Observation Camera	Shoot Camera
0	6
2	5,7
3	5,7
4	4

Table 1 shows task and function of each camera. Task field in the table means the object the camera is to observe or shoot.

As mentioned before, the direction of the observation cameras are set so that the viewing volume of the each observation camera does not overlap its viewing volume against those of other observation cameras.

Based on the tasks assigned to the cameras, the camera control relation is defined by Table 2.

We show an example of video sequence taken by this system in Figure 7. The sequence in the figure last for 40 seconds.

Since video clips of a lecturer is considered to be more important than those of students, we set $\tau_0 = 3.0$[sec] and $\tau_2 = \tau_3 = \tau_4 = 1.0$[sec] in the experiment so that the camera No.0 generates selection requests more frequently than the rest cameras.

3 Video Skimming for Single Camera

3.1 Video Skimming

Different from video on TV, video taken by a static camera which places in a scene is boring for most of time because events which people are interested in seldom happen in front of the camera.

So the first step to handle such kind of scene watching is to remove the subsequences within which there are no events. We have developped a video skimming program that makes long video sequence shorter. It estimates area of pixels which are different from those of previous frames, and only when the area is larger than a threshold value, the frame is stored and is used to generate a digital video.

As our method assumes there is usually no movement,

Figure 7: An Example of Generated Video

we cannot apply this method to a crowded corridor or pathway in downtown where a lot of people go by. Continuous movement of flow can be eliminated from the image plane by introducing analysis of optical flow, so we think we can improve our method for that situation in our future research.

3.2 Experiment

We have installed this protoype system in several places and have operated the system.

Table 3 shows the number of generated videos for the past three years and Figure 8 shows an example of generated video. We set two systems, one is at the entrance of our laboratory and the other is on the wall inside our laboratory. The videos were generated for every 30 minutes starting at each o'clock in 1997 and 1998, and every 60 minutes starting at each o'clock in 1999. The authors put the results on WWW[2].

With this experiment data, let us show the skimming rate of videos by taking up 495 videos stored in September 1999. The camera shot the entrance of our laboratory and each video was taken for 1 hour in which the program processed 35549.4 frames on the average. This means the current system processed 9.875 frames/second at image size of 320 pixel by 240 pixels. The average number of stored frames was 92.37 per video, that is only 0.26 % of the whole video sequence.

When a camera is set in a building, most popular

Table 3: Skimmed Movies

Period	Place	Movies
1997/06	Entrance	250
1997/07	Entrance	362
1997/08	Entrance	308
1997/09	Entrance	175
1997/10	Entrance	7
1997/11	Entrance	24
1997/12	Entrance	208
1998/01	Entrance	90
1998/02	Wall	52
1998/02	Entrance	74
1998/03	Entrance	163
1998/04	Wall	488
1998/05	Wall	597
1998/06	Wall	518
1998/07	Wall	463
1998/08	Wall	550
1998/09	Wall	623
1998/10	Wall	424
1998/11	Wall	193
1999/08	Entrance	57
1999/09	Entrance	495

Figure 8: Example of Skimmed Video

events that can be imaged with static camera environment is passing of a person in front of the camera. The videos taken within the experiment include a lot of passing persons.

For example, results taken on September 20th in 1999 are shown in Figure 9. Number of persons who passwd in front of the camera is also shown on Figure 10. From this result, one person took 11.61 frames to pass in average, that means 1.18 seconds.

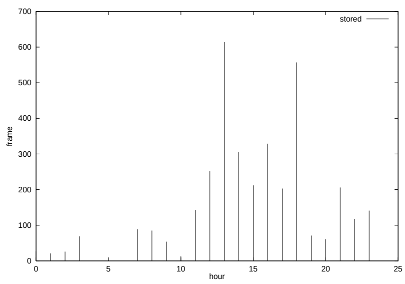

Figure 9: Stored frames on Sep.20, 1999

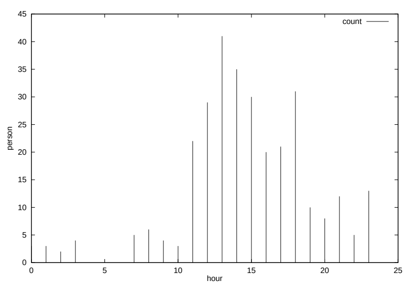

Figure 10: Passing Persons on Sep.20, 1999

4 Conclusion

We described two reseaches related visual medium which are aimed to generate new WWW contents.

Our system for multiple cameras is now under operation for distance learning classes and we are collecting data and experimental results. We are discussing how to archive videos of the lectures after distance learning classes are over.

We plan to install the video skimming system in a certain building where we will prepare hundreds of cameras inside rooms, entrances, and corridors. A future work is video analysis for various scenes so as to classify video subsequences and retieve them in video database.

Acknowledgment

This work has been partly supported by "Cooperative Distributed Vision for Dynamic Three Dimensional Scene Understanding (CDV)" project (JSPS-RFTF96P00501, Research for the Future Program, the Japan Society for the Promotion of Science).

References

[1] Yoshinari Kameda, Hideaki Miyazaki, and Michihiko Minoh, "A Live Video Imaging for Multiple Users," Proceedings of International Conference on Multimedia Computing and Systems (ICMCS'99), Vol.2, pp.897-902, 1999.

[2] http://www.imel1.kuis.kyoto-u.ac.jp/members/kameda/rec/peeping_archive/