Distance Learning Environment based on the Interpretation of Dynamic Situation of Lecture Room

Michihiko Minoh and Yoshinari KAMEDA
Center for Information and Multimedia Studies
Kyoto University
e-mail: `minoh@media.kyoto-u.ac.jp`, `kameda@media.kyoto-u.ac.jp`
`http://www.imel1.kuis.kyoto-u.ac.jp/`

Abstract

Based on the research of CDV project, we are constructing a distance learning environment which will be in practical use. The purpose is to evaluate our imaging method, particularly the dynamic situation, in the context of the distance learning and to improve the method. Practical problems are also discussed.

1 Introduction

People usually gather together and do various activities in a certain fixed space. For instance, there are conferences, amusement events such as sports and concerts, lectures in schools, and business activities. It is technically becoming possible to watch or even participate in such an on-going activity from a remote place thanks to the power-up of computers and speed-up of the transmission lines.

When a person in a remote place participates in such an activity, he/she will enter a kind of *information media environment*. Users not only watch objects in the real space but also obtain related information about them in the information media environment.

The synchronized virtual space is a space which reflects the changes of the geometric and photometric information of the real space. Since the real space changes in real-time, the synchronized virtual space has to change accordingly. The only difference between the real space and the synchronized virtual space is the time delay which is caused by the transmission and process of the data in recovering shape and surface information of the objects.

It is not necessary to construct the synchronized virtual space completely to construct the information media environment. Conceptually, consider one axis one end of which is the real space and the other end is the synchronized virtual space (Figure 1). Between these two spaces, there exist many spaces which are mixture of these two spaces. In other words, there is a seamless transition from the real space to the synchronized virtual space.

Video image is considered to be a useful medium of the information media environment and is suitable to convey visual information of what happens in the real space. In this

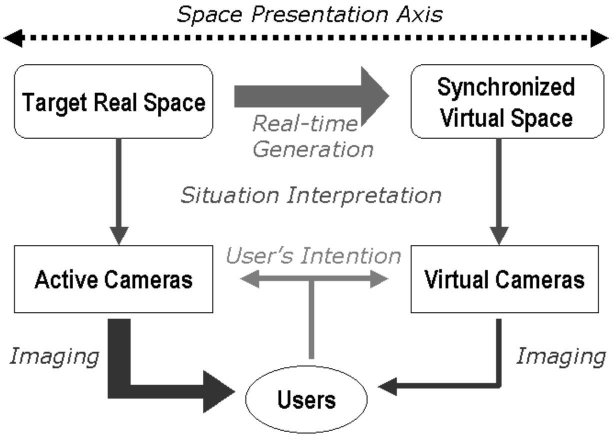

Figure 1: Information media environment

paper, we focus on the video image generation of these spaces. For the synchronized virtual space, it is relatively easy to realize the function because we can put virtual cameras wherever we like. However, for the real space, it is difficult because the camera locations are fixed and restricted. When the information media environment is presented to the users with the video image, active video cameras are used to observe the target real space. The cameras are set to surround the target real space to sense the activities and can change their direction and zooming parameter dynamically. One of the interesting topics in our research is how to control the cameras and generate interesting video for the users.

We set a lecture room in the Kyoto university as the target real space. This has two reasons. One is that the activities in the real space are relatively simple, though we have many people there. The other is that lecture is one of the most important activities in the university, and there is a strong request to watch a lecture from remote places.

2 Imaging method based on dynamic situation

2.1 Imaging framework

Video image is considered to be a typical presentation media and is suitable to convey visual information of what happens in the real space. Video cameras are set to surround the real space to sense the activities and can change their direction and zooming parameter dynamically. Therefore, dynamic camera work is a key issue in our research.

Formulated camera works in movies[1] and computer supported movie makers[2][3] have been proposed, but they assume that screenplays or shooting scripts are given in advance and they do not support live camera control. As our case does not allow us to have such scripts, we use concept of *dynamic situation* instead.

A dynamic situation is a description of what happens in the real space. Therefore, an activity is described by a sequence of the dynamic situations. The dynamic situation is linked to the camera works in imaging rules, and is described by situation features derived from sensor data, i.e. video images, as is shown in Figure 2. Camera works and situation features are mentioned in detail in the later section.

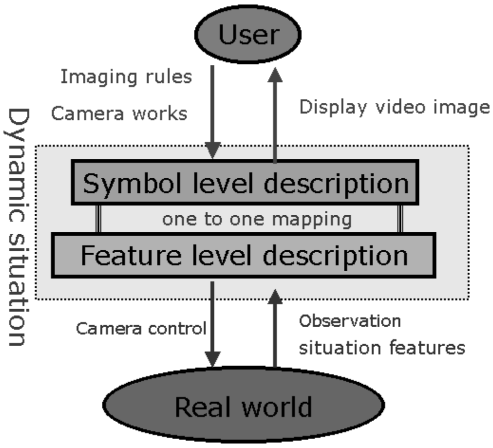

Figure 2: Imaging framework

2.2 Dynamic situation

A dynamic situation is the key concept for the video imaging. It indicates a status of the real space at a time and is defined by the strategy of

- what kind of video is desired
- how it is generated

Hence it depends on the activities or the contents.

The dynamic situation is represented in two ways; the symbolic representation for the user and the feature representation for the system. The user takes advantage of the symbolic representation, while the system takes advantage of the feature representation. To have a mapping between the two layers is a typical pattern recognition problem.

In the symbolic representation, the dynamic situation is described by a set of actions of the active objects. A description of one action is called action component and is denoted by *A-component* for short.

On the other hand, in the feature representation, the dynamic situation is defined by the combination of the features extracted from the sensor data. We call these features *situation features*.

2.3 Symbolic representation of the dynamic situation

An A-component is a symbolic representation of the dynamic situation and consists of one active object, one verb, one target object, and supplemental target objects. The active object is a dynamic object. The verb describes what the active object does. Both target object and supplemental target objects are objective to the verb. Whereas a target object can be subjective, supplemental target objects could not be subjective, i.e. static objects.

On describing the A-components, one A-component that has both an active object and a target object could be written in either way; in the active form or the passive form. To avoid confusion, we prohibit describing the A-component in the passive form.

For example, in the case of lectures in the lecture room, the A-components are listed in Table 1. The A-component ID 1 and 2 should be written in active form. A 'student group' in Table 1 consists of several students who are sitting in a certain part of the lecture room.

Table 1: A-components

ID	active object	verb	target object	supplemental target object
1	lecturer	talk	whole students	-
2	lecturer	talk	whole students	blackboard
3	lecturer	write	-	blackboard
4	student group	become active	-	-
5	student group	stay calm	-	-

2.4 Feature representation of the dynamic situation

The dynamic situation is also described by the situation features. Feature extraction methods would differ if the activities in the real space are different. The extraction method of the situation features must not affect the human activities in the real world.

Therefore suitable sensors are image sensors and voice sensors. As the dynamic situation varies in real time according to the activities, the situation features should be extracted in real time.

With respect to the lectures, we use three kinds of situation features; lecturer's location, lecturer's voice level, and activation degree of student group. The situation feature description of the dynamic situation defined in Table 1 is shown in Table 2.

Table 2: Situation features and dynamic situations

Situation ID	Lecturer Location	Lecturer Direction	Lecturer Voice Level	Student Group Activation
1	-	student	positive	-
2	-	blackboard	positive	-
3	blackboard	blackboard	-	-
4	-	-	-	positive
5	-	-	-	zero

2.5 Camera control method

A camera work describes how to image an object at a time. It consists of three fields; label of an object, direction, and range.

$$w(objectlabel, direction, range) \tag{1}$$

Direction field indicates the relation between the direction of the object and of the camera. Note that it implies the camera location which are assumed to be fixed. Range field tells the size of the object in the image.

We adopt eight directions and five discrete range values. See Figure 3 and Figure 4.

2.6 Imaging rules

An imaging rule is a set of functions which map a dynamic situation to camera works. Hence, the imaging rule $\mathcal{I}$ consists of two-tuples, A-component and camera works. The A-component includes several objects and the camera work is designed for only one object. Therefore, each A-component generally has a set of camera works.

Changes in the real space are detected by the sensors, and the situation features are extracted. Referring to the feature level representation of dynamic situation, i.e. situation features, the dynamic situation is detected.

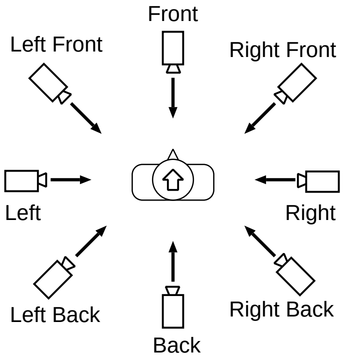

Figure 3: Direction of camera work

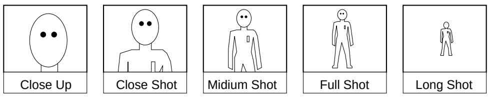

Figure 4: Range of camera work

Let us explain by an example. Suppose the imaging rule $\mathcal{I}$ is like Table 3 against the A-components shown in Table 1. The empty rows in the Table 3 mean that the object (lecturer of the A-component No.3 for example) is not necessary to be imaged even if the corresponding A-component is detected.

When the A-component No.1 and No.4 are detected in the real space by estimating the situation features at time t, the imaging rule $\mathcal{I}(t)$ that consists of three camera works w_1, w_2 and w_8 in Table 3 becomes active and at least one camera work is realized.

There are three ways of who decides the imaging rule.

- Transmitter: A Certain imaging rule is defined by a director who can transmit the video streams.
- Receiver: A user can request the imaging to be received. In this case, the user has to write the imaging rule he/she likes. In the information media environment, many users can request his/her own imaging at the same time as is realized in MULVIS[4] (Figure 5).
- System designer: The designer defines some events to be imaged, if the activities are simple enough and the way of the imaging is fixed. For example, if only the moving objects are considered to be interesting, they are imaged.

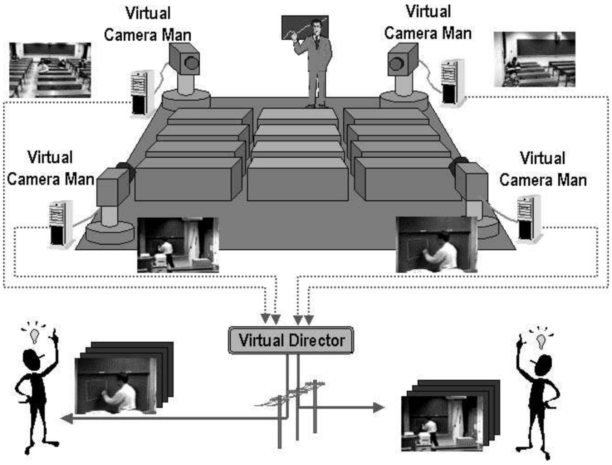

Figure 5: MULVIS

Since the way of imaging depends on the activities to be imaged, we have to fix an activity and analyze it.

Table 3: An example of imaging rule

A-component	object	camera work
1	lecturer	w_1(lecturer, front, closed shot)
	whole students	w_2(whole students, front, full shot)
2	lecturer	w_3(lecturer, front, full shot)
	whole students	w_4(whole students, right front, full shot) w_5(whole students, left front, full shot)
	blackboard	w_6(blackboard, front, full shot)
3	lecturer	-
	blackboard	w_7(blackboard, front, full shot)
4	student group	w_8(student group, right side, medium shot)
5	student group	-

3 Distance learning system

3.1 Concept of the distance learning system

Distance learning system is not only the system for audio/video transmission as is advertised widely but also includes contents, lecture, video imaging, presentation and students support system. There is a layer structure in the distance learning system as is shown in Figure 6. The lowest layer is network layer. The network has to assure data transmission, particularly streaming video data transmission. It includes encoding and decoding the video data, IP routing or PVC/SVC on ATM, etc..

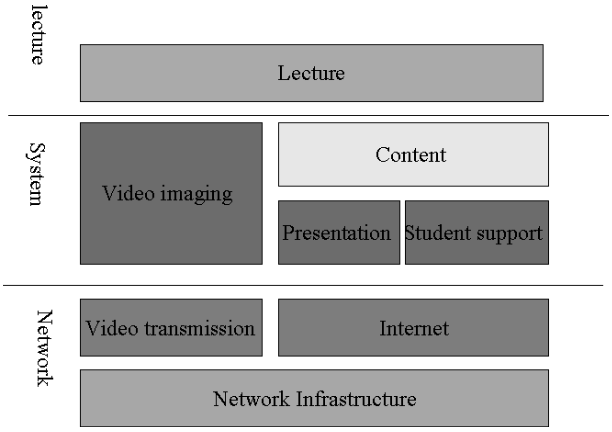

Figure 6: Layer Structure

Above the network layer, we have system layer in which the video imaging system, presentation system and student support system exist. The video imaging system hopefully works automatically, while the presentation system has to be operated by the lecturer. The key issue of the presentation system is the synchronization of the operation among the classrooms. The student support system is operated by the students before/after the classes. Content displayed in the both presentation and student support systems is supposed to be in the form of the digital multimedia data and presented by the computer. How to make the content is very important from the viewpoint not only of the academic but also of the business.

The top layer of the distance learning system is the real lecture, a kind of communication between the lecturer and the students. The communication is via electric media so the style of the lecture and the way of teaching have to be changed accordingly. This problem is considered to be important in the field of communication and/or education.

The whole class must be operated by a lecturer in the classroom as the ordinary, not distance learning, class does. In any case, this strategy has to be held to design the distance learning system technically. Among the systems in the system layer, the video imaging system has to be autonomous and to have a kind of intelligence. In other words, the system has to recognize what happens in the classrooms and control the cameras accordingly. In this context, we will consider the video imaging system.

Here, we assume the distance learning system with two classrooms, which have the bidirectional single video transmission facilities with CODECs and a synchronized presentation system. Each classroom has the video imaging system with several pan/tilt/zoom controllable cameras, with two projectors and an electric blackboard as is shown in Figure 7.

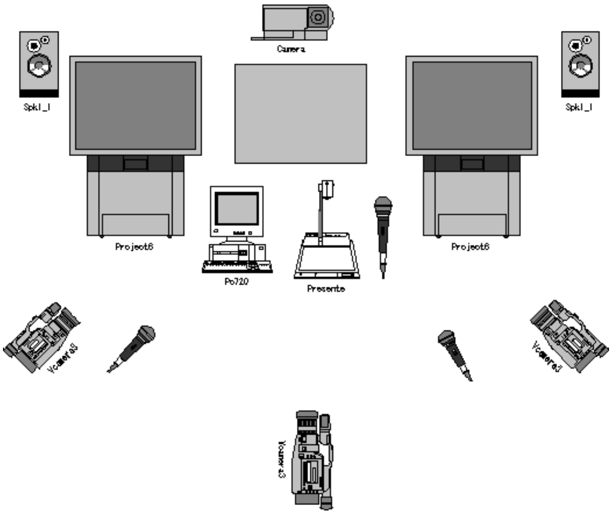

Figure 7: Equipment in one classroom

4 Video imaging in the distance learning system

The camera system is designed for the purpose to realize easy to use distance learning system. To produce cinemas and TV dramas, many people are involved in general. Considering the situation of universities, it is almost impossible to assign a camera man to each class in order to image the class unless it brings the universities a lot of profits. Though automated camera system may not produce better video than human operators, it is worth imaging the classroom automatically because the lecture is rather static from the viewpoint of camera control.

In order to shoot a lecturer and students in a classroom at the same time and select one video camera that is suitable to send to the remote classroom, the system should know: "Which object is to be shot now ? And how ?"

Because this is based on a dynamic situation in the classroom at that time, then it should know: "What kind of situation is now ?"

In this context, we will divide the cameras in the classroom into two categories; observing cameras and shooting cameras. The observing cameras are to obtain the situation features for discriminate the dynamic situation, while the shooting cameras are to generate the video streams for the remote users to watch.

In the distance learning system, the situation features are defined as follows.

- Location of the lecturer
- Location of active students

These locations are estimated by computer vision technique. The lecturer is detected in the image by checking motion region in the video images. Then, the location of the lecturer is calculated with single camera by introducing the constraint that the lecturer always locates on a certain plane in the 3D world and by giving the camera parameters of location, direction, focal length, screen aspect, and so on in advance. The location of the active students is also estimated by almost the same approach.

For an example, we can define a dynamic situation which represents "The lecturer is talking to the students" by the parameters like:

- The lecturer is close to the teacher's table
- The students are calm (No active students found)

After the dynamic situation is obtained, the system controls the pan/tilt/zoom values of the cameras in order to shoot the object. We have develop the method to declare the way of camera control linked to the detected dynamic situation[4].

Since we have several shooting cameras all of which generate the video streams, the system has to select one video stream to the CODEC, which is transmitted to the remote classroom. As the system knows the position of the lecturer and the cameras, the geometric relation and the present camera parameters will give the key to select one video stream. Also, the duration time of the specific video stream may be the key as well. With these keys, the system selects one video stream and send it to the CODEC.

In practical case, we avoid the pattern recognition problem. Instead, the dynamic situation is defined by the situation features of the location. These locations will be calculated automatically. To cope with the estimation error of the location, the shooting cameras do not zoom in so much. In the real environment, we do not have a blackboard in the classroom, so the dynamic situation becomes very simple. In the implementation of the practical system, simpler way is considered to assure the reasonable imaging than the best video imaging.

4.1 Automatic Camera Control Method

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

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

Although the observing camera is like cameraman's eye and the shooting camera is like his video camera, we have to determine which shooting camera should be used when an object is observed at certain observing camera. We introduce "camera control relation" which is presented by two-tuple. One element is observing camera, and the other is a list of controlled shooting cameras. We draw an example relation in Figure 8 in which one observing camera uses two shooting cameras.

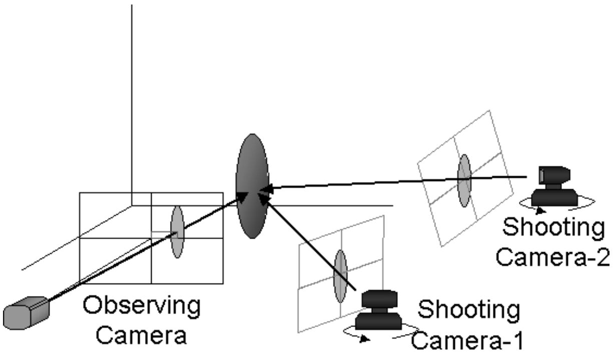

Figure 8: Camera control relation

When the observing camera detects the movement of an object, the shooting cameras under control of that observing camera change their direction so as to place the object in the center of their video image. Note that one observing camera can have multiple shooting cameras, and the observing cameras can share the same shooting cameras in their camera control relations.

By sharing the shooting cameras, each observing 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 shooting camera is forced to shoot two (and more) objects at the same time if different objects are observed simultaneously at the two observing cameras and they share the shooting camera. This situation is inevitable, but we can reduce the possibility of the situation to practically low level by assigning the shooting cameras under the rule below.

- Do not share shooting cameras among the observing cameras which observe different objects.

For example, it is acceptable to share the shooting cameras among the observing 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 with the viewing volumes of the observing cameras

A viewing volume is a subspace inside which an observing 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 9.)

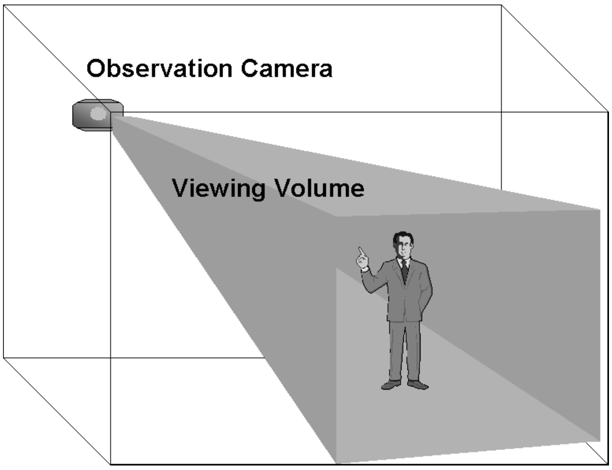

Figure 9: Viewing volume

Object detection is done by inter-frame subtraction. Therefore, the observing cameras cannot change their direction during the detection. The pan and tilt parameters for the shooting cameras are determined according to the location of the detected movement on the image planes of the observing cameras[5].

4.2 Automatic Camera Selection Rule

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 observing camera sends a request to video selector to select one of the shooting 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 observing 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 10. In this example, there are 4 observing cameras. The horizontal axes mean the time flow, and the vertical axes indicate the motion detection level at each observing 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 10, the camera No.4 cannot generate a request after the video under 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.

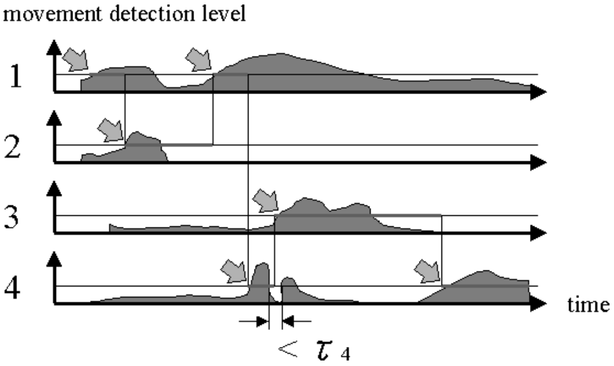

Figure 10: Video selection

5 Experiment

5.1 Tide Project

We have begun the project called TIDE(Trans-Pacific Interactive Distance learning Environment) with UCLA and NTT. The main purpose of TIDE project is to find a way to realize time and location free lecture environment. In this sense, collaboration between Kyoto-U and UCLA intrinsically includes the main problem, i.e. we are far away in the location and have 7 or 8 hours of time difference.

Outstanding features of our project includes;

1. Cultural exchange between Japan and USA via lecture
2. The first experimental trial for the whole class distance learning
3. Fully interactive classes
4. Automatic camera system for video imaging
5. Communication support by WWW and E-mail
6. High quality audio/video transmission over Pacific ocean

We are aiming the system which is operated only by the lecturer, but in practice, we need at least one person in the remote room to cope with unexpected situation. In our project now, we have a lecturer in each class because of the credit problem of both universities.

5.2 Classroom System

The camera system we have designed[5] has two functions:

- **[Automatic Mode]**
  Produce one video stream of the lecture automatically with several pan/tilt/zoom cameras
- **[Manual Mode]**
  A human operator can manually operate several cameras in case special camera control is required

The manual mode is useful for back-up particularly in the practical use.

The followings are the outstanding features of our camera system in the automatic mode.

1. Track a lecturer in the classroom.
2. Shoot him/her as he/she moves around.
3. Observe students in the classroom.
4. Shoot some of the students when they ask questions.
5. Select the best camera shooting the lecturer/students and send its video to the CODEC.

The functions an operator can do in the manual mode are as follows.

1. Pan/Tilt/Zoom control for shooting cameras with only one mouse.
2. Watch all the videos from the shooting cameras.
3. Select the best video to send by just pushing one button.

We have a prototype system for serving the distance learning lectures. Figure 11 shows overview of the classroom in Kyoto University where the system is installed. There are 8 cameras inside the classroom and 4 among them are used as the shooting cameras. Figure 12 is a snapshot of the classroom.

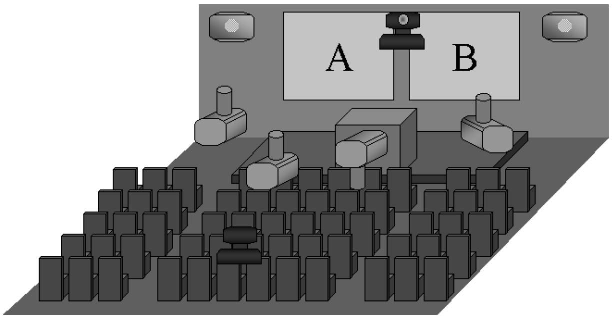

Figure 11: Classroom overview

Figure 13 is top-view of the classroom. In the Figure 13, 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.

Table 4 shows task and function of each camera. Task field in the table means the object the camera is to observe or shoot.

Figure 12: Classroom snapshot

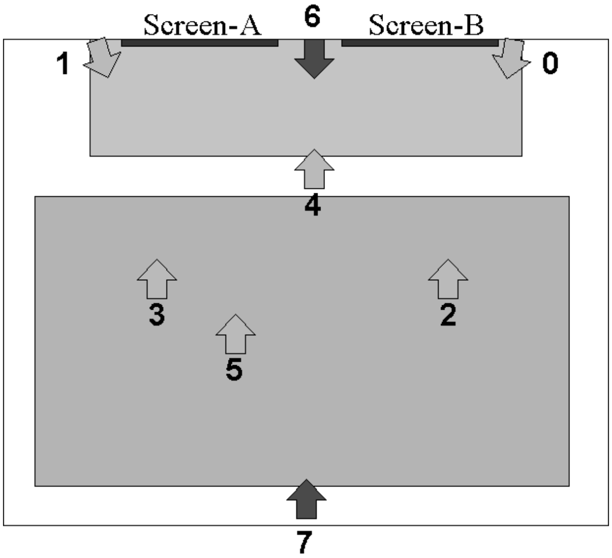

Figure 13: Camera layout

Table 4: 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 5: Camera control relations

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

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

Based on the tasks assigned to the cameras, the camera control relations are defined by Table 5.

5.3 Experimental Result

We show an example of video sequence taken by this system at Kyoto University in Figure 14. The sequence in the figure last for 40 seconds.

Since the video clips of the 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 located at the rightmost position of the screen-B generates selection requests less frequently than the rest cameras.

The class is continuing and we are collecting the evaluation forms from not only the

Figure 14: An example of generated video

students but also the teachers. They include items of the system itself, the lecture style and the way of teaching with the system.

6 Conclusion

The true evaluation of information media environment has to be done in practical use. In this sense, the TIDE project will give us a great evaluation of our information media environment. During the experiments not a few technical problems remain in future.

First of all, the observation part has to be improved. Voice signal has to be processed to sense the location where the voice comes from. Image sensor suffers noise which cause the position estimation errors. Hence, the cameras can not zoom up which makes the video image degrade. Also, we will consider to use a position sensor with the image and voice sensor to verify the estimation.

We avoid the pattern recognition problem either by keeping the system inside the signal level or by define one to one mapping between the two description of the dynamic situation. However, the system is not flexible enough to cope with various situation in the real classroom. With more sensors, the system has to recognize the dynamic situation.

The most important aspect of the cooperative distributed vision research is how to cooperate the two video imaging systems. The dynamic situation is defined only in the physical classroom. We have to extend the concept of the dynamic situation in the virtual classroom which consists of two or more physical classrooms.

Fortunately, we have developed the experimental environment, so we will continue to make our system progress.

Thanks

We thank Professor Maha Ashour-Abdalla of UCLA and Dr. Makoto Mizukawa of NTT for their contribution to TIDE project.

References

[1] D.Arijon: "Grammar of the Film Language," Focal Press Limited, 1976.

[2] L. He, M.F. Cohen, D.H. Salesin: " The Virtual Cinematographer: A Paradigm for Automatic Real-Time Camera Control and Directing," SIGGRAPH'96, pp.217-224, 1996.

[3] D.B. Christianson, S.E. Anderson, L. He, D.H. Salesin, D.S.Weld, and M.F. Cohen: "Declarative Camera Control for Automatic Cinematography," Proceedings of AAAI '96, pp.148-155, 1996.

[4] 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.

[5] Yoshinari KAMEDA, Michihiko MINOH, and Katsuo IKEDA, "Studies of Automatic Video Generation from Real World," Proceedings AEARU's Second Web Technology Workshop, pp.57-62, 1999.