# Proceedings of the Fourth International Workshop on Cooperative Distributed Vision

March 22–24, 2001

Kyoto, Japan

Sponsored by

Cooperative Distributed Vision Project
Japan Society for the Promotion of Science

All rights reserved. Copyright©2001 of each paper belongs to its author(s).
*Copyright and Reprints Permissions: The papers in this book compromise the proceedings of the workshop mentioned on the cover and this title page. The proceedings are not intended for public distribution. Abstraction and copying are prohibited. Those who want to have the proceedings should contact with* cdvws@vision.kuee.kyoto-u.ac.jp

# Contents

**1 Imaging a 3D lecture room by interpreting its dynamic situation** **3**
*Michihiko Minoh and Yoshinari Kameda*

# Notice on This Electric Document

This paper includes only the part of Minoh and Kameda in the proceedings of the 4th international workshop of Cooperative Distributed Vision. Layout of figures is changed from the original paper.

If you want to read the full proceeding, please contact `cdvws@vision.kuee.kyoto-u.ac.jp`. Thank you.

# Imaging A 3D Lecture Room by Interpreting Its Dynamic Situation

Michihiko Minoh and Yoshinari Kameda

Center of Information and Multimedia Studies,
Kyoto University
e-mail: minoh@media.kyoto-u.ac.jp kameda@media.kyoto-u.ac.jp
http://www.mm.media.kyoto-u.ac.jp/

**Abstract**

We envision a new computer supported environment named information media environment. Users can watch what happens in a fixed real space where people get together to do something through raw/synthesized video images in real-time with this environment.

On watching video image of the scene, people want to see and understand what happens there. In this sense, we should know not only what is happening in the scene but also what they want to see and how. Therefore, on imaging the scene, we have to not only consider the situation of the scene but also the intension of persons who watch video images. There are two ways of generating images. One is to obtain good images by controlling active cameras according to the situation. The other is to reconstruct whole 3D shape models and enable persons watch them from any arbitrary viewpoint. In the first approach, we introduce imaging rules which bind the way of controlling active cameras and dynamic situation of the scene. We also show how the situation should be transformed into the camera-works, which describe the way of controlling active cameras.

We select a lecture room for distance learning lectures in our experiment. Interests of remote students should be understood and the active cameras should be cooperated so as to image the focused object in the 3D lecture room. We have implemented a prototype system and served it for several regular courses between UCLA and Kyoto University, which shows soundness of our proposed method.

We also propose fast 3D shape reconstruction method which can be used for providing free views of the focused object in the information media environment.

# 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 event 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 event, he/she will enter a kind of *information media environment*. In the target real space, there are discrete objects which are involved in the events. Users are able to not only watch the objects in the real space but also obtain related information about them in the information media environment.

In this sense, video image is the most important medium in the information media environment in order to let users understand and join the events in the real space. We have to provide video images in real-time because the information media environment is used in an interactive manner.

People watch video images of the events, because they want to understand what happens there. The motivation of watching the images affects on their favorites of shooting the objects in the events. In this sense, on imaging the scene and providing its video images, we should know not only what is happening in the scene but also what they want to see and how. Therefore, we have to understand not only the situation of the scene but also the intension of the persons who watch video images.

The ideal imaging is that all the robot cameras move around and shoot the focused object from the viewpoint where each user likes as an activity goes on. However, this is neither possible nor preferable because the moving cameras may interfere the activity and the number of the cameras is usually smaller than that of the users. Hence, we prepare two ways of generating video images of the scene instead of the moving robot cameras. One is to reconstruct whole 3D shape models and enable the users watch them from an arbitrary viewpoint. The other is to obtain good images by controlling multiple active cameras according to the situation and the intension of the viewer. The cameras can rotate around its pan and tilt axis and zoom, but do not change their position so they do not interfere the activity. Figure 1 shows our approaches. The first imaging approach is shown as the top center thick arrow, whereas the second imaging approach is represented by the lower left thick arrow. We call the fully reconstructed 3D virtual space "synchronized virtual space."

Although the first approach enables us realize any requests on imaging the events because there is no restriction to set and move the virtual camera in the synchronized virtual space, there are lots of problems to provide high quality video. We have proposed

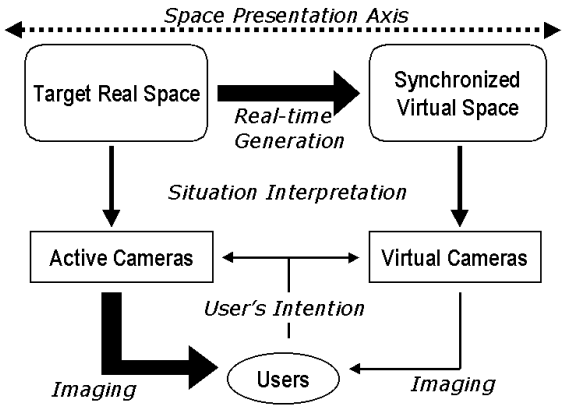



Figure 1: Information Media Environment

a fast 3D shape reconstruction method[1][6][2] and describes it in Section7.

On the contrary, the second approach with pan/tilt/zoom cameras can provide video images the quality of which is as fine as that of TV. The conditions that this approach are that the number of the cameras is limited and that they cannot move because we do not let the active cameras interfere the activities. Without careful camera handling, such cameras cannot always provide the video images that are exactly the same as what the users want to see. However, we show several methods which deal with this problem and succeeded in providing video images suitable to the users. The key of our proposed methods is to bind the intension of the users and description of the situation of the scene in imaging rules. We first mention the general formulation of imaging methods with multiple pan/tilt/zoom cameras in Section2. We have proposed three imaging methods on this formulation through CDV project. The first one is based on the intension of the common human behavior which is discussed in Section3. The second one is designed for broadcast imaging philosophy and is described in Section4. In addition, we report one application in which we implement and use these two methods for actual distance learning courses in Section5. The third one is to provide multiple, different video images for multiple users who have different requests on imaging. We describe it Section6 in detail.

We conclude our total research in Section8.

# 2 Imaging Method Based on Dynamic Situation

## 2.1 Imaging Philosophy

We consider that there are three styles of imaging the lectures. These are characterized by the way of viewing presentation information.

- Common attention (Surveillance)
- Lecture understanding (Broadcast Video)
- Remote user's interest (Customized Video)

The first style is for an event which happens in the classroom but is not related to the lecture[17]. As no one can predict such kind of unexpected event, there is no systematic imaging way against it. Therefore, we should model common reaction when people encounter the unexpected events.

The second style is the most suitable style for distance learning because it aims to let the remote students understand the lecture[8]. Although several cameras are used in this style, only one video stream is generated to carry visual information due to the communication cost of the network. We call this style a broadcast imaging style which is usually programmed in broadcast studio where a director controls cameramen by considering the audience who receive the video images.

The third style is acceptable when the interests of the remote students take precedence over the context of the lecture[16].

Our main interests are on the second and third style in this paper. Since movement of the lecturer and behaviors of the students are very important presentation information in the lecture, our method deals with them on shooting them based on the situation of the lecture.

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 control is a key issue in our research. We call a description that defines dynamic camera control a *camera-work*.

Formulated camera-works in movies[9] and computer supported movie makers[10][11] have been proposed, but they assume that screen-plays or shooting scripts are given in advance and they do not need to support live camera control. As our case does not allow us to possess such scripts, we use concept of *dynamic situation* instead.

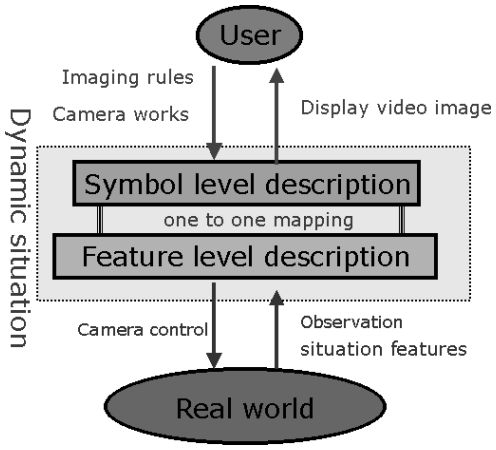



Figure 2: Framework of Imaging

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. A dynamic situation is linked to several camera-works through *imaging rules*, and can be 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.

## 2.2 Dynamic Situation

A dynamic situation is the basic concept on video imaging. It indicates a status of the real space at a time and is defined by the notation of

- what kind of video image is desired
- how it should be generated

Hence it depends on the activities or the contents.

The dynamic situation is represented in two layers; the symbolic representation for the user and the numerical 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 active objects. Active object plays a important role of the event in the scene. A description of one action called action component is denoted by *A-component* for short.

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 | - | - |

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

## 2.3 Symbolic Representation of The Dynamic Situation

An A-component is a part of symbolic representation of the dynamic situation and consists of one active object, one verb, one target object, and supplemental target objects. Active object is an object that triggers an event and plays an important role in it. 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. In this sense, target objects are thought as 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 active form or passive form. To avoid redundant expression, we prohibit describing the A-component in passive form.

For example, in the case of lectures in a lecture room, A-components can be listed like 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. Note that multiple A-components may occur at the same time.

## 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 space. Therefore one of the most suitable sensors is image sensor. Voice sensor is also good as supplemental sensor. As the dynamic situation varies in real time according to the

Table 2: A-components Defined by Situation Features

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

activities, the situation features should be extracted in real time.

Values of situation features are displayed for each A-component because a dynamic situation consists of several A-components.

With respect to the lectures, we will use at most three kinds of situation features; lecturer's location, its direction, its voice level, and activation degree of student group. Situation feature description we defined here for the A-components in Table 1 is shown in Table 2.

## 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 relation between direction of the object and that of the camera. Note that it implies the camera location which is considered to be fixed in our approach. Range field tells the size of the object in the image.

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

## 2.6 Imaging Rule

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

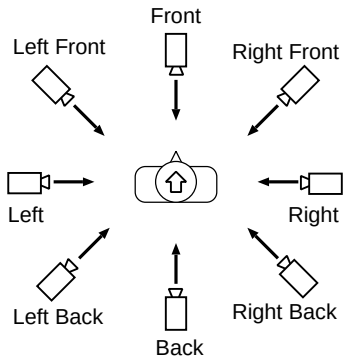



Figure 3: Direction of camera work

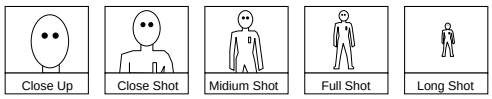



Figure 4: Range of camera work

Changes in the real space are detected by the sensors and the corresponding values of the situation features are extracted. Referring to the feature layer representation of dynamic situation, current dynamic situation is obtained by imaging system.

Let us explain by an example. Suppose an 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 required 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.

Figure 5.

Because there are a lot of strategies in establishing imaging rules and choosing situation features, we have to remind the imaging philosophy. As mentioned before, there are three ways of thinking that define the imaging rule.

- Common attention (Surveillance) : The system designer defines the objects in the events to be imaged based on the common attention of people's behavior. This is useful if the activities are simple enough and the way of the imaging is fixed. For example, if only the moving object are considered to be interesting, simply it is

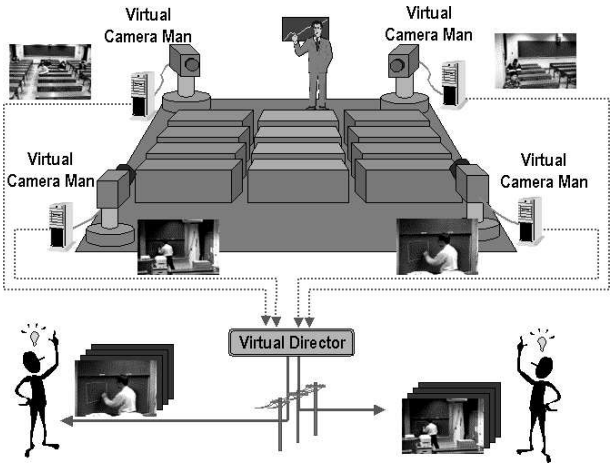



Figure 5: Imaging A Lecture Room

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)<br>$w_5$(whole students, left front, full shot) |
| | blackboard | $w_6$(blackboard, front, full shot) |
| 3 | lecturer | - |
| | blackboard | $w_7$(blackboard, front, close shot) |
| 4 | student group | $w_8$(student group, right side, medium shot) |
| 5 | student group | - |

imaged. This is also useful when the context of the activity is understandable or unexpected because this imaging rule does not depend on what is happening in the scene.

- Lecture understanding (Broadcast Video) : A Certain imaging rule is defined by a director who has the intension to transmit the video streams so as to let remote users understand what is going in the scene. Usually a lecturer is considered as the director, but you can regard the summarized intention of the remote users as the director too.

- Remote user's interest (Customized Video) : Each remote user can request the style of imaging to be received. In this case, the user has to write the imaging rule that he/she prefers. Some of the remote users may have similar intention and some are not. This is the hardest issue to solve as the number of the users increases.

In each imaging philosophy, the way of imaging depends on the activities in the scene very much. In the following 4 sections, we clarify the activity we focus on and show how the imaging rules should be defined and what kind of situation features should be used.

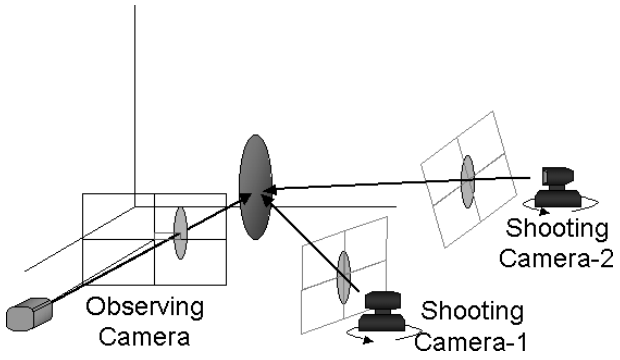



Figure 6: Camera Control Relation

# 3 Imaging Under Surveillance Philosophy

With the surveillance philosophy, there is only one set of common rules for imaging the scene. We assume that there is only one dynamic situation saying there is merely a moving object, and that users want to see the the moving object in a certain size in the video image.

This is very simple and therefore this approach does not understand the situation and cannot reflect it on imaging although it can make decision to move active cameras even when unexpected events happens.

## 3.1 Automatic Camera Control

In this method, we use two kinds of cameras. One is observation camera and the other is shoot camera.

The observation camera is like an human camera-man'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 lecture room.

On the other hand, 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 video images against the target object.

While the observation camera is like camera-man's eye and the shoot camera is like his video camera, we have to determine which shoot camera should be used when an object is observed at certain observation camera. We introduce "camera control relation" which is presented by two-tuple. One element is observation camera, and the other is a list of controlled shoot cameras. We draw an example relation in Figure 6 in which one observation camera uses two shoot cameras.

When the observation camera detects the movement of an object, the shoot cameras under the control of that observation camera change their direction so as to place the object in the certain location in their video image such as the center of the video image. In this sense, the binding between A-component and camera-work is directly done. The location of a moving object which is extracted on the observation camera in Figure 6 corresponds to situation feature. As the location is given, the two camera-works immediately invoked on the two shoot cameras.

Note that one observation camera can have multiple shoot cameras, and the observation cameras can share the same shoot cameras in their camera control relations. By sharing the shoot cameras, each observation camera can generate wide variety of video images because each of which is taken from different camera locations although they image the same object. On the other hand, as a drawback of camera sharing, there may be a situation that one shoot camera is forced to shoot two (or more) objects at the same time if different objects are observed simultaneously at the two observation cameras and they share the same shoot camera. This situation is inevitable, but we can reduce the possibility of the situation to practically low level by assigning the shoot cameras under the rule below.

- Do not share shoot cameras among the observation cameras which observe different subspace in the target real space.

For example, it is acceptable to share the 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 lecture room.

That is :

- Cover the space of the lecture room as large as possible with the viewing volumes of the observation cameras

A viewing volume is a subspace inside which an observation camera can see 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 7.)

In our approach, object detection (extraction of the situation features) is done by inter-frame subtraction. Therefore, the observation cameras should be fixed during the detection. The pan and tilt parameters for the shoot cameras are determined according to the detected location of the object on the image planes on each observation camera[17].

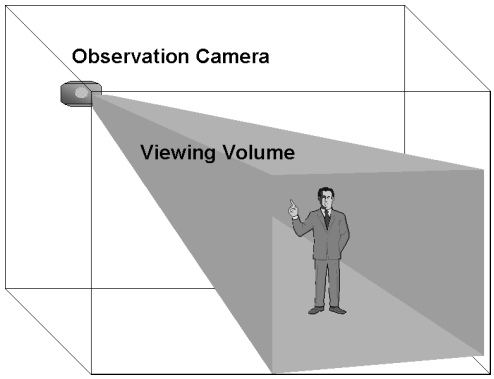



Figure 7: Viewing Volume

## 3.2 Automatic Camera Selection

After multiple video images from multiple shoot cameras are obtained against current event, remaining problem is the selection of video images because the users see only one video image.

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 shooting cameras it is controlling when it detects movement of the object. Once after the movement has been detected and sent the request, it would not send a selection request while the object keeps moving. To detect the beginning of the movement, we introduce time parameter $\tau_i$ 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 exists no-motion period longer than $\tau_i$ seconds. An example of video selection function is shown in Figure 8. In this example, there are 4 observation cameras. The horizontal axes mean the time flow, and the vertical axes indicate the 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 submitted. In the middle of the sequence in Figure 8, the camera No.4 does not submit a request after the video under the camera No.3 is selected because the beginning of the second movement is too close (smaller than $\tau_4$) from the end of the previous movement.

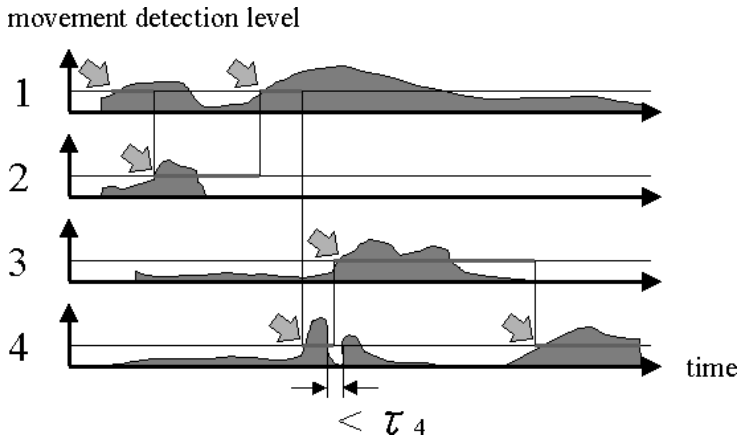



Figure 8: Video Selection

# 4 Imaging under Broadcast Philosophy

Broadcast philosophy is the most typical in many shooting environment such as TV studio, stages, lecture rooms, and so on. Events held on these spaces are usually intended to share a certain, sometimes planned, experience with audience. Therefore, it is natural to consider that there is always a certain focused point that everyone who wants to share the experience should watch and anything else is not worth seeing.

One particular application is distance learning in which a director aim to let the remote students understand the lecture. Although multiple shoot cameras are also used in this style, only one video stream is generated to carry visual information.

Note that the director in this approach does not indicate a particular person but a abstract personality which has appropriate intension to watch appropriate object along with the activity in the scene. Usually a lecturer is considered as this abstract director, but you can regard the summarized intention of the remote users as the abstract director too.

In distance learning, the main objects to be imaged are persons who take part in discussion in the lecture. They are a lecturer and a part of students in the lecture room. Hence, A-components which are part of dynamic situation should include the behavior of the lecturer and the students.

For an example, a dynamic situation which represents "The lecturer is talking to the students" could be detected by checking two A-components like:

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

With respect to situation features, we find that A-components in lectures has strong relation with the 3D location of the active objects and so we use location dependent features in this approach. They are :

- Location of the lecturer
- Location of active students

These locations are estimated by computer vision technique. The lecturer is detected in the image of an observation camera by checking motion region. Then, the location of the lecturer is calculated with monocular observation 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 a similar approach.

## 4.1 Automatic Camera Control

We prepare the cameras some of which can pan, tilt, and zoom in/out and some are static as the shoot cameras. Although only one shoot camera is needed at a time, we use multiple shoot cameras in the lecture room.

One of the reasons is that the lecturer may walk around in the lecture room and the students in any seats may become active. In this case, the object to be shot may easily go out of the visible area of the static shoot camera. The other reason is that the lecturer may walk faster than the maximum speed of pan/tilt/zoom of the active camera control. Therefore, the system should control multiple shoot cameras simultaneously not to miss the object from their frame.

When one A-component is detected, the corresponding camera-work start using a free shoot camera and will keep using so that it follows the object as long as the A-component is being detected.

## 4.2 Automatic Video Selection

As our automatic camera control method can let the shoot cameras image the object, the remaining issue is to select the most appropriate video stream for the users, who are in remote lecture room in this case.

Our video selection method consists of two rules. The first rule is processed based on the dynamic situation in the lecture room. The second rule is based on human intrinsic feature which will imposed as the intention of the abstract director.

1. Select one of the detected A-components in the dynamic situation

2. Change camera under time constraint

The detail of the video selection process is as follows.

First, if multiple A-components are detected at the same time, the A-component of which the object is a group of students takes precedence over the other A-components. This is because we implement prototype system in Japan and conclude that Japanese students rarely become active during the lecture and so it is a noteworthy activity rather than anything else. If the behavior of the students varies, this rule should be changed accordingly.

Then, the shoot cameras that are assigned to the camera-works of the selected A-component become candidates to be used. Since there may be multiple camera-works (it means multiple shoot cameras) for one A-component, we have to choose one of them under the time constraint we introduce above. There are two criteria for it; one is that

a man/woman needs certain time to understand what he/she watches. Based on the criterion, we introduce "glance time" which limits the shortest time of keeping the video stream from a single camera. The other is that a person become bored if he/she watches the same object from the same camera location for a long time. To avoid this, we introduce "release time" which limits the longest time of keeping the video stream from the same camera.

Therefore, if there is a camera-work that lasts less than the glance time, it is always adopted. On the other hand, if there is no change of dynamic situation even after the release time is over and if there are the other camera-works corresponding to the same A-component, another camera-work among them is newly selected.

# 5 Application for Distance Learning

We not only propose these two imaging methods in Section3 and Section4 but also implement prototype systems and serve them in regular distance learning lectures held between UCLA and Kyoto University.

In this section, we first describes the overall distance learning system because we believe its explanation helps readers to understand our motivation and our goal well. Then, two experimental results are shown for both imaging methods of surveillance philosophy and broadcast philosophy.

## 5.1 Concept of 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 9. 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..

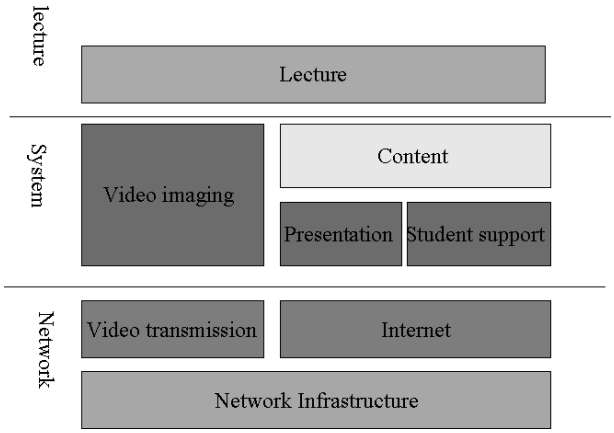



Figure 9: Layer Structure

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 lecture rooms. 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 lecture room 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 lecture rooms and control the cameras accordingly. In this context, we will consider the video imaging system.

Here, we assume the distance learning system with two lecture rooms, which have the bi-directional single video transmission facilities with CODECs and a synchronized presentation system. Each lecture room has the video imaging system with several pan/tilt/zoom controllable cameras, with two projectors and an electric blackboard as is shown in Figure 10.

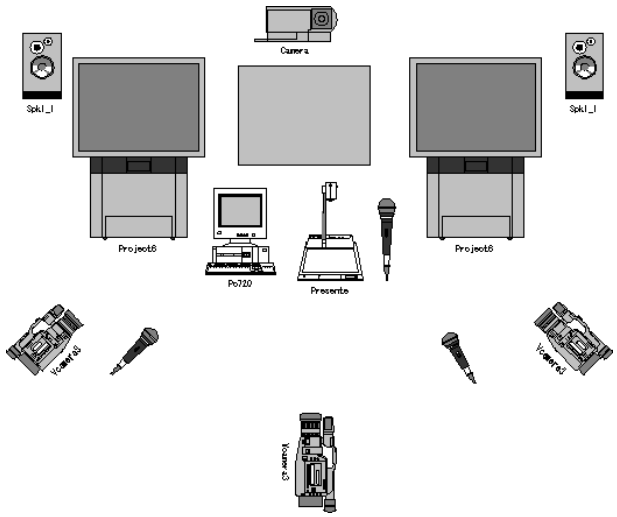



Figure 10: Equipment in One Lecture Room

## 5.2 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.

## 5.3 Lecture Room System

The camera system we have designed[17] 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 lecture room.

2. Shoot him/her as he/she moves around.

3. Observe students in the lecture room.

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 lecture room in Kyoto University where the system is installed. There are 8 cameras inside the lecture room and 4 among them are used as the shooting cameras. Figure 12 is a snapshot of the lecture room.

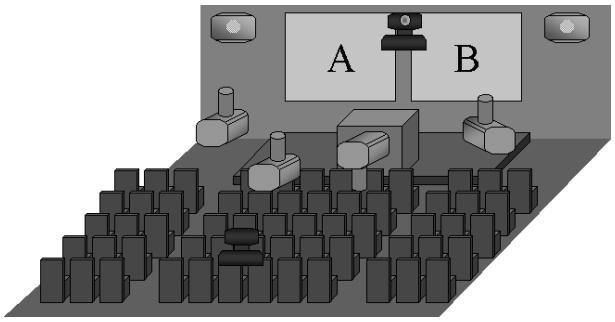



Figure 11: Lecture Room Overview

Figure 13 is top-view of the lecture room. 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 lecture room.

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: A Snapshot of The Lecture Room

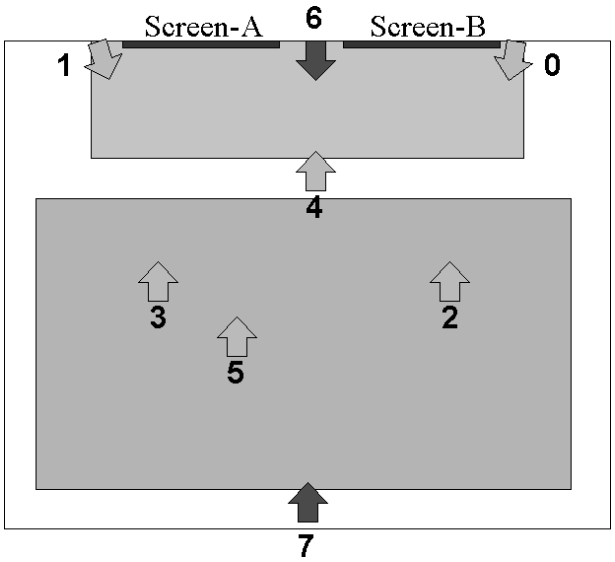



Figure 13: Camera Layout

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.

## 5.4 Experimental Result – Surveillance Philosophy

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

## 5.5 Experimental Result – Broadcast Philosophy

We installed a distance learning system based on our method with 4 shoot cameras and 4 observation cameras.

We conducted an experiment on actual distance learning lectures which were held between Kyoto university and UCLA. The prepared A-components are as follows:

| | |
|---|---|
| $A$ | The lecturer is pointing on the screen-T. |
| $B_{1-3}$ | The lecturer is talking to the students. |
| $C_{1-12}$ | Some of the students are active. |

Figure 14: A Result of Imaging with Surveillance Philosophy

$B_{1-3}$ are the same A-component but are found on three different places. We divide audience area into 12 blocks so we need 12 A-components for $C_i$. As $A, B_1, B_2$ and $B_3$ represent the lecturer and $C_1, \cdots, C_{12}$ represent the students, they may occur at the same time.

The set of imaging rules given to the system in advance is shown in Table 6, and the mapping function is shown in Figure 15. If a moving object is found in each area, the corresponding A-component is detected. We use three static shoot cameras in the imaging rules.

Figure 16 shows a sequence of generated lecture video transmitted to the remote lecture room. Two kinds of arrows indicate video selection rule in the sequence.

Table 6: Imaging Rules

| A-component | camera-work | |
|---|---|---|
| | object | camera |
| $A$ | lecturer | 7 |
| $A$ | screen-T | 5[fixed] |
| $B_1$ | lecturer | 7 |
| $B_1$ | lecturer | 4[fixed] |
| $B_2$ | lecturer | 7 |
| $B_3$ | lecturer | 7 |
| $C_{1-12}$ | student(1-12) | 6[fixed] |

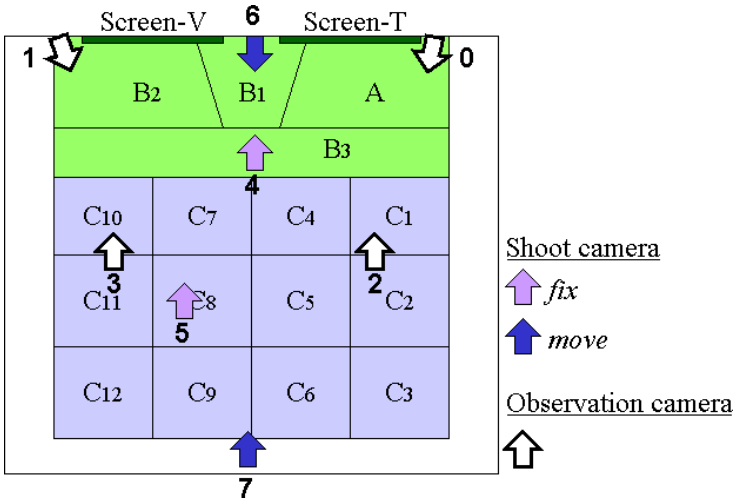



Figure 15: Mapping Function of Situation Features

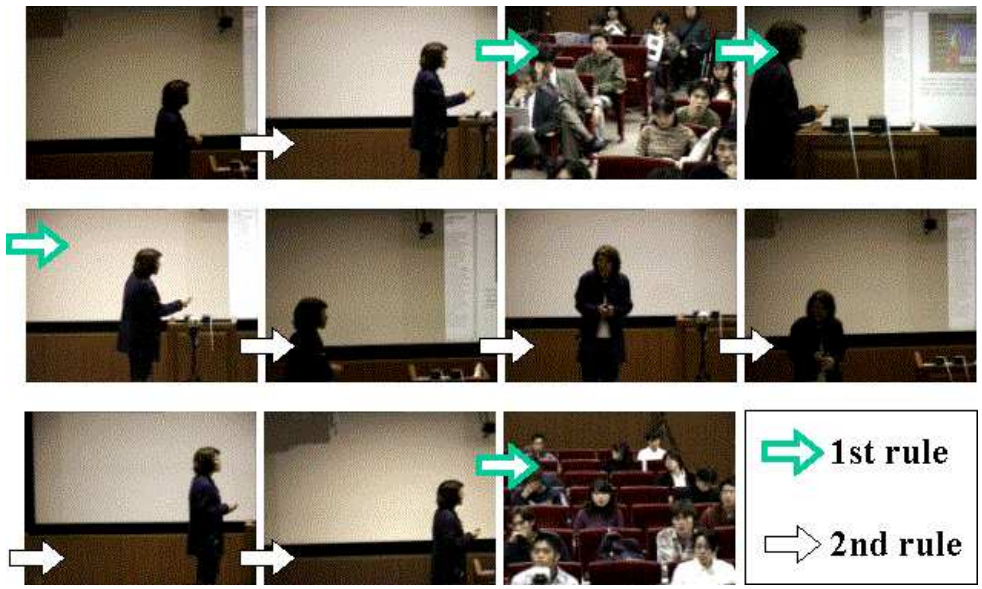



Figure 16: A Generated Video Images

# 6 Imaging under Customizing Philosophy

Although various events and activities are usually held in a fixed real space for a certain purpose under an abstract director, The users sometimes want to see different objects in different viewing angle apart from the imaging way that the abstract director recommends. Since it is impossible to infer their preference for imaging, we have the users proclaim their preference in advance and we consider to build an information media environment where each remote user can receive the customized video images. As some of the remote users may have similar intention and some are not, The purpose here is to maximize the number of the users who receive the desired video images.

We have proposed a new method that can handle this problem and constructed a prototype system named MULVIS (Multi-User Live Video Imaging System)[16][18].

In Section 6.1, we first introduce *imaging method* that represents personal preference of imaging the scene. and then mediation of the camera-works which are derived from the imaging methods of multiple users is discussed. Dynamic situations are detected in the way explained in Section 6.2. After those discussions, we examine process role assignment on constructing the information media environment in Section 6.3. and show experiments in Section 6.4.

## 6.1 Mediation among Imaging Methods of Multiple Users

### 6.1.1 Imaging Method

An imaging method is a set of camera-works defined for each dynamic situation by one user. If user differs, the imaging method would be different for the same dynamic situation in the real space.

Consider a user $j$. An imaging method $\mathcal{I}^j$ consists of the A-components, and each A-component has a set of the camera-works. The user $j$ proclaims the camera-works for the objects which appear in each A-component. The user can proclaim multiple camera-works for the same object in a certain A-component if he/she wants to. Hence the number of the camera-works in $\mathcal{I}^j$ is at most the summation of the number of camera-works that the A-components have. For example, if $\mathcal{I}^j$ consists of two A-component $p, q$ each of which the user $j$ proclaims two camera-works, $\mathcal{I}^j$ has four camera-works.

Note that the user $j$ does not need to define all the camera-works for the objects in the A-components. For example, suppose the same example shown above and the user does not have a desire to see the active object in the A-component $p$, the user is required to

proclaim only three camera-works for $\mathcal{I}^j$. Since multiple A-components would occur in the real space at time $t$, the imaging method the user desires may be represented by a set of the camera-works $\mathcal{I}^j(t)$.

Let us explain by an example. Suppose the targe real space will has a set of A-components shown in Table 1, And the user $i$ defines the imaging method $\mathcal{I}^i$ shown in Table 3. (Table 3 is denoted as imaging rule because in the former section it is shown in order to present a sample imaging preference. However, consider it as the table of the user $i$ here.)

The user $j$ defines the imaging method $\mathcal{I}^j$ like Table 7 different from the user $i$. As the camera-work ID is incremented sequentially throughout all the imaging methods and the same camera-work has the same ID, the ID in Table 7 does not seem to be incremental.

If the A-component No.1 and No.4 are detected in the real space, three camera-works $w_9, w_2$ and $w_12$ in Table 7 are submitted which represent the requirement of the user $j$ at that time, while $w_1, w_2$ and $w_4$ are submitted by the user $i$.

Table 7: An Example of Imaging Method $\mathcal{I}^j$ Proclaimed by User $j$

| A-component | Object | Camera-work |
|---|---|---|
| 1 | lecturer | $w_9$(lecturer, front, long shot) |
| | whole students | $w_2$(whole students, front, full shot) |
| 2 | lecturer | $w_10$(lecturer, right front, close shot) |
| | whole students | $w_4$(whole students, right front, full shot)<br>$w_5$(whole students, left front, full shot) |
| | | $w_2$(whole students, front, full shot) |
| | blackboard | $w_6$(blackboard, front, full shot) |
| 3 | lecturer | $w_11$(lecturer, front, close up) |
| | blackboard | $w_7$(blackboard, front, close shot) |
| 4 | student group | $w_12$(student group, front, medium shot) |
| 5 | student group | - |

In the human activities played in the real space, certain A-component sometimes occurs repeatedly, or one A-component may last in quite a long period. The users feel boring if there is no change in the video images due to the same A-components in the real space. To keep the video images interesting in these cases, it is a good approach to change the way of imaging the object even in the same A-component. To achieve this idea, the second and third field of the camera-work would be modified within a long interval. This is called

"camera work deviation".

### 6.1.2 Mediation And User Satisfaction

The request of the user $j$ is satisfied if one camera-work in $\mathcal{I}^j(t)$ is taken by a certain camera. Suppose there are $u$ users and $c$ cameras. A camera can realize only one camera-work at a time, so at most $c$ camera-works can be realized at time $t$. As a consequence, mediation of user requests is to select at most $c$ camera-works among $\mathcal{I}^j(t)$ where $j = 1 \cdots u$. Let us denote $n(t)$ that means the number of the camera-works in $\mathcal{I}^j(t)$ for all $j$. If it includes the same camera-works, they are counted as one camera-work. We can describe the mediation by the mediation matrix $M(t)$ which has $n(t)$ rows and $u$ columns where each component $m_{ij}(t)$ is either 1 or 0. A column represents $\mathcal{I}^j(t)$ and a row corresponds one camera-work $i$ and so $m_{ij}(t) = 1$ means that user $j$ realizes this request of the camera-work $i$. Hence, the next equation ought to be true.

$$\sum_{i=1}^{n(t)} m_{ij}(t) = 1 \tag{2}$$

Note that at each column $j$, $m_{ij}$ is always 0 if camera-work $i$ is not included by $\mathcal{I}^j(t)$. We introduce $a_i(t)$ that indicates whether the corresponding camera work of the $i$th row is selected or not. The number $s_i(t)$ indicates the number of the users who support the camera-work $i$.

$$s_i(t) = \sum_{j=1}^{u} m_{ij}(t) \tag{3}$$

$$a_i(t) = \begin{cases} 1 & (\text{if } s_i(t) \geq 1) \\ 0 & (\text{otherwise}) \end{cases} \tag{4}$$

The constraint of the number of the cameras is formulated by

$$\sum_{i=1}^{n(t)} a_i(t) \leq c \tag{5}$$

The mediation process is finding the mediation matrix $M$ that satisfies Equation (2), (4), and (5). The mediation matrix $M(t)$ is found in the case $u \leq c$, but there might be a dynamic situation where $u > c$ and so $M(t)$ does not exist. So, we select camera-works in the following procedure so as to satisfy as many users as possible.

#### Mediation Procedure

1. Set the number of the remaining cameras $r = c$.
2. Calculate $s_i(t)$ for all $i$.

3. Select one camera-work $k$ where

$$s_k(t) = max_{i=1}^{n(t)} s_i(t) \tag{6}$$

4. Reduce the number of the rest cameras $r$ to $c-1$.

5. If $r$ is 0, mediation procedure is finished.

6. Eliminate the users whose $\mathcal{I}^i$ contains the camera-work $k$. Go back to the step 2.

Once the camera-works are selected, the next problem is to assign the camera suitable to each selected camera-works under the constraint of the camera location, direction and range. The criteria to solve this problem is to satisfy the requirement of the users as much as possible. Our strategy is : the selected camera-works are ordered by $s_i(t)$ and choose the best camera in this order.

Therefore the video image user $j$ watches is taken by the camera assigned to a selected camera-work in the mediation matrix $M(t)$ if his/her imaging method is realized. On the other hand, there might be a case where some users cannot find their camera-works in the selected camera-works if $u > c$. Those who cannot satisfy with them select the second best camera-work among them which is similar to one of the camera-works in $\mathcal{I}^j(t)$ and watch the corresponding video image.

A user might have no submitted camera-works in a certain dynamic situation. In this case, the user watches an arbitrarily chosen video image.

## 6.2 Detecting Dynamic Situation

A dynamic situation is practically defined by the combination of the features extracted from the sensor data. We call these features the situation features. The feature extraction methods would differ if the real space and the activities are changed. Most of the situation features are extracted via image processing because the image sensor has an advantage that it does not affect the human activities played in the real space.

As the imaging methods are proclaimed by A-components and the A-components consists of the objects, the situation features ought to be extracted for each object. They should be extracted with little delay because they reflect the real-time dynamic situation in the real space.

With respect to the lectures, we use three kinds of the situation features; lecturer's location, lecturer's voice level, and activation degree of student group. The extraction methods are explained in Section 6.4.1.

## 6.3 Process Role Assignment

The functions in this system are classified into three categories. One is for imaging objects based on the camera-works, and another is for detecting the dynamic situation, and the other is for mediating the requirements of the multiple users. We design three types of processes for each function.

A process that controls an active camera and images an object is called an imaging process. Its purpose is to realize a camera-work and generate video of the object. The number of the imaging processes is the same as that of the selected camera-works at that time.

A process that detects the dynamic situation is called an observation process. Unlike the imaging processes, the observation processes have different functions one another because each observation processes extracts different situation features. The number of the observation processes is determined by the number of the situation features that are needed to detect the dynamic situation.

The last kind of the processes is designed mainly to mediate the requirements of the multiple users. We call this kind of process a mediation process. While imaging processes and observation processes are device (camera or sensor) dependent, the mediation process is device independent. Currently, we build one mediation process that interprets the information of the dynamic situation from situation features and selects the camera-works according to the mediation procedure.

In the framework of the cooperative distributed vision[7], a process has three functions such as perception, action, and communication. In our framework, the imaging process plays an action role and the observation process does a perception role. Both processes use the active cameras, but the observation processes do not exchange their cameras because the perception process should not miss what occurs in the real space and so they hold the cameras as their continuous sensors. On the other hand, the imaging processes can change their cameras one another because their purpose is to obtain the desired video image, and the exchange may lead them to achieve the specified camera-works more precisely.

## 6.4 Experiment

### 6.4.1 Situation Features

At the implementation for the lectures in the lecture room, the system uses three kinds of the situation features to detect the A-components shown in the imaging methods of the users. They are lecturer's location, voice level, and activation degree of student group.

The lecturer's location indicates the location of the lecturer in the real space. The lecturer's voice level shows the loudness of his/her voice. The activation degree of the student group becomes large as they move their bodies.

The lecturer's location is measured by the image based triangulation by two active cameras. As the lecturer walks around in the lecture room, the active cameras track the lecturer and calculate the location based on the camera parameters and the subtraction regions. We assume that the lecturer is the only wandering object.

To obtain the lecturer's voice level, the lecturer is asked to equip a wireless microphone and the input level of the A/D converter is used directly.

We divide the student desks into six groups and call the students in one desk group the student group in this experiment. The activation degree of the student group is presented by the area of the subtraction region in the image in which the student group is framed from their front view.

### 6.4.2 Experimental Result

Here we present some snapshots of two video images generated for a lecture held at a lecture room(Figure 17). Camera (b) and (c) in Table 8 were used for extracting the situation features related to the lecturer. Camera (a) and (d) were used for measuring the activities of the student groups. The imaging processes used four active cameras which are located as shown in Table 9.

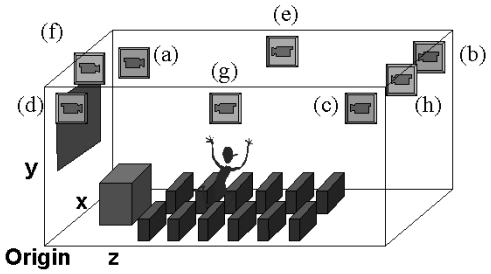



Figure 17: Camera Layout in A Lecture Room

The video images in Figure 18 and Figure 19 are generated for the different users. User (A) proclaimed three camera-works whereas user (B) proclaimed five camera-works. The horizontal axis indicates the time flow for about four minutes and C1 to C5 are the labels of the selected camera-work. Note that both users sometimes watch the same video image because their desired camera-works are overlapped at that time.

Table 8: Location of Observation Cameras

| Camera ID | $X$ [m] | $Y$ [m] | $Z$ [m] |
|---|---|---|---|
| (a) | 6.47 | 2.77 | 1.64 |
| (b) | 6.45 | 2.80 | 10.42 |
| (c) | 0.53 | 2.80 | 10.41 |
| (d) | 0.70 | 2.80 | 1.65 |

Table 9: Location of Shoot Cameras

| Camera ID | $X$ [m] | $Y$ [m] | $Z$ [m] |
|---|---|---|---|
| (e) | 6.84 | 2.17 | 3.63 |
| (f) | 4.59 | 2.43 | 0.41 |
| (g) | 0.64 | 2.32 | 7.37 |
| (h) | 3.741 | 2.16 | 10.73 |

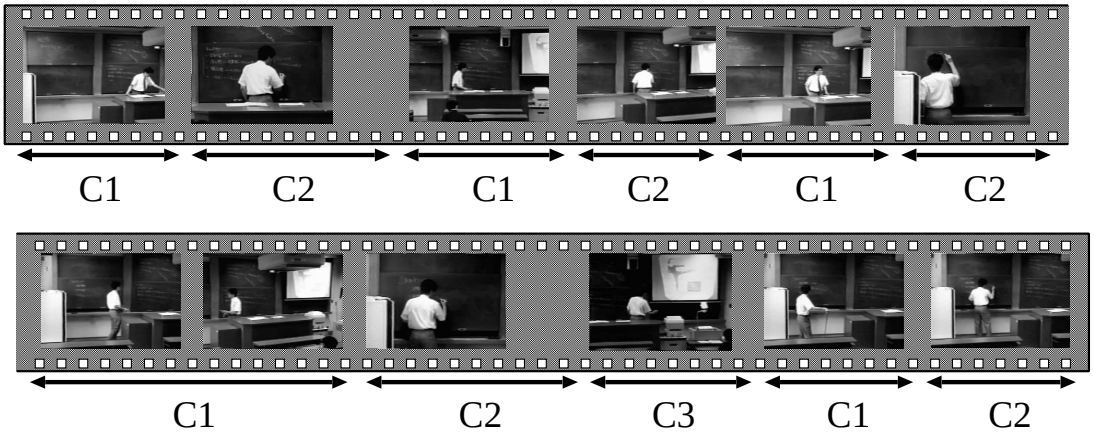



Figure 18: Video Images for User (A)

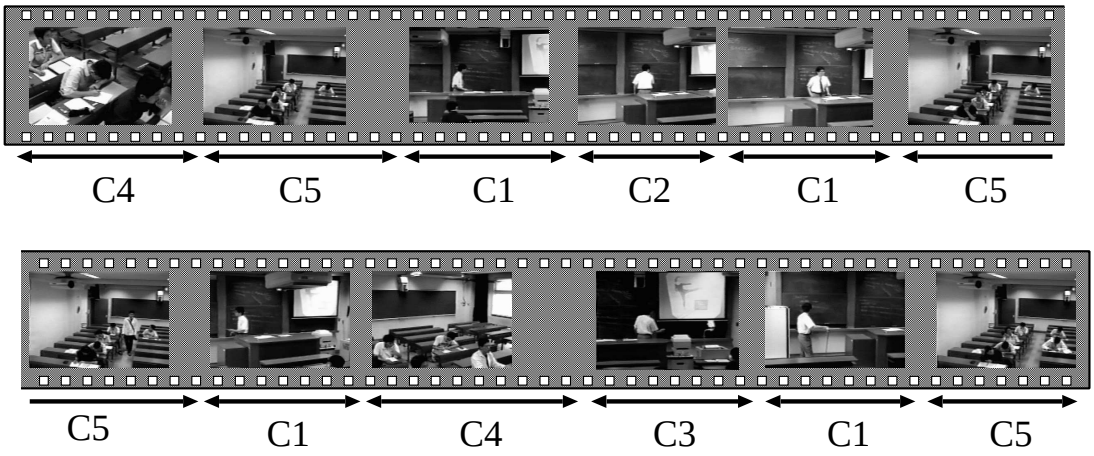



Figure 19: Video Images for User (B)

# 7 Generation of Synchronized Virtual Space

With improvement in processing speed of computers and with increase of their storage size, it may come true to synchronize a virtual space in computers with a real 3D space[14]. Our final goal is to construct the synchronized virtual space which displays real-time human activities occurred in the real space (see [6]).

Once the synchronized virtual space is constructed, anyone outside the real space can observe the human activities in the real space from any viewpoint with a little delay.

Realizing the synchronized virtual space needs two technical issues. One is geometric reconstruction and the other is photometric reconstruction. We have done mainly on geometric reconstruction method which is described in this section.

Slit light projection methods and structured light projection methods achieve real-time 3D reconstruction, but these methods require active sensing which disturbs human activities in the real space. On the contrary, passive vision based approach[12, 13] does not affect the activities. Stereo vision methods achieve real-time 3D reconstruction though they cannot reconstruct backside shapes that cannot be seen by stereo cameras. Therefore, cameras have to be placed so as to surround the real space. Realistic 3D reconstruction methods[3, 4] have been proposed which use over ten cameras, but their approaches need sophisticated way of calculation to reconstruct one scene and are not suitable for getting support of hardware acceleration.

The main problem of 3D reconstruction with such camera surrounding layout is that it requires much calculation time because there are many images at each frame. This problem is resolved by distributed computing in our approach. We reconstruct the real space by preparing one computer for each camera to execute image processing, and other computers to calculate 3D reconstruction. All the computers are connected one another with 100baseT Ethernet and 155Mbps ATM LAN.

We present reconstructed space by voxel representation. In our method, we improve throughput by dividing video processing into some stages and forming them as the pipeline processing, and decrease latency by dividing a real 3D space into some subspaces and reconstructing each subspace simultaneously with several distributed computers. We can also control throughput and latency by changing the pipeline formation in the system and satisfy the requirements of the applications.

In the last part of this section, we show not only a result of 3D shape reconstruction mentioned here in detail but also a result of photometric reconstruction method we are working.

## 7.1 3D Reconstruction Method

The reconstruction algorithm has to be suitable for the distributed computing, so that the algorithm has the following two characteristics.

- It is possible to equalize processing time of each process by dividing program and data.
- The amount of communication among processes is not so much.

The "shape from silhouette" method (SSM in short) satisfies these two characteristics.

With SSM, we reconstruct the real space in real-time by generating voxel data from several images taken at the same time. We call the part of the real space which can be imaged by the cameras the *target space*.

### 7.1.1 Static Object Occupation Subspace (SOOS)

Since our objective is to reconstruct the fixed real space, it is reasonable to have a knowledge of static objects in advance. As the static objects do not change their locations and shapes, we can exclude the subspace where the static objects occupy. We call the subspace as *static object occupation subspace (SOOS)* denoted by $\mathcal{S}$. For example, see Figure 20 where a circle is a dynamic object and a rectangle is a static object. $\mathcal{S}$ corresponds to the rectangle region.

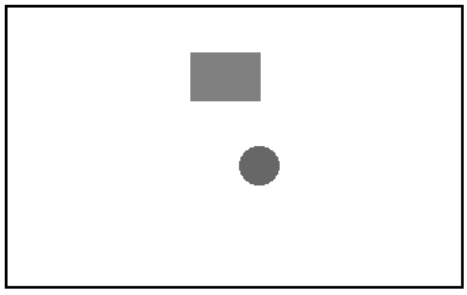


Figure 20: Dynamic Object (Circle) and Static Object (Rectangle)

If we see the target space from the viewpoint of camera $i$, some subspaces cannot be seen because $\mathcal{S}$ occludes them. We merge these occluded subspaces into $\mathcal{S}$ and call it *static object influence subspace* (SOIS) $\mathcal{S}_i$. Figure 21 shows two $\mathcal{S}_1, \mathcal{S}_2$ for two cameras.

From now on, we concentrate on reconstructing the voxels which represent dynamic objects in the target space, i.e. dynamic situation.

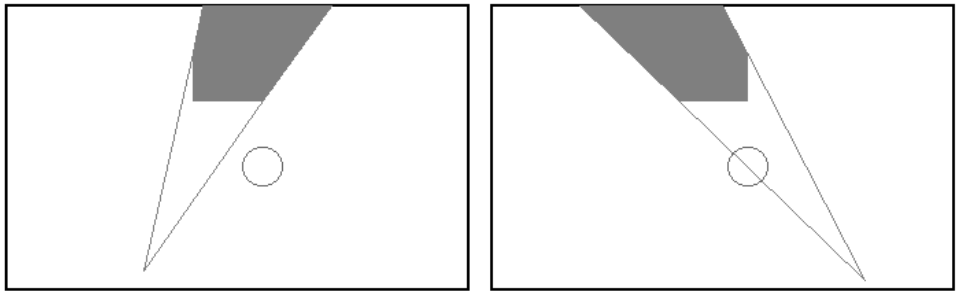


Figure 21: Static Object Influence Subspace $\mathcal{S}_1, \mathcal{S}_2$

### 7.1.2 3D Reconstruction

When the dynamic objects are imaged by a camera $i$, they exist within a viewing pyramid that circumscribes their projected regions on the image and whose apexes are focus point of the camera. We call all the projected regions in the same image together a *dynamic region* $D_i$ and let us denote a subspace consisting of these viewing pyramids by $\mathcal{V}_i$. Figure 22 shows two $\mathcal{V}_1, \mathcal{V}_2$ for two cameras.

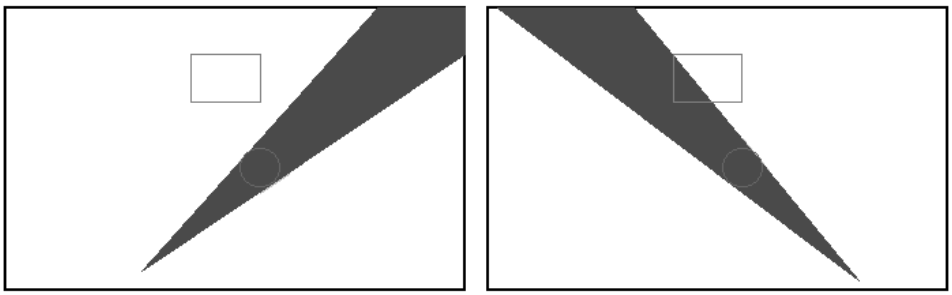


Figure 22: Viewing Pyramid $\mathcal{V}_1, \mathcal{V}_2$

As the dynamic objects can exist only outside $\mathcal{S}_i$, we only care a subspace named *existence shadow subspace* (ESS) $\mathcal{U}_i$ defined by Equation (7).

$$\mathcal{U}_i = \mathcal{V}_i \cap \overline{\mathcal{S}_i} \tag{7}$$

The dynamic objects exist somewhere inside $\mathcal{U}_i$. Figure 23 shows two $\mathcal{U}_1, \mathcal{U}_2$.

In the case where the dynamic objects are imaged by $n$ cameras, they exist within the product of all of these pyramids. We denote this subspace as $\mathcal{U}$ where

$$\mathcal{U} = \bigcap_{i=1}^{n} \mathcal{U}_i \tag{8}$$

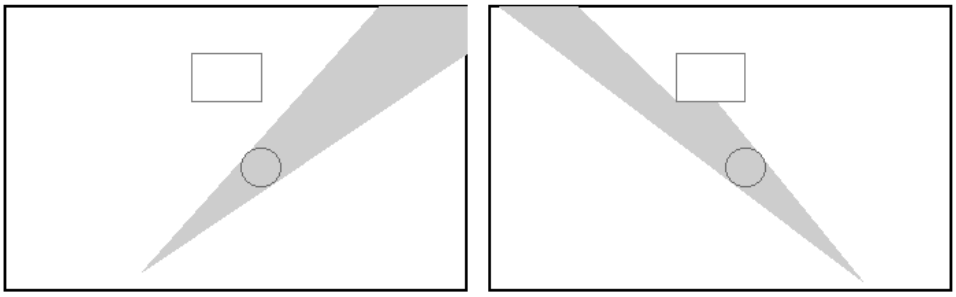


Figure 23: Existence Shadow Space $\mathcal{U}_1, \mathcal{U}_2$

Figure 24 shows the result $\mathcal{U}$. If the number of the cameras is small, there exists shape difference between $\mathcal{U}$ and the shape of the dynamic object, but it comes small as the number of the cameras increases.

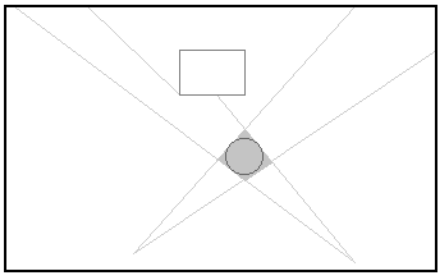


Figure 24: Reconstructed Shape $\mathcal{U}$

Suppose there are $n$ cameras in the real space and the cameras capture images simultaneously. We call this set of images *a frame*. 3D reconstruction process named *3D composer* can generate $\mathcal{U}$ at each frame under the condition that $D_1, D_2, \cdots, D_n$ are given in advance, because $\mathcal{S}$ and the focus locations of the cameras are known. Since $D_i$ is described by a binary image, the amount of transferred data is quite small.

This 3D reconstruction calculation in Equation (8) is easily expanded to parallel distributed computing because several 3D composers can reconstruct different subspaces simultaneously. Thus we achieve spatio-division of the 3D reconstruction process based on the locality of 3D reconstruction calculation.

In the actual implementation, 3D composer generates $\mathcal{U}$ by voxel representation.

### 7.1.3 Pipeline Processing

Our 3D reconstruction system named SCRAPER can be divided into three stages.

1. image capture
2. extraction of dynamic region
3. ESS calculation by SSM method

Let us call this sequence of stages a *path*. If the 3D reconstruction is done in this order sequentially, some parts of the system always idle. For example, when images are being captured, extraction and ESS calculation cannot be done. As a result, throughput is low and that is not desirable for real-time applications. To improve the throughput, we propose to activate several paths simultaneously in the pipeline architecture. We prepare three kinds of processes: *image captor*, *extractor*, and *3D composer* and increase the number of these processes to support multiple paths.

In our prototype system, an image is captured by the video capture card for which CPU power is not necessary whereas an extractor needs CPU power because it extracts $\mathcal{D}_i$ by detecting regions where the pixel values differ from its background image taken beforehand, so the two processes need only one CPU to work together. In addition, captured image data which is transferred to the extractor is not small and so it is not desirable to use physical network device to transmit the data between them. Therefore, we assign one video image captor and one extractor on the same workstation. As a result, the number of video image captors and that of the extractors are the same as that of the cameras.

On the contrary, the number of 3D composers can be increased because the calculation on the 3D composer is completely localized. The system can improve the throughput by preparing the 3D composers on different workstations distributed in a LAN.

As a consequence, the throughput is improved by preparing the multiple paths in the pipeline architecture, which means temporal division of 3D reconstruction process. The number of the paths are subjected to the number of the 3D composers the system can offer. Figure 25 shows the process timing chart when the system has three cameras and four 3D composers and assigns two 3D composers at each path.

We introduce a process named *scheduler* to synchronize the processes in the pipeline architecture.

## 7.2 Experiment

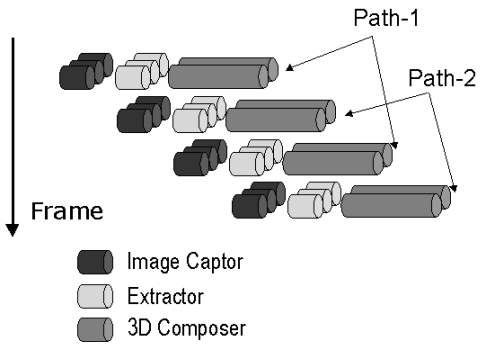



Figure 25: Pipeline Architecture

### 7.2.1 Geometric Reconstruction

We implemented a 3D reconstruction system named SCRAPER. We experimentally reconstructed a part of a lecture room (Figure 17) in the graduate school of informatics in Kyoto University.

The target space is imaged by four SONY EVI-G20 video cameras fixed at the corners of the lecture room. Table 8 shows the camera location in the room coordinate system.

In the experiment, we prepared four image captors and four extractors, and used four SUN Ultra2 200MHz workstations for them. We prepared four 3D composers and assigned them to four SUN Ultra1 170MHz workstations. A scheduler runs on a different workstation. All the workstations are connected on a LAN. The scheduler makes a synchronization among the image captors, the extractors and the 3D composers via 100 base-T Ethernet and 155Mbps ATM LAN. The dynamic region data from the extractors to the 3D composers are transferred on ATM LAN.

Figure 27 shows SOOS defined by the static object database given in advance. The SOISs from each camera in Figure 17 are shown in Figure 28. These subspace have been calculated before the SCRAPER system starts the reconstruction.

The system reconstructed the target space which was imaged more than three cameras. Hence, several parts of the target space were observed by four cameras, and the other parts were observed by three cameras. In the case four cameras imaged the subspace, $n$ in Equation (8) should be four, and in the other case, if a camera $j$ could not observe the subspace, $\mathcal{U}$ is the product of $\mathcal{U}_i, (i = 1, 2, \cdots, n,\ i \neq j)$. Figure 29 displays the target space which is visible by at least three cameras in the lecture room.

In the experiment, the image captor takes images with the size of 320 × 240 pixels. The camera which locates the furthest location from the target space images a cubic

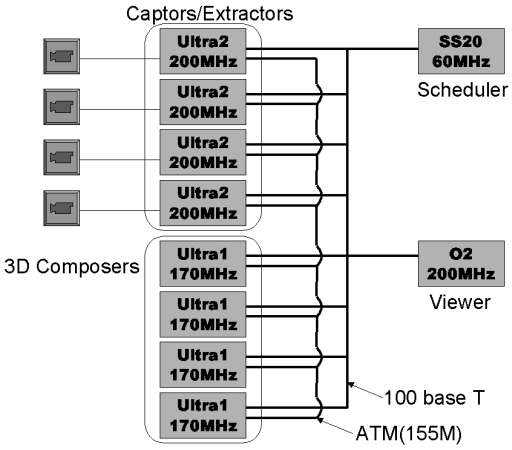



Figure 26: LAN Layout

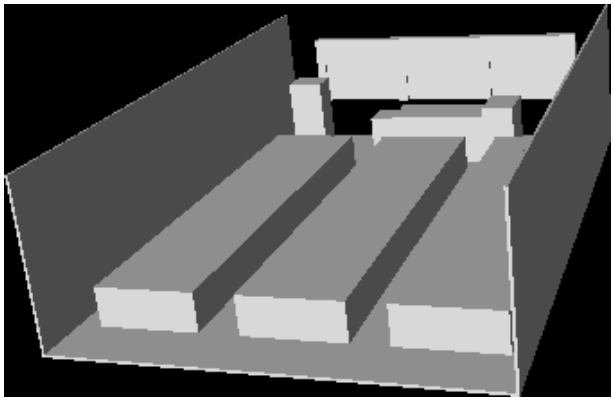
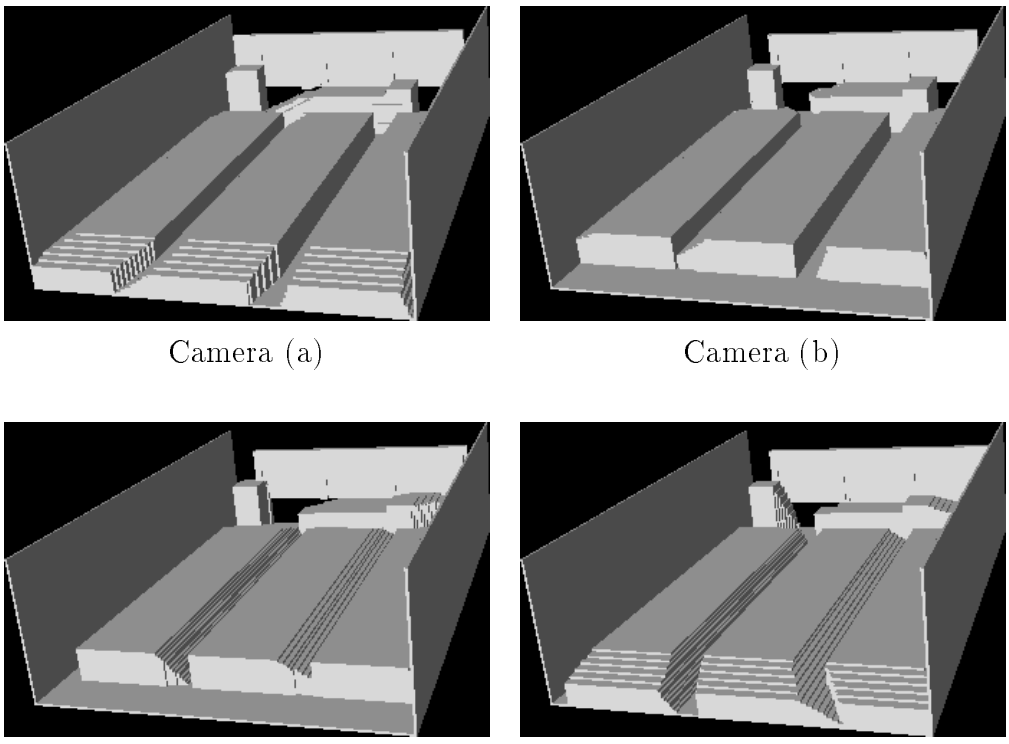


Figure 27: SOOS



Camera (a) Camera (b)

Camera (c) Camera (d)

Figure 28: SOIS of Cameras

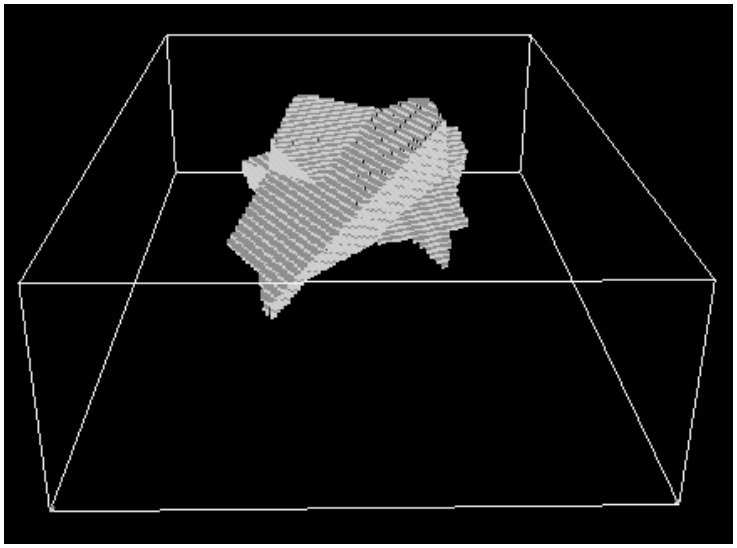


Figure 29: Target Space

subspace of 5 centimeters on a side in the target space onto one pixel in the captured image. Therefore, we set the voxel size as a cube of 5 centimeters on a side. The target space shown in Figure 29 corresponds to 96,769 voxels.

We conducted an experiment to measure the throughput and the latency of our prototype system. We put a box as a dynamic object whose size is $55cm \times 55cm \times 25cm$. The result of using four 3D composers are shown in Table 10. A variable $r$ indicates number of 3D composers served in each path and $s$ indicates number of paths in the system. We also conducted an experiment with only one 3D composer just for comparison and its throughput is 2.2 $fps$ and its latency is 1,384 $msec$.

Table 10: Throughput and Latency

| 3D comp. per path : $r$ | 1 | 2 | 4 | 1 |
|---|---|---|---|---|
| Number of paths : $s$ | 4 | 2 | 1 | 1 |
| Latency [msec] | 730 | 560 | 490 | 1,384 |
| Throughput [fps] | 7.3 | 7.2 | 6.1 | 2.2 |

The required throughput and latency differ according to the applications. One good feature of our method is that we can change the formation suitable to the applications by changing $r$ and $s$. The result indicates that the case of two 3D composers at two paths is good because the throughput is almost the same as four 3D composers at one path and the latency is as short as that of the case of the four paths.

We implemented a virtual space viewer which displays the reconstructed real space as a set of voxels in real-time. This viewer is an implementation of the information media environment. It displays not only the dynamic objects but also the static objects given to the system in advance, so a user can walk around the lecture room and observe the real space from any viewpoint with a little delay.

An example of a captured image is shown in Figure 30 . Figure 31 shows the reconstructed space displayed by the viewer. The voxels displayed in the center corresponds to $\mathcal{U}$, which were transmitted from the SCRAPER system.

### 7.2.2 Photometric Reconstruction

We have been working on photometric reconstruction that can reconstruct not only 3D geometric shape but also color information using multiple cameras. Our method also uses voxel based approach, which enables us hardware acceleration by splitting reconstruction process by voxel unit.

Camera (a)

Camera (b)

Camera (c)

Camera (d)

Figure 30: Input Video Images

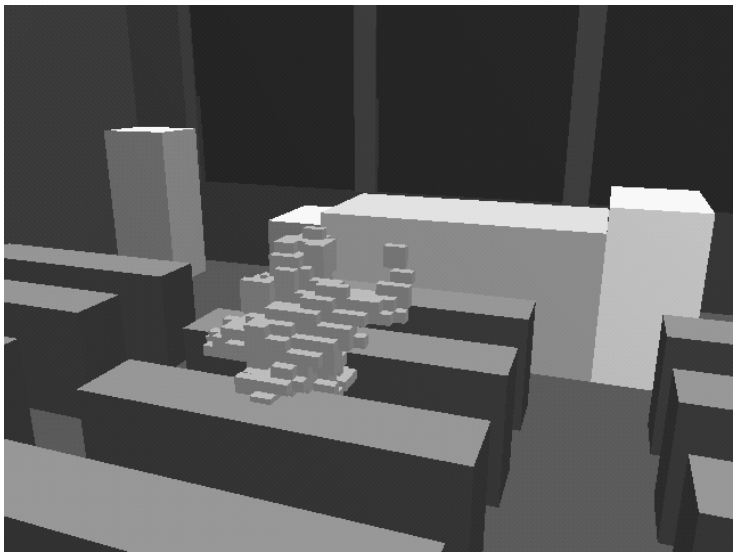
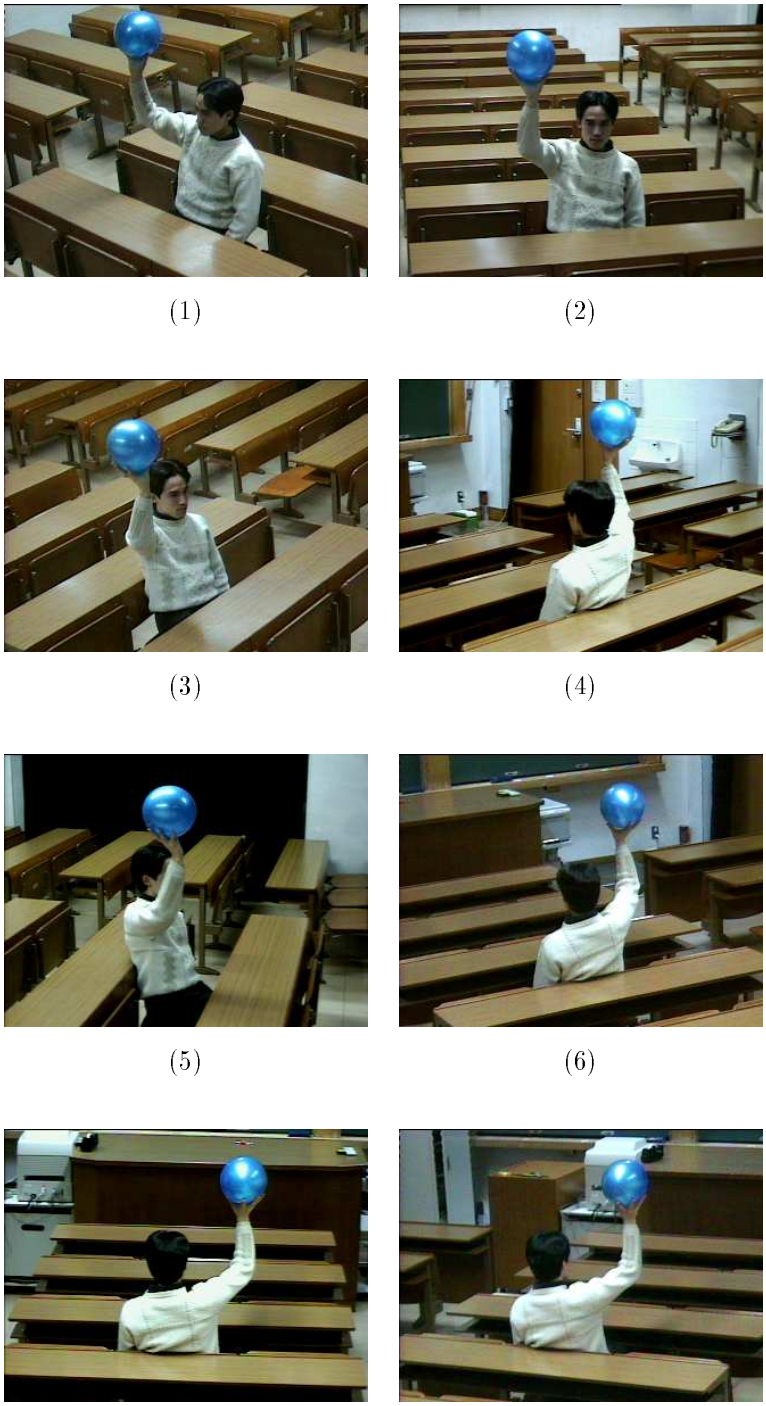


Figure 31: Reconstructed Space from Virtual Viewpoint

With respect to color information on the focused object, our proposed method extracts specular color reflectance which is considered to be difficult with other method. We show an example result in Figure 32 and Figure 33. A blue ball which has high specular reflectance is reconstructed by our method. Figure 32 shows 8 input images. In Figure 33, (i) shows the reconstructed shape and (ii)(iii) are the reconstructed images with diffuse and specular reflectance information. (iv) is same as (ii), but it is reconstructed without specular information. The images shown in (v)(vi)(vii)(viii) are the reconstructed images from virtual 4 viewpoints between (ii) and (iii). You can see the high light moving on the surface of the blue ball.



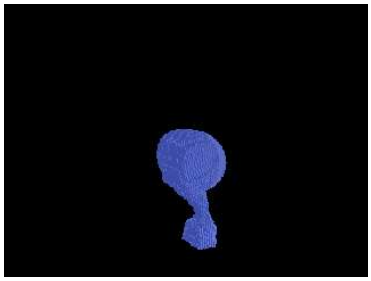
(1) (2)

(3) (4)

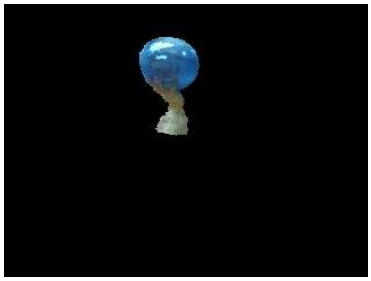
(5) (6)

(7) (8)

Figure 32: Input Images



(i) Reconstructed Shape

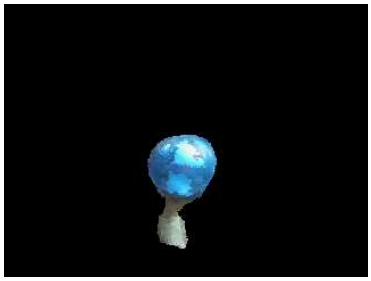


(ii) View from Camera(5)

(iii) View from Camera(6)

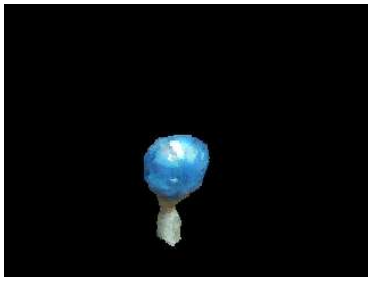
(iv) Diffuse Color from Camera(5)

(v) Virtual Viewpoint1



(vi) Virtual Viewpoint2

(vii) Virtual Viewpoint3



(viii) Virtual Viewpoint4

Figure 33: Photometric Reconstruction Results

# 8 Conclusion

We proposed two types of imaging method for 3D lecture room.

The first method that produces raw video images as final product is realized in three different ways; surveillance, broadcast, and customizing video philosophy. Multiple pan/tilt/zoom cameras are controlled according to the detected dynamic situation that consists of several A-components. We clarified that both the imaging philosophy and the understanding of the situation in the scene are essential to provide good video images to the users.

The frame work of the cooperative distributed vision gives a good process role design especially on imaging under the customizing philosophy.

We also proposed fast 3D shape reconstruction method because it enables us to view the scene from arbitrary eye-point without any physical limitation. Distributed processing is a key technology to archive the fast reconstruction.

With respect to the raw video generation approach, we have a plan to let the imaging system cooperate acoustic process, on-line teaching material, and WWW interface on imaging a lecture room. Application for sports or plays is an interesting future study for us.

On making the synchronized virtual space, finer shape reconstruction is required to provide good synthesized video images. In addition, not only fast geometric reconstruction but also fast photometric reconstruction should be studied in the future.

# References

[1] Yoshinari Kameda, Takeo Taoda, Koh Kakusho, Michihiko Minoh: "High Speed 3D Reconstruction by Pipeline Video Image Processing and Division of Spatio-Temporal Space," IPSJ Journal, Vol.40, No.1, pp.13-22, 1999. (Japanese)

[2] Yoshinari KAMEDA, Takeo TAODA, Michihiko MINOH: "High Speed 3D Reconstruction by Spatio-Temporal Division of Video Image Processing," The Transactions of the IEICE D, J83-D, No.7, pp.1422-1428, 2000.

[3] Takeo Kanade and Peter Rander, "Virtualized Reality: Constructing Virtual Worlds from Real Scenes," IEEE MultiMedia, vol.4, no.1, pp.34-47, 1997.

[4] Arun Katkere, Saied Moezzi, Don Y. Kuramura, Patrick Kelly, and Ramesh Jain, "Toward Video-based Immersive Environments," Multimedia Systems, vol.5, No.2, pp.69-85, 1997.

[5] KAMEDA Yoshinari and MINOH Michihiko, "A Human Motion Estimation Method using 3-successive video frames," Proc. of Int. Conf. on Virtual Systems and Multimedia(VSMM)'96, pp.135-140, 1996.

[6] Y. Kameda, T.Taoda, M.Minoh: "High Speed 3D Reconstruction by Video Image Pipeline Processing and Division of Spatio-Temporal Space," IAPR Workshop on Machine Vision Applications, pp.406-409, 1998.

[7] T. Matsuyama: " Cooperative Distributed Vision – Dynamic Integration of Visual Perception, Action, and Communication –," Proc. of Image Understanding Workshop, Monterey CA, Nov, 1998.

[8] Yoshinari KAMEDA, Kentaro ISHIZUKA, Michihiko MINOH: "A Study for Distance Learning Service - TIDE Project -," IEEE International Conference on Multimedia and Expo, Vol.3, pp.1237-1240, 2000.

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

[10] 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.

[11] 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.

[12] T.Kanade, A.Yoshida, K.Oda, H.Kano, and M.Tanaka, "A Stereo Machine for Video-rate Dense Depth Mapping And Its New Applications," Proc. CVPR, pp.196-202, 1996.

[13] K.Sato, A.Yokoyama, and S.Inokuchi, "Silicon range finder-a real-time range finding VLSI sensor," Proc. IEEE 1994 CIC, pp.339-342, 1994.

[14] R.Raskar, G.Welch, M.Cutts,A.Lake,L.Stesin, and H.Fuchs, "The Office of the Future: A Unified Approach to Image Based Modeling and Spatially Immersive Displays," SIGGRAPH98 Conference Proceedings, Annual Conference Series, pp.179-188, 1998.

[15] R.Grzeszczuk, D.Terzopoulos, G.Hinton, "NeuroAnimator: Fast Neural Network Emulation and Control of Physics-Based Models," SIGGRAPH98 Conference Proceedings, Annual Conference Series, pp.9–20, 1998.

[16] 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.

[17] 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.

[18] Hideaki Miyazaki, Yoshinari Kameda, and Michihiko Minoh, "A Real-time Method of Generating Lecture Video for Multiple Users Using Multiple Cameras," The Transactions of the IEICE D-II, J82-D-II, No.10, pp.1598-1605, 1999.