A Live Video Imaging for Multiple Users

Yoshinari Kameda
Center for Information and Multimedia Studies, Kyoto University
kameda@media.kyoto-u.ac.jp

Hideaki Miyazaki
Graduate School of Engineering, Kyoto University
miyazaki@kuis.kyoto-u.ac.jp

Michihiko Minoh
Center for Information and Multimedia Studies, Kyoto University
minoh@media.kyoto-u.ac.jp

Abstract

Many activities are held in a certain fixed space in our society. It becomes possible to participate in such activities from remote places. Video is one of the best media for this purpose, and it is important to consider how to image such activities. In this paper, we propose a new video imaging method with multiple cameras for multiple users. Dynamic situations in an activity are defined for describing an imaging rule of each user, and they are detected by processing sensor data. Concretely, the imaging rule is described by a user with camera-works linked to each dynamic situation. With this description of the imaging rules, many users can request their own favorite camera-works, while the system mediates these requests to satisfy as many users as possible. We have built a prototype system and conducted an experiment in which several users could participate in lecture.

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 technically becomes possible to watch such an ongoing activity from a remote place thanks to power-up of computers and speed-up of transmission lines.

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 there and can change their direction and zooming parameter dynamically. Therefore, the dynamic camera work is a key issue in our research.

Formulated camera works in movies[2] and computer supported movie makers[3][4] 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.

We propose a method of live video imaging for multiple users, which is in a field of multimedia content generation from the real space. This research is located as a visualization part of the CDV framework[1].

In our method, each user is asked to write his/her *imaging rule* which represents his/her favorite way of imaging an activity. To describe an imaging rule, we define the dynamic situation of the real space and *camera-works*. The representation of them is the point of this paper.

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 the imaging rules of the users and is described by the situation features derived from sensor data, i.e. video images.

The camera-work is concerned how to image an object. It consists of three elements; object, imaging direction, and zooming degree of the camera. With this mechanism, the live video imaging is realized.

In addition, for multi-user environment, every user has his/her own favorite video imaging, which conflicts with a constraint of the location and the number of the cameras. Therefore, *mediation process* is necessary.

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

We explain the imaging rule and the representation of dynamic situation in Section 2. In Section 3, our proposing multi-user live video imaging method is presented. Section 4 describes experiments and shows the result.

2. Description of imaging rule

As the imaging rules are different among users, every user can proclaim his/her imaging rule for each dynamic situation. We describe the way of representing the dynamic situation and how to proclaim the imaging rules.

2.1. 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 and 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 advantage of this representation is to avoid the pattern recognition problem. In other words, a user takes advantage of the symbolic representation, while the system takes advantage of the feature representation. Hence the description has two layers. The bridge between the two layers is the two way representation of the dynamic situation.

In the symbolic representation, the dynamic situation is described by a set of actions of the active objects. A description of this 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 the *situation features*.

2.2. A-component

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 the confusion, we prohibit describing the A-component in the passive form.

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.

2.3. Situation features

The dynamic situation is also described by the situation features. Feature extraction methods would differ if the real space and the activities are different. Most of the situation features are extracted via image processing because image sensor does not affect the human activities performed in the real space. 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 3. These tables are, what we call, knowledge representation of the lecture.

The extraction methods for the situation features are explained in Section 4.2.

2.4. Camera-work

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

$$w(objectlabel, direction, range) \quad (1)$$

Direction field indicates the relation between the direction of the object and the camera direction. Note that it implies the camera location because the locations of the cameras are fixed in our environment. Range field tells the size of the object in the image.

In our prototype system, we adopt eight directions and five discrete range values. See Figure 1 and Figure 2.

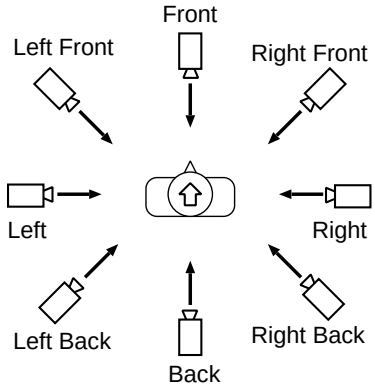

Figure 1: Direction of camera-work

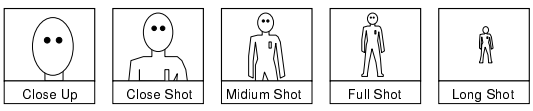

Figure 2: Range of camera-work

2.5. Imaging rule

An imaging rule is a set of functions from a dynamic situation to camera-works . Hence, the imaging

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	-	-

rule $\mathcal{I}$ consists of the two-tuples, A-component and the 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 the camera-works.

The dynamic situation in the real space varies as time is passed. This is detected by the sensors, and the situation features are extracted. Referring to the other representation of dynamic situation, i.e. situation feature representation, the dynamic situation is detected.

Let us explain by an example. Suppose the user defines the imaging rule $\mathcal{I}$ like Table 2 against the A-components shown in Table 1. The empty rows in the Table 2 mean that the user does not want to image the object (lecturer of the A-component No.3 for example) even if the corresponding A-component is detected.

If the A-component No.1 and No.4 are detected in the real space at time t, the imaging rule $\mathcal{I}(t)$ consists of three camera-works w_1, w_2 and w_8 in Table 2 and represents the request of the user at that time.

Whereas a certain imaging rule is defined by one director in conventional remote lecturing system or ordinary multimedia video generation, our approach allows all the users to define imaging rules so that they can proclaim their favorite imaging ways.

3. Multi-user live video imaging

3.1. Video imaging for multiple users

When a person participate in the activities in the real world, he/she can watch the activities in his/her own way. In the case a user is in a remote place, the user would do the same as if he/she were in the place. This means that each user in the remote place is assigned to an active camera in the target space. If we can reconstruct the target space in 3D space, this could be possible with virtual cameras[5]. However, considering image quality, this is impractical. Therefore the problem here is how to get the users' favorite video images under the constraint that the number of the active cameras is limited.

Since it is difficult to see multiple video images simultaneously and it needs broad band width of the network in transmitting multiple video images for one user, the user is assumed to watch only one video image. A video image selection process for each user should be prepared which satisfies the request of the user under the physical constraint of the active cameras such as locations, pan/tilt angles, and camera ranges.

Each user proclaims his/her favorite imaging rule in advance by specifying several camera-works at each dynamic situation. Since the number and position of the cameras are limited, the mediation is necessary to satisfy as many users as possible.

3.2. Mediation and user satisfaction

Request of an user j is satisfied if a camera-work in $\mathcal{I}_j(t)$ is selected and assigned to a certain camera at time t. Suppose there are u users and c cameras. As a camera can realize only one camera-work at a time, at most c camera-works can be realized at time t. As a consequence, mediation for all the user requests is necessary to select at most c camera-works among $\mathcal{I}_j(t)$ where $j = 1, \cdots, u$. Let us denote $n(t)$ be the summed 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. The mediation is described by the mediation matrix $M(t)$ of $n(t)$ 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. $m_{ij}(t) = 1$ means that the user j realizes his/her request of the camera-work i. Hence, the next inequality ought to be true for all j.

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

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 ith row is selected or not. $s_i(t)$ indicates the number of the users who support the camera-work i.

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

Table 2: 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	-

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

The constraint of the number of the cameras is formulated by

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

The mediation process is finding the mediation matrix M that maximizes $satisfied\_user$ under the constraints of Inequality (2) and (5).

$$satisfied\_user = \sum_{j=1}^{u} g_j(t) \quad (6)$$

where $g_j(t)$ is defined as:

$$g_j(t) = \begin{cases} 1 & (\text{if Inequality (2) is true}) \\ 0 & (\text{otherwise}) \end{cases} \quad (7)$$

The mediation matrix $M(t)$ that satisfies $satisfied\_user = u$ is always found in the case $u \leq c$, but there might be a dynamic situation where $u > c$ and so $satisfied\_user < u$.

We use greedy algorithm to solve this problem. Although it does not give us the best solution that maximizes $satisfied\_user$, it reaches locally optimal solution with less amount of calculation, which is essential in real-time processing.

After the camera-works are selected, the next problem is to assign the camera suitable to each selected camera-work under the constraint of the camera location, direction and range. The criteria to solve this problem is to satisfy the requests of the users as much as possible. Our strategy is as follows.

1. Order the selected camera-works by $s_i(t)$.
2. Choose the best camera for each camera-work in the order.

In this way, the video image the user j watches is taken by assigning the camera the selected camera-work in the mediation matrix $M(t)$ if his/her request is realized. On the other hand, there might be a case where some users cannot find their camera-works in the selected camera-works. In this case, the user watches an arbitrary chosen video image.

4. Experiment

4.1. Agent design

We implement our method in multi-agents system proposed in the Cooperative Distributed Vision (CDV) framework[1].

Functions in this system are classified into three categories. One is to detect the dynamic situation by extracting the situation features , another is to image objects by the camera-works based on the dynamic situation , and the other is to mediate the requests of the multiple users . We design three types of agents for each function.

An agent that extracts the situation features is called an *observation agent* . The observation agents have different functions one another because each observation agent extracts different situation feature. The number of the observation agents is determined by the number of the situation features that are needed to detect dynamic situations .

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

The last kind of the agents is designed mainly to mediate the requests of the multiple users. We call this kind of agent a *mediation agent*. While imaging agents and observation agents are device (camera or sensor) dependent, the mediation agent is device independent. Currently, we build one mediation agent that interprets

the information of the dynamic situation from the situation features and selects the camera-works based on the mediation procedure.

4.2. Calculation of situation features

We implemented a prototype system at a lecture room in the graduate school of informatics in Kyoto University.

The target space is imaged by four Hi8 video cameras which are set on a pan/tilt mount at the center of the walls, and four SONY EVI-G20 video cameras fixed at the corners of the lecture room (Figure 3).

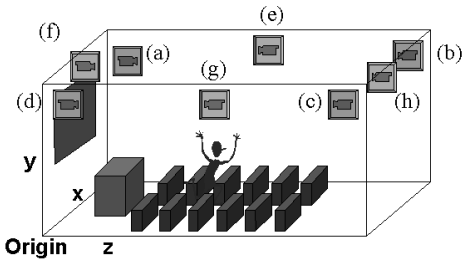

Figure 3: Camera layout in the lecture room

The system uses three kinds of the situation features to detect the five dynamic situations shown in Table 1; lecturer's location, voice level, and activation degree of student group. They are shown in Table 3.

Table 3: Situation features and A-components

A-component ID	lecturer location	lecturer voice level	student group activation
1	-	positive	-
2	blackboard	positive	-
3	blackboard	-	-
4	-	-	positive
5	-	-	zero

The lecturer's location is measured by the image based triangulation by the two active cameras (b) and (c) in Figure 3.

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

We divide the student desks into six groups and call the students in one desk group the student group. The activation degree of the student group is represented by area of the subtraction region in the image of the student group from their front view. The cameras (a) and (d) in Figure 3 are assigned to measure this situation feature.

4.3. Experimental result

We conducted a experiment on a lecture. The imaging agents used four active cameras (e) - (h) in Figure 3. We prepared 10 users and they proclaimed 141 camera-works in their imaging rules in the experiment, and 43 out of them are different camera-works.

Figure 4 shows the detected dynamic situations in the example lecture. The line indicates the periods that the dynamic situation labeled at the left is detected (see also Table 1). Multiple dynamic situations are detected from time to time simultaneously.

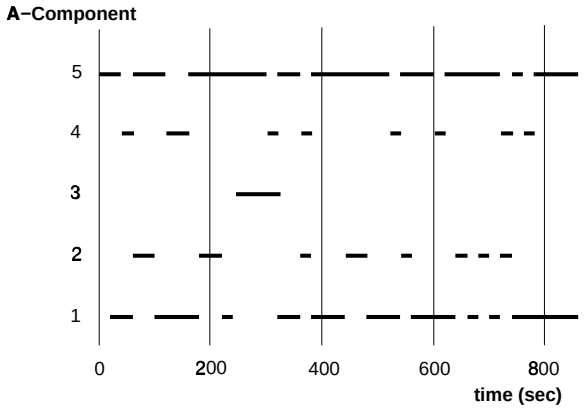

Figure 4: Detected A-components

Figure 5 shows the lasting period of the requested video image by each user. The line along the time axis represents the duration when the user watched the video image taken by his/her own imaging rule. As the duration is longer, the user's imaging rule is realized for longer time. The right figures show the percentage of the realized period against the whole period. Although there are only four cameras for the imaging agents, 10 users could watch their favorite video images in 69.5% period. An example snapshot of the camera-works is shown in Figure 6. It corresponds to the camera-work of w(lecturer, right front, medium shot).

We show the two snapshot sequences of the video images taken from the lecture. The video images in Figure 7 and Figure 8 are generated for the different users for about five minutes. User (A) proclaimed seven camera-works whereas user (B) proclaimed seven different camera-works . The horizontal axis indicates the time flow and C1 to C11 are labels of the selected camera-works. Note that both users sometimes watch the same video image because their desired camera-works are overlapped at that time.

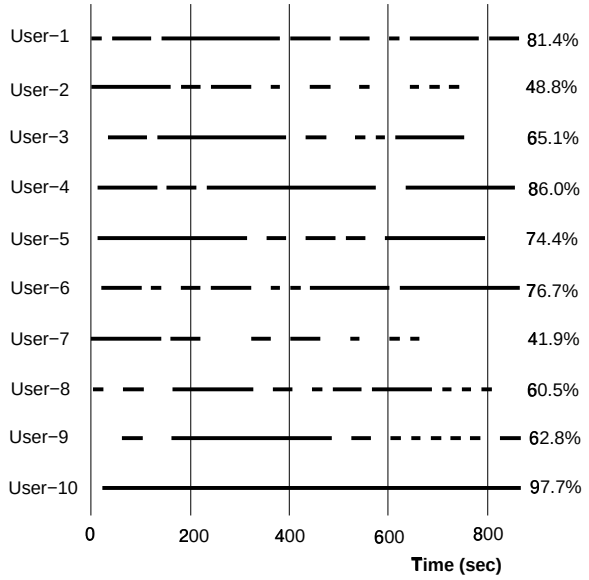

Figure 5: Duration of realized imaging rules

Figure 6: w(lecturer, right front, medium shot)

5. Conclusion

We have proposed the video imaging method that allows the multiple users to watch the activities in the real space in their favorite ways with assistance of the computers.

Our proposed approach can make the system to distribute live video images generated by the user's favorite imaging rules according to the snatched dynamic situation in the real space. We represented the dynamic situation in two ways to bridge the imaging rules of the users and the situation features, and formulated the mediation between the users' imaging requests and the physical constraint of the cameras so that most users can satisfy what they watch.

Further evaluation of the dynamic situation for the lecture is necessary in future. Also we would like to apply our method to other activities.

Figure 7: Video images for user (A)

Figure 8: Video images for user (B)

Acknowledgments

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

References

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

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

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

[4] 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*, 1996, pp.148-155.

[5] Yoshinari Kameda, Takeo Taoda, and Michihiro Minoh: "High Speed 3D Recognition by Video Image Pipeline Processing and Division of Spatio-Temporal Space", *Proceedings of MVA '98*, 1998, pp.406-409.