データベースシステム 116－56
（1998. 7. 9）

カメラの時空間記述情報を利用したビデオデータの空間ブラウジング

有川正俊
広島市立大学 情報科学部

抄録

大量のビデオデータの管理や問い合わせは，ビデオデータに人手で記述情報を付加し，類似性およびキーワードによって問い合わせする方法が主流として研究されてきた．一方，撮影時のカメラの状態を，各種センサーを用いて時空間記述情報として自動的に生成することが可能である．本稿では，このの時空間記述情報を実時間３次元コンピュータグラフィックスとして可視化を行うことより，大量のビデオデータを容易に管理・検索ができる３次元空間ハイパーメディアシステムを提案し，その有用性と基本構成を議論し，同時に実現したプロトタイプシステムの紹介を行う．

キーワード：空間質問，３次元ハイパーメディア，センサー，ビデオ

Spatial Browsing for Video Data Using Spatio-Temporal Descriptions of Cameras

Masatoshi ARIKAWA
Faculty of Information Sciences, Hiroshima City University

Abstract

Time-series description data concerning a camera can be automatically generated by various sensors, such as a camera's time-series positions, directions and zoom ratios so as to provide a rich environment for retrieving and browsing video data spatially. We also have used real-time 3D CG (three dimensional computer graphics) for user interfaces to browse videos in a virtual space corresponding to an existent space in the real world. Cameras and video sequences are represented as 3D icons in a virtual space. If we click one of the 3D icons, the corresponding video sequence will be replayed in a virtual space. This paper presents a basic principle of the 3D spatial hypermedia for video data browsing and reports some demonstrations of our prototype system.

KEY WORDS: Spatial Query, 3D Hypermedia, Sensors, Videos

1 Introduction

It is important to retrieve our intended scenes easily and naturally from a large amount of video data using their "description information." Many research efforts have been made for creating description information for video data by hands and for making use of the hand-made description information in order to retrieve our intended video sequences from large scale video data [2]. For example, keywords are attached to frames of videos and users can use the keywords for their queries to retrieve their intended video sequences. Narrations of video data can also be used as description information. The keywords and narrations have been created by persons for news videos on television, but it is difficult to create them for all videos, including home made videos etc., by human's hands.

Data of a camera's condition may be useful for the description information of video data. The data of the camera's condition such as position, direction and zoom ratio can be automatically gen-

erated by up-to-date sensors. The position can be measured by high precision GPS (Global Positioning System). The precise direction can be available using optical fiber gyros. The zoom ratio can be taken from a digital camera itself. This paper discusses how to make use of these automatic generated data of the camera's condition to browse and retrieve video data spatially.

2 An Overview of Spatial Browsing for Video Data

This chapter overviews the use of spatial description information of video data for browsing and retrieving video sequences of users' interests from the viewpoint of spatial queries. In this paper, a video sequence is considered a collection of time-series images. First, we discuss the use of some spatial description information of "photo pictures," that is, images, then extend it to the use for videos later. In the remainder of this paper, a picture means a photo picture which is one image. One picture corresponds to a camera at a certain moment. The camera at a certain moment is represented by some spatial attributes, such as its position, direction, zoom ratio and so on. 2D (dimensional) map data are also used for making their relations to some spatial attribute data of a camera. If the time-series positions of a camera for some duration are visualized on a 2D map, the visualization can show the distribution of the camera's movement which corresponds to a collection of time-series pictures. The position data can also be used to create clickable icons representing time-series pictures. If we click one of the position icons, the picture corresponding to the clicked icon will be shown on a screen. Also, we use the direction of the camera as well as its position to represent cameras or pictures (Figure 1). In that case, a camera may be visualized as an arrow which can provide the information of the direction in addition to the position. The arrow on a 2D map enables a user to understand what direction's scene can be viewed in a picture. If the position, direction, height and zoom ratio of the camera are available, we can tell what region in the real world can be taken or viewed in a picture at a certain moment(Figure 2). The region can also be used as a clickable icon for users' interactions. If we click the region icon, the corresponding picture will be displayed on a screen.

The spatial data, such as position and direction of a camera, can be used for spatial queries. The spatial data are also used as clickable icons for hypermedia 2D maps as we mentioned before. For example, if we want some pictures which view

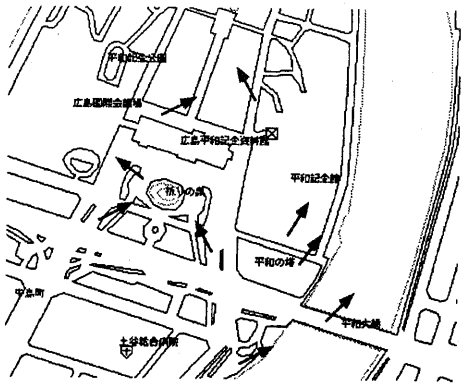


Figure 1: 2D arrow icons representing positions and directions of cameras on a 2D map.

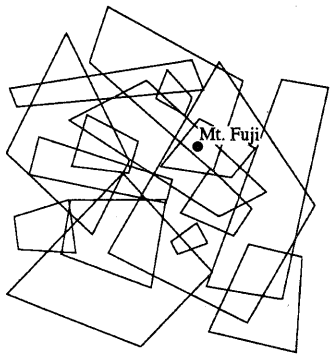


Figure 2: A set of 2D regions viewed in a picture or a video frame by a camera at a certain moment.

Mt. Fuji, we can realize it by making a spatial query to find regions containing a point (Figure 2). The regions represent pictures and the point represents Mt. Fuji in Figure 2. Thus, we can indirectly retrieve pictures by queries on contents in pictures using spatial data corresponding to pictures. We can extend this idea to applications of video data. A video data can be considered a collection of images or pictures. The unit images comprising a video sequence are called "frames" of the video sequence. Each frame of a video sequence can correspond to a camera at a certain moment and it has its spatial description information (Figure 3). Figure 3 shows the correspondence between an image and its momentary camera's condition as well as some temporal relations of time-series images. We also make some spatial queries for video sequences. For example, we can

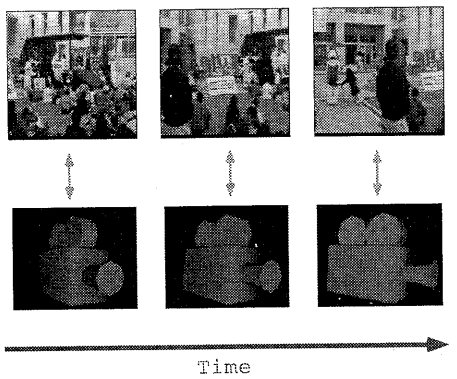


Figure 3: A sequence of time-series video frames and moving cameras in the real world.

make a query to select some video sequences which show Mt. Fuji. For cameras taking videos, their time-series positions and directions can be automatically recorded by using sensors. The movement of the camera for the duration of playing a video sequence can be simulated. We can see the camera's movement and replay the corresponding video at the same time. If we visualize all spatial description data for all frames of a video sequence as arrow or region icons on a 2D map, the number of icons or regions becomes too large to browse and click. We must simplify such many icons for pieces of video sequences. For example, many arrow icons can be represented by a small number of arrow icons. The representative arrow icons should be considered to represent video sequences or camera's movements. A video sequence should be replayed when the corresponding icon is clicked by a user.

There is a problem how to divide a video sequence into multiple video sub-sequences which should be represented by arrow icons. For example, arrow icons may represent segments of every 5 minute video sequences. It is useful to divide a video sequence into more meaningful subsequences, but it will be much more difficult than dividing them into a pieces of constant duration sub-sequences. A simple method of generating a representative arrow icon for a video sequence is to use the average values of the positions and directions of a set of time-series momentary cameras. The method often fails because it is generally difficult that an average value represents all values for any case. Selecting methods of generating representative arrow icons should be adaptable for various cases.

We can extend the idea of spatial browsing of videos on a 2D map to the one for a 3D CG space. All 3D CG spaces discussed in this paper can correspond to existent spaces in the real world. Such 3D CG spaces corresponding to parts of the real world are called "3D virtual spaces" in this paper. We can walk through a 3D virtual space, and can browse or retrieve our intended video sequences which should view similar to the current views in the current 3D virtual space. This user's requests of interactions can be interpreted as spatial queries which find arrows or regions closer to user's intention. The view in a 3D virtual space means the user's intended region and can serve as the selective condition of a spatial query. Furthermore, while walking through a virtual space, a user can click an arrow icon representing a video sequence so as to replay it in another window on a screen (Figure 4). It is promised to incorporate replaying videos into a virtual space as components of the virtual space (Figure 5). Users appreciate past real-world videos in a 3D virtual space. This kind of application is called Augmented Virtuality. It can provide users with more spatial experience. Another application of using spatial queries for 3D virtual spaces is to retrieve and replay videos which show a user's clicked objects in a 3D virtual space. This application uses a spatial query to select regions, which correspond to video sequences, including or intersecting the user's clicked objects.

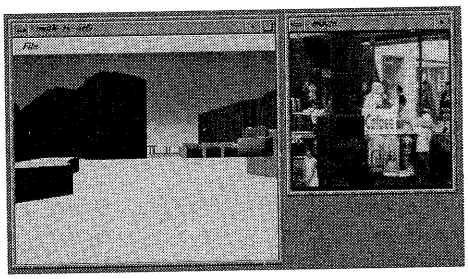


Figure 4: A moving camera in a 3D virtual space and the corresponding video replayed in another window.

3 A Model of Spatial Browsing for Video Data

3.1 Pictures and cameras

A picture can be represented by condition of a camera. The condition includes position, height, direction, zoom ratio, characteristics of the lens, time and so on. The term "image" is used for

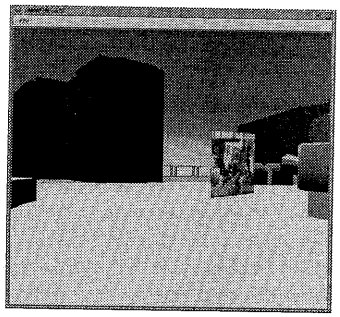


Figure 5: Replaying a video projected on a rectangle plane in a virtual space with the corresponding moving camera icon.

a picture in order to construct some definitions in this paper. An image can be identified by an image identifier, id_{image}. An image of id_{image} is denoted by $image(id_{image})$ and has two attributes concerning time, t and camera identifier, id_{cam}.

[Definition of "image" taken by a camera]

$$image(id_{image}) : (t, id_{cam})$$

t : time or duration
id_{cam} : camera identifier

A camera at a certain time or duration is defined as follows.

[Definition of time-series "camera"]

$$cam(id_{cam}, t) : (id_{image}, id_{space}, p, d, zoom, reg)$$

id_{space} : space identifier
p : position
d : direction
$zoom$: zoom ratio
reg : $region$ which covers the area viewed in the image of id_{image}

A camera at a certain time has attributes of id_{image} which corresponds to one image, and id_{space} which addresses a space such as the real world, a 2D map and a 3D virtual space. The attributes p, d, $zoom$ and reg respectively denote the camera's position, its direction, its zoom ratio, and the region which is viewed through the camera.

3.2 Spatial objects and spaces

A spatial object (SO) at a certain time or duration is defined as follows.

[Definition of "Spatial Object(SO)"]

$$SO(id_{SO}, t) : (id_{space}, geo, p, d, s, graf)$$

id_{SO} : spatial object identifier
t : time or duration
id_{space} : space identifier
geo : geometry
p : position
d : direction
s : scale factor
$graf$: graphic properties

Since meaning of attributes described above are apparent, we will explain only new attributes. id_{SO} is a unique identifier for a spatial object as a component of a space such as a real space and a virtual space. geo represents the geometry for a spatial object and is used in a spatial query. p, d, s are used to define the geometry in a local coordinate. $graf$ means a set of graphic properties such as its material color and texture. A space is defined as a set of spatial objects.

[Definition of "Space"]

$$Space(id_{space}, t) : \text{a set of Spatial Objects}$$

For example, a 2D map of Hiroshima City in 1998 is a visualization of a space composed of spatial objects such as Hiroshima Baseball Stadium and Peace Park in Hiroshima in 1998. The time attribute values for the spatial objects are all the same, and can be considered that their time attribute values are inherited from the time attribute of the space.

It is useful to visualize cameras on a 2D map to access to pictures of our interests. For example, if we visualize a camera's position where a picture was taken from, it helps us realize the place where the picture was taken. Many positions are visualized on 2D maps as the places of pictures, and clicking one of the visualized positions of cameras can cause popping up the corresponding picture on a screen. If the direction of a camera is visualized as an arrow icon on a 2D map as well as its position, we can get more information from the visualization. Pictures can be selected from the viewpoint of the direction in addition to the position. It is useful to put some annotations beside the arrows. The examples of the annotations are the time when pictures were taken and names of persons who took the pictures. Furthermore, the pictures correspond to regions which are viewed in the pictures (Figure 2). The region can also be a representative symbol of the picture. If we want to see some pictures of the regions of our

interests, we just click one of the regions so that the corresponding picture will be displayed on a screen. Thus, a camera can be visualized as a spatial object in a 2D map. The following is an example of visualizing a camera as an arrow.

$$SO(id_{SO}, t_i) : (\ id_{space_j} = SPACE_j,\ geo = \text{ARROW\_GEOMETRY},\ p = cam(id_{cam}, t_k).p,\ d = cam(id_{cam}, t_k).d,\ s = \text{DEFAULT},\ graf = \text{DEFAULT}\ )$$

3.3 Spatial query for spatial data for pictures and real world objects

First, we discuss the use of only position data for spatial retrieval. For example, we are looking for some pictures which were taken from the bridge "A." The bridge "A" may be interpreted as a region and the query is made for finding a set of points of camera's positions covered by the region. The query is realized in the following query formulation.

$$\{\ c \mid c \in CAM \text{ AND } c \in Space(id_{Hiroshima}, 1998) \text{ AND } SO(id_{Bridge"A"}, 1998).reg \textbf{ contain } c\ \}$$

"$Space(id_{Hiroshima}$" means a set of spatial objects in Hiroshima in 1998. CAM means a universal set of cameras. $SO(id_{Bridge"A"}, 1998).reg$ defines the region of the bridge "A" in 1998. **contain** is a spatial comparative operator which returns TRUE if its left side geometry contains the right one, and FALSE if not.

If direction data of a camera become available in addition to its position data, we can make more complex spatial query. For example, we can retrieve some pictures by a spatial query which formulates that the pictures were taken from the bridge "A" and show the north direction from the bridge. Furthermore, we can bind pictures with their corresponding "regions" which cover scenes of the pictures. We can make much more practical spatial query by using the region data for pictures. For example, we search for some pictures which show the building "B." To realize this query, we have to use spatial data of the building "B," that is, the location of it or the region of it. The position of the building can be addressed by specifying the name of the building "B." The spatial query will find a set of regions containing the position of the building "B." The selected set of regions are converted into a set of pictures which view the building "B." Then, the set of pictures viewing the building "B" can be displayed on a screen as a result of a user's request.

3.4 Video data extensions

This section discusses video data with a 2D map. Video data are considered as a sequence of pictures which are called "frames" of videos. Each frame has its time attribute. For example, a video is composed of 30 frames of pictures per second. The corresponding spatial data for all frames of a video sequence may become large. If we put the corresponding representative symbols for all frames on 2D map, the number of representative symbols for them becomes too large to browse and click. If the motion of the camera is very fast or we use only a short-duration video sequence, it will be better to use all representative symbols for all frames. In the case of a long duration video sequences such as 10 or 30 minutes videos, the number of the symbols becomes too large and the visual result becomes too dense or unreadable. For example, even if the camera is still or does not move, a large number of the same symbols for all frames of the video sequence are generated. It may be useless. In the case that the camera is still, it is reasonable to use only one representative symbol for many same symbols corresponding to a certain duration video sequence.

It is difficult to decide the duration of video sequence which only one symbol should represent. It is naive to use only one symbol for a time-series video sub-sequence. We can use a certain constant duration video sequence. For example, we can use a point icon for a set of frames of every 5 seconds video sequence in a virtual space. We can appreciate the path of the camera's movement through the time-series point icons. The movement of a camera can be represented by animating point symbols in real-time on a 2D map. The point symbol can be extended to the arrow symbol which represents direction of the camera as well as its position. If users can click one of the arrow symbols on a 2D map, the corresponding video sequence can be played in another window on a screen. Region symbols are also useful and can correspond to all frames of a video sequence. It is good to use some of all region symbols for all frames. For example, representative region symbols can stand for segments of every 5 seconds video sequence in order to decrease the number of symbols on a 2D map. The regions can be animated in real time on a 2D map for representing the movement of the camera.

A video sequence should be divided into multiple meaningful segments of video sub-sequences. There are many ways to make representative symbols for the segments of video sub-sequences. For example, a line symbol can be used as a collection of point symbols, each of which corresponds to

each frame of a video sub-sequence. If the camera was panned frequently while taking a video, it is difficult to use only one still arrow to represent all directions of the camera for a certain duration. In that case, many arrows may represent the panning camera. It is also useful to make special symbols or animated symbols for a certain duration movement of a moving camera.

Previous paragraph has discussed the ways of visualizing the movement of a camera as components of a hypermedia 2D map to access to video sequences of our interest. Here, we discuss some use of spatial query for spatial data for video sequences. The basic principle of retrieving video sequences of our interests is almost the same as the one for retrieving pictures. For example, if we search for some video sequences which were taken from the bridge "A" and show the building "B," we should make some queries to find some frames which were taken from the place near the bridge "A" and view the building "B." The videos which include the frames selected by the spatial query should be replayed on a screen as a result of video data retrieval.

3.5 3D virtual space extensions

If a 3D virtual space data are available, the movement of the camera should be visualized using 3D CG techniques for browsing and retrieving video data. Representative 3D icons can be used for video sequences. The basic principles of browsing and retrieving video sequences in a 3D virtual space are almost the same as the case of a 2D map. We walk through a virtual space and can execute some spatial queries to select video sequences which could show the same views as the current views in the virtual space. Videos can be incorporated into a 3D virtual space as components of it. It means videos can be replayed in a 3D virtual space in real time. We can experience the direction and position of the camera which took a video, and appreciate the video more spatially. It allows users to immerse themselves in a 3D virtual space with realistic video browsing.

4 An Experimental System

For our experiment, we used two sensors, a gyro and a GPS (Global Positioning System) to generate time-series spatial data for video sequences which were taken by a digital video camera. We collected some spatial data for videos taking some scenes of campus festival of Hiroshima City University held in October 18th and 19th, 1997 (Figure 6, 7). The gyro was precise enough to record the direction of the camera. The GPS was not precise to measure the position of the camera. We compensated the position data of the camera by plotting every one second positions of it on a 2D campus map by hands. The zoom ratio can be obtained automatically from the current style of digital camera, but we could not prepare the procedure to read the internal condition of the digital camera. Instead, we created the zoom ratio for every one second by hands by means of watching recorded video images and measuring the focal points by eyes. To simulate the continuous movement of the camera as real-time 3D computer graphics animation, we use the method of linear interpolation for the discrete time-series spatial data such as position, direction and zoom ratio.

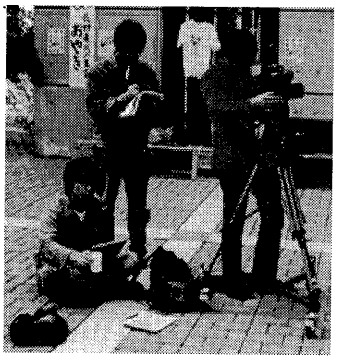


Figure 6: A crew taking a video of scenes of campus festival of Hiroshima City University with a digital video camera and sensors.

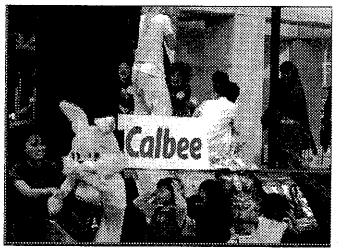
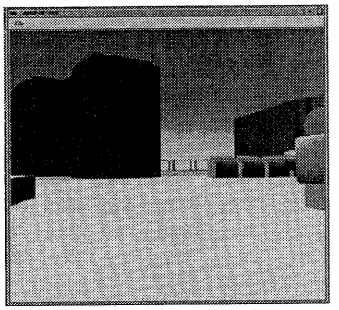


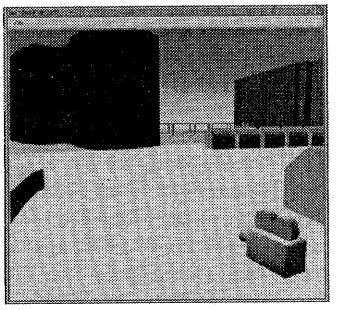
Figure 7: A scene of campus festival of Hiroshima City University in October 1997 which was taken by a digital video camera.

We have implemented two typical applications for spatial browsing video data. One is to simulate the movement of the camera as real-time 3D CG animated camera icons in a virtual space.

Users could appreciate the movement of 3D CG animated camera icons in the virtual space from arbitrary views (Figure 8). If we click a 3D CG animated camera icon, the video corresponding to the current 3D CG animated camera icon can be replayed on a CG rectangle plane in front of the 3D CG animated camera icon in a virtual space (Figure 9). The zoom ratio of the camera is used for the distance between the camera and the CG plane for replaying a video. Thus, we can appreciate spatially both the movement of the camera and replaying videos in a 3D virtual space.



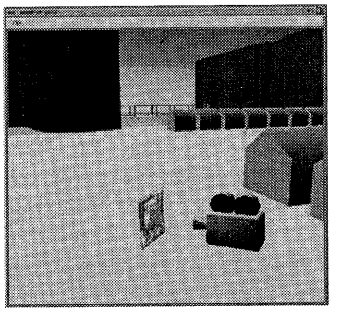
(A)



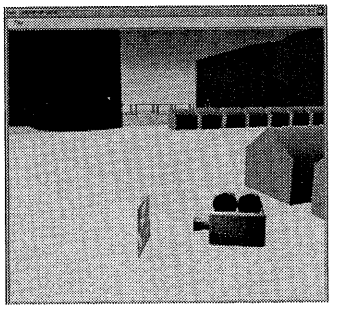
(B)

Figure 8: Simulation of the movement of a real-world camera from two different user's views.

The other application is to use 3D CG still arrows, each of which represents each of segments of video sequences. In the experiment, a video sequence has been divided into meaningful segments of video sub-sequences by hands. 3D CG still arrow icons representing the segments of the video sequences were automatically generated by visualizing the average values of the position, direction and zoom ratio. The length of the arrow represents the zoom ratio of the camera (Figure 10). We can walk through a virtual spaces with the 3D CG still arrow icons which address video



(A) A video plane at the time "t"



(B) A video plane at the time "t+Δt"

Figure 9: Replaying a video on a rectangle plane with a moving camera icon in a 3D virtual space.

sequences and indicate the average values of the position, direction and zoom ratio of the movement of the camera (Figure 11). If we click one of the arrow icons, we can appreciate videos being played in a virtual space (Figure 12). In this application, the camera to view the virtual space is restricted to be set in the same position, direction and zoom ratio of the real-world camera. It guarantees that we can see the video in the middle of the scene in a virtual space. We could have a wider view compared to only a video being played in another window on a screen. The video can be augmented as a wider view and be imposed of a virtual space so as to enable a user to experience the video more spatially.

5 Concluding Remarks

Sensors can be affordable and become precise enough to generate spatial description data for objects in the real world. The spatial description data of a video camera's condition measured by the spatial sensors is also useful for browsing and retrieving video data spatially. The spatial data of the video camera enables richer visual in-

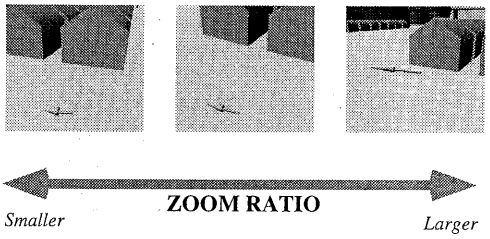


Figure 10: Zoom ratio of a camera is represented as the length of an 3D arrow icon.

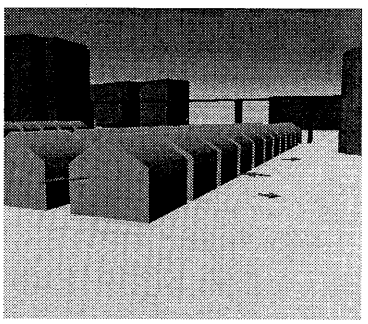


Figure 11: 3D CG arrow icons representing video sequences in a 3D virtual space.

dex environment. We can retrieve both pictures and videos in 3D virtual spaces as well as on 2D maps. Also, real-time 3D computer graphics hardware enables us to walk through a 3D virtual space and to appreciate replaying videos in the 3D virtual space. This kind of environment can provide users with spatially browsing videos in a virtual space. We can experience virtual spaces in more realistic ways because real world videos can be replayed in a virtual space. While we walk though 3D virtual spaces which have their corresponding existent real world spaces, we can be aware of the existence of videos in the virtual spaces with watching 3D icons for video sequences, and appreciate replaying videos more spatially and naturally in virtual spaces. This application may promise to manage video data for some application domains, such as disaster management systems, simulations of the real world, and virtual sight-seeing tours of the real world.

Though there are many problems in managing a large amount of ideas, we believe our technology in this paper could give some ideas to solve some of the problems presented in this paper. We have

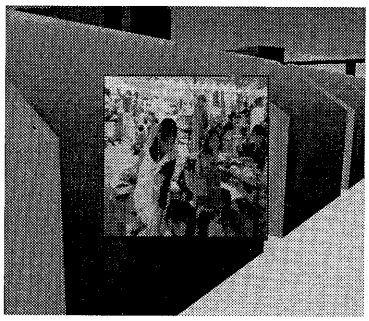


Figure 12: A replaying video plane positioned in the center of a window because of synchronization of a virtual camera and a real camera.

been researching on making a theoretical model of these applications. We also plan to evolve our prototype system to more practical one from the viewpoint of functionality, easiness of use, and salability for the size of real video data.

Acknowledgements

We would like to thank Mr. Tetsu Kamiyama, Hohchi Shinbun Co., for his great effort to advance this research and help the implementation of our prototype system. We appreciate supports by staff at Database Systems Laboratory, Department of Intelligent Systems, Faculty of Information Sciences, Hiroshima City University. This work was supported in part by the Grant-in-Aid for Scientific Research on Priority Areas "Advanced Databases" of the Ministry of Education, Science, Sports and Culture of Japan, Research for the Future Program of Japan Society for the Promotion of Science under the Project "Researches on Advanced Multimedia Contents Processing (JSPS-RFTF97P00501)," and the grant "Specified Research" at Hiroshima City University. Our virtual campus data was created by Miss Hiromi Michiyori, a student of Faculty of Art at Hiroshima City University, using SOFTIMAGE, Microsoft Inc.

References

[1] ART+COM, http://www.artcom.de/.

[2] K. Uehara, M. Oe, K. Maehara, "Knowledge Representation, Concept Acquisition and Retrieval of Video Data," Proceedings of the International Symposium on Cooperative Database Systems for Advanced Application, Kyoto, Japan, December 5 - 7, 1996, pp. 218 – 225.