Interactive Video Description on the Network

— Interactive Video Representation of Real World based on Digital City Map —

Tomoyuki YATABE, Hiroshi KAWASAKI, Masao SAKAUCHI
Institute of Industrial Science, University of Tokyo
7-22-1 Roppongi, Minato-ku, Tokyo, 106-8558, Japan
yatabe@sak.iis.u-tokyo.ac.jp

Abstract

Owing to the recent development of digital broadcasting technology such as Satellite, CATV, the Internet and the expansion of the network systems, the amount of video data we can obtain is increasing day by day. In this paper, we propose Advanced Database TV (ADTV), an interactive video database system which lets us share video data on the network, as an application for future Interactive TV service. We propose an interactive video database system to process real-world video data as a further example of ADTV. On its implementation, we propose an automatic organization method focusing on video objects in each frames to describe video data in an efficient way. The prototype system provide three basic services , "Description", "Question and Answer" and "Image and Information Retrieval", about buildings as real-world video objects linked with digital city maps.

1 Introduction

Recently we can obtain more and more video contents which are provided by TV Broadcasting through CATV, Satellite and the Internet day by day. However, we consider video contents won't be enough in the future, because they are distributed in many programs by broadcasting at the same time. In the environment, "Real-time Video" which is not edited and processed will be more distributed than edited or processed one in broadcasting programs. The video has too many video objects such as street scenes, buildings, trees, cars, for us to access needed video data easily. For example, it is so hard to search video frames including some building and some one. Thus content based access methods are required. It is an effective solution that people not only watch video passively but also enjoy interactive service actively and process information which are needed. We propose an interactive system that users can join broadcasting program services actively.

In this paper, we discuss an automatic annotation method of video objects in real-world video using digital maps. Furthermore we propose a system that users can obtain annotations and images needed for themselves interactively.

2 Interactive Video System

It is under review that Digital TV will provide additional information about video on each programs and frames using MHEG[6] and HTML[2]. Video data streams distributed through the Internet can have more information and links to other contents by use of SMIL[9]. If the accomaniments to video are provided, we will be able to construct more advanced video database than current one. Viewers can select and retrieve information using it.

However, even if information is provided by not so many information providers probably, problem that lack of just needed information remain when viewer select and retrieve information. So, a mechanism that not only information providers but also users can distribute their own information is needed.

2.1 Advanced Database TV

We proposed a system, ADTV(Advanced Database TV): Many users who watch TV can describe content in real-time and interactively. Collected information from users and providers is share among users through the broadcasting or computer network and users can retrieve, question and answer about video objects by index of the information.

The system provides several services on video objects and information given interactively to users as follows.

- Question and Answer about video objects
- Retrieval about video objects and information
- Complement structure data

The system allows that not only interactive information of users but also precise and rich information given by providers and service providers, which are called "mediator" about video objects. Information given about video object is kept in database system.

In related works, video is structured by physical unit like "cut" or "scene", but we propose a new method which make structure based on video objects in frames.

2.2 Sharing data on the network

Users can share broadcasted video source because it is same, but data given interactively can't be shared in current broadcasting system.

Then collected data from users using network is managed in database, so information is provided to other one. It is widely known that DBMS(Database Management System) is designed for multi users use, therefore this function is realizable based on DBMS.

3 Automatic video structuring

We don't consider that all video objects can be described even so many users join to describe. In this situation, incomplete indexing for them doesn't generally satisfy functions: "Question and Answer" and retrieval function. It is needed that a function of the system make any reply to query of users as correctly as possible using incomplete indexes.

If a user describe some video objects at a frame, the system can't provide replies about the same objects at the other frames or the other objects at the same frames for users.

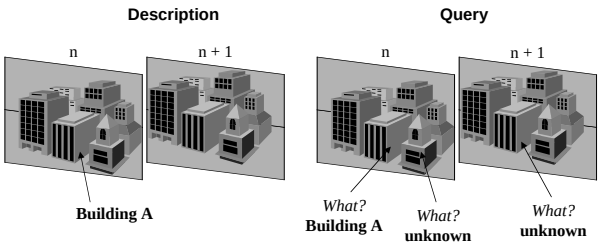

Figure 1. Query for frame or object beyond description range

For example, as Fig.1 shows, when users give information about a building, one of video objects, in the n th frame, other users can take an information about the building at the n th frame. However users can not take information about the building at the $(n+1)$th frame without spatiotemporal assumption, and about neighbor video objects at the n frame without spatial assumption. Because database of the system don't have information which is matched to query. This problem is solved by assuming video objects which are not described from information described by users spatially and spatiotemporally.

It is applied to spatiotemporal assumption to track object motion. In this process, motion tracking is done automatically. For example, this process is refer to a method proposed in [5].

On the other hand, it is hard to assume neighbor video objects spatially and automatically from information given for video objects. However, if there is obvious models of video objects particularly, they can be assumed from models spatially and automatically.

4 Automatic structuring methods of real-world video based on maps

When users give information of buildings in real-world video, we discuss a method of automatic structuring information of the other buildings using map models and description.

4.1 Process methods

Users and information providers give information about video objects at first. Therefore real-world video is matched with models based on digital maps, so users can obtain information with as follows process.

1. Making collect position of buildings in the real-world
2. Making models from digital maps
3. Matching between models and read-world video

Many buildings are recorded in sequence of real-world video frames. In Proc.1, buildings are tracked spatiotemporally and relative positions between them in the real-world are calculated from match models. Details are described in **4.2**.

Relative positions between buildings are found out using geographical data on maps. In Proc.2, models are created from calculated relative positions.

4.2 Real-world Video processing

We process the following video.

Target video: Recorded by camera which is turn sideways, against direction of moving perpendicularly on a moving car. There are buildings along the road in the video.

First, motion vector is calculated by block-matching method. We consider that the camera is moved straight at

some speed. Secondly, overlapped images based on motion vector can be obtained. Obtained images as these processes are handle as namely “Panorama images”. As results of these processes, Fig.2 shows panorama images are created actually based on camera motion using video recorded as the above.

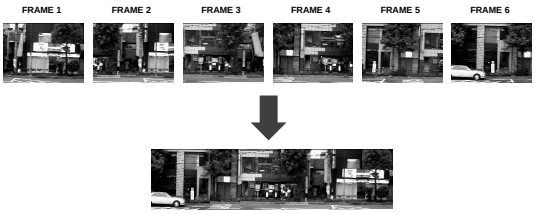

Figure 2. Panorama images based on camera motion

Buildings are tracked by calculating relative motion vector of camera. They are considered as solid models, thus real-world video is structured spatiotemporally and automatically by tracking them.

4.3 Making models from digital maps

Buildings along the road are recorded on the target video for process. Thus it is necessary to make models matched to video sequence. Boundary pattern of buildings is obtained from calculating digital map data geometrically.

When it assumes that there is two buildings along the road, boundary pattern of buildings which are exist between two buildings are obtained by as follows.

Two buildings are matched to two polygons on the digital map, when they are described as video objects.

As Fig.3 shows, boundary pattern of buildings which are existing from one described buildings to the other one models is determined as models of real-world video from digital maps.

4.4 Matching between models and real-world video

Fitting scaling and positioning by DP matching can make matching between boundary pattern models obtained from digital maps and processed image of real-world, namely “panorama image” which is representing position of buildings. It is enough to find out a pair of each correspondent points in order to solve the problem.

As Fig.4 shows, segments of building in real-world video are assumable, when position of each points are correspondent and boundaries of buildings are matched.

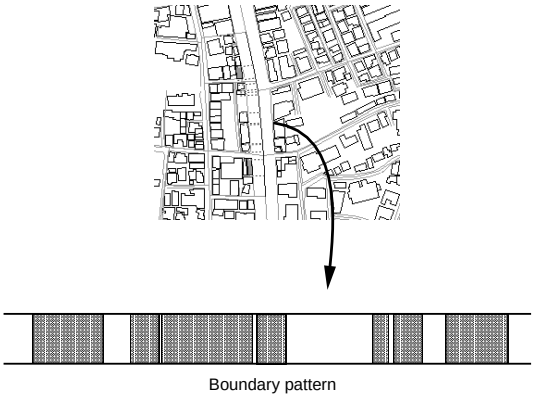

Figure 3. Boundary pattern of buildings

Thus the system make a reply to query and retrieval of users about unknown buildings if matched between models and images.

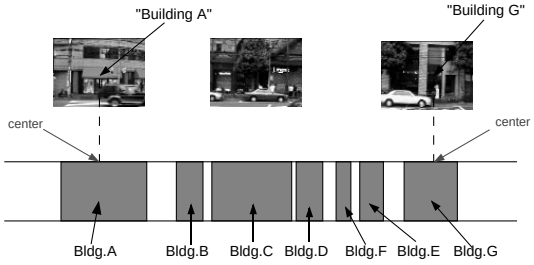

Figure 4. Matching between models and real-world images

5 Implementation of prototype system

As the above, we implement a prototype system which can provide three basic services ,“Description”, “Question and Answer” and “Retrieval” about real-world video for users.

The specification of application is as follow: The application runs on Silicon Graphics Inc. O_2 using $1/2500$ scale map and video stream which are encoded by SGI Movie format at 6 frames/second.

5.1 Description of buildings

Fig.5 shows an example image that users describe buildings. The left-bottom window of the figure, called video window, displays video sequence which is played or broadcasted. Map is displayed in the left-top window, called map window.

Users register information of known building in video as follows process.

1. Indicate segment of known buildings in the video window.
2. Indicate correspond polygon to the indicated building on the map window.
3. Thus matching segment of buildings and models is obtained from map.

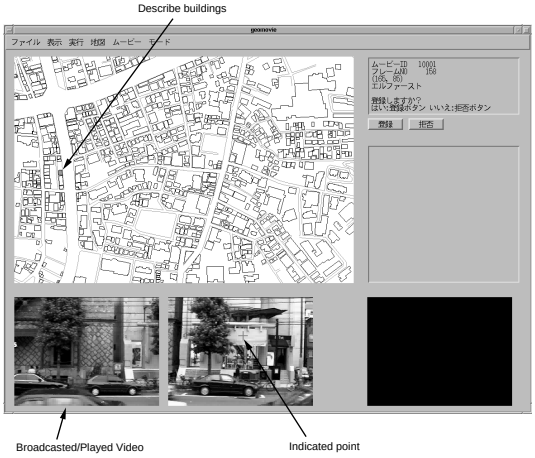

Figure 5. A snapshot of description

5.2 Question and Answer about buildings

Fig. 6 shows an example image that users make a Question and Answer about buildings. Users can query attribution about an building displayed in the played video.

Users query attribution of buildings displayed in video as follows process.

1. Indicate a point of buildings which users want to know in the video window. They can be both described one and the other one.
2. Assume the indicated building by a method which is described in Sec.**4**
3. The system present answers to user. They are name of building and position of map in this case.

5.3 Retrieval of building images

Fig.7 shows an example image that users retrieve images which have specific building.

Users retrieve images of buildings needed as follows process.

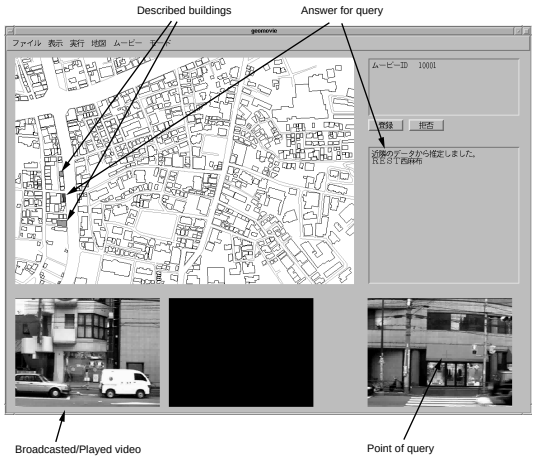

Figure 6. A snapshot of question and answer

1. Indicate a polygon of buildings which users want to retrieve in the map window.
2. Assume the indicated building by method described in Sec.**4**
3. The system present image and attribution building to user in the right-bottom of the figure. This image is center of sequential frames which have specific building partly or entirely in this case.

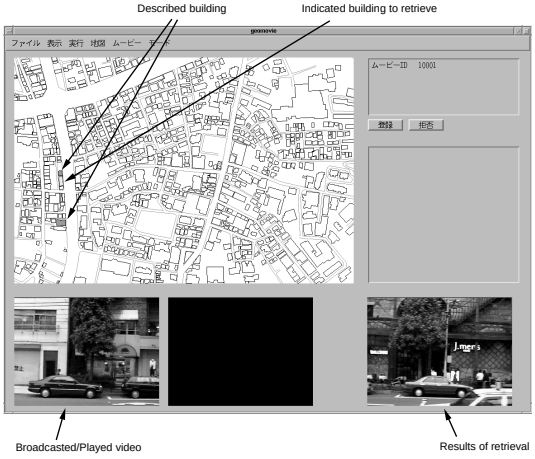

Figure 7. A snapshot of retrieval

6 Evaluate and Rusults

Buildings along the road are recorded by camera which is turn sideways, against direction of moving perpendicu-

larly on a moving car. Actually, we use a video which has 18 buildings in 280 video frames and is recorded by 6 frames per second for test.

In this situation, users query and retrieve about buildings which are between two described buildings by someone.

Thus precision of the question: "What is this building in the video?" is examined, when the nunber of building between two described building is changed to 4, 6, 8, 10 or 12. Precise rate is ratio of collect answer which users can get for buildings when users indicate near center of segment which have each building. Since users don't always indicate center of buildings exactly, error occurs at the assumption of building.

As the broken line in Fig.8 shows, the system can provide answer more exactly than when the number of buildings which are between two building described by users are small. Therefore if many video objects are described, it is easy to assume from nearest buildings. The method is effective if there are ten buildings or less between described ones. However, it is hard to annotate buildings which are far from marked object.

6.1 Revise method using edge information

Real-world video includes both buildings and the other video objects such as automobile, human and trees, thus it is hard to extract buildings from video. Therefore we have better results when video is pre-processed before separating buildings as follows: vertical edges obtained from video are including not only boundary of buildings but also frames of windows. They must be reduced by pre-processing.

- Vertical edge is not boundary of building in the section with strong horizontal edge
- Grouping vertical edge and they approximate lines
- Some gaps between buildings have many vertical edges, so both ends of them are considered as boundaries of buildings.

As the line in Fig.8 shows, it can be considered that the method as mentioned above make an improvement in all cases compared with original method. Because we consider that correct segmentation of buildings can be processed using edges.

7 Conclusions

We proposed a system, ADTV(Advanced Database TV): Many users who watch TV can describe contents in real-time interactively. Collected information is share among users each other through the network and users can retrieve,

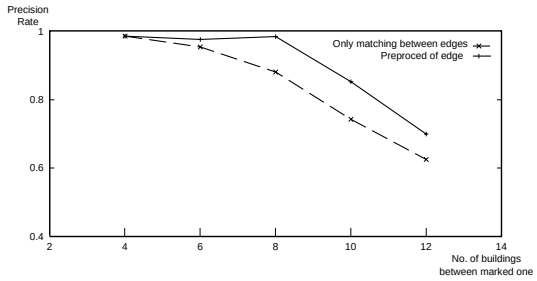

Figure 8. Comparison of precision rate between original and revice method

question and answer about video objects by such incomplete index of the information.

There is several problem in the ADTV, as one of them, we describe that it is hard to distribute only raw information given by users to the others. We discuss needs of automatic video structuring.

Concretely we propose a method of automatic structuring of real-world video based on digital maps and evaluate the method in prototype system. "Question and Answer" and "Retrieval" prepared function of ADTV are effective. Furthermore we propose revise method for automatic structuring using image information, show advancement of precision in "question and answer".

References

[1] H. Arisawa and T. Tomii. Design of multimedia databese and a query language. In *Proceedings of MULTIMEDIA '96 IEEE*, pages 462–467, Jun. 1996.

[2] ATVEF. Advanced television enhancement forum specification (atvef), 1999. `http://www.atvef.com/atvef_spec/TVE-public-1-1r26.htm`.

[3] M. Flickner, H. Sawhney, et al. Query by image content:the QBIC system. *IEEE Computer*, pages 23–32, Sept. 1995.

[4] A. Hampapur, R. Jain, and T. Weymouth. Digital video segmentation. In *Proceedings of Second Annual ACM Multimedia Conference*, pages 357–364, Oct. 1994.

[5] P. Kauff, B. Stefan, S. Rauthenberg, U. Gölz, J. L. P. D. Lameillieure, and T. Sikora. Functional coding of video using a shape-adaptive DCT algorithm and an Object-Based motion prediction toolbox. *IEEE Transctions on Circuit and Systems for Video Technology*, 7(1):181–195, Feb. 1997.

[6] T. Meyer-Boudnik and W. Effelsberg. MHEG explained. *IEEE MultiMedia*, 2(1):26–38, Spring 1995.

[7] E. Oomoto and K. Tanaka. OVID: Design and Implementation of a Video-Object Databese System. *IEEE Transactions on Knowledge and Data Engineering*, 5(4):629–643, Aug. 1993.

[8] R. Szeliski. Video mosaics for virtual envrionment. *IEEE Computer Graphics and Applications*, pages 22–30, 1996.

[9] W3C. Synchronized multimedia integration language (SMIL) 1.0 specification, June 1998. `http://www.w3.org/TR/REC-smil/`.