MULTI LAYER VIDEO OBJECT DATABASE BASED ON INTERACTIVE ANNOTATION AND ITS APPLICATION

Tomoyuki Yatabe Hiroshi Kawasaki Hiroshi Mo Masao Sakauchi

Institute of Industrial Science University of Tokyo
Roppongi 7–22–1, Minato-ku, Tokyo 106–8558, JAPAN
{yatabe,h-kawa,sakauchi}@sak.iis.u-tokyo.ac.jp, mo@nii.ac.jp

ABSTRACT

As the use of applications employing video contents becomes widespread, it is crucial for representation systems to be able to handle video effectively and flexibly. In this paper, the authors propose a method that has enriched video contents and an application for future interactive TV service, called Advanced Database TV (ADTV).

The system requires three basic methods: one is a video handling method based on video objects; another is a construction method of a video object database achieved by feature extraction and tracking of objects; and the third is a protocol for delivery annotations of objects which are given by users. The objective is to provide a multi-layer video object database based on users' interactive description of video objects. The database will have the ability to organize incomplete descriptions of video objects in each frame, mainly annotation, position and motion. We have implemented the prototype system, which provides users functions such as semi-automatic enrichment of annotations, similarity retrieval of objects, and semantic indexing, based on the video object database.

1. INTRODUCTION

Due to the huge amount of video content which has become obtainable through TV Broadcasting by Satellite, CATV and the Internet, multimedia applications using video have become popular, especially non-linear video editing systems and PC video recorders like VCRs. When using these applications, we are always aware that, while the amount of raw video data is very large, it nevertheless has no context or indexes. As a result, people experience difficulties in accessing necessary.

In future TV broadcasting, additional information on each frame will be given as enhanced electronic program guides (EPGs). It is natural to use the information for making a database; however, such information is not always sufficient for a video database. In the future broadcasting enviroment, it is expected that real-time video which is not edited or processed will be more widely distributed than ever before. And we assume that obtaining additional information from these real-time videos will be still difficult in the future.

In this paper, we describe how we concentrate on the objects in video frames. The objects are sets of regions which are extracted from the video stream because of their spatio-temporal feature similarity. Up to now, the achievement of automatic object segmentation methods has been a challenging task, and the methods' performances have not been robust; however some of them can segment and track objects[2] with reasonable accuracy. And furthermore, the newly established MPEG–4 standard[3] has proposed a new object-based framework for efficient multimedia representation.

For more effective representation it is necessar to construct some type of object-based video database[4]. But doing so is not easy, because automatic indexing of objects is not usually performed completely. Therefore, we propose a construction method for a video database which is not fully automatic, in that it needs the assistance of human interaction such as annotation, for example.

In this paper, we discuss the automatic structuring method of the description of objects in rally video sources. Furthermore, we describe how we constructed a prototype of our proposed interactive system to demonstrate how it allows users to participate in the broadcasting services and obtain information and images interactively.

2. VIDEO OBJECT DATABASE

Generally, video sources consist of sets of physical video clips, namely shots, and logical video segments, namely scenes. Much work has been done already in decomposing and classifying video data based on shot analysis[5, 7]. Shots consist of sequences of frames. These frames contain collections of regions, namely video objects. Fig.1 shows that all of the layers have some physical features and some semantic ones. Features at the various layers are used according to the types of applications. For instance, some au-

thoring systems use some features at the Object Layer, some editing systems must use features such as the cutting point at the Segment Layer.

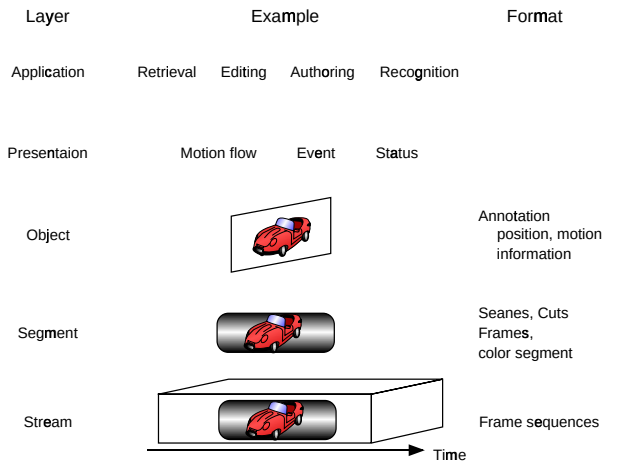


Figure 1: Multi layer representation of video structure

In our work, we focus on video objects which have spatio-temporal continuous structure in video frames. An efficient object-based representation has been proposed in the newly established MPEG–4 standard. This representation enables semantic object segmentation, called "Video Object" and content-based object searching for general sources. Furthermore, MPEG–7 standard defines an video object description interface[1].

2.1. Video object model

In this paper, we discuss video objects which have the following three features and their respective parameters:

Static: colors, texture, contour, edge

Dynamic: position, motion, transformation, co-relation

Semantic: description

However, it is difficult to extract these features with high accuracy, especially from general target objects rather than from specific ones. Therefore, we consider that, even if the methods extract features with errors, the video object database should allow them as features of objects and make corrections as appropriate as possible.

We propose a video object database based on the video object model. The model is described by *Bounding Box*, *Motion Vector of centroid*, and *Annotation* as shown in Fig. 2. All objects have such *Bounding Volume*(V) with annotation that is expressed as

$$V(oid) = \big(Bounding\ Box(t),\ Motion\ Vector(t)\big), \quad (1)$$

where,

$Bounding\ Box = (x,\ y,\ width,\ height)$
oid: object identifier assigned by the system.
t : time from starting one of the objects' appearance.

In addition, *Bounding Box* and *Motion Vector* are sometimes calculated by some image processing method; however, annotation is almost always described by users; in a few cases, it is automatically described.

Each set of video objects (VO) is described as

$$VO(oid) = \big\{\big(V_i(oid), Annotation_{i,j}(oid)\big)\big\}, \quad (2)$$

where,

i : shot identifier when object appears.
j : annotation identifier when each object has a different one.

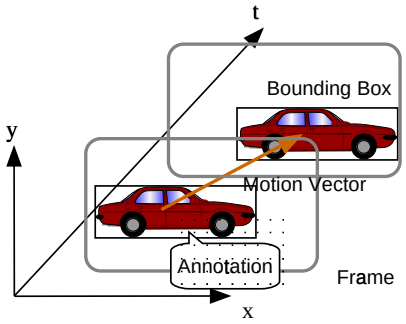


Figure 2: Proposed video object description model

2.2. Query and Retrieval function

In this way, we construct a video object database based on motion and annotation. The database provides many possibilities for the selection of query functions.

For example, we use a simple function to select information about an indicated object; this is one of the most basic functions of interactive TV. If the user's indicated point P_i is defined as (x_i, y_i, t_i), the system first provides corresponding *Bounding Box* and oid with P_i, although there are some cases of missing corresponding objects or taking some oids. In the former case, the system provides a nearest *Bounding Volume* or information in a corresponding frame as the answer. In the latter case, users can select from some objects provided as the answer.

The system also provides some functions of similarity retrieval. If the user indicates the *Bounding Box* area, the system extracts properties, image features and annotations from this area. Then in the next step, objects can be retrieved from the database using either annotation, or similarity retrieval using the image.

3. INTERACTIVE VIDEO SYSTEM

In the near future the Digital TV will provide additional information about videos on each program and frame. The other hand, Video data streams distributed through the Internet can have more information and links to other contents by use of SMIL.If the accompaniments to video are provided, we will be able to construct a more advanced video database than the current one. Viewers can select and retrieve information using it.

If the video information is not provided by so many information providers, the lack of the necessary information will be an important problem for viewers selecting and retrieving the information. So, it is an essential mechanism for TV that not only information providers but also users can distribute their own information.

3.1. Advanced Database TV (ADTV)

We proposed an interactive video system, called Advanced Database TV (ADTV)[8]. The system has the capability to provide functions such as real-time querying, indexing and describing content information for users who watch TV. In a previous paper, we chose video sources from whichwe recorded some buildings from a car as an indexing target. Video indexing is performed based on digital maps and spatio-temporal structure.

In this paper, we describe how we developed the system. Among the server functions are: to collect annotations about video objects from users and providers; to send back the result to users; and to broadcast fundamental annotations and index which are needed at client side interaction. Each client has the ability to display information received from the server at the indicated point and to make queries of manipulating objects. This information about video objects is kept in database system.

The system accepts information about video objects not only from individual users but also from information providers or a service providers, called “mediator” who give precise and rich information. As mentioned above, since video sources consist of some physical units like “cut” or “scene”, we propose a new method which makes a structure based on video objects in frames.

3.2. Protocol of database access

Users can share broadcasting video sources, but in the current broadcasting system, data attached to video cannot be shared. Fig.3 shows the architecture of ADTV. The information server makes indexes for objects in all frames which have some features, e.g., colors, region, location, motion and annotation. We construct the video object database based on these indexes.

In the query processing at each client, a user indicate objects about which he or she requires information. Some requests are sent to the information server; the server processes them as described above and sends back results of queries. The server distributes some information or a part of the database to each client in response to other request. Finally, all the clients display the result to the user.

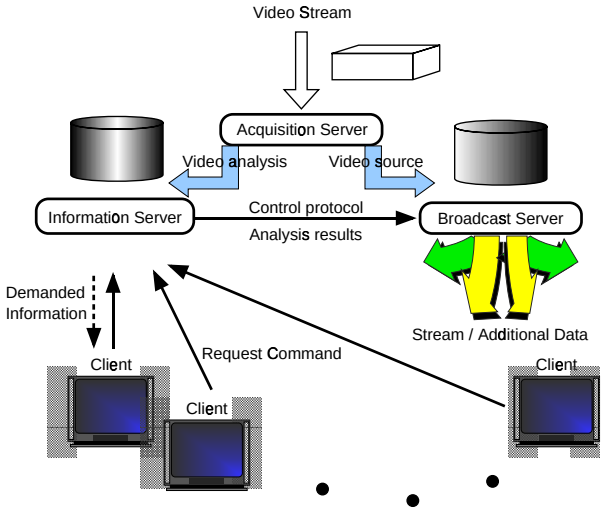


Figure 3: Architecture of ADTV

We use original format in protocol between client and server now. We consider using description based on XML such as SMIL at the protocol, because it is easy to display and process for clients. In this case, query parser and description generator server is needed to parse lient requests and convert queries for the database and to then generate descriptions for clients as a result of their queries.

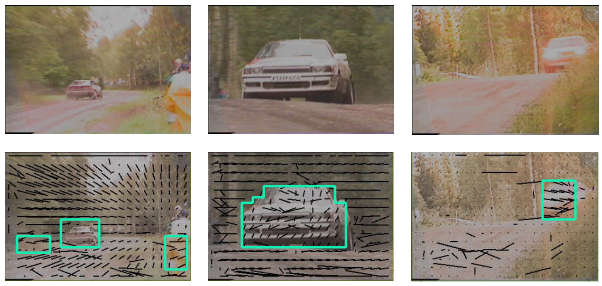
3.3. Database for real-world video objects

As mentioned above, we propose a video object database using feature extraction, tracking and description. We discuss a construction method of an object database for real-world video sources. It is well known that almost all real-world objects have 3-dimensional (3D) structure. In other words, if we prepare specific models fitted for target objects, it is effective to track them meeting the conditions of spatio-temporal continuity. In this process, motion tracking is performed automatically. For example, this process is refered to in a method proposed in [6].

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

We have developed a prototype system that handles video

sequences in which moving automobiles are recorded. We consider that spatio-temporal assumption methods also apply to general sports video sequences. Fig.4 shows 3 cases of extraction according to camera operations; motion compensation vectors which look like optical flow are characteristic in each case. We can extract some video objects from video sources using feature of vectors with high speed and some accuracy.



(a) Zooming (b) Pan (c) Fixed

Figure 4: Automatic video object extraction using Motion Compensation value in MPEG stream

3.4. Implementation of prototype system

As mentioned in the previous section, we have implemented a prototype system which provides basic functions: "taking annotation about video objects", "video annotation base on similarity retrieval". (In outdoor scenes, we consider the automobile to be a typical example of a video object; cars and buildings in the frames are looked upon as noise.) In this case, the target is some video sequences of "The World Rally Championship (WRC)". Some video shots are recorded by the same cameras at some points; therefore, video shots have similar frames with different automobiles.

The top of Fig.5 shows a video frame with a rally car, called "LANCIA DETLA". For example, if the video object database has the car name as *Annotation* and *Bounding Volume* of the car, someone takes information of car indicated in the frame. The procedure is the same in another frames using similarity retrieval.

In addition, the same car appears in another shots, so new VO can be created from V of same oid.

4. CONCLUSIONS

In this paper, we proposed a framework of a video object database for a interactive video system, Advanced Database TV (ADTV). We defined the video object as *Bounding Volume* which is described by *Bounding Box* and *Motion Vector* with annotation. Furthermore, some basic functions of queries and retrieval were defined in the video object database.

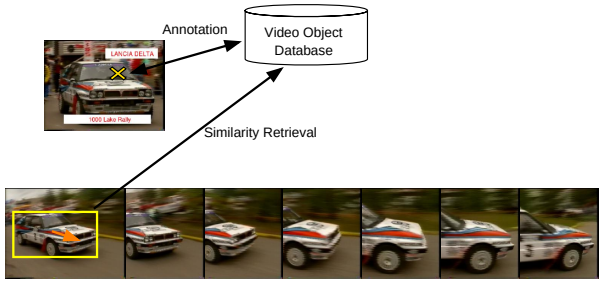


Figure 5: Users' interaction in the implementation system

There remain several problems in the ADTV; for example, it is hard to distribute only raw information given by users to the others. We discussed an indexing method of video sequence by feature extraction and tracking of objects. Concretely, we proposed a method of automatic structuring of real-world video, especially rally video, based on object features in the prototype system. We have presented some functions that performed well in the system.

We plan to develop the technique of feature extraction and tracking method not only for automobiles but also for other objects in general video sources. Furthermore, we will make a representation of the video object based on the acquired VO.

5. REFERENCES

[1] MPEG-7 overview. ISO/IEC JTC1/SC29/WG11 N3158, http://drogo.cselt.stet.it/mpeg/standards/mpeg-7/mpeg-7.htm, Dec. 1999.

[2] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong. VideoQ: An automated content based video search system using visual cues. In *ACM Multimedia97 Proc.*, pages 313–324, Nov. 1997.

[3] L. Chiariglione. MPEG and multimedia communications. *IEEE Transactions on Circuits and Systems for Video Technology*, 7(1):5–18, Feb. 1997.

[4] Y. F. Day, A. Khokhar, S. Dagtas, and A. Ghafoor. A multi abstraction and modeling in video database. *Springer–Verlag, Multimedia Systems*, 7(5):409–423, 1999.

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

[6] 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.

[7] K. Manske, M. Mühlhäuser, S. Volg, and M. Goldberg. OBVI: Hierarchial 3d Video-Browsing. In *ACM Multimedia98 Proc.*, pages 369–374, Sept. 1998.

[8] T. Yatabe, H. Kawasaki, and M. Sakauchi. Interactive Video Description on the Network. In *Proceedings of IEEE Multimedia Computing and Systems99*, volume 2, pages 194–198, Firenze ,Italy, June 1999.