MVA'94 IAPR Workshop on Machine Vision Applications Dec. 13-15, 1994, Kawasaki

Video Retrieval Method Using a Sequence of Representative Images in a Scene

Akio Nagasaka, Takafumi Miyatake and Hirotada Ueda
Central Research Lab., Hitachi, Ltd.

1-280, Higashi-Koigakubo, Kokubunji-shi,
Tokyo, 185, Japan

ABSTRACT

We describe a new method that can quickly search an entire video. This method is used to first detect every shot change of a target video in real time. Then unique frame sequence features at the shot changes are found. The method we have developed and described enables real-time scene retrieval. It may also be an appropriate basis for an information filter that automatically selects appropriate videos or scenes among newly input video.

1. INTRODUCTION

A large amount of video is produced for various purposes, including news, entertainment, and presentations. The volume and variety make it difficult to efficiently find videos which are of interested in. Some ways do exist to retrieve videos stored in database systems by means of an appropriate tag or annotation [1, 2]. In most cases, however, videos are released to the market without any helpful information for later retrieval. This is most typical of broadcast videos that is provided by mass media everyday. Advanced digital video production technology and network infrastructure will someday allow us to purchase full-annotated or structured videos but this day has yet to come. A video retrieval function that investigates the video data itself is needed in the mean time. For this, image recognition technologies are needed.

Understanding the meaning of a natural motion picture is quite difficult with current machine vision or voice recognition technologies. Model-based video parsing approach has been proposed [3] but this is not sufficient. We have taken a different approach. The method developed and described does not need any semantic information, such as annotations or keywords. It quickly searches for a video that contains a particular scene, or a short string of video, by comparing image features. For example, this function can find specific television programs containing similar scenes as the registered one, for example, an opening scene of television programs or the credits of movies. This can be used as a kind of information filter that automatically selects appropriate videos or scenes among newly input videos once an interest is specified.

The next section describes a real-time algorithm to detect one or more specific scenes in a broadcast video by using the sequences of frame image features at every shot change. Section 3 reviews experimental results.

2. SCENE RETRIEVAL

2.1 THE SCENE DETECTION METHOD CONCEPT

This section describes how to detect the scenes. The first step is to make a dictionary that describes the relationships between a pattern and the unique title of one scene. If one of the patterns in the dictionary matches the pattern extracted from a target video, the title corresponding to the pattern in the dictionary is output. To achieve this, a simple and quick matching algorithm is required. This is because videos have considerably more volume than other media, such as print Moreover, it is essential that the method is robust in regard to video fluctuations due to equipment characteristics or signal noise. The matching algorithm we have developed is based upon those design features.

(1) A scene is divided into several shots and only the first frames of each shot are compared,

(2) Frame features are compressed to compact parameters or codes, and

(3) Redundancy is given to feature matching.

To date, exact scene identification has obliged to pixelwise comparison of each frame. Design feature (1) dramatically reduces the processing time because only the specific frames in the scene need to be compared. This can be done by real-time shot change detection technology [4, 5, 6]. A shot is defined as a single continuous part of a video made by one camera. A sequence of shots makes a video program. The shot change is a boundary between two consecutive shots. The number of shots in a scene and the frame images at every shot change are scene specific. The sequence of such frame images is unique enough for scene specification.

Design feature (2) reduces the image matching process to simple code matching. Thus, quick matching of video scenes is possible because it is not necessary to calculate the complex resemblance between image features. Once the frame image codes at shot changes in a target video is produced and stored, a video search is as rapid as a text search. Our prototype converts an image feature into a character string of about 10 characters. In some cases, the new techniques to speed-up text retrieval may be applicable to video matching.

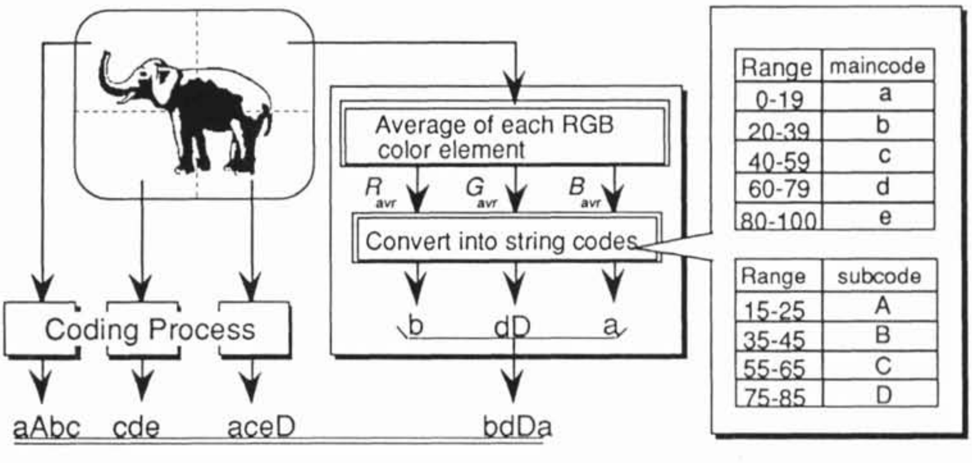
Design feature (3) prevents matching errors caused by video fluctuations. It may seem difficult to simultaneously maintain design policies (2) and (3). However, since a scene often consists of two or more shots, the sequence of frame features at each shot change are sufficient for scene specification.

2.2 IMAGE FEATURE ENCODING METHOD

A frame has many features. These include image characteristics, such as a color histogram and the area of a specific color, the time between representative frames and strength of the sound. In addition to these features, it is also possible to use other features if they are robust against video fluctuations and represent a certain portion of a video. Even if the feature of one frame is insufficient for unique specification, the combination of features provides sufficient representative information to be uniquely associated with a specific scene. Our prototype uses the average of each RGB color element of a frame as a feature because this is quickly and easily obtained.

Figure 1 shows the encoding process flow. The color average taken with the entire frame is still an insufficient feature. The algorithm divides a frame into 2 by 2 rectangular regions and calculates the average in each region. The average values of each color element are standardized within a range from 0 to 100 and converted into a main code of one character based on the conversion correspondence table shown on the right side of Figure 1. At this time, if the value is in a boundary neighborhood within the ranges shown in the correspondence table, a supplementary code is added. This forms two characters.

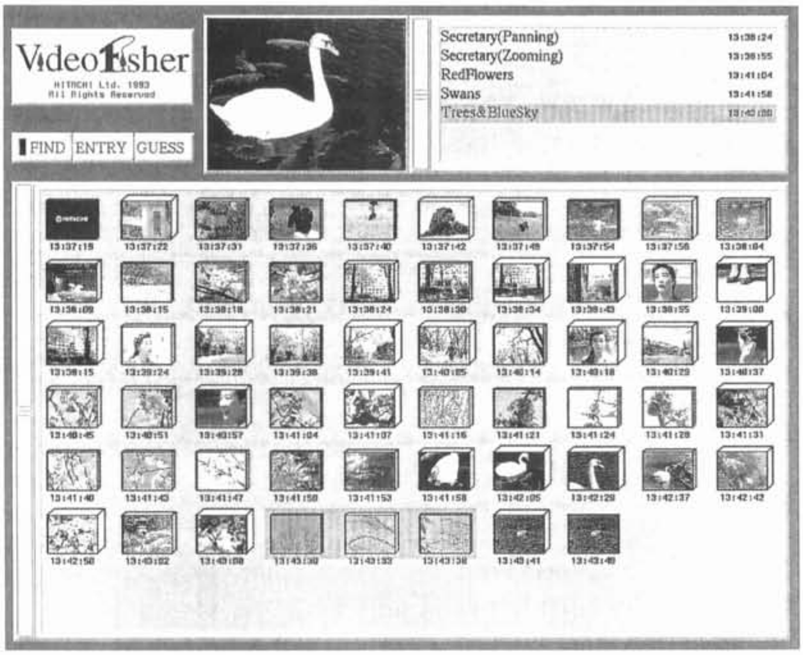
2.3 CODE MATCHING METHOD



Range	maincode
0-19	a
20-39	b
40-59	c
60-79	d
80-100	e

Range	subcode
15-25	A
35-45	B
55-65	C
75-85	D

Figure 1 Converting a feature of a representative frame into string codes



Secretary(Panning) 13:38:24
Secretary(Zooming) 13:38:55
RedFlowers 13:41:04
Swans 13:41:58
Trees&BlueSky 13:40:00

Figure 2 Scene detection system display

Comparison between the code of the target scene and the reference code registered in the dictionary is based on main code comparison. When a main code is different, a supplementary code is examined. Main codes may vary by fluctuation of the video if the value of the feature at the main code generation is in the boundary neighborhood. Thus, even if the two main codes to be compared are different, the presence of a supplementary code is examined. The main code is compared with the other main code close to the same boundary of the code. If matched, the next codes in the reference code are further compared one by one. If all codes which form the reference code in the dictionary match the code of the target scene, the detection of the scene is completed. The matching of the remainder of the dictionary ceases. However, when code disagreement is detected, matching with the next reference code in the dictionary begins. The above process is repeated for the number of reference codes in the dictionary.

3. EXPERIMENTS USING COMMERCIAL FILMS

Figure 2 is a snapshot of the prototype used for scene detection. It processes the video displayed in the upper middle window in real time. The lower window shows the icon list of the reduced first frame of each shot and adds a new icon at every shot change. If the scene registered in the dictionary is found, the upper right window shows the titles of the scene.

Table 1 provides results for detection of commercial films in a consecutive television broadcast of 315 minutes. There are 123 kinds of reference code registered in the dictionary. There were made from a television broadcast the previous day. The same commercial films appeared two or more times. There were 45 kinds of commercial film and these appeared a total of 71 times. The results approximated a practical detection performance,

Table 1 Scene detection results

Total Time		315 min (3,160 shots)
Kinds of scenes		45
Correct Detection / Total Appearances		71/ 71
Mistakes	Overdetections	25
	Detection Miss	0

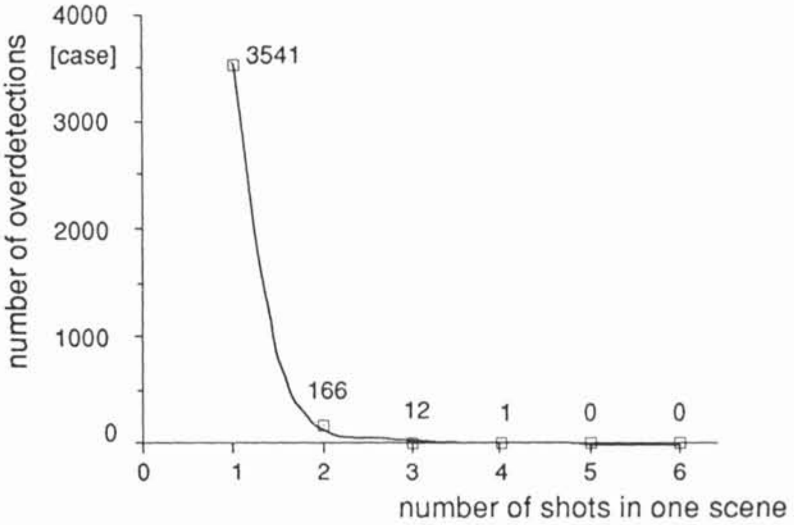


Figure 3 Relationship between number of shots and overdetections

as can be seen in the table. Overdetections refer to the number of times that an irrelevant scene was detected. Twenty-one of overdetections occurred at retrievals of dark and colorless commercial film. These were mistaken for night scenes in dramas. Thus, it appears necessary to use characteristics in addition to color for specification of such commercial films.

Moreover, because of the characteristics of the algorithm overdetection occurs frequently for scenes with few shots. Change in the number of overdetections as a function of number of shots in a scene is plotted in Figure 3. It appears that this technique is appropriate for scenes having more than three shots.

We found that broadcast video scenes could be detected in real time when the prototype was used with an IRIS 4D310/VGX. It could detect scene about 10,000 times faster than real time, once the frame code at every shot change was obtained. The table of such codes can be used to retrieve any scene in a video. Thus, this table permits the scene detection of five hours of video in about five seconds.

4. CONCLUSION

We have described a basic technique for efficient video retrievals. This technique can be used as a kind of filter that automatically selects an appropriate scene among newly input videos. It quickly identifies a scene by means of a table between the pattern of the scene feature sequence and its name. The approach described in this paper enables real-time scene identification by comparing frames only at every shot change.

REFERENCES

[1] Davenport G., Smith T.A., Pincever N.: "Cinematic Primitives for Multimedia," IEEE Computer Graphics & Applications, 11(4), pp. 67-75 (1991)

[2] Weiss R., Duda A., Gifford K.D., "Content-Based Access to Algebraic Video", Proc. The International Conference on Multimedia Computing and Systems, IEEE, pp.140-151 (1994)

[3] Zhang H., Gong Y., Smoliar W.S., Tan Y.S., "Automatic Parsing of News Video", Proc. The International Conference on Multimedia Computing and Systems, IEEE, pp.45-54 (1994)

[4] Ueda H., Miyatake T., Yoshizawa S.: "IMPACT: An Interactive Natural-Motion-Picture Dedicated Multimedia Authoring System," CHI'91, ACM, New Orleans, pp. 343-350 (1991)

[5] Nagasaka A., Tanaka Y.: "Automatic video indexing and full-video search for object appearances," IFIP Transactions A-7, Visual Database Systems, II, pp. 113-127, North-Holland (1991)

[6] Otsuji K., Tonomura Y.: "Projection Detecting Filter for Video Cut Detection," Proceedings of ACM Multimedia 93, pp. 251-257 (1993)