18

A CONTENT-BASED VIDEO RETRIEVAL METHOD USING A VISUALIZED SOUND PATTERN

Katsunobu Fushikida, Yoshitsugu Hiwatari, Hideyo Waki
Tokyo Waterfront Research Center, TAO Japan
Telecom. Center Build. 2-38, Aomi, Koutou-ku, Tokyo, Japan
{fusikida,hiwatari}@tokyobay.tao.or.jp

Abstract

A content-based video scene retrieval method using a multimodal index is proposed. Representative images of video scenes and corresponding visualized sound patterns are used as the multimodal index. Color-coded patterns of the sound spectrogram are adopted as the sound index. An image search engine is used not only for image retrieval but also for sound pattern retrieval. The results of parallel query experiments using the multimodal index suggest that it is more effective in improving query accuracy than single query methods. The results of video index browsing experiments indicate the effectiveness of the image-sound combined index method for efficient video queries and understanding the content.

Keywords

Multimodal index, Visualized sound, Parallel retrieval, Video browsing

1. Introduction

For multimedia database queries, the development of a powerful indexing, retrieving, and browsing scheme for large video databases has become essential.

Visual Database Systems 4 Y. Ioannidis & W. Klas (Eds.)

Content-based multimedia database queries is one approach being looked at and much research is currently being done in this area (Aigrain, 1996).

The efficient querying of a video scene requires a multimodal index including text, images, sound and any combination of these modes to be interactively used. The effectiveness of sound information in video retrieval or video handling has been pointed out (Hauptmann, 1995 and Brown, 1996). In order to fully understand the video content, each piece of index information should be presented in an effective manner to allow rapid video browsing.

We propose a multimodal retrieval scheme using a multimedia index, namely images and a corresponding visualized sound pattern that we call a "multimodal index" in this paper.

To improve search accuracy and reduce search time, we have discussed a parallel retrieval scheme on the distributed image search sites using image attributes, shape and color in a previous paper (Fushikida, 1998). In this paper, we extended this scheme to multimodal retrieval using image and sound attributes.

2. Multimodal retrieval for video database

Typical video indexes such as text indexes, image indexes and sound indexes have been discussed for content-based retrieval. Each index has its own particular features. A text index is an effective index for video retrieval. However, manual text indexing for video scenes is extremely time-consuming. Existing automatic text indexing techniques using speech recognition techniques such as word spotting or analyzing the meaning of a video scene are not yet sufficiently refined for practical application.

Although an image index can be viewed simultaneously on a screen and the video rapidly browsed we can not hear the sound indexes of different scenes simultaneously. To do this, we propose a visualized sound index which has the following two advantages.

(1) We can "glance" at the sounds of different scenes simultaneously.

(2) An image search engine can be used for sound retrieval.

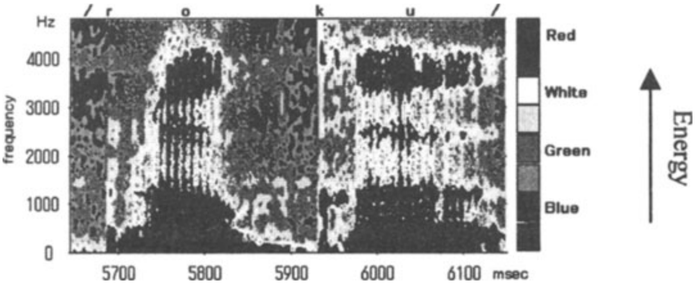

Fig. 1 Visualized sound index (sound spectrogram)

- Sound visualization method

We adopted a time-frequency pattern for speech, the so-called sound spectrogram, because of its robustness and universality compared with the other parametric methods. Figure 1 shows a sound spectrogram for word speech. A Fourier transformation is calculated and the energy values for each frequency component are quantized into eight levels. The energy level is represented by a color from dark blue (low energy) to red (high energy).

3. Experimental system

Figure 2 shows the experimental video retrieval system used for the multimodal retrieval. It consists of an image search engine, a multimodal index database and a video database. The input picture to the image search engine can be selected from images, visualized sound images, or a combined image of the preceding two images.

The image search engine in the experimental system uses a region-based matching method which was developed by Hirata (Hirata, 1993).

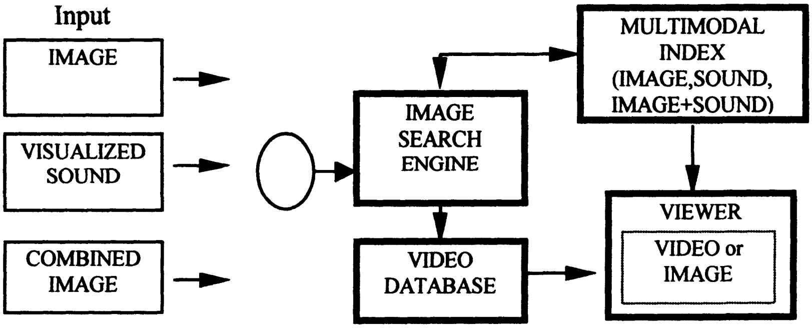

Fig. 2 Video retrieval experimental system using multimodal indexes

4. Experiments and results

4-1. Spoken word retrieval

We carried out preliminary spoken word retrieval experiments to investigate the ability of the image search engine to retrieve a visualized sound pattern. We used a 261 sound database containing 190 spoken words, 50 bird songs and 21samples from soccer video. As input words, 10 numbers and 10 Japanese nouns were spoken by four male speakers. The average recall rank of 20 words was 1.3. This is reasonable compared with conventional speech recognition results.

4-2. Retrieval using image indexes

Retrieval experiments using image indexes were conducted on a database of 600 frames (180 frames of 63 bird species, 250 soccer frames, etc.) manually extracted from colored video streams.

We selected eight target pictures from eight bird species and eight input pictures beforehand. These input pictures had different backgrounds, figures, and daylight conditions from the target pictures. The average recall rank of retrieval was 232. Thus, if the video scene changed significantly, it was difficult to search the target picture efficiently. Figure 3 (a) shows a pair of an input picture and a target picture. (Bird video from NHK software)

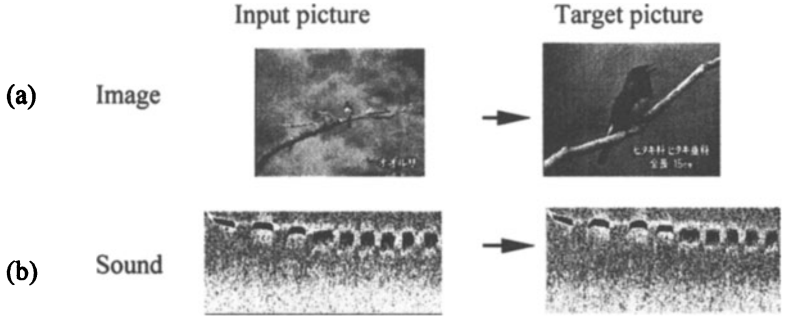

Fig.3 Examples of input pictures and target pictures for (a)image and (b)sound

4-3. Retrieval using visualized sound indexes

Retrieval experiments using visualized sound indexes were conducted on a database of 300 sound patterns. These contained 90 patterns from 38 bird species, 190 spoken word patterns and 20 patterns from a soccer video. The sound data was obtained manually by viewing the sound spectrogram on the display. The duration of the sounds was about 1 sec. We selected 12 input patterns from eight bird species and searched the target patterns for the same bird.

The average recall rank of retrieval was 2.0. This result indicates the superiority of the visualized sound index method. Figure 3(b) shows a pair of a visualized sound input and a target picture.

4-4. Parallel retrieval using image and sound indexes

To evaluate the performance of parallel retrieval for the multimodal index, we practiced retrieval using the image index and the sound index simultaneously.

The database was the same as that in 4-2 and 4-3. The input pictures were 25 pictures and their associated visualized sound (birdsong) which were selected at random. Table 1 shows the results of single retrievals and a parallel retrieval.

They suggest that image retrieval and sound retrieval can compliment each other and retrieval accuracy can be improved compared with the single retrieval method. However, it should be noted that the sound retrieval method is only effective when the video scene contains sound.

Table 1 Results for single retrieval and parallel retrieval

	Single		
Retrieval by	Sound index	Image index	Parallel
Retrieval accuracy (%)	96	68	99

4-5. Understanding video content through retrieval and browsing

The image-sound combined index can easily be used by the user to discriminate whether a scene is interesting or not. Figure 4 shows examples of the combined multimodal indexes. In the soccer game, the spectators are shouting enthusiastically and the reporter is talking excitedly. Consequently, the sound spectrogram for the highlight indicates high intensity and most part of the spectrogram is red (black in the sound index of Fig. 4).

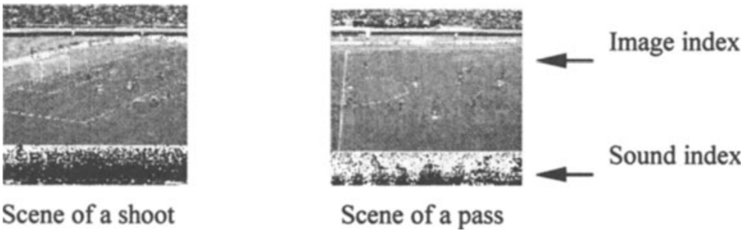

Fig. 4 Examples of combined multimodal indexes

Query experiments using a combined multimodal index were conducted on a soccer video with database of 44 scenes (highlighted scenes: 14, others: 30).
Figure 5 depicts browser screen of the experiments for highlight scene retrieval. We used the sketch of green and red as the input picture. All the candidate pictures that ranked in the top 10 were the highlight scenes.

A user can browse scenes with several sound images simultaneously. It helps the user to browse the content of the video database efficiently. The user can also listen to the actual sounds by clicking the sound image on the display.

Our multimodal index browsing experiments proved the effectiveness of presenting sound indexes in parallel. However, the experimenters needed a short period to understand the meaning of the sound index to obtain the characteristics of the sound spectrogram. We were able to glance not only the landscape but also the "soundscape " simultaneously.

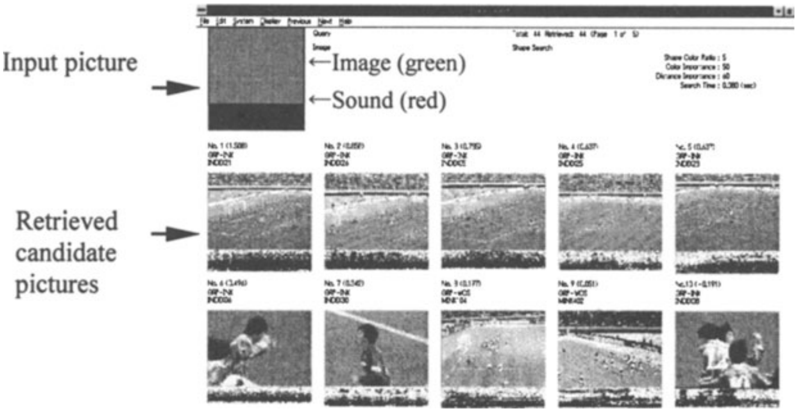

Fig. 5 An example of highlight scene retrieval and browsing

5. Conclusion

This paper proposed a multimodal retrieval scheme using a visualized sound index that was associated with an image index. Multimodal query experiments for visualized spoken words, bird video scenes and soccer video scenes, which contained sounds, resulted in the following.

- The image search engine could be effectively used to retrieve visualized sound.
- Parallel queries using the image indexes and sound indexes could improve the accuracy of retrieval. This suggests that the multimodal index method can improve the efficiency of a interactive video query system.
- Browsing the combined multimodal index is an effective way of understanding the video content or in picking out a desired scene rapidly .

6 References

Aigrain, P., Zhang, H. and Petkovic, D. (1996) Content-based representation and retrieval of visual media : A state-of-the-art review, *Multimedia Tools and Applications* 3, 179-202.

Brown, M. G., Foote, J. T., Jones, G. J. F. Jones, K. S., and Young, S. J. (1996) Open-vocabulary speech Indexing for voice and video mail retrieval, *Proc. ACM Multimedia 96*, 307-316, Boston, ACM.

Fushikida, K., Hiwatari, Y. and Waki, H. (1998) Content-based image query method using parallel retrieval scheme , *ICCIMA 98*, 830-835.

Hauptmann, A.G. and Smith, M. (1995) Text Speech, and Vision for Video Segmentation: The Informedia Project, *AAAI Fall 1995 Symposium on Computational Models for Integrating Language and Vision.*

Hirata, Y., Hara, K., Shibata, N. and Hirabayashi, F. (1993) Media-based Navigation for Hypermedia system, *ACM Hypertext*, 157-173.