Audio-based Copy Detection in the Large-scale Internet Videos

Hongliang Bai[†], Lezi Wang[‡], Chong Huang[‡], Wei Liu[†], Chengbin Zeng[†], and Yuan Dong[†‡]

[†]France Telecom Research & Development - Beijing, 100190, P.R.China
[‡]Beijing University of Posts and Telecommunications,100876, P.R.China
{hongliang.bai,wei.liu,chengbin.zeng,yuan.dong}@orange.com
{wanglezi.bupt,huangchong661100}@gmail.com

Abstract. With the large-scale internet video data explosion, the content-based copy detection (CCD) related application and research are significant and necessary. Beside the image-based CCD, the audio-based method has the advantage in its simpleness and efficiency. The article improves the recent methods on the audio-based copy detection. Three improvements are introduced in the study. Firstly, the CEPS-like feature is proposed to satisfy the different audio scale requirements in the feature extraction. Then, the flexible hash-based searching algorithm is presented to strengthen the querying robustness. Finally, the results-based fusion is introduced to take the advantages of the different features. The actual NDCR performances of the balanced profile vary in 0.223∼0.460 in the TRECVID2011 copy detection database. The results outperform any single feature.

Keywords: CCD, CEPS-like, Flexible Searching, Fusion, TRECVID

1 Introduction

With the growth of images and videos in the internet, the retrieving requirement from users has increased enormously. They can record videos or take photos by the mobile phones, video camcorders, or directly download from the video webs, and then distribute them with some modifications. More than 13 million hours of video were uploaded during 2010 and 35 hours of video are uploaded every minute, and YouTube reached over 700 billion playbacks in 2010 [5]. Among these huge volumes of images and videos, the large number of them are duplicate or near duplicate.

Based on a sample of 24 popular queries from YouTube, Google Video and Yahoo! Video, on average there are 27% redundant videos which are duplicate or nearly duplicate to the most popular version of a video in the search results [12]. Nearly 30% videos are duplicated in one-day Orangesport videos[1]. Users always feel frustrated when they see many duplicate sequences and don't find what

[1] http://sports.orange.fr/

they are interested. So the copy detection is one of very important techniques to retrieve and delete the videos. It also can reduce the large disk storage for the video website.

The video and audio information can be used to implement the copy detection. The audio-based methods can well solve the difficulty, especially when the audio information is consistent with the variable video frames. The audio-based copy detection is to find the corresponding copy sequences of one query from the video database, and the query maybe have different compression style, and mix with speech. Usually, the framework is composed by preprocessing, feature extraction, searching engine and postprocessing.

For the feature extraction, a Weighted Audio Spectrum Flatness (WASF) is presented to extend the MPEG-7 descriptor-ASF by introducing human auditory system functions to weight audio data [1]. The feature is robust to several audio transformations, but tuning the parameters is one hard work. The HAAR filters are influenced by the training data [8]. Mel-Frequency Cepstral Coefficients (MFCC) is a feature used in the speech recognition and copy detection [7]. Energy Differences Feature (EDF) is widely used in [3, 4, 11], and the good performance is achieved in the large-scale video database. However, EDF can only consider one scale property of the frequency. For the video retrieving, the hash function [2] is used for the accurate searching with the higher efficiency. But, the hash-based searching can not deal with the near duplicate audio clips. Locality-Sensitive Hashing (LSH) [6] is not suitable in the low-dimensional audio feature space.

So three improvements have been introduced to solve the above problems in the study. The system framework is described in Section 2. The multi-scale audio feature extraction is proposed in Section 3. Section 4 presents the flexible Hash-based retrieval algorithm. The fusion of searching results is introduced in Section 5. Section 6 shows some experimental results as well as its limitation. Finally, the conclusions and future works are listed.

2 System Overview

In this section, the audio-based copy detection system framework is introduced in Fig. 1. Firstly, the querying audio signal is separated from the videos. Then the audio signal is processed by the Butterworth and Hamming window filtering. After the Fast Fourier Transform (FFT) analysis, the 17 sub frequent bands are selected in the mel-frequency space. 16-bit EDF and 16-bit CEPS-like feature are extracted respectively. The two types of features are used to query in the reference database. The different searching results from the above features are fusioned finally.

3 Feature Extraction

3.1 Butterworth and Hamming Window Filtering

In the reference video database from the internet, the audio' sampling rates vary in a large range. The first step is to normalize the sampling rates into a constant

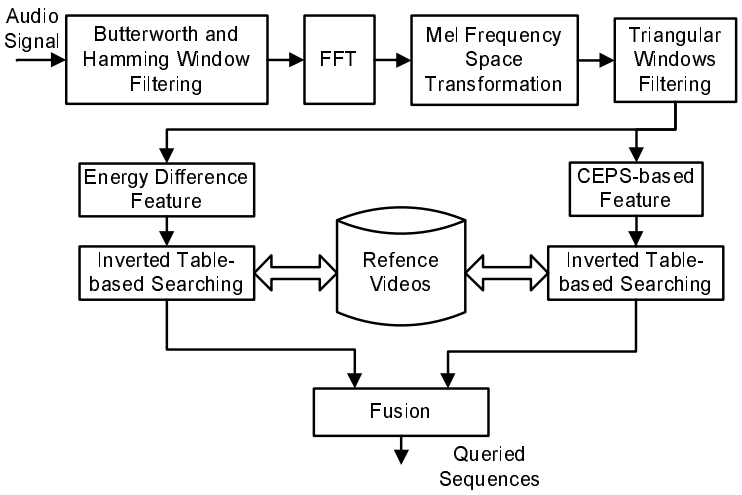

Fig. 1. Audio-based copy detection system framework

value F_N, here F_N is set 44100 Hz. Then the normalized signals S are lowpass filtered to 4000 Hz by a Butterworth filter. The magnitude-squared response of a N-order analog lowpass Butterworth filter is $|\mathbf{H}(j\Omega)|^2 = 1/(1 + (\Omega/\Omega_c)^{2N})$, where the cutoff frequency Ω_c is 3dB. Through the filter, the top 100 coefficients are used to convolve with S in the time domain.

Then the hamming window filtering is applied to every frame in order to keep the continuity of the first and the last points in the frame before FFT. The hamming window filtering is $\mathbf{H}(i) = 0.54 - 0.46 * \cos(2\pi i/(N-1))$, where N is the sample number in each frame and set 2048. The inter overlapping is 1024 samples ($23.2ms$).

3.2 FFT and Mel-frequency Space Transformation

After the Hamming window filtering, the 1-D audio signals are transformed into 2-D spectrograms by FFT. The spectrum between 300 Hz and 4000 Hz is equally divided into 17 sub bands in the mel-frequency space. The mel-frequency can reflect similar effects in the human's subjective aural perception. The relation of the mel-frequency and natural frequency is $Mel(f) = 2595 * \log(f/700 + 1)$, where f is the natural frequency.

3.3 Energy Difference Feature

A triangular filtering is used in the magnitude frequency response to compute the energy of each sub band. The number of the filters is equal to that of the sub bands. The coefficients of the filter are defined by

$$w(n) = \begin{cases} \frac{2n}{N-1} & n = 0, 1, ..., \frac{N-1}{2} \\ 2 - \frac{2n}{N-1} & n = \frac{N-1}{2}, ..., N-1 \end{cases} \tag{1}$$

EDF features between the sub-bands are used to generate the fingerprint of each frame, which is calculated by Equation 2.

$$EF_n(m) = \begin{cases} 1 & EB_n(m) > EB_n(m+1) \\ 0 & otherwise \end{cases} \tag{2}$$

where $EB_n(m)$ represents the energy value of the n-th frame at the m-th sub-band, and $m \in [1 \cdots 16]$. The 15-bit and 32-bit fingerprints are used in [3, 4] respectively. After considering the storage size of *short int* and robustness of the searching algorithm, the 16-bit fingerprint $EF_n(m)$ is selected. The feature is demonstrated in the Fig. 2(a).

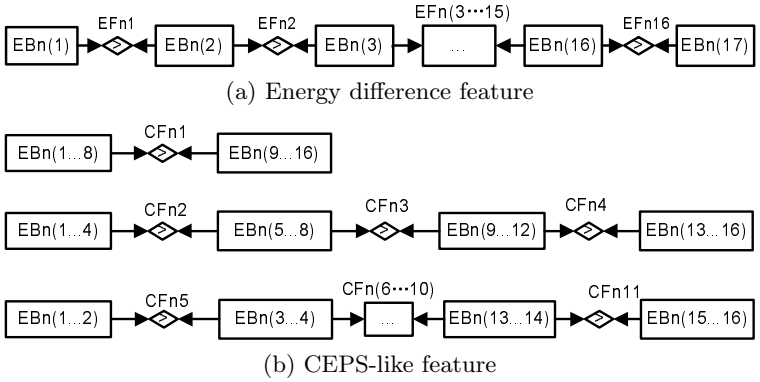

Fig. 2. Extraction of two types of audio features, which describe the energy property of the different scales

3.4 CEPS-like Feature

The cepstrum is the information about the rate of the change in the different spectrum bands and the result of taking Fourier Transform (FT) of the log spectrum. The EDF feature only considers the energy difference in the low level. The CEPS-like feature is proposed to combine the multi-scale energies into one feature.

In Fig.2(b), $CF_n(1)$ is the highest-scale feature, which used all information of 16 sub bands. $CF_n(2 \cdots 4)$ are in the second level and the difference of four adjacent sub bands. $CF_n(5 \cdots 11)$ are in the third level. $CF_n(12 \cdots 16)$ are the same with $EF_n(1)$,$EF_n(4)$,$EF_n(7)$,$EF_n(10)$ and $EF_n(13)$ respectively. $EB_n(m_1 \cdots m_2)$ is the energy sum from the m_1-th sub band to the m_2-th sub band.

4 Flexible Hash-based Searching

The hash-based searching is a very important and widely used technology. The searching performance is improved by two aspects: (1)one-bit modification in the hash matching. If the hamming distant of a querying and reference feature is one, they are regarded as a matching pair; (2)matching time can tolerate some time errors because of the frame losing or noise interference. The above algorithms can improve the searching flexibility.

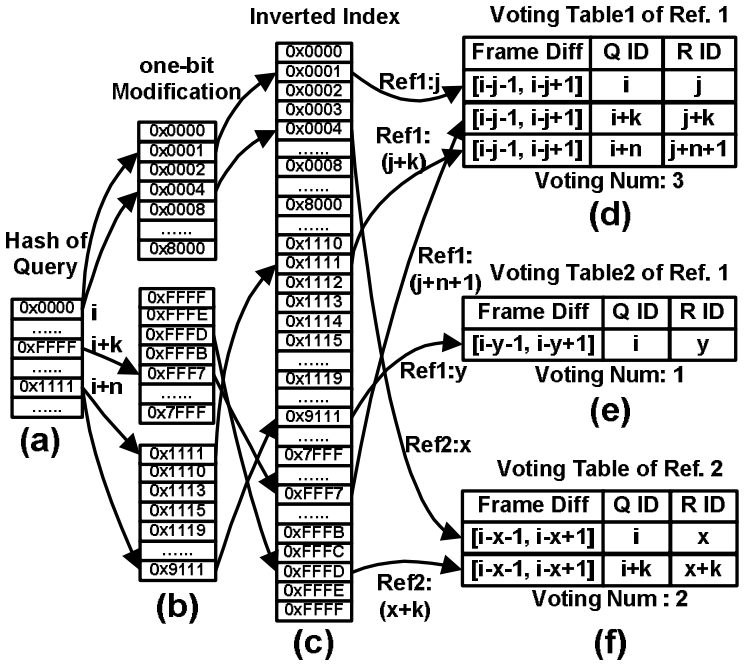

Fig. 3. Flexible hash-based searching with one-bit modification and time redundance

Fig. 3 demonstrates the audio fingerprint matching strategy. Fig. 3(a) shows the sequence of hash values in a querying clip. And Fig. 3(b) can make the hash matching more robust by modification of any one bit of a hash value. Seventeen different values are generated for a 16-bit feature. These modified hash values are matched with the ones from reference data in the inverted index table, shown in Fig. 3(c). The voting tables are used in references, which is related to the matched hash values from the inverted table. The voting number are the hitting values in some time difference between indexes of the matched reference and querying. The voting strategy is illustrated in Fig. 3(d)(e)(f). The largest voting results N_{vote}(Voting Number 3) occurs in Fig. 3(d). The time duration of the queried sequence is $[j, j+n+1]$ in the reference database.

$$N_{vote} \triangleq \arg\max_{\tau} \sum_{r,q \in N} \delta(\tau - |r - q|) \tag{3}$$

where r and q are the time indexes of the matching sequence of the querying and reference. If N_{vote} is greater than the predefined threshold T, the queried reference sequences will be regarded as the querying results.

5 Result-based Fusion from Different Features

The fusion algorithm can be used in the stages of the feature extraction or searching results. The fusion of the searching results are proposed from the different features, shown in Fig. 4. For the retrieving results from every feature, the higher precision is generated if the threshold T is set with higher values. In Fig. 4, G_1 and G_2 are the above reliable querying results, and G_3 is the logical "AND" operation results from EDF and CEPS-like features. Both the advantages of EDF and CEPS-like are taken in the G_3. The querying results are more reliable if the outputs of above two features are same. The final results are the logical "OR" of G_1, G_2 and G_3. The parameters TH_1 and TH_2 will be discussed in the experimental section.

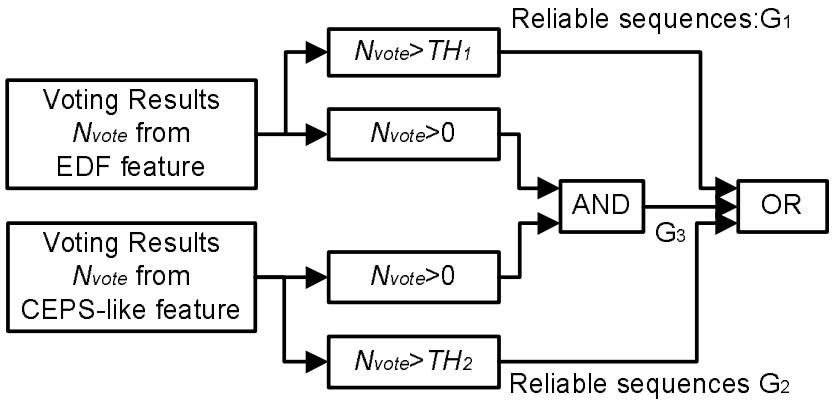

Fig. 4. Fusion of the searching results from EDF and CEPS-like features

6 Experiments

In this section, the experiments are conducted to demonstrate the effectiveness of the proposed method.

6.1 Database Description

TRECVID [10] has a well-known CCD task. The reference data is identical to 400 hours and 12000 files in the 2011 test and training data. Each query has 8-type video and 7-type audio transformations. In the audio-related task, the original audio clips are transformed into the following seven types, namely, ($T1$) do "nothing"; ($T2$) mp3 compression; ($T3$) mp3 compression and multiband companding; ($T4$) bandwidth limit and single-band companding; ($T5$) mix with speech; ($T6$) mix with speech, then multiband compress; ($T7$) bandpass filter, mix with speech, compress. Many evaluation metrics are used in the CCD task. Normalized Detection Cost Rate (NDCR) is defined in following

$$NDCR = P_{miss} + \beta.R_{FA} \tag{4}$$

where P_{miss} and R_{FA} are the conditional probability of a missed copy and the false alarm rate respectively, and β is profile related.

6.2 Parameter Selection and System Performance

In the audio-based copy detection system, the parameters TH_1 and TH_2 are main parameters, which can influence the system performance. After the parameters are trained from TRECVID CCD 2010 training and testing data, then they are used in the TRECVID CCD 2011 testing data. The performances of the balanced profile are shown in Table 1 after the different values has been selected. In the tables, EF-10 means the threshold TH_1 is set 10 for the EDF feature, and CF-15 means the threshold TH_2 is set 15 for the CEPS-like feature.

For the NDCR metric, the best performance of EDF-based querying is 0.338 in the transform $T1$, and that of CEPS-like-based querying is 0.576. After the fusion , the NDCR is improved to 0.233, which is better than any of features querying results. In the most complex transform $T7$, the performance is also improved from 0.577 and 0.515 to 0.323.

The comparison has also been done with other groups submitted into the TRECVID CCD 2011. Only INRIA-LEAR submitted their audio-only detection results in Fig. 4. The INRIA-LEAR's descriptor is constructed by concatenating several filter banks[9]. From Table 1, our auido-based querying results outperform the INRIA-LEAR's. For example, in the *T1* case, our and INRIA-LEAR's actual NDCRs are 0.223 and 0.634, respectively.

From the experiments, after TH_1 and TH_2 are selected as 22 and 30, the best fusion results can be archived. From the above table, the worst performance occurred in the speech-related transformation because their frequencies focus on the higher frequency.

7 Conclusions and Future Works

In the copy detection system, the feature extraction and querying methods are the most important sections. The article focuses on the audio-based copy detection. In the feature extraction, the novel CEPS-like feature is proposed. The

Table 1. Actual NDCR of the different TH_1 and TH_2, and the fusion results from different features

	T1	*T2*	*T3*	*T4*	*T5*	*T6*	*T7*
EF-10	1.837	1.623	1.501	1.562	0.973	0.698	0.943
EF-22	**0.338**	**0.454**	**0.454**	**0.385**	**0.485**	**0.500**	**0.577**
EF-30	0.377	0.485	0.569	0.438	0.562	0.600	0.692
CF-15	2.709	2.319	1.501	1.982	0.767	0.561	0.675
CF-30	0.744	0.690	**0.730**	0.583	**0.515**	**0.608**	**0.515**
CF-40	**0.576**	**0.592**	0.822	**0.423**	**0.515**	0.669	0.638
FUSION	0.223	0.460	0.384	0.338	0.384	0.323	0.323
INRIA	0.634	0.520	0.507	0.520	0.540	0.642	0.455

flexible hash-based searching algorithm can improve the querying performance. The querying results-based fusion is also presented. After the experiments are conducted, the proper parameters are selected and the fusion performance outperforms any feature. In the future, the speech mixture querying is our research topics.

References

1. Chen, J., Huang, T.: A robust feature extraction algorithm for audio fingerprinting. In: Pacific Rim Conference on Multimedia(PCM). pp. 887–890 (2008)
2. Döhring, I., Lienhart, R.: Mining tv broadcasts for recurring video sequences. In: Conference on Image and Video Retrieval(CIVR). pp. 1–8 (2009)
3. Haitsma, J., Kalker, T.: A highly robust audio fingerprinting system. In: Music Information Retrieval(ISMIR) (2002)
4. Heritier, M., Gupta, V., Gagnon, L., Cardinal, P.: Crim's content-based copy detection system for trecvid. In: Content-Based Multimedia Indexing (CBMI) (2010)
5. http://www.youtube.com/t/press_statistics:
6. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: symposium on Theory of computing(STOC). pp. 604–613 (1998)
7. Jegou, H., Douze, M., Gravier, G., Schmid, C., Gros, P.: Inria lear-texmex: Video copy detection task (2010)
8. Ke, Y., Hoiem, D., Sukthankar, R.: Computer vision for music identification: Video demonstration. In: CVPR (2005)
9. Serra, J.: Identification of versions of the same musical composition by processing audio descriptions. Ph.D. thesis, Universitat Pompeu Fabra (2011)
10. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: Multimedia Information Retrieval(MIR). pp. 321–330 (2006)
11. Ton, J.H., Kalker, T.: Robust audio hashing for content identification. In: Content-Based Multimedia Indexing(CBMI) (2001)
12. Wu, X., Ngo, C., Hauptmann, A.G., Tan, H.: Real-time near-duplicate elimination for web video search with content and context. IEEE Tran. on Multimedia 11(2), 196–207 (2009)