CONTENTED-BASED LARGE SCALE WEB AUDIO COPY DETECTION

Lezi Wang[†], *Yuan Dong*[†], *Hongliang Bai*[‡], *Jiwei Zhang*[†], *Chong Huang*[†], *Wei Liu*[‡]

[†]Beijing University of Posts and Telecommunications,100876, P.R.China
[‡]France Telecom Research & Development - Beijing, 100190, P.R.China
{wanglezi, yuandong, zhangjiwei, huangchong661100}@bupt.edu.cn
{hongliang.bai, wei.liu}@orange.com

ABSTRACT

The exponential growth of web videos brings content based copy detection into a crucial issue. Besides the image information, audio also plays an important role in copy detection. In this paper, the audio-based copy detection framework is introduced. Three contributions are presented: (1) the band energy difference based feature is improved by adding multi-scale information, which extends the candidate feature sets; (2) a conditional entropy based method is used to select 16 ordinal relations to generate a more compact and robust feature combination among the random $C_{91}^{16} \approx 2.6 \times 10^{17}$ combinations; (3) the result-based fusion strategy is introduced to recall the missed true positives. The proposed algorithm outperforms the traditional coarse fingerprints, shown by experiments conducted in the TRECVID 2011 Content-based Copy Detection (CCD) database.

***Keywords*-Conditional Entropy; Fusion, Audio; TRECVID; Contend-based Copy Detection;**

I. INTRODUCTION

The audio copy detection aims at determining if a given query audio clip has its copies in the database and the detection system is required to return the time stamps where the copies are located in the reference audio stream. The exponential growth of the searchable videos in the web, e.g. YouTube, brings the task to a new crucial issue. There are over 700 billion playbacks on YouTube in 2010 [1] and videos of more than 13 millions are uploaded in the web. Web users can record videos by digital cameras, mobile phones or directly downloading from the Internet and publish data after modifying, which leads to a lot of redundant videos with the same content. For instance, there are 27% duplicate videos in the samples of 24 popular queries from YouTube, Google Video and Yahoo! Video [2]. Generally, image and audio based information can be both used to implement copy detection. And most existing copy detection systems focus more on finding the schemes to handle the various image transformations. The audio-based algorithm may achieve better performance when the audio content is relatively consistent over the video frames of severe transformations, e.g. camera-coding with depth variation, which is difficult to deal with by image based algorithm.

Fingerprint system is widely used in audio copy detection due to its efficiency, robustness, reliability and compactness [3]. Recently, a certain researches study extraction of audio fingerprints. J. Chen et al.[3] present an algorithm to extract weighted ASF (WASF) based on a MPEG-7 descriptor-Audio Spectrum Flatness (ASF) and Human Auditory System(HAS); several effective filters are adopted to improve the robustness of WASF. The similarity between two WASF features is measured by Euclidean distance, which is not suitable for efficient indexing due to high computation cost. And tuning the filter parameters is also difficult. H.Jegou et al. introduce compound descriptors into copy detection task[4], which obtains high accuracy via approximate nearest searching algorithm. However, it is time consuming to compare descriptors based on Euclidean metric and solve the high dimension problem, e.g 144 dimensions for each descriptor. The energy difference between two consecutive frequency bands is adopted to extract binary audio fingerprints [5], [6], [7], [8], which performs well taking the trade-off between detection accuracy and indexing efficiency into consideration. Y.Ken et al. [9] introduce computer vision based method to extract a 32-bit binary audio fingerprint, which guarantees the efficient indexing but is vulnerable to distortions. And we observe that there are more than one half bits describing the specific frequency bands differences among the 32-bits binary fingerprint. Therefore band energy difference based fingerprint representation is a good choice for achieving effective and efficient searching in large-scale database.

The performance of bands energy difference based algorithm is improved by three contributions: (1) adding more band information; (2) selecting a strong features combination from candidate feature sets; (3) presenting the result-based fusion strategy. Adding the information of multi-scale and multi bands combinations extends the feature set. In section III, a conditional entropy based strategy is adopted to get the discriminative sub feature combination from the set, which is stable to several audio transformations, such as mp3 compression, single or multi-band companding, mix

978-0-7695-4711-4/12 $26.00 © 2012 IEEE
DOI 10.1109/ICME.2012.17

with speech and the combinations. The feature combination models an audio frame into a binary fingerprint,which enables efficient searching by exact inverted indexing. The result-based fusion strategy is introduced to improve the system performance by recalling the missed correct detection due a single threshold. Experiments on TRECVID content based copy detection database show our method achieve high accuracy and efficient searching.

The rest of paper is arranged as follows. Section II describes three different feature extraction algorithms. Section III presents the conditional entropy based feature selection strategy. The efficient inverted indexing algorithm is described in Section IV. The result-based fusion strategy is presented in section V. In section VI, the performances of different features extraction and fusion schemes are evaluated on TRECVID 2011 copy detection database. Section VII draws a conclusion and lists the further work.

II. FEATURE EXTRACTION

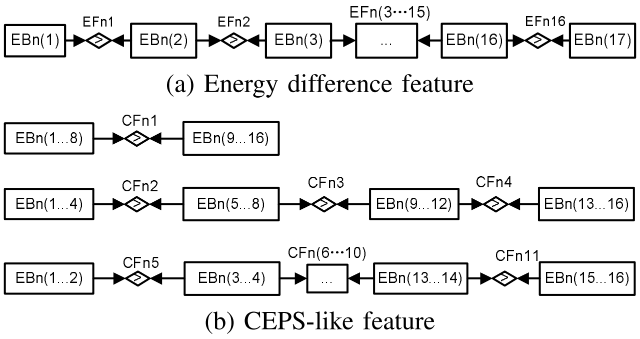
It is a challenge to extract features that meet two criteria of robustness to severe audio transform and matching efficiency. The binary representation of fingerprint performs excellent efficiency through exact searching. Coarse audio fingerprints introduced in [5], [6], [7], [8] describe the energy difference between two consecutive frequency bands. And we observed that 19 bits out of the 32-bit binary feature [9] describe the energy difference over bands. The experiments show those features are vulnerable to audio transformation. It motivates us to improve the coarse fingerprints to be robust to the severe distortions. This section describes the preprocessing of audio streams and three different audio fingerprint strategies based on the energy difference over specific bands.

II-A. Pre-processing

The preprocessing contains two main steps: sampling rate normalization and transformation the temporal audio signal into frequency domain representation. The sampling rates of web audio data vary in a large range. Commonly, the system normalizes data into audio streams with uniform sampling rate $F_N = 44100Hz$. Generally, frequency domain representation of audio reflects the signal variation which is more stable to distortion than the temporal audio data. The basic step of fingerprints extraction is transforming the temporal audio signal to its frequency representation. The frequency spectrum is generated by Fast Fourier Transformation (FFT) on 2048 samples with 50% overlapping increment. The normalized audio streams are convolved with Butterworth low-pass filter before FFT, so that there is no alias in frequency domain.

II-B. Energy difference over consecutive bands

The Energy Difference Feature (EDF) over consecutive bands is used in [5], [6], [7], [8]. The frequency spectrum of $300Hz$ to $4000Hz$ is divided into fixed N bands in equally mel-frequency space. A triangular filter is applied to the frequency response of each band before computing the energy. The filter coefficients is defined as:



(a) Energy difference feature

(b) CEPS-like feature

Fig. 1. Extraction of two types of audio features

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

The energy difference between two consecutive bands is used to compute the $N-1$ bits binary fingerprint for each audio frame, defined by:

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

where $EB_n(m)$ represents the energy value of the nth frame at the mth sub-band, and $m \in [1 \cdots 17]$. The 15-bit and 32-bit fingerprints are used in [6], [7] respectively. After considering the storage size of $short\ int$ and robustness of the searching algorithm, the 16-bit fingerprint $EF_n(m)$ is selected.

II-C. CEPs-like fingerprint

The fixed scale problem is a limitation of the EDF, which only considers the energy difference in the low level. In TRECVID 2011 copy detection task, we propose CEPS-like feature to combine the multi-scale energies into one feature based on cepstrum. The cepstrum is the information about the rate of the change in the different spectrum bands and the result of taking the Fourier Transform (FT) of the log spectrum. In Fig.1(b), $CF_n(1)$ is the highest-scale feature, which use all information of 16 sub bands. In the second level, $CF_n(2 \cdots 4)$ are 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.

II-D. Bands energy difference with multi-scale

The proposed audio fingerprint is inspired by the above two extraction schemes. The EDF and CEPs-like features both only consider the ordinal measure between two consecutive bands. CEPs-like feature doesn't consider ordinal

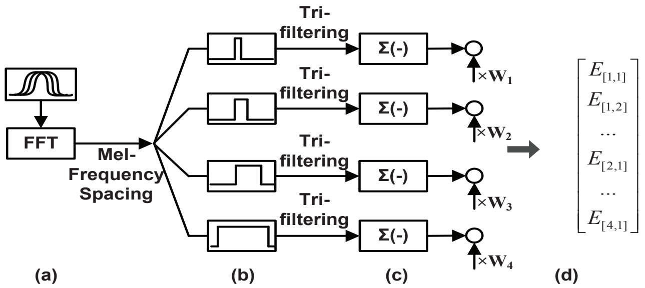


Fig. 2. Multi-scale Band Energy Computation

relations between different scales. And we observe that the frame number distribution of CEPs-like fingerprint in database is very sparse and non-uniform. It means that some subset of bits carry little information and this kind of bit combinations is not discriminative to specific distortions. The proposed fingerprint is expected to examine the audio characteristics that are resistant to distortions, based on the assumption that audio fingerprints of the large-scale web videos are uniformly random.

Figure 2 illustrates the procedure of computing the band energy with multi-scale. Firstly, the preprocessing of audio stream and the frame frequency spectrum is generated by FFT, Figure 2(a). Then, the spectrum between $300Hz$ and $4000Hz$ is divided into N sub-bands with equally mel-frequency spaced, Figure 2(b). Fifteen bands are generated in four different scales, where $N \in \{8, 4, 2, 1\}$. A corresponding triangular filter, defined in Equ. (1), is applied to samples of one frequency band, in Figure 2(c). The band energy values in different scales are normalized by multiplying with specific weights. The weighted band energy vector of 15 dimensions is shown in Figure 2(d), where $E_{[i,j]}$ indicates the weighted energy of jth band in ith scale.

The fingerprint describes energy distribution via the energy ordinal measure of two-by-two bands. The ordinal measure is represented by relations between two sub-bands. Totally $105(C_{15}^{2})$ relations are generated in 15 bands, where 14 relations occur twice. Therefore the audio fingerprint can be represented as a 91 dimensional feature set. The set is represented as $X = \{X^1, \cdots, X^{91}\}$, where the X^i denotes the ith value of the feature set X. A sub feature set $X^S = \{X^{S(1)}, \cdots, X^{S(M)}\}$ needs to be selected as the compact and discriminative feature. The conditional entropy based binary feature selection method is presented in the next section. In the rest of this paper, "CE_EDF" is used to represent the conditional entropy based binary audio fingerprint.

III. CONDITIONAL ENTROPY BASED FEATURE SELECTION

The binary feature selection method aims at selecting a subset from the high dimensional feature set $X = \{X^1, \cdots, X^{91}\}$ to form a strong and compact fingerprint which enables the efficient exact searching with inverted indexing, as described in section IV. The candidate subset feature combinations is very large. For example, if M is chosen to be 16, the number of feature combinations is $C_{91}^{16} \approx 2.6 \times 10^{17}$. Therefore, an efficient feature selection method is needed. The conditional entropy based informative ordinal relations selection scheme is the obvious choice, which is introduced by L. Shang et al. to select image CE-based spatiotemporal features in [10]. The main goal of this scheme is to select a subset of features which carry the information as much as possible.

The procedure of feature combination selection is presented in Algorithm 1. The training data is the sequence of N feature samples, $X_1, \cdots, X_N$, where N is the number of extracted audio frames. X_i^j denotes the jth relation value for ith frame. The algorithm returns a kind of sub-set features combination, $X^S = (X^{S(1)}, \cdots, X^{S(M)})$, which is informative and two-by-two weakly dependent. First, the most informative relation is selected as the $X^{S(1)}$, namely $H(X^{S(1)}) \geq H(X^n), 1 \leq n \leq 91$. $H(X^n)$ denotes the entropy of relation X^n in the database, defined as:

$$\begin{aligned} H(X^n) &= p(X^n = 1) \log p^{-1}(X^n = 1) \\ &\quad + p(X^n = 0) \log p^{-1}(X^n = 0) \end{aligned} \tag{3}$$

where $p(X^n = 1)$ denotes the probability of the relation $X^n = 1$ occurring in the database:

$$p(X^n = 1) = \frac{\#(frames\ having\ X^n\ value\ of\ 1)}{N} \tag{4}$$

and $p(X^n = 0) = 1 - p(X^n = 1)$ indicates the probability of the relation $X^n = 0$. The feature set of $X = \{X^1, \ldots, X^{91}\}$ is updated by subtracting the $X^{S(1)}$. In the second step, the rest $M - 1$ relations are selected iteratively. The mth selected relation satisfies the criteria:

$$S(m) = \arg\max_{n} \{ \min_{k \leq m-1} H(X^n | X^{S(k)}) \}, X^n \in X \tag{5}$$

where $H(X^n | X^{S(k)})$ denotes the conditional entropy of X^n on $X^{S(k)}$, defined as:

$$H(X^n | X^{S(k)}) = H(X^n, X^{S(k)}) - H(X^{S(k)}) \tag{6}$$

The minimization of conditional entropy guarantees the weak dependency between two selected features. The maximization of $min_{k \leq m-1} H(X^n | X^{S(k)})$ ensure that the newly selected relation is informative in the feature set.

Table I lists the top 16 ordinal relations, $X^{S(n)} n \in [1, 16]$, picked by the selection algorithm, where $E_{[i,j]}$ indicates the energy of jth band in ith scale. These selected relations are used to generate the M dimensional fingerprint $F = \{I_1, I_2, ..., I_M\}$, where $I_k = 1_{X^{S(k)}}, k \in [1, M]$, $1_{X^{S(k)}}$

Algorithm 1 Informative Fingerprints Selection

```
Input: The sequences of $N$ features samples with 91 dimensions,
    $X_1 \leftarrow (X_1^1, ..., X_1^{91})$, ..., $X_N \leftarrow (X_N^1, ..., X_N^{91})$,
    and initialize the number of selected features $M$
Output: The combination of $M$ informative and two-by-two weakly dependent CE_EDF feature $X^S \leftarrow (X^{S(1)}, ..., X^{S(M)})$
    Step 1:
    Find the first feature element of $X^S$ that maximizes $H(X^n)$, $n \in [1, 91]$
    Update the feature set: $X \leftarrow X - X^{S(1)}$
    Step 2: for $m = 2 \cdots M$
    Find the $m$th element of $X^S$ that maximize the set of conditional entropies $min_{k \leq m-1}\{H(X^n|X^{S(k)})\}, X^n \in X$
    Update the feature set: $X \leftarrow X - X^{S(m)}$
```

Table I. The top-16 selected band energy ordinal relations

Feature	Relation	Feature	Relation
$X^{S(1)}$	$E_{[1,5]} > E_{[2,3]}$	$X^{S(9)}$	$E_{[1,4]} > E_{[2,2]}$
$X^{S(2)}$	$E_{[1,2]} > E_{[3,1]}$	$X^{S(10)}$	$E_{[1,5]} > E_{[2,4]}$
$X^{S(3)}$	$E_{[1,3]} > E_{[2,2]}$	$X^{S(11)}$	$E_{[1,8]} > E_{[2,4]}$
$X^{S(4)}$	$E_{[1,7]} > E_{[2,4]}$	$X^{S(12)}$	$E_{[1,3]} > E_{[1,5]}$
$X^{S(5)}$	$E_{[2,2]} > E_{[4,1]}$	$X^{S(13)}$	$E_{[1,4]} > E_{[1,6]}$
$X^{S(6)}$	$E_{[1,6]} > E_{[1,7]}$	$X^{S(14)}$	$E_{[1,6]} > E_{[2,3]}$
$X^{S(7)}$	$E_{[1,1]} > E_{[3,1]}$	$X^{S(15)}$	$E_{[1,6]} > E_{[1,8]}$
$X^{S(8)}$	$E_{[1,4]} > E_{[1,5]}$	$X^{S(16)}$	$E_{[1,2]} > E_{[2,3]}$

is the indicator function with value of 1 if the relation $X^{S(k)}$ is true and 0 otherwise.

M is set to be 16 in experiment so that it enables efficient exact indexing by representing the fingerprints with an ***unsigned short integer***. And the storage requirement for a average length of 600s audio file is approximately 62KB.

IV. INVERTED INDEXING

The exact searching scheme is proposed to achieve efficient indexing. The inverted indexing with direct hashing is a obvious choice. All the fingerprints of reference audio are hashed into the inverted indexing table, keyed by the fingerprint value. The value can be represented by a unsigned short integer for the 16-bit fingerprint. The indexing strategy is shown in Fig. 3. Fig. 3(a) shows the sequence of hash values in a querying clip. And these values directly hash the ones in the inverted index table, shown in Fig. 3(b). The voting tables record the the voting number of the query and a specific reference clip. The voting number are the hitting value with the same frame difference between indexes of the matched reference and querying. The voting strategy is illustrated in Fig. 3(c)(d)(e). The voting candidate with the maximum value of N_{vote} is selected as candidate, e.g. (c) is the candidate. The time duration of the queried sequence is $[j, j+n+1]$ in the reference database.

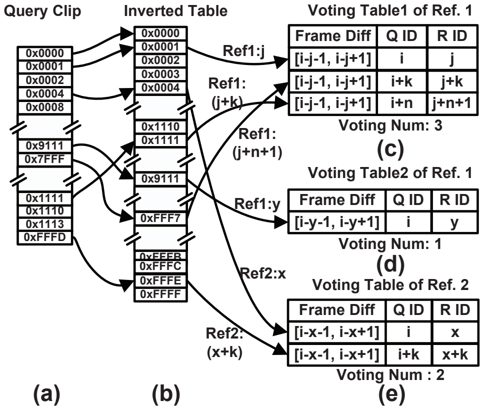


Fig. 3. Samples of the Inverted Indexing

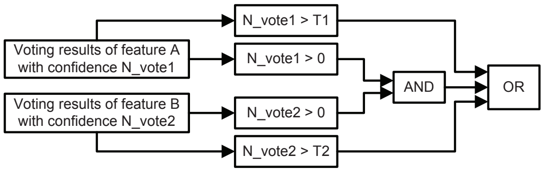


Fig. 4. The result-based fusion of two different fingerprints.

$$N_{vote} = \arg\max_{\tau} \sum_{r,q \in N} \delta(\tau - |r - q|) \tag{7}$$

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 reference sequences are regarded as the final results.

V. RESULT-BASED FUSION

Figure 4 illustrates the fusion of results given by two different audio fingerprints. The results of one fingerprint are reserved if the confidences are higher than a relatively high threshold. And the results with lower confidence are reserved if the same reference file are detected by two different fingerprints. The final results are generated by applying the "OR" to those reserved results.

The fusion results outperform any single fingerprints, described in experiment section; it indicates that the several correct detection results are missed due to their confidences are lower than the specific threshold.

VI. EXPERIMENTS

In this section, the performance of the proposed algorithm is evaluated on the TRECVID 2011 copy detection database.

VI-A. Database Description

The well-known TRECVID content-based copy detection database [11] consists of two parts: the reference datasets and query clips. The reference data contains 11187 video files, totally 400 hours. There are 8-type video and 7-type audio transformations in query data. The seven transformation types of audio based detection task are described as:(*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.

The reference 11187 audio streams contains total 70 million frames. The non-silent 1.7 million frames are randomly picked to train the 16 informative ordinal relations as described in section III.

VI-B. Parameter setting and performance

Three methods of band energy difference based fingerprint extraction are compared in this section: EDF, CEPs-like and CE_EDF. The performance are evaluated on different parameter T settings, where T is the decision threshold mentioned in section IV. Empirically, the T varies in 10, 22, 30 for EDF; 15, 20 ,40 for CEPS-like fingerprint; 40, 50, 65 for CE_EDF. Fu-1 indicates the result fusion of EDF and CEPs-like fingerprint; Fu-2 is the fusion of EDF and CE_EDF;and Inria is used to describe the compound descriptors, used by INRIA-LEAR, introduced in [4]. The Actual Normalized Detection Cost Rate(NDCR) and F1-Measure are used to measure the detection performance.

Table II shows the detection results measured by actual NDCR metric. Generally, the performance of CE_EDF is better than other features. The NDCR of fusion are lower than detection result of using any single feature. The best NDCR value of this system is 0.321 given by fusion of EDF and CE_EDF features. It is due the recalling several correct detection results which are missed, where the confidences are lower than the predefined threshold T. And for any single fingerprint, merely reducing the value of T to recall the missed correct detections leads to the NDCR increasing with more false alarms. In table II, our audio-based querying results outperform the compound descriptors. For example, in the T1 case, our and Inria's actual NDCRs are 0.321 and 0.634, respectively.

Table III shows results of different fingerprints and fusion measured by actual F1-Measure. The F1-Measure reflects the performance of time localization. The F1-measure is evaluated only for time stamps of correct detected reference copies. The missed and false detected videos have no impact on the value of F1-measure. As shown in the table, opposite to the results of NDCR, the F1-measure of CE_EDF is generally worse than the other features. It can be explained that more true positive copies ,missed by other fingerprints due to its severe transformations, are detected by CE_EDF, however the time localization of those copies are not accurate. Table III also indicates that the F1-measures of our algorithm, with lower NDCR, are comparable to the measurements of INRIA-LEAR's compound descriptors.

Table II. Actual NDCR of each fingerprint and fusion varied with threshold T

	T1	T2	T3	T4	T5	T6	T7
EDF-10	1.808	1.175	1.533	1.602	1.081	1.240	1.307
EDF-22	**0.343**	**0.448**	**0.500**	**0.480**	**0.657**	**0.672**	**0.761**
EDF-30	0.410	0.584	0.597	0.577	0.709	0.776	0.813
CEPS-15	2.007	1.915	0.838	1.481	1.006	**0.637**	0.776
CEPS-30	**0.428**	**0.517**	0.823	**0.599**	**0.679**	0.724	**0.724**
CEPS-40	0.465	0.599	**0.761**	0.746	0.679	0.791	0.746
ceEDF-40	0.465	0.639	0.940	0.823	0.604	**0.621**	0.778
ceEDF-50	**0.396**	**0.515**	**0.575**	**0.381**	**0.597**	0.753	**0.679**
ceEDF-65	0.463	0.537	0.694	0.463	0.724	0.714	0.721
Fu-1	0.321	0.527	0.396	0.520	0.562	0.500	0.545
Fu-2	0.321	0.420	0.396	0.413	0.562	0.500	0.545
Inria	**0.634**	**0.520**	**0.507**	**0.520**	**0.540**	**0.642**	**0.455**

Table III. F1-measure of each fingerprint and fusion varied with threshold T

	T1	T2	T3	T4	T5	T6	T7
EDF-10	0.901	0.891	0.912	0.884	0.866	0.875	0.867
EDF-22	0.906	0.910	0.926	0.899	0.896	0.876	0.932
EDF-30	**0.920**	**0.932**	**0.933**	**0.920**	**0.932**	**0.913**	**0.946**
CEPs-15	0.887	0.879	0.914	0.916	0.855	0.863	0.880
CEPs-30	0.886	0.899	0.923	0.931	0.900	0.889	0.917
CEPs-40	**0.903**	**0.921**	**0.930**	**0.951**	**0.902**	**0.906**	**0.919**
ceEDF-40	0.906	0.909	0.894	0.909	**0.891**	0.887	0.893
ceEDF-50	0.908	0.921	0.900	0.923	0.885	0.913	0.891
ceEDF-65	**0.921**	**0.925**	**0.911**	**0.938**	0.879	**0.932**	**0.92**0
FU-1	0.901	0.890	0.917	0.885	0.865	0.874	0.869
Fu-2	0.901	0.890	0.917	0.885	0.865	0.874	0.869
Inria	**0.939**	**0.937**	**0.904**	**0.939**	**0.923**	**0.853**	**0.923**

VI-C. Performance of fusion strategy

In table II, Fu-1 indicates fusion result of EDF and CEPs-like and Fu-2 is the fusion of EDF and CE EDF. The NDCR of fusion is lower than any single feature. The best NDCR value of this system is 0.321 given by fusion of EDF and CE_EDF. Two types of missed true positive detections are recalled, so fusion of fingerprint A and B gets the better performance: (1) As correct detections with confidences higher than TA are absence of the Bs and vice versa; (2) the detections of A and B have the same reference video ID, of which the confidences are lower than TA and TB respectively. The first one indicates that mutual complementation occurs in the result region of high confidence, and the second shows that it is able to combine difference features to form a strong voting scheme to classify the the region of low confidence. And we observed that much more missed true positive detections belong to second type. In this case, the F1-measures of fusion decrease due to recalling missed severe transformed copies in which the time localization is inaccurate, shown in table III.

VI-D. Distribution of fingerprint

The proposed feature extraction algorithm aims at mapping the audio data into fingerprints with uniform distribution based on the assumption that videos over web are uniformly random. Fig.5 shows that the distribution of different fingerprints in database is plotted in 128 contiguous buckets. The bucket is generated by equally dividing the 2^{16}

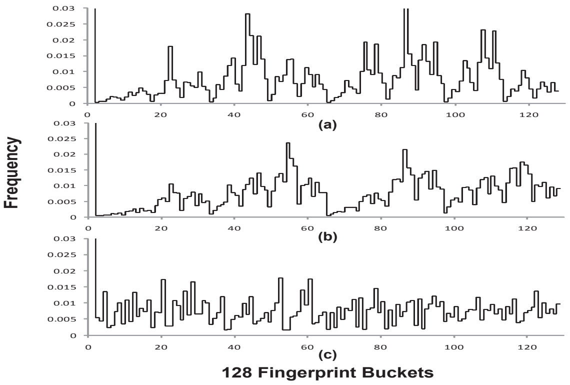


Fig. 5. Fingerprints distribution in database.

fingerprints space. The three fingerprints database distribution is more or less uniform. The exitance of some impulses is due to several reasons:(a)the a large amount of silent audio frames lead to the largest frequency value occurring in the first bucket; silent audio frame has the fingerprint value of 0; (b)there are several fingerprints value of which the frequency in database is 0; (c)a lot of audio streams have similar content in reference database.

The CE_EDF fingerprints have more uniform distribution than EDF and CEPs-like shown in Fig. 5(c), due to two advantages of the feature selection algorithm. First, the algorithm weakens the dependency of any two dimensions; the uncertainty of determine unknown bits is larger if some bit values are given. Second, the most informative relations in database are selected; for a set of specific random variables, the distribution is more uniform, the entropy is larger.

VII. CONCLUSIONS AND FURTHER WORK

In this study, two steps are adopted to generate the robust and compact audio fingerprint CE_EDF: adding multi-scale information to extend band energy difference based features and selecting the strong feature combination from the extended candidate feature set. The better performance given by fusion of two different features shows that a certain amount of correct detection results are missed due to single threshold problem. Many researches need to be done to improve the quality of detection are listed as follows:

A. The number of audio frame in the database is near 70 million. It is very time consuming if using all the frames to do feature selection with conditional entropy based algorithm. We randomly picked up 2.4% frames to select ordinal relations. Perhaps we should improve the selection algorithm so that it can handle the large amount of training data.

B. Is it possible to select the strong audio feature combination with classical algorithms such as Adaboost and SVM? And the performance comparison is needed.

C. Perhaps, for a single feature, a result refinement algorithm are needed to classify the correct and error detection results of which the confidences are low.

D. A modified indexing method is need to increase the accuracy and F1-measure without impact on the indexing efficiency.

ACKNOWLEDGMENT

The authors would like to thank Graduate Innovation Fund of SICEBUPT2011. And this work is also sponsored by National Natural Science Foundation of China (90920001), and collaborative Research Project (SEV01100474) between Beijing University of Posts and Telecommunications and France Telecom R&D.

VIII. REFERENCES

[1] “http://www.youtube.com/t/press_statistics,” .

[2] X. Wu, C. Ngo, A. G. Hauptmann, and H. Tan, “Real-time near-duplicate elimination for web video search with content and context,” *IEEE Tran. on Multimedia*, vol. 11, no. 2, pp. 196–207, 2009.

[3] J. Chen and T. Huang, “A robust feature extraction algorithm for audio fingerprinting,” in *Pacific Rim Conference on Multimedia(PCM)*, 2008, pp. 887–890.

[4] H. Jegou, Ma. Douze, G. Gravier, C. Schmid, and P. Gros, “Inria lear-texmex: Video copy detection task,” 2010.

[5] Ahmet Saracoglu, Ersin Esen, Tugrul K. Ates, Banu Oskay Acar, Unal Zubari, Ezgi C. Ozan, Egemen Ozalp, A. Aydin Alatan, and Tolga Ciloglu, “Content based copy detection with coarse audio-visual fingerprints,” in *Proceedings of the 2009 Seventh International Workshop on Content-Based Multimedia Indexing*, Washington, DC, USA, 2009, pp. 213–218, IEEE Computer Society.

[6] M. Heritier, V. Gupta, L. Gagnon, and P. Cardinal, “Crim’s content-based copy detection system for trecvid,” in *Content-Based Multimedia Indexing (CBMI)*, 2010.

[7] J. Haitsma and T. Kalker, “A highly robust audio fingerprinting system,” in *Music Information Retrieval(ISMIR)*, 2002.

[8] J. Haitsma and T. Kalker, “Robust audio hashing for content identification,” in *Content-Based Multimedia Indexing(CBMI)*, 2001.

[9] Y. Ke, D. Hoiem, and R. Sukthankar, “Computer vision for music identification: Video demonstration,” in *CVPR*, 2005.

[10] L. Shang, L. Yang, F. Wang, K. Chan, and X. Hua, “Real-time large scale near-duplicate web video retrieval,” in *ACM MM*, 2010.

[11] A. F. Smeaton, P. Over, and W. Kraaij, “Evaluation campaigns and trecvid,” in *Multimedia Information Retrieval(MIR)*, 2006, pp. 321–330.