Coherent Bag-of Audio Words Model For Efficient Large-Scale Video Copy Detection

Yang Liu†
Institute of Automation,
Chinese Academy of Sciences
Beijing, China
liuyang@nlpr.ia.ac.cn

Wan-Lei Zhao‡
Dept. of Computer Science
City University of Hong Kong
Kowloon, Hong Kong
wzhao2@cs.cityu.edu.hk

Chong-Wah Ngo‡
Dept. of Computer Science
City University of Hong Kong
Kowloon, Hong Kong
cwngo@cs.cityu.edu.hk

Chang-Sheng Xu†
Institute of Automation,
Chinese Academy of Sciences
Beijing, China
csxu@nlpr.ia.ac.cn

Han-Qing Lu†
Institute of Automation,
Chinese Academy of Sciences
Beijing, China
luhq@nlpr.ia.ac.cn

ABSTRACT

Current content-based video copy detection approaches mostly concentrate on the visual cues and neglect the audio information. In this paper, we attempt to tackle the video copy detection task resorting to audio information, which is equivalently important as well as visual information in multimedia processing. Firstly, inspired by bag-of visual words model, a bag-of audio words (BoA) representation is proposed to characterize each audio frame. Different from naive single-based modeling audio retrieval approaches, BoA is a high-level model due to its perceptual and semantical property. Within the BoA model, a coherency vocabulary indexing structure is adopted to achieve more efficient and effective indexing than single vocabulary of standard BoW model. The coherency vocabulary takes advantage of multiple audio features by computing co-occurrence of them across different feature spaces. By enforcing the tight coherency constraint across feature spaces, coherency vocabulary makes the BoA model more discriminative and robust to various audio transforms. 2D Hough transform is then applied to aggregate scores from matched audio segments. The segements fall into the peak bin is identified as the copy segments in reference video. In addition, we also accomplish video copy detection from both audio and visual cues by performing four late fusion strategies to demonstrate complementarity of audio and visual information in video copy detection. Intensive experiments are conducted on the large-scale dataset of TRECVID 2009 and competitve results are achieved.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Retrieval models

General Terms

Algorithms, Experimentation, Performance

Keywords

Audio words, copy detection, coherency vocabulary

1. INTRODUCTION

Stimulated by the growing availability of large amount of audiovisual data to the public via different ways of distribution (*e.g.*, Web-TV, video blogs, video sharing websites, etc.), which contain both professionally and home made videos, an increasing need to provide efficient access of desired content in large-scale database has emerged. Considerable research efforts [1][2] have been invested in developing the theories for content-based copy detection(CBCD) and significant improvement has been achieved. Current approaches mostly focus on visual cues which view the videos as image sequences. However, we argue that the potential contribution of audio content analysis to these tasks should not be ignored. Fig.1 shows two examples where visual cues fail while audio turns out more reliable. In news videos shown on the left part of Fig.1, anchor frames are removed as useless materials, because it is difficult to distinguish different topics which are reported by the same anchorperson in the same TV station. In contrast, different topics can be distinguished by audio cues easily. The right part of Fig.1 shows three different live show versions in different years of the song 'Dangerous' of Michael Jackson on YouTube website. Due to the fact that visual information are different from each other significantly, it is difficult or even impossible to associate them together, whereas we are able to easily identify that they belong to the same song of the same person using audio information.

Inspired by the text (key) words in classic text document, several attempts have been tried to apply audio (key) words in the multimedia content analysis. Xu *et al.*[3] proposed an audio keywords to detect key audio effect such as whistling and audience sound to assist event detection in soccer video. Lu *et al.*[4] proposed an audio content analysis approach to retrieve audio clips based on audio keywords extracted using several acoustic features. However, the studies on keyword-

CIVR '10, July 5-7, Xi'an China

That is our report on “world news tonight.” later on “nightline” - they are graphic, disturbing, and apparently effective. They are some of the newest anti - smoking ads. I m peter Jennings. Have a good evening.

President Clinton said today the U. S. must try to limit the international financial crisis before it gets worse in Latin America.

In Berlin today one of the world’s most famous places has been rededicated dedicated after more than 50 years.

Michael Jackson Dangerous (Live in MJ&Friends Seoul 1999)
Michael Jackson's Dangerous performance in Seoul concert which aim was helping the children in 1999. When you watch this performance, you'll see ...
★★★★★ 3 years ago 253,503 views Erkal

Michael Jackson Dangerous (Live in Munich 1997)
Michael Jackson's Dangerous performance of Munich concert which is conceded as his best HIStory tour concert.This concert is one of the most ...
★★★★★ 3 years ago 4,663,912 views Erkal

Michael jackson - dangerous live 2002 (latest performance)
Dangerous live
★★★★★ 1 year ago 671,046 views mjytaps

Figure 1: Two examples show the contribution of audio information in audiovisual data analysis. The left example is the case of similar visual appearance (the same anchorperson) with different audio information (different news topics), while right example is the case of different visual appearance (different live shows) with the same audio information (the same song).

based audio copy retrieval/detection are get to be fully addressed in real application. There are basically two challenging issues for audio feature representation. Firstly, previous researchers have proven that single audio feature is insufficient to represent audio word, which could be music, speech and other sound such as applause, cheer, cry, noise and so on. Therefore, the diversity of audio content makes it necessary to characterize audio words with appropriate audio features. One direction of existing approaches attempted to concatenate several different features, which represent the same audio word, into a long feature vector, which did not consider the correspondence between audio descriptor and audio content. Therefore, it is desirable to come up with a solution to characterize the audio keywords with appropriate audio features. Secondly, a lot of existing works applied the simplest matching strategy for audio retrieval, which directly compares query audio with candidate audio clip with a fixed length sliding window without any indexing structure. In order to achieve efficient index and effective retrieval, Haitsma *et al.* [5] used binary audio fingerprint and hash table to realize fast similarity measure. However, the binary fingerprint is sensitive to the hash function and noise, even a flip with change of two bits will make the binary fingerprint change largely. Hence it is necessary to find out an efficient and robust indexing structure for audio information retrieval.

In this paper, we propose a novel Bag-of Audio words (BoA) based video copy detection approach. This approach is based on the standard BoW representation and inverted file retrieval. We firstly segment the audio signals into frames. Each frame is viewed as a bag of audio words analogous to bag of visual words in image. The fixed time durations with overlap in one frame are viewed as audio words. In order to maintain the robustness and distinctiveness for each audio word, instead of concatenating several audio descriptors into a long vector, we apply coherency vocabulary indexing structure to effectively take advantage of multiple audio descriptors and make the retrieval process more efficient. The coherency vocabulary uses co-occurrence vocabulary of multiple audio descriptors instead of audio vocabulary of single descriptor to characterize each audio word. The BoA based coherency vocabulary offers several advantages.

- **Effective**: Coherency vocabulary enforces a temporal constraint across multiple audio feature spaces to avoid false matches, which are introduced by the quantization error.
- **Robust**: Co-occurrence of multiple audio descriptors are able to characterize different audio signals sufficiently, which makes the BoA model robust to various audio transforms.
- **Efficient**: Coherency vocabulary is able to reduce the indexing time significantly. If dimension of audio vocabulary is N, traditional BoA model should compute N times to search the nearest neighbor, while coherency vocabulary only needs $M \times \sqrt[M]{N}$ comparisons with M audio descriptors characterizing each audio word.

After obtaining the sets of matched audio frame pairs by inverted file retrieval, 2D Hough transform is applied to aggregate the vote score and locate the audio copy clip. By accumulating vote score of each matched audio frame pairs, 2D Hough transform can find the optimal time offset to accurately locate copy segment. In order to investigate complementarity between visual and audio information in video copy detection, we also attempt different late fusion strategies between audio-only copy detection result and our video-only copy detection result to generate audio-video detection results.

The rest of this paper is organized as follows. In section 2, we give a brief overview of related works on audio copy detection. Section 3 introduces the audio features used in this paper. The BoA representation and coherency vocabulary are presented in section 4. Section 5 describes the framework of our BoA based copy detection. Section 6 presents the late fusion of audio and visual cue for video copy detection with audio and visual cues. Experiment results and analysis are reported in section 7. Finally we conclude the paper with future work in section 8.

2. RELATED WORK

The definition of ‘copy’ is that a segment of video (audio) derived from another video(audio) by means of various visual/audio transformations such as addition, deletion, modification of aspect, color, contrast (compression, multiband companding, mix with speech) etc[1]. It is worth noting that CBCD is different from identical and near-duplicate detection. As pointed out in [2], one major challenge in CBCD is that a copy can be visually dissimilar, which means it is not necessarily an identical or a near-replication, but rather

[1] http://www-mlpir.nist.gov/projects/tv2009/tv2009.html#4.4

a transformed video sequence. In this work, an audio copy is a segment of audio derived from another with some audio transformations such as compression, multiband companding, mix with speech etc.

Before introducing the related work using audio information to resolve video copy detection, we give a brief overview of copy detection approaches using visual cues. Existing CBCD approaches are mostly related to near-duplicate detection, since most approaches apply the duplicate or near-duplicate methods to copy detection task. Usually the general framework of CBCD with visual information can be roughly divided into two categories: frame-level and clip-level. Most frame-level approaches extract local interest point descriptor as visual feature, which is more discriminative than global feature and has achieved significant result [6][7]. Due to the heavy computational cost of local descriptors, it is necessary to index them with efficient indexing structure. Ngo *et al.* [8] addressed the speed issue by indexing the PCA-SIFT keypoints with LIP-IS technique. Similarly, a Z-grid based probabilistic retrieval approach is proposed in [9] to realize efficient indexing. In order to improve the effectiveness and efficiency of approaches based on local interest points, BoW representation combined with inverted file retrieval approach is proposed in [10]. Especially in [11], Hamming Embedding (HE) and weak geometric constraint (WGC) have been proposed to impose visual and geometric verification on the visual word matches. Due to the robustness of such verification, it shows satisfactory performances in last year video-only copy detection task. However, this approach needs additional storage space for binary signatures and scale, orientation information. For clip-level approaches, Law-To *et al.* [1] proposed a copy detection approach and achieved promising result. They built the trajectories along with the temporal line by means of tracking local interest points and detecting the copy in terms of robust voting function. Nevertheless, trajectory extraction is a very expensive process due to the need for tracking keypoints over frames, especially for online retrieval.

Video copy detection using audio cues has attracted more research interests recently. TRECVID has tried to explore audio-only copy detection for large-scale video copy detection. Most existing audio copy detection approaches are signal-based labeling/modeling audio retrieval, which extract audio waveform and compute fingerprint from time and frequency domains. Haitsma *et al.* [5] proposed a highly robust audio fingerprinting system for audio retrieval. Audio signal is segmented into frames and a 32-bit sub-fingerprint is extracted for every frame from 33 non-overlapping frequency energy bands after Fourier transform. Hash table is then adopted for efficient online retrieval. It is considered as one of the most famous audio fingerprint techniques, therefore, some variants have been proposed to improve it. Li *et al.* [12] chose five most stable bits from each 32-bit sub-fingerprint to represent every frame instead of using the original 32-bit fingerprint. To enhance the robustness of the audio fingerprint, Liu *et al.* [13] proposed a multiple hashing algorithm by means of DCT coefficients of the time sequence of band energies in each band instead of band energies themselves. Besides the audio fingerprint, audio keywords are created from low-level audio features by using SVM learning in [3]. The audio keywords can be used to detect semantic events in sports video by applying a heuristic mapping. Lu *et al.* [4] proposed an audio word based audio analysis and retrieval approach, directly concatenating several audio features, a 29-dimensional feature vector is extracted for each audio frame as an audio word.

3. AUDIO FEATURE EXTRACTION

We briefly describe the commonly used audio features and discuss the one we adopted in this work. Audio Fingerprint has attracted attention since it allows monitoring of audio independently of its format without the need of metadata or watermark embedding. Audio fingerprint techniques perform quite well due to the property of fast computation, memory efficient and easily undatable. Such features do not need to have anything to do with human perception or music semantic, while they just need to be unique and robust against distortions. Moreover, an alternative way of comparing audio recordings in a meaningful way is to extract an abstract description which can reflect the perceptional and semantical aspects of the audio. Audio is usually segmented into short, possibly overlapping frames. The features which can be derived from time and frequency domains will be extracted from each frame.

- **Features derived in the time domain:** Time domain representation is the most basic signal representation, where a signal is represented as amplitude varying with time. For example, average energy, zero crossing rate and silence ratio are time domain audio features.

- **Features derived in the frequency domain:** The frequency component and distribution of audio can be shown in the frequency domain. Most audio representations are derived from frequency domain, such as loudness, pitch, tone, Mel-frequency Cepstral Coefficients(MFCCs), LPCs, etc.

The first step in any pattern recognition system is the feature selection that is how exactly to represent the audio signal, in order to facilitate audio copy detection. Audio signals usually contain various contents characterized by diverse audio features. For instance, MFCCs is the most popular feature representation in speech recognition system, while music retrieval systems usually adopt pitch, harmony and rhythm features. In our work, we select two common audio features: MFCCs and Relative Spectral Transform-Perceptual Linear Prediction (RASTA-PLP) to represent audio words.

MFCCs: Through more than 30 years of speech recognition research, many different feature representations of the speech signal have been suggested and researched. The most popular feature representation currently used is the MFCCs. The MFCCs are coefficients representing the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

RASTA-PLP: Another popular audio feature representation is known as RASTA-PLP, which was originally proposed by Hynek Hermanky [14] as a way of warping spectra to minimize the differences between audio signals. RASTA-PLP replaces a conventional critical-band short-term spectrum in PLP speech analysis with a spectral estimate in which each frequency channel is band-pass filtered by a filter with a sharp spectral zero at the zero frequency.

4. BAG-OF-AUDIO WORDS REPRESENTATION

4.1 Audio words generation

Inspired by the BoW model used in visual information retrieval, in this paper, we propose the bag-of audio words (BoA) representation to accomplish audio copy detection. Similar to video, audio should be segmented into frames. However, different from the video that similar visual scenes can persist a few seconds, even several minutes, audio signal may change largely during a slight duration. In other words, audio signal may be only stable within milliseconds. In retrieval system with BoA model, there are two factors that will affect the system efficiency. One is the number of the bags and the other is the number of audio words in each bag. In BoA based audio retrieval system, the word length plays a trade-off role between system effectiveness and efficiency. Decreasing the number of words could improve the retrieval speed, but it could also impair the discriminative power of the words. Taking into account the large scale dataset, in order to make a compromise between effectiveness and efficiency, the proposed BoA extraction scheme extracts one frame per second. Within one frame, one audio word is extracted every 40ms with an 50% overlapping between consecutive words. The high overlapping rate assures that the query audio clip is still similar to the copy clip in the database under strong distortion. As a result, one second audio segment is represented as BoA which contains 50 audio words. The MFCCs and RASTA-PLP feautres are extracted in each 40ms audio signal with 20ms overlap, one second audio clip is equivalent to one frame of a video sequence.

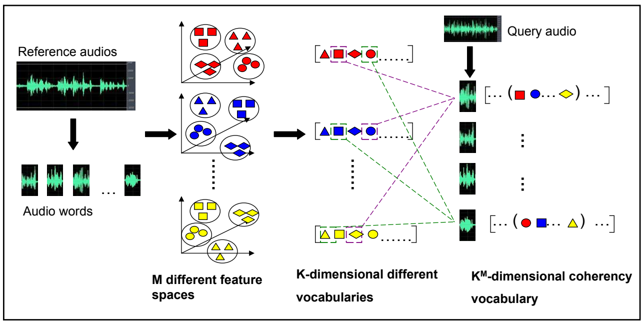
Reference audios
Audio words
M different feature spaces
K-dimensional different vocabularies
Query audio
[... (...) ...]
[... (...) ...]
K^M-dimensional coherency vocabulary

Figure 2: The demonstration of the coherency vocabulary indexing structure.

4.2 Coherency vocabulary indexing structure

Vocabulary (also called codebook) generation and indexing is a necessary step of video retrieval with BoW model. Recently, many researchers have noticed the importance of the vocabulary indexing in BoW model and several novel algorithms are proposed to improve the performance of vocabulary. Jegou *et al.* [11] proposed a two-layer vocabulary tree and the Hamming Embedding (HE) to prune false alarms in each cluster of the vocabulary. HE achieves good performance with small vocabulary size. Yeh *et al.* [15] proposed an adaptive vocabulary forests to improve recognition and indexing performance. An ensemble vocabulary forest consists of several vocabularies and each vocabulary can dynamically and incrementally adapt with new data added. Following the similar idea described in [16], which proposed a coherent phrase of different visual descriptors to perform image near-duplicate detection, we simplify and extend the coherent phrase on audio copy detection, which only considers the co-occurrence between complementary audio features and generates coherent vocabulary for efficient indexing and retrieval.

As aforementioned in section 3, different audio contents should be analyzed with corresponding audio features. Due to the limitation of the single audio features, in this paper, we propose an audio word based coherency vocabulary indexing structure, which attempts to characterize each 40ms audio clip with multiple audio descriptors. Fig.2 shows the coherency vocabulary generation. For each audio word (40ms audio clip) M audio features are extracted, and M single vocabularies are generated by clustering algorithm with the size of K.

$$V_i = \{V_{i1}, V_{i2}, ..., V_{iK}\} \tag{1}$$

where $i \in [1, 2, ..., M]$ indicates the series number of single vocabulary, V_{iK} indicates the K_{th} audio word in i_{th} vocabulary. Given an audio word, the nearest neighbor audio word center is searched by measuring the metric distance such as cosine distance and Euclidean distance in each single vocabulary. Then the combination of the M single vocabularies with dimension of K can form a K^M dimensional coherency vocabulary, where each dimension denotes the co-occurrence of different audio features in feature spaces. The advantage of the coherency vocabulary is that it sufficiently takes into account the reaction of audio signal to various audio features and effectively avoids the limitation of the single audio features, because the co-occurrence enforces a tight coherency in different audio feature spaces. In addition, the coherency vocabulary indexing structure significantly improves the efficiency of the offline indexing and online retrieval. For example, usually one million dimensional vocabularies need 1,000,000 comparisons to find the nearest neighbor for each audio word. Nevertheless, it only needs 2000 comparisons $M = 2, K = 100$, even 300 comparisons with $M = 3, K = 100$.

5. AUDIO COPY DETECTION FRAMEWORK

In this section, we describe the framework of coherent bag-of audio words based audio copy detection, which is shown in Fig.3. The procedure generally consists of two parts: the offline indexing (the pink blocks connected by red arrows)and online retrieval (the yellow blocks connected by blue arrows). In offline phrase, audio signals are extracted from reference videos and converted to mono PCM with fixed sampling rate. Afterwards, each audio signal is segmented into frames every 1s without overlap. Within a frame, we extract an audio word every 40ms with 50% overlap between consecutive audio words. For each audio word, we extract 39-dimensional Mel-Frequency Cepstrum Coefficients (MFCCs) and 27-dimensional RASTA-PLP respectively. As a result, one audio frame (1s audio duration) contains 50 audio words (40ms with 50% overlap) and each frame is characterized by two bags of audio words. For each audio feature, we generate a vocabulary with N clusters, which results in a N^2 dimensional coherency vocabulary by computing the co-occurrence of two audio features across feature spaces. Ultimately, in order to compute the similarity between audio frames efficiently, the entire set of coherent descriptors of audio dataset is stored in an inverted file, which is composed of N^2 lists of coherent descriptors. At the online phrase, the queries are undergone the same processing as the reference set. The online retrieval is conducted on the inverted file structure. The output of the online retrieval are a set of matched audio frame pairs between the query and reference. We apply a 2D Hough transform to estimate the optimal temporal offset and localize the copy segment in reference audio.

Given a query, the output of the BoA retrieval is a candidate matching audio frame list for each query audio frame. Each candidate matching audio frame is associated with a similarity score between query frame and itself. The aim of 2D Hough transform is to estimate the optimal temporal offset between query and reference audio based on these audio frame pairs and their similarity score. We denote the m_{th} audio frame of query i to be $f_{i,m}$ and the n_{th} audio frame of reference j $f_{j,n}$. We consider the difference of frame ID m and n as the temporal offset $\delta_{m,n}$ between query and reference audio. The two parameters of 2D Hough Transform are reference video ID R and temporal offset δ. The combination of R and δ is viewed as the bin of 2D Hough Transform, which also corresponds to a potential audio copy segment. The accumulated similarity score for each bin $[R, \delta]$ is the weight of hypothesis that the ID R reference audio contains copy segment, which has δ time shift with copy segment in query. Therefore, via 2D Hough transform, we can aggregate matched audio frame pairs and accurately locate the copy segment in reference audio.

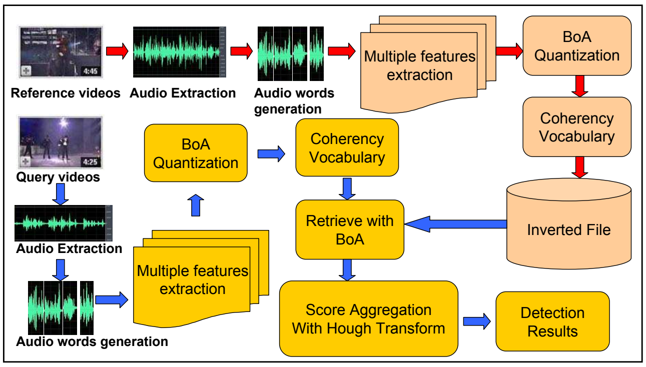

Figure 3: The framework of coherent BoA based copy detection.

6. AUDIO-VIDEO COPY DETECTION

6.1 BoV based Video-only Copy Detection

Our video-only copy detection algorithm is based on single feature BoW representation and inverted file retrieval [17]. The framework is composed of two phases: offline indexing phase and online retrieval phase. In offline indexing phase, firstly we sample the frames from reference videos with a fixed sampling rate and each frame is sampled every 2.5 seconds. Then we adopt DoG detector and SIFT descriptor to extract local interest point descriptors. Subsequently, a two-layers hierarchical visual vocabulary with 2,000 visual words in top layer and 20,000 in bottom layer is generated up-bottom for BoW frame representation, and inverted file is employed to index the reference frames. It is worth noting that multiple assignment strategy is employed in indexing step. An input descriptor searches 20 nearest neighbors in top layer of two-layer indexing structure, which results in that this input descriptor falls into 200 visual words in bottom layer. In online retrieval phase, because BoW model ignores the spatial information and results in the ambiguity of visual word, enhanced weak geometric consistency checking (EWGC) is adopted for geometric verification. In addition, we also attempt to use hamming embedding (HE) and pattern entropy checking (SR-PE [18]) to prune false alarms prior to and after the EWGC.

6.2 Comparisons between bag-of audio words and bag-of visual words

In this paper, we perform the video copy detection from both audio and visual cues respectively. Although we concentrate on the BoA based video copy detection, we attempt to give a comparison between BoA and BoV for following.

- **Memory & storage consumption**: The memory and storage consumption relates to many factors, such as the number of features (words), the dimension of features, vocabulary size and the sparsity of the vocabulary after feature indexing. Compared with visual information, audio signals are able to persist shorter time duration, hence wwe sample audio frames 2.5 denser than the video frames. However, there are average 400 SIFT descriptors extracted from one video frames as the visual words, while audio words in each audio frame is a constant number 50. Due to low feature dimension and smaller size of the bag of feature, although we sample audio frames 2.5 times denser than video frames, the total storage consumption for audio word is still much lower than BoV in video-only copy detection[2]. Table 1 lists more details for this. Both the number of audio words in one frame and dimension of the audio feature are much fewer than visual words. Therefore, BoA model consumes little memory than BoV model.

Table 1: Comparisons of BoV and BoA in terms of memory & storage consumption.

	BoV	BoA
sample rate	2.5s	1s
N_{frame}	567,056	1,417,640
N_{fea}/frame	400	50
dim_{fea}	128 (SIFT)	66 (MFCC+RASTA-PLP)
storage	277.12G	43.3G

N_{frame} denotes the total number of the frames. N_{fea}/frame means number of features extracted in one frame.

- **Efficiency**: Time consumption of system consists of three parts: feature extraction, indexing and online retrieval. The time of each parts is listed in Table 2. Among three parts, BoA is more efficient than BoV. Because the MFCCs and RASTA-PLP extraction is relatively simple compared with SIFT descriptor. In addition, the coherency vocabulary make feature indexing more efficient For instance, 250k dimension coherency vocabulary with two audio features only needs 1000 comparisons, while 20k two-layer hierarchical BoV

[2] In BoV, additional information such as the location, scale and orientation also takes up memory space.

vocabulary needs 2200 comparisons. In TRECVID 2009, the mean length of query is 178 seconds. Both our BoV and BoA model, which address video copy detection task from different views, are able to carry out real-time detection.

Table 2: Efficiency comparisons between BoV and BoA.

	BoV	BoA
T_{ex}	5.3ms	0.98s
V_{size}	top 2k bottom 20k	500×500=250,000
T_{index}	1.7ms	0.2ms
T_{retri}	1.5ms	0.23ms
T_{query}	134s	15s

T_{ex}: time of one feature extraction; V_{size}: vocabulary size; T_{index}: time of indexing one feature; T_{retri}: time of one feature retrieval; T_{query}: the average time for retrieving one query.

6.3 Late fusion strategies of audio-video copy detection

After obtaining the audio-only and video-only retrieval results via BoA and BoV models respectively, we normalize the scores of audio-only and video-only respectively. Linear fusion is employed to generate the final audio-video results. We summarize several fusion strategies in our late fusion as follows:

- Average fusion: the weights of audio and video scores are 0.5 in linear combination, $S_{fusion} = 0.5 \times S_{BoA} + 0.5 \times S_{BoV}$
- Max operation: The larger score between audio and video results is accepted as fused score. $S_{fusion} = \max\{S_{BoA}, S_{BoV}\}$
- Multiply operation: The fusion score is determined by the product of audio and video scores. $S_{fusion} = S_{BoA} \times S_{BoV}$
- Logistic regression: It is used for prediction of the optimal fusion weights (p_v, p_a) by fitting data to a logistic curve. $S_{fusion} = p_v \times S_{BoV} + p_a \times S_{BoA}$

7. EXPERIMENT

In order to verify our BoA based audio copy detection approach, we conduct the experiment on the 2009 TRECVID dataset corpus. In this section, we firstly introduce the experimental dataset, then give the details of the experiment setup. The experiment results are reported mainly for two tasks: audio-only and audio-video. Finally, experiment analysis is elaborated based on the experimental result observation.

7.1 Dataset and experimental setup

The size of dataset for copy detection task of TRECVID 2009 is one time more than TRECVID 2008 [19]. The reference dataset contains 838 videos, which mainly refer to news videos and films. The total duration are about 380 hours with 343.4 GB storage consumption. 201 queries are given for video and audio only detection. Those 201 queries are undergone 7 different visual transforms and 7 different audio transforms respectively, which results in 1407 video-only queries and 1407 audio-only queries. The transforms are listed in Table 3.

The audio-video queries consist of the aligned versions of transformed audio and video queries, which are various combinations of transformed audio and transformed video from a given base audio-video query. Therefore, in this way, 9849 audio-video queries are generated according to 49 transforms.

For audio-only copy detection, we first convert the audio signals into mono PCM (16 bits) with $44.1KHz$ sampling rate. Afterwards, each audio signal is segmented into frames every 1s without overlap. Within a frame, we extract an audio word every $40ms$ with 50% overlap between consecutive words. For each audio word, we extract 39-dimensional MFCCs and 27-dimensional RASTA-PLP respectively. As a result, one frame (1s audio duration) contains 50 audio words (40ms with 50% overlap) and each frame is characterized by two bags of audio words. For each audio feature, we generate a vocabulary with 500 clusters, which results in a 250,000 dimensional coherency vocabulary by computing the co-occurrence.

For video-only copy detection [17], due to the large volume of reference dataset, we sample the frames from reference dataset and queries asymmetrically. One frame is sampled every 2.5s in reference dataset, the query is processed by sampling one frame per 1.25 second. Keypoints are extracted by DoG and described with SIFT descriptor. A two-layers hierarchical visual vocabulary of 2K top words and 20K bottom words is generated in a top-down manner for BoV frame-based representation. Inverted file is then employed to index the set of extracted frames.

7.2 Evaluation

There are three parameters for CBCD performance evaluation.

- Normalized detection cost rate (NDCR): NDCR is defined by NIST to evaluate the precision and recall, which can be formulated as follows:

$$NDCR = P_{Miss} + \beta \cdot R_{FA}$$

  where P_{Miss} and R_{FA} are the conditional probability of a missed copy and the false alarm rate respectively, $\beta = (C_{FA}/(C_{Miss} \cdot R_{target}))$, $R_{target} = 0.5/hr$, $C_{Miss} = 1$, $C_{FA} = 1000$ for 'no false alarm' profile and $C_{FA} = 1$ for 'balanced' profile
- F_1: it indicates that when a copy is detected, the asserted and actual extents of the copy in the reference data will be compared using precision and recall and these two numbers will be combined using the F_1 measure.
- Process time: mean time to process a query.

7.3 Audio-only copy detection

In this section, the copy detection results only based on audio cue is presented. In order to verify the superior ability of coherency vocabulary related to BoW model with single vocabulary, we extract MFCCs (39 dimension) and RASTA-PLP (27 dimension) to represent each audio word respectively. Then we conduct the comparison between coherency vocabulary and single vocabulary. Table.4 shows the experimental performance. R-PLP, M+R and MFCC use traditional single BoW approach with three single features. M+R denotes that we concatenate the MFCCs and RASTA-PLP features directly to form a 66 dimension features. Based

Table 3: video and audio Transformations.

Audio transforms		Video transforms	
T1	nothing	T2	Picture in picture
T2	mp3 compression	T3	Insertions of pattern
T3	mp3 compression and multiband companding	T4	Strong reencoding
T4	bandwidth limit and single-band companding	T5	Change of gamma
T5	mix with speech	T6	Decrease in quality
T6	mix with speech, then multiband compress	T8	Post production
T7	bandpass filter, mix with speech, compress	T10	random combination

Table 4: Audio-only performances of different video-only runs.

	Opt.NDCR					**Opt.F1**				
	R-PLP	M+R	MFCC	coherency	best	R-PLP	M+R	MFCC	coherency	best
T1	0.493	0.204	0.204	0.142	0.052	0.462	0.489	0.534	0.547	0.921
T2	0.455	0.418	0.313	0.306	0.06	0.457	0.515	0.523	0.515	0.936
T3	0.697	0.731	0.629	0.629	0.067	0.436	0.491	0.544	0.545	0.924
T4	0.674	0.532	0.515	0.485	0.06	0.428	0.505	0.537	0.537	0.89
T5	0.634	0.664	0.537	0.530	0.06	0.406	0.475	0.495	0.58	0.92
T6	0.791	0.799	0.726	0.697	0.075	0.364	0.447	0.466	0.481	0.90
T7	0.963	0.978	0.948	0.918	0.082	0.256	0.376	0.415	0.425	0.90

on the Table 4, the coherency vocabulary is best among four implementations. The fact is that the discriminative power of simply concatenating features does not increase as the feature dimension increasing, even it is worse than original single MFCCs feature, because the RASTA-PLP impairs the discriminative power of the MFCCs. In addition, we also conduct a comparison between our BoA model with the CRIM's approach [20] which achieved the best performance in TRECVID 2009. This approach is based on [5] with some variation of fingerprint and matching strategy. Its performance is much better than other participants' as well as ours, although the processing time is slightly longer than ours.

In order to give an intuitive expression of the precision and recall, we list the details of detection results in Table 5 and the overall performance in Table 6. We evaluate the performances of our approach on each transformation in two ways. One is that we only consider the returned reference ID, the other considers both the reference ID and the time location of the detected copy. The performance shown in Table 5 is consistent with Table 4, where the coherency vocabulary achieves the best performance. However, the difference between two cases is obvious. If only consider the result ID, T1, T2 and T4 can achieve good performance, but consider both ID and time location, the performance will be droped significantly, about by 25%. The reason is that the discriminative power of features such as MFCCs, RASTA-PLP of audio is not as well as SIFT, SURF of image. Another reason of the inaccurate location is that audio signal can change significantly within a very short time duration, therefore, the case that large amount matching lines between reference and query which usually appear in videos retrieval seldom happens in audio retrieval. In our experiments, most frame pairs only share 2-3 correct matching lines while video frame pairs are usually more than 10.

Among the seven audio transforms, the detection difficulty increases from T1 to T7, which can be also confirmed by the performance of TRECVID 2008. Just like the performance of audio fingerprint approach which employed in TRECVID 2008, our approach cannot handle T6 and T7 well. We also check the miss detection of T1 performance and found that most of them are silence signals or other sound without speech, such as crying, wind sound which mostly appear in tv7.sv.devel and tv7.s.test datasets. MFCCs are robust to the speech and usually are applied to speech recognition, however, it does not work for silence signals. The overall precision and recall are shown in Table 6.

Table 5: The truth positive number for each transform.

	GT	R-PLP		M+R	
		TP_{id}	TP_{loc}	TP_{id}	TP_{loc}
T1	134	96	69	122	87
T2	134	89	64	107	76
T3	134	69	50	88	65
T4	134	88	64	98	66
T5	134	74	53	77	55
T6	134	52	36	59	44
T7	134	24	18	25	17

	GT	MFCC		coherency	
		TP_{id}	TP_{loc}	TP_{id}	TP_{loc}
T1	134	121	89	128	96
T2	134	113	82	125	94
T3	134	85	63	92	70
T4	134	104	74	115	85
T5	134	93	69	108	83
T6	134	66	46	83	63
T7	134	34	26	52	44

GT denotes the number of ground truth for each transform; TP_{id} indicates the truth positive number of the correct results only based on ID, regardless of the time location; TP_{loc} indicates that besides correct ID, the truth positive number of the results with correct time location.

7.4 Audio-video copy detection

In this section, we attempt to accomplish video copy detection by lately fusing BoA and BoV results. We choose four late fusion strategies which are elaborated in section 6.3. Due to that there are only two cases to be fused, the fused performance of these four fusion strategies dose not show large difference as reported in Table 7. Fusion result based on logistic regression is slightly better than others. Based on

Table 6: Overall of the precision and recall for audio-only detection.

	R-PLP		M+R	
	TP_{id}	TP_{loc}	TP_{id}	TP_{loc}
Precision	43.77%	31.49%	61.41%	43.71%
Recall	52.45%	37.74%	58.60%	41.71%
Time	13.649 s/query		17.006 s/query	
	MFCC		coherency	
	TP_{id}	TP_{loc}	TP_{id}	TP_{loc}
Precision	60.93%	44.41%	75.10%	57.16%
Recall	65.67%	47.87%	74.95%	57.04%
Time	13.717 s/query		8.107 s/query	

our observation, compared with performances of our video-only, the fusion between audio and video wins 10% improvement of recall while at the cost of 13% precision dropping. However, the fusion results are much better than audio-only. The precision improves about 15% and recall also increase 6%. This indicates that audio information and visual information is relatively complementary with each other, thus helpful for solving to solve copy detection task.

Table 7: Overall of the precision and recall for audio-video detection.

	Precision	Recall
Average fusion	78.32%	74.54%
Max operation	76.53%	72.42%
Multiply	73.89%	68.58%
Logistic regression	81.16%	76.36%

8. CONCLUSIONS

In this paper, we propose a bag-of audio word (BoA) model for efficient large-scale video copy detection from the audio cue. BoA model combined with inverted file retrieval can realize real-time detection. Another novelty of this paper is coherency vocabulary indexing structure, which not only improves the efficiency of whole detection process, but also enforces a strong coherency constraint across feature spaces to make the BoA model more discriminative and robust to audio transforms. In addition, we also employ the late fusion to perform the video copy detection with both audio and visual information. Intensive experiments are conducted on TRECVID 2009 and explanations are given in terms of various evaluations. Our work can be extended for the early fusion of BoA and BoV. Due to the nature that audio and visual cues are synchronized, early fusion is allowed.

9. ACKNOWLEDGMENTS

The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 119508), a grant from City University of Hong Kong (Project No. 7002438), NSFC Grant No.60833006, 90920303 and 973 project No.2010CB32905.

10. REFERENCES

[1] J. Law-To, O. Buisson, V. Gouet-Brune, and N. Boujemaa. Robust voting algorithm based on labels of behavior for video copy detection. *ACM MM 2006.*

[2] J. Law-To, L cheng, A. Joly, I. Laptev, O. Buisson, V.G. Brunet, N. Boujemaa, and F. Stentiford. Video copy detection: a comparative study. *CIVR 2007.*

[3] M. Xu, Changsheng Xu, M. Kankanhalli, and Q. Tian. Creating audio keywords for event detection in soccer video. *ICME 2003.*

[4] L. Lu and A. Hanjalic. Audio keywords discovery for text-like audio content analysis and retrieval. *IEEE Transaction on Multimedia*, 10, 2008.

[5] J. Haitsma and T. Kalker. A highly robust audio fingerprinting system. *IRCAM 2002.*

[6] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, pages 91–110, 2004.

[7] B. Hebert, E. Andreas, T. Tinne, and G.L. Van. Surf: Speeded up robust features. *CVIU*, pages 346–359, 2008.

[8] C.-W. Ngo, W.-L. Zhao, and Y.-G. Jiang. Fast tracking of near-duplicate keyframe in broadcoast domain with transitivity propagation. *ACM MM 2006.*

[9] S. Poullot, O. Buisson, and M. Crucianu. Z-grid-based probabilistic retrieval for scaling up content-based copy detection. *CIVR 2007*, pages 348–355.

[10] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. *ICCV 2003.*

[11] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. *ECCV 2008.*

[12] Q. Li, J. Wu, and X. He. Content-based audio retrieval using perceptual hash. *ICIIHMSP 2008.*

[13] Y. Liu, K. Cho, H.S. Yun, J.W. Shin, and N.S. Kim. Discrete cosine transform based multiple hashing technique for robust audio fingerprinting. *ICASSP 2009.*

[14] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Transaction on speech and audio processing, 1994.*

[15] T. Yeh, J. Lee, and T. Darrell. Adaptive vocabulary forests br dynamic indexing and category learming. *ICCV 2007.*

[16] Y.-Q Hu, X.-G. Cheng, L.-T. Chia, X. Xie, D. Rajan, and A.H. Tan. Coherent phrase model for efficient image near-duplicate retrieval. *IEEE Transaction on multimedia, 2009.*

[17] C.-W. Ngo, Y.-G. Jiang, X.-Y. Wei, W.-L. Zhao, Y. Liu, S.-A. Zhu, J. Wang, and S.-F. Chang. High-level feature extraction, automatic video search, and content-based copy detection. *notebook of TRECVID 2009.*

[18] W.-L. Zhao and C.-W. Ngo. Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection. *IEEE Transaction on Image processing, 2009.*

[19] Wessel Kraij, George Awad, and Paul Over. Trecvid 2009 content-based copy detection task. *TRECVID 2009 notebook.*

[20] M. Héritier, V. Gupta, L. Gagnon, G. Boulianne, S. Foucher, and P. Cardinal. Crimąäs content-based copy detection system for trecvid. *notebook of TRECVID 2009.*