Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 5 (2011) 149–156

Procedia Computer Science

The 2nd International Conference on Ambient Systems, Networks and Technologies (ANT-2011)

A Novel Approach to Video Copy Detection Using Audio Fingerprints and PCA

R. Roopalakshmi[a,*], G. Ram Mohana Reddy[a]

[a] *Information Technology Department, National Institute of Technology Karnataka, Surathkal, Mangalore, India 575025.*

Abstract

In Content-Based Copy detection(CBCD)literature, numerous state-of-the-art techniques are primarily focusing on visual content of video. Exploiting audio fingerprints for CBCD problem is necessary, because of following reasons:audio content constitutes an indispensable information source;transformations on audio content is limited compared to visual content. In this paper, a novel CBCD approach using audio features and PCA is proposed, which includes two stages:first, multiple feature vectors are computed by utilizing MFCC and four spectral descriptors;second, features are further processed using PCA, to provide compact feature description. The results of experiments tested on TRECVID-2007 dataset, demonstrate the efficiency of proposed method against various transformations.

Keywords:
Content-based video copy detection, audio fingerprints, MFCC, PCA, Spectral Descriptors.

1. Introduction

The exponential growth of multimedia and web technologies have increased Internet based video publishing and sharing activities tremendously. Controlling the copyright of huge number of uploaded videos is a critical challenge for popular web servers. Hence, video copy detection is compulsory to reduce copyright violations. In general, a video copy is defined as, a transformed video sequence, derived from master video. There are two approaches for detecting copies of a digital media: digital watermarking and content based video copy detection. The primary task of any CBCD system is, to detect video copies by utilizing the content based features of the media [1]. The CBCD approaches are preferred compared to watermarking techniques [2], because of the following key features: i) The video signature generation will neither destroy nor damage video content, ii) CBCD techniques are more robust than fragile watermarking techniques, iii) Signature extraction can also be done after the distribution of digital media and iv) Capable of detecting copies, even if the original document is not watermarked. The key challenge of any CBCD system is, to provide accurate matching of a copy clip with its master clip. CBCD techniques can be roughly classified into global descriptors and local descriptors techniques. Global descriptors like Ordinal measure [3], Color histograms [4] are compact and easy to extract, but they are less robust against region based attacks. SIFT [5], SURF

*Corresponding author.Tel.: +0-080-23245586; fax:+0-080-28362393.
Email address: roopanagendran2002@gmail.com (R. Roopalakshmi)

Selection and/or peer-review under responsibility of Prof. Elhadi Shakshuki and Prof. Muhammad Younas.
doi:10.1016/j.procs.2011.07.021

Table 1: List of transformations used in proposed CBCD task

Category	Type	Description
Transformations-Level 1 (TL1)	T1: Brightness change	Increase brightness by 15% -25%
	T2: Noise Addition	Adding 15% random noise
	T3: Rotation	Rotating up to 90°
	T4: Blurring	Blurring by 20%
	T5: Horizontal flip	Horizontal mirroring up to 90°
	T6: Vertical flip	Vertical mirroring up to 100°
	T7: Color change	Changing color spectrum
	T8: Pattern insertion	Pattern is inserted into selective frames
	T9: Moving caption insertion	Entire video includes moving caption
	T10: Slow motion	Halve the video speed
	T11: Fast forward	Double the video speed
	T12: Zooming in	Zoom in by 15%
Transformations-Level 2 (TL2)	T13: Combination of 3 transformations of TL1	Applying 3 transformations amongst T1-T5
	T14: Combination of 5 transformations of TL1	Applying 5 transformations amongst T1-T4, T6-T8
	T15: Combination of 8 transformations of TL1	Applying 8 transformations amongst T1-T5, T7, T8, T10 and T12
	T16: Combination of 10 transformations of TL1	Applying 10 transformations amongst T1-T12

[6] , PCA-SIFT [7], and CS-LBP [8] are some of the popular local descriptors, which use local interest points for feature extraction. Kim et.al [9], proposed spatio- temporal feature descriptors for their copy detection task.

Visual words based feature descriptors are proposed by Poullot et.al [10], in order to detect pirated video contents. Hampapur and Bolle [11] made a comparative analysis of color histograms and edge-based methods for detecting video copies. Law-To et al. [12] performed a comparative study of various global and local descriptors. The local descriptors are more robust against region based transformations, but their computational cost is high compared to global descriptors. In [13], authors used facial shot matching, MPEG-7 descriptors and activity subsequence matching techniques for their copy detection task. Sarkar et al. [14] used MPEG-7 color layout descriptors and proposed a non-metric distance measure to search for duplicate videos in high-dimensional space. Chiu et al. [15] proposed a sliding window based time series linear search method for detecting video copies.

In general, audio content is a significant information source of any video sequence and in most of the CBCD cases, it is unaffected. Hence it is desirable to detect illegal videos using their audio features,even the visual content is badly distorted. The video transformations used in our CBCD task, is given in Table 1. Fig.1. illustrates all transformations with some example frames, extracted from the transformed query videos. The main contributions of this paper are as follows:

a) Novel copy detection method by exploiting audio fingerprints, compared to the visual content based state-of-the art techniques.
b) Construction of multi-feature vectors, by concatenating various spectral feature sequences.
c) Dimensionality reduction of multi-feature vectors using PCA.

The rest of this paper is organized as follows: Section 2 introduces framework of proposed scheme along with feature extraction and matching techniques; Section 3 shows the experimental setup and results of the proposed scheme, followed by conclusion in section 4.

2. Proposed framework

The block diagram of proposed copy detection framework is shown in Fig.2. and the relevant symbols are explained in Table 2. The proposed framework consists of two main components: Off-line (Master video processing)

Fig.1. Example frames from transformed query videos

stage and Online (query video processing) stage. In the off-line stage, audio based features are extracted from individual frames of master video files. These intra-frame features are concatenated into high-dimensional Multi- Feature (MF) vectors of predefined window size. Since MF vectors combine raw features (includes intra & inter frame features), they effectively represent frame-level and clip-level information of video files. Principal Component Analysis (PCA) is performed on high-dimensional MF vectors, in order to get compact & low dimensional representation. The sequence of principal components are subsequently combined and stored as fingerprints of video files. In the online stage, MF vectors are calculated, after extracting audio features from query frames. Then principal components of query video are calculated from MF vectors, and compared against the fingerprints of master video files. The L2 distance based comparison gives the output of proposed copy detection task.

2.1. Fingerprint extraction

The audio signal is down sampled to 22050 Hz, in order to reduce the size of data to be processed. In case of 10-30 ms of window length, the magnitude spectrum of audio signal is assumed to be stationary. Hence, the down sampled audio signal is segmented into 11.60 ms windows with an overlap factor of 86% using Hamming window function. The most important perceptual audio features exist in the frequency domain. Therefore spectral representation of each analysis window is computed by applying FFT (Fast-Fourier transform). From the spectral decomposition, two sets of features are extracted: Mel-Frequency Cepstral Coefficients (MFCC) and spectral distribution descriptors.

2.1.1. MFCC extraction

MFCC are dominantly used by the audio processing community to give good discriminative performance, with reasonable noise robustness [16], [17]. The MFCC are based on the discrete cosine transform of the log amplitude Mel-frequency spectrum. In the proposed scheme, FFT spectrum is divided into 24 bands and 40 triangular band pass filters are placed using Mel-scale. First 15 MFCC are calculated, to capture short term spectral features of video frames.

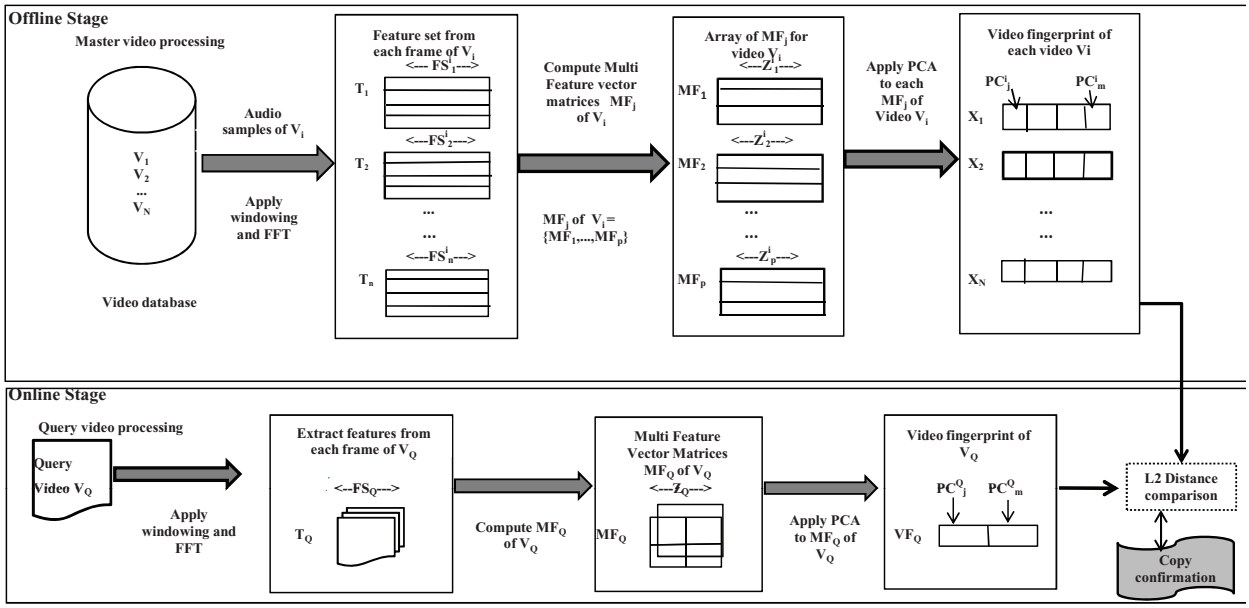

Fig.2.The framework of proposed scheme

2.1.2. Spectral distribution descriptors

Let $X_i(k)$ represents k-th FFT coefficient of i-th frame of length N, then spectral distribution descriptors can be calculated as follows:

Spectral Centroid

Spectral centroid is the center of gravity of the spectrum, which is a measure of spectral brightness [18]. The spectral centroid of i-th frame is given by,

$$Centroid(i) = \frac{\sum_{k=1}^{N} k * X_i(k)}{\sum_{k=1}^{N} X_i(k)} \tag{1}$$

Signal energy

This descriptor estimates the signal power at a given time [19], which is given by,

$$Energy(i) = \frac{1}{N} \sum_{k=1}^{N} |X_i(k)|^2 \tag{2}$$

Spectral Roll-off

The spectral roll-off, is the frequency below which 85% of the magnitude distribution is concentrated [20] and it measures the spectral shape. The roll-off of i-th frame is given by,

$$Roll - off(i) = 0.85 * \sum_{k=1}^{N} X_i(k) \tag{3}$$

Spectral Flux

Flux is the squared difference between the normalized magnitudes of successive spectral distributions. It measures

Table 2.Glossary of Notations

Notation	Definition	Notation	Definition
N	Number of master videos	T_Q	Number of frames of query video V_Q
V_i	i-th master video in the database	X_i	Audio fingerprint of i-th video V_i where $X_i = \{PC^i_j, ..., PC^i_m\}$
n	Total number of frames of video V_i	PC^i_j	j-th Principal component of V_i
T_i	i-th frame of video V_i, where $V_i = \{T_1, T_2, T_3, .., T_n\}$	FS_Q	Feature set extracted from Query video V_Q
FS^i_n	Feature Set extracted from n-th frame of video V_i	MF_Q	MF vector matrix of video V_Q
MF_j	j-th MF vector matrix of V_i,where $j = \{1, 2, .., p\}$	Z_Q	Dimension of MF_Q of V_Q
Z^i_j	Dimension of j-th MF vector matrix of video V_i	PC^Q_j	j-th principal component of video V_Q where $j = \{1, 2, .., m\}$

the amount of local spectral change [18]. Flux of i-th frame is given by,

$$Flux(i) = \sum_{k=1}^{N} |X_i(k) - X_{i-1}(k)| \tag{4}$$

The output of feature extraction process results in conversion of 11.60ms frames in to a stream of feature vectors with 6 feature values. The resulting feature sequences are concatenated into MF vectors of length 580ms. Since the dimension of MF vector is very high (50*50*6 = 15000) , it is not feasible to perform any computations. In order to convert MF vector into low dimensional compact vector, the following two techniques are used:

a) Instead of using all 15 MFCCs of frames, only MFCC means and variances are included in the feature set of frames.
b) Application of PCA to get principal components of MF vectors.

2.1.3. Principal Component Analysis

Given d-dimensional MF vectors MF_i ,such that i={1, 2, 3, ..., N}, the mean vector M [20]is given by,

$$M = \frac{1}{N} * \sum_{i=1}^{N} MF_i \tag{5}$$

The mean subtracted data set is given by $B = MF_i$ - M. The covariance matrix is given by,

$$Cov = \frac{1}{N-1} * B * B^T \tag{6}$$

where B^T represents transpose of B. Finally, the eigenvectors V and eigen values λ are calculated directly from the covariance matrix by solving the generalized eigenvector problem [20] for,

$$C.V = \lambda.V \tag{7}$$

In our experiments only K eigenvectors with largest eigen values are considered as fingerprints, where K varies between 2 to 8.

2.2. Fingerprint matching

In this proposed CBCD task, similarity matching is performed using weighted L2 Euclidean distance calculations. If P_1 and Q_1 are master and query video files, f_p and f_q are their corresponding video fingerprints. The components of f_p includes p_i eigenvectors and the corresponding λ_i eigen values. The query video fingerprint f_q contains q_j eigen vectors and corresponding σ_j eigen values. The similarity between p_i and q_j [21] is given by,

$$Dist\ (i, j) = |p_i - q_j|^2 \tag{8}$$

In general, eigen vectors with large eigen values specify most significant relationships between data dimensions. The inclusion of eigen values in similarity calculations improves the performance of CBCD system. Hence we have considered a weighting factor in our experiments, which is given by,

$$W(i,j) = \frac{1}{\sqrt{\lambda_i^2}} * \frac{1}{\sqrt{\sigma_i^2}} \tag{9}$$

The similarity between two video files S (P1,Q1), is defined as the weighted sum of similarity between their fingerprints, given by

$$S(P_1, Q_1) = \sum_{i=1}^{f_p} \sum_{j=1}^{f_q} W(i,j)\, Dist(i,j) \tag{10}$$

3. Experimental setup

3.1. Reference data set & query construction

The proposed CBCD system is evaluated on Sound & Vision data set used in TRECVID 2007 [22] tasks. The video database includes 25 hours of video covering a wide variety of content. The format of reference video clips is 352*288 pixels and 30 frames/ sec. In our experiments, seven video clips are selected from reference dataset. One video clip collected from Open Video Project [23] serves as non-reference video stream. The sixteen types of transformations listed in Table 1 are applied to the eight query video clips, and duration of these clips varies from 30 to 45 seconds. The resulting 128 (16*8) video sequences are used as queries for proposed CBCD task.

Table 3: Precision and Recall Rates for T1-T8 Transformations

Transformations		Cao's Method	Baseline Method	Proposed Method
T1	P	0.70943	0.72775	0.85714
	R	0.69604	0.81861	0.96428
T2	P	0.71621	0.82901	**1.00000**
	R	0.71542	0.80142	0.96825
T3	P	0.69654	0.83675	**1.00000**
	R	0.67554	0.78864	0.92461
T4	P	0.62910	0.71076	0.91541
	R	0.64839	0.67785	0.96923
T5	P	0.74472	0.88675	**1.00000**
	R	0.69843	0.73652	0.88405
T6	P	0.76871	0.81843	**1.00000**
	R	0.57983	0.54908	0.79602
T7	P	0.64911	0.71453	**1.00000**
	R	0.60152	0.62303	0.87341
T8	P	0.63301	0.78994	**1.00000**
	R	0.58973	0.61952	0.83554

3.2. Evaluation Criteria

To measure the detection accuracy of proposed scheme, we used standard metrics, which are given by,

$$Precision = TP/(TP + FP) \tag{11}$$

$$Recall = TP/(TP + FN) \tag{12}$$

True Positives (TP) are positive examples correctly labeled as positives. False Positives (FP) refer to negative examples incorrectly labeled as positives. False Negatives (FN) refer to positive examples incorrectly labeled as negatives.

We have compared the results of our method with Cao's method, stands for approach [24] and Baseline method. In Cao's method, authors have used mean of YCbCr components as the feature descriptors for their copy detection task. Baseline method uses only MFCC means and variances as feature descriptors. Table 3 lists the PR rates of baseline, Cao's and proposed methods for first eight transformations of type TL1.

For T8 transformation (Pattern insertion), Cao's Method gives poor recall rate (0.58973), when compared to that of proposed method (0.83544). The reason for the poor performance of Cao's method is, limited capability of global descriptors.The results from Table 3 shows that, the proposed method yields good precision rates compared to baseline method, especially for T8 and T2 transformations. For Flipping transformation (T6) baseline method gives poor recall rate (0.54908) compared to that of proposed method (0.79602). Hence, results from Table 3 proves that the proposed method yields better detection rates compared with that of Cao's method and baseline method.

The precision and recall rates of Cao's method, baseline and proposed methods for T9 -T16 transformations are shown in Table 4. Since TL2 transformations include multiple video editing tasks, the overall detection rates are slightly less compared to that of TL1 transformations. Although T16 transformation includes ten types of complicated video editing activities, still the proposed method manages to give better precision rates (0.90679), compared to that of Cao's and baseline methods. For T15 and T16 transformations,the detection rates of Cao's method is poor,because YCbCr values are significantly affected by combined visual distortions.

Table 4: Precision and Recall Rates for T9-T16 Transformations

Transformations		Cao's Method	Baseline Method	Proposed Method
T9	P	0.60812	0.73564	**1.00000**
	R	0.43867	0.61762	0.87342
T10	P	0.49972	0.54921	0.74332
	R	0.49889	0.63544	0.83747
T11	P	0.61367	0.65990	0.83875
	R	0.40175	0.59211	0.79002
T12	P	0.61832	0.72178	0.99642
	R	0.53761	0.65156	0.85714
T13	P	0.68120	0.78805	0.99218
	R	0.40961	0.52865	0.69543
T14	P	0.63592	0.79664	0.97564
	R	0.50183	0.65271	0.85285
T15	P	0.64883	0.78853	0.96605
	R	0.54241	0.64400	0.81824
T16	P	0.59971	0.69904	0.90679
	R	0.48762	0.52743	0.79775

4. Conclusion

In this article, a novel duplicate video detection method using audio fingerprints is proposed. The proposed algorithm includes two steps: First MFCC and four spectral descriptors are utilized to capture audio based features; next, multiple feature vectors are further processed using PCA in order to provide compact feature representation. The detection results demonstrate the efficiency of proposed method against different video editing and transformations.

Our future work will be targeted at,

i) Incorporation of visual, audio features to improve detection performance of proposed copy detection system.
ii) Using effective indexing methods, to enhance the detection accuracy of existing copy detection framework.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and suggestions.

References

[1] C. Y. Chiu, H. M. Wang, and C. S. Chen, Fast min-hashing indexing and robust spatio-temporal matching for detecting video copies, ACM Trans. Multimedia Comput. Commun. Applicat., vol. 6, no. 2, 123, (2010).
[2] R. Roopalakshmi and G. Ram Mohana Reddy, Recent Trends in Content Based video copy detection, in: Proceedings of IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), India,(2010), doi:10.1109/ICCIC.2010.5705802.
[3] Xian-Sheng Hua, Xian Chen , Hong-Jiang Zhang, Robust video signature based on ordinal measure, in: Procedings of IEEE International Conference on Image Processing (ICIP), vol. 1, pp. 685-688,(2004).
[4] H. T. Shen, X. Zhou, Z. Huang, J. Shao, and X. Zhou, UQLIPS: A real-time near-duplicate video clip detection system, in: Proceedings of VLDB, pp. 1374-1377, (2007).
[5] David G.Lowe, Distinctive image features from scale-invariant key points, International Journal of Computer Vision, 91–110, (2004).
[6] Herbert Bay, Tinne Tuytelaars, Luc Van Gool, SURF: Speeded Up Robust Features, Computer Vision and Image Understanding, 346-359, (2008).
[7] Y.Ke and R. Sukthankar, PCA-SIFT: A More Distinctive Representation for Local Image Descriptors, in: Proceedings of IEEE Computer Vision and Pattern Recognition conference (CVPR-04), pp. 506513, (2004).
[8] M. Heikkila, M. Pietikainen, and C. Schmid, Description of interest regions with local binary patterns, in: Proceedings of Pattern Recognition, vol. 42, no. 3, pp. 425-436, (2009).
[9] C. Kim and B. Vasudev, Spatiotemporal sequence matching for efficient video Copy detection, IEEE Trans. on Circuits and Systems for Video Technol. 15, 1,(2005).
[10] Poullot S., Crucianu M., and Buisson, Scalable mining of large video databases using copy detection, in: Proceedings of the ACM International Conference on Multimedia (MM),pp. 61-70, (2008).
[11] A. Hampapur, R. Bolle, Comparison of distance measures for video copy detection, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME-01), pp. 737-740,(2001).
[12] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, Video copy detection: a comparative study, in: Proceedings of sixth ACM International Conference on Image and Video Retrieval (CIVR-07), pp. 371-378, (2007).
[13] O. K*ücü*ktunc, M.Bastan, U. G*üdü*kbay, *Özgü*r Ulusoy, Video copy detection using multiple visual cues and MPEG-7 descriptors, Visual Communication and Image Representation , 21, 125-134, (2010).
[14] Anindya Sarkar, Vishwarkarma Singh, Pratim Ghosh , Bangalore S. Manjunath, and Ambuj Singh, Efficient and Robust Detection of Duplicate Videos in a Large Database, IEEE Trans. on Circuits and Systems for Video Technology, vol. 20, no. 6, (2010).
[15] Chih-Yi Chiu and Hsin-Min Wang ,Time-Series Linear Search for Video Copies Based on Compact Signature Manipulation and Containment Relation Modeling, IEEE Transactions on Circuits and Systems for Video Technology, vol.20, no.11,(2010).
[16] D. Nistr and H. Stewnius, Scalable recognition with a vocabulary tree, in: Proceedings of IEEE CVPR, 2161-2168, (2006).
[17] Pedro Cano, Eloi Batlle, Ton Kalker, Jaap Haitsma, A Review of Audio Fingerprinting, Journal of VLSI Signal Processing , 41, 271-284, (2005).
[18] A. Pikrakis, T. Giannakopoulos, and S. Theodoridis, An overview of speech/music discrimination techniques in the context of audio recordings, vol. 120, 81-102, (2008).
[19] Kris West, Novel techniques for Audio Music Classification and Search , Doctoral Thesis, (2008).
[20] Zak Burka,Perceptual Audio Classification Using Principal Component Analysis, M.S. Thesis, (2010).
[21] Jing Gu, Lie Lu, Rui Cai, Hong-Jiang Zhang, and Jian Yang, Dominant Feature Vectors Based Audio Similarity Measure, in: Proceedings of PacificRim conference on Multimedia(PCM),pp.890–897, Tokyo,Japan,(2004).
[22] TRECVID 2010 Guidelines [Online]. Available: http://www.nlpir.nist.gov/projects/tv2010/tv2010.html.
[23] Open Video Project , www.open-video.org
[24] Zheng Cao, and Ming Zhu, An Efficient Video Copy Detection Method Based on Video Signature, in: Proceedings of International Conference on Automation and Logistics , Shenyang, China ,(2009).