Towards a New Approach to Video Copy Detection Using Acoustic Features

R. Roopalakshmi
Information Technology Department
National Institute of Technology Karnataka (NITK)
Surathkal, Mangalore, India - 575025
Email: roopanagendran2002@gmail.com

G. Ram Mohana Reddy
Information Technology Department
National Institute of Technology Karnataka (NITK)
Surathkal, Mangalore, India - 575025
Email: profgrmreddy@gmail.com

***Abstract*—Acoustic features are robust and powerful in video description, but not fully exploited for the emerging Content-Based video Copy Detection (CBCD) methods. To solve this discrepancy, this paper proposes a new CBCD approach using audio spectral features compared to existing visual content based methods. The proposed method incorporates three stages: 1) Extraction of spectral descriptors including centroid and energy; 2) Integration of resultant features to compute highly informative spectral descriptive words; 3) Utilization of clustering approach to speed up the similarity matching process. The results tested on TRECVID-2008 dataset, demonstrate the improved detection accuracy of proposed method (up to 27.845%) compared to reference methods against various transformations such as fast forward, slow motion, mp3 compression, and multiband companding.**

***Index Terms*—Content based video copy detection, spectral centroid, signal energy, spectral roll-off, spectral flux.**

I. Introduction

The massive growth of media streaming activities have increased the amount of duplicate videos and resulted in huge piracy issues. Hence, copy detection is compulsory to reduce the copyright violations.

In general, a video copy is defined as, a transformed video sequence, derived from a master video [1]. There are two standard approaches for detecting copies of a digital media: digital watermarking and content based copy detection [2]. The primary task of any CBCD system is to detect video copies by utilizing content based features of the media [3]. The CBCD approaches are preferred compared to watermarking techniques 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) Can detect copies, even if the original document is not watermarked.

In CBCD literature, the existing techniques are based on global and local features. Global features such as Ordinal measure [4], Color histograms [5] are compact and easy to extract, but they are less robust against the region based attacks such as cropping. SIFT [6] and SURF [7] are some of the popular local descriptors, which use interest points for feature extraction. Local features are more robust against region based transformations, but their computational cost is high,when compared to global features. Itoh et al. [8] used average power of audio signals as feature descriptors for their copy detection task. Although this method is fast, the computation time required is more, which may degrade the performance of CBCD system. In [9] authors used both visual and audio features for detecting duplicate videos, but the performance of this method is limited to global transformations.

Since audio content is a significant information source of video sequence, they are widely used in video parsing, indexing and scene categorization approaches [10], [11]. Also, past acoustic investigations [12], [13] prove that, the most important perceptual audio features exist in the frequency domain. Hence, the main objective of this article is to show that, the audio spectral features are robust descriptors and can be efficiently utilized for the copy detection task. If these audio fingerprints are integrated with other visual features, a completely robust CBCD system can be developed. The main contributions of this paper are as follows:

a) New copy detection method using audio features, instead of state-of-the art visual content based methods.
b) Computation of compact spectral descriptive words, by combining robust features such as centroid, signal energy, roll-off and flux.
c) Clustering based pruned searching rather than direct searching of video signatures.

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

II. Proposed Scheme

Fig. 1 shows the block diagram of the proposed copy detection framework. The framework consists of two main stages: Master video processing stage (off-line) and Query video processing stage (on-line). In the off-line stage, spectral descriptors including centroid, energy, roll-off and flux are extracted from master video frames. These features are further processed and SPectral Descriptive (SPD) words are computed. SPD words combine raw spectral features, thus they summarize the overall audio profile of a given video sequence. K-means clustering approach is utilized, in order

978-1-4577-1328-6/11/$26.00 2011 IEEE

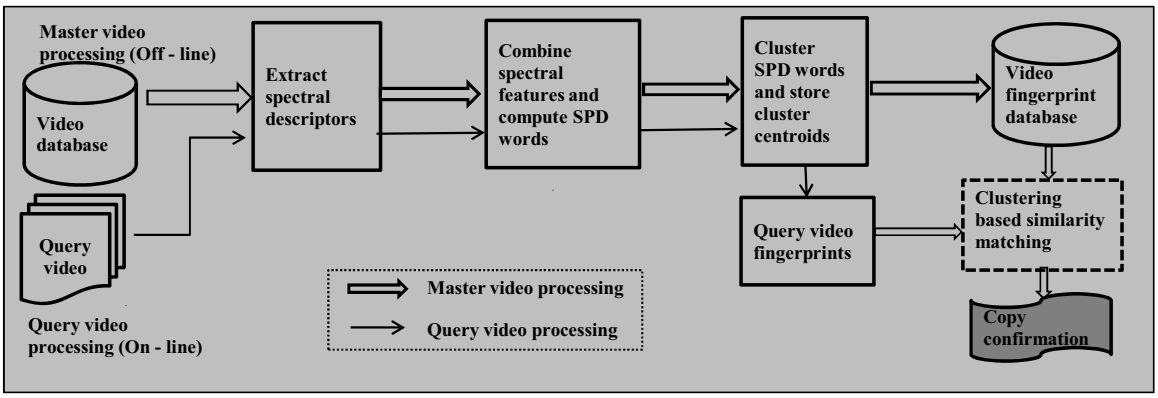

Fig. 1. Framework of the the proposed CBCD system

to get low-dimensional representation of SPD words and the cluster centroids are stored as video fingerprints of master video files.

In the on-line stage, SPD words are calculated, after extracting spectral descriptors from query video frames. The resulting SPD words are clustered and cluster centroids are stored as video fingerprints. Finally clustering based similarity matching is performed for detecting video copies.Table I lists the audio and visual transformations used in the proposed CBCD task.

TABLE I
LIST OF TRANSFORMATIONS USED IN THE CBCD TASK

Type	Transformations
T1	Blurring
T2	Color change
T3	Slow motion
T4	Fast forward
T5	Pattern insertion
T6	Moving caption insertion
T7	Cropping
T8	Picture-inside-picture
T9	Mp3 compression
T10	singleband companding
T11	Multiband companding
T12	Combination of 3 transformations (Cropping, pattern insertion and mp3 compression)

A. Spectral Descriptors Extraction

In order to reduce the size of data to be processed, first audio signal is down sampled to 22050 Hz. The magnitude spectrum of audio signal behaves almost stationary for 10-30 ms of window length. Hence the down sampled audio signal is segmented into 11.60 ms windows using Hamming window function with an overlapping factor of 86% [14]. Then the spectral descriptors such as centroid, energy, roll-off and flux are extracted from the short term power spectrum of audio signals.

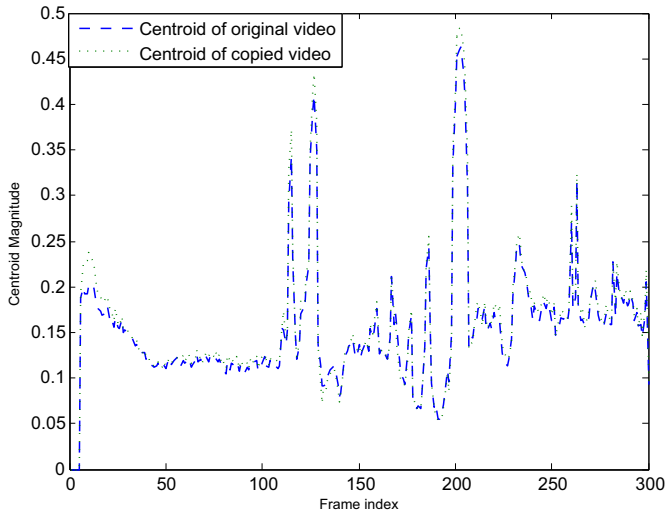

Fig. 2. Similarity of spectral centroid plots of original and copied videos

1) Spectral Centroid (SC): This centroid is a timbral feature, which describes the brightness of a sound signal [15]. In general, sound with brighter quality contains more amount of high frequency components, compared to sound with dark quality. In the literature of sound synthesis techniques, spectral centroid feature is proved to be an important descriptor [13], [15], that specifies the center of gravity of the signal spectrum. The centroid is computed as [15],

$$SC = \frac{\sum_{k=1}^{N-1} k * x^d[k]}{\sum_{k=1}^{N-1} x^d[k]} \quad (1)$$

Where $x^d[k]$ represents the magnitude of k-th frequency bin and N is the frame length.

The statistical properties of spectral centroid such as mean, standard deviation and log amplitude are used in various speech analysis and recognition algorithms [15], [16]. We

used average frequency distribution values as centroids for the proposed CBCD task. Fig. 2 shows an example spectral centroid plot of original and copied video files. Here the copied video is created by applying T12 type of transformation. The centroid plots indicate a very high similarity (up to 98.7%) between original and copied video files, thus prove the robust nature of this spectral descriptor used in the proposed CBCD task.

2) Spectral Energy: This descriptor measures average short term power of the input signal [16]. In this proposed work, the sum of squared magnitude of samples is utilized to calculate the spectral energy of signal. Fig. 3 shows an example spectral energy plot of original and copied video files, that indicates a very high similarity (up to 95.8%) between the two video features. Here the copied video is created by applying T12 type of transformation.

3) Spectral Roll-off (SR): This feature is commonly referred as skew present in the shape of power spectrum. This roll-off point defines the frequency boundary, where 85% of the total energy of power spectrum resides. This descriptor is widely used to differentiate constant and highly transient sounds [17]. The spectral roll-off can be calculated as [15],

$$SR = \sum_{k=0}^{R} x^d[k] = 0.85 \sum_{k=0}^{N-1} x^d[k] \tag{2}$$

where N is frame length and $x^d[k]$ indicate magnitude components of k-th frequency bin and R is the frequency roll-off point with 85% of energy.

4) Spectral Flux (SF): Generally, speech signals change at a faster rate, compared to music signals [15]. Spectral flux defines the amount of energy difference between consecutive analysis frames [17], which is extracted as follows,

$$SF = \mid x_f^d[*] - x_{f-1}^d[*] \mid \tag{3}$$

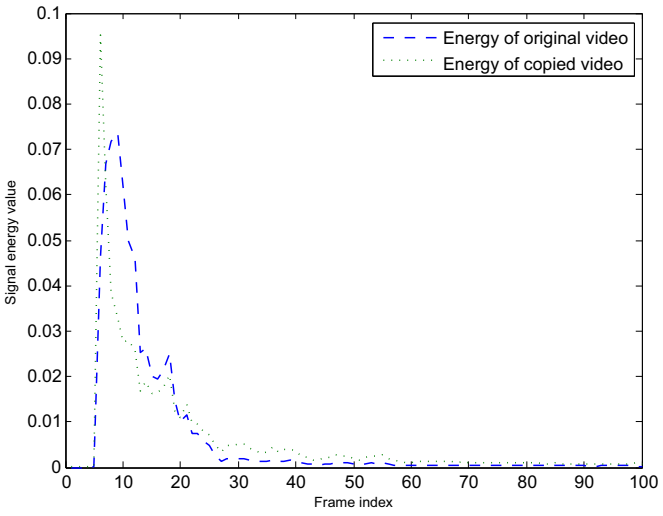

Fig. 3. Similarity of signal energy plots of original and copied videos

where $x^d[*]$ represents magnitude of frequency components and $f, f-1$ indicate current and previous frames respectively. This flux is mainly used to compare speech and musical signals.

The dimension of extracted spectral features is high (in terms of 10240/sec) and hence direct processing of raw features is computationally expensive. The resultant feature descriptors are combined into highly informative SPD words. K-means clustering is used to get the compact representation of SPD words. In experiments, it is observed that the number of clusters for video files range from 47-413 based upon individual video contents.

B. Fingerprint Matching

In the proposed copy detection system, L1-norm Manhattan distance is used to compute the similarity between two video clips. If M_k is k-th master video and Q is query video clip, then f_m and f_q are their corresponding video fingerprints. The similarity score (Sim) between M_k and Q is computed as,

$$Sim(M_k, Q) = \sum_{i=1}^{m} \sum_{j=1}^{n} abs(f_m(i) - f_q(j)) \tag{4}$$

where m and n indicate the size of master and query video signatures. The Sim scores are compared against predefined confidence measure and copy detection results are reported. We experimented various confidence measures ranging from 0.55-0.73 and the results lead us to conclude 0.70 confidence measure provides better accuracy for the proposed copy detection task.

III. EXPERIMENTAL RESULTS

A. Reference Database & Query Construction

We used TRECVID-2008 Sound & Vision data set [18] for evaluating the proposed method. The video database includes 75 hours of video covering a wide variety of content. In our experiments, seven video clips are selected from reference data set. Two video clips from Open Video Project [19] are used as non-reference video streams. The transformations listed in Table I are applied to the nine query video clips, and duration of these clips vary from 20 to 25 seconds. The resulting 108 (12×9) video sequences are used as query video clips for the proposed copy detection task.

B. Evaluation Metric

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

$$Precision = TP/(TP + FP) \tag{5}$$

$$Recall = TP/(TP + FN) \tag{6}$$

$$F - Measure(FM) = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

TP (True Positives) are positive examples, FP (False Positives) are negative examples and FN (False Negatives) are incorrectly labeled positive examples. F-Measure indicates the robustness and discrimination ability of a system.

TABLE II
DETECTION RESULTS FOR T1-T6 TRANSFORMATIONS

Transformations		Ordinal	Itoh's	Proposed
Type	**Metric**	**Measure [4]** (%)	**Method [8]** (%)	**Method** (%)
T1	P	76.24	79.09	**100.00**
	R	70.58	69.08	79.48
	F-M	73.30	73.74	88.56
T2	P	65.35	81.57	97.89
	R	79.14	78.34	96.37
	F-M	71.58	79.92	97.12
T3	P	61.22	71.58	99.39
	R	58.45	61.97	99.40
	F-M	59.80	66.42	99.39
T4	P	74.29	79.59	99.69
	R	70.11	70.16	**100.00**
	F-M	72.13	74.57	99.84
T5	P	68.36	81.62	99.09
	R	69.16	79.31	98.86
	F-M	64.07	80.44	98.13
T6	P	59.69	80.66	97.42
	R	69.16	74.31	98.86
	F-M	64.07	77.35	98.13

C. Copy Detection Results

We have compared the proposed detection method with Ordinal measure [4] and Itoh's [8] methods.The ordinal measure is extracted as follows: partitioning the image into N blocks; sorting the blocks according to their average intensity values and the ranking order of blocks are considered as ordinal signatures.

Itoh's method [8] uses significant points in acoustic data for detecting illegal videos, which is implemented as follows: first acoustical power envelopes of input signal are computed; from the power envelopes, the significant points denoting local minimum/maximum values are extracted and used for the copy detection task.

Table II lists the detection results of proposed and reference methods for T1-T6 transformations. The results from Table II demonstrate that, the proposed method improves detection performance by 29.78%, when compared with the reference methods.

For T3 (slow motion) transformation, Ordinal measure gives poor recall rate (58.45%), when compared to that of Itoh's method (61.97%). The reason for this poor performance is its global descriptive nature. Although Itoh's method performs better than ordinal measure for T1-T6 transformations, still the proposed method outperforms Itoh's method for all six transformations. The robust nature of spectral features considered is the reason for the better results of the proposed method. The proposed method achieves better recall rate (100%), when compared to that of ordinal measure (70.11%) and Itoh's method (70.16%) for T4 (fast forward) transformation.

Table III lists the detection rates of the proposed and reference methods for T7-T12 transformations. The results from Table III demonstrate that, the proposed method improves the detection accuracy by 25.91%, when compared with the reference methods.

TABLE III
DETECTION RESULTS FOR T7-T12 TRANSFORMATIONS

Transformations		Ordinal	Itoh's	Proposed
Type	**Metric**	**Measure [4]** (%)	**Method [8]** (%)	**Method** (%)
T7	P	74.24	85.61	99.00
	R	68.86	80.25	92.66
	F-M	71.44	82.84	95.72
T8	P	72.69	88.19	94.44
	R	71.10	80.27	90.26
	F-M	71.88	84.04	92.30
T9	P	73.65	60.24	97.22
	R	72.58	62.33	97.29
	F-M	73.11	61.27	97.25
T10	P	80.06	74.31	93.44
	R	73.64	72.59	90.28
	F-M	76.71	73.44	91.83
T11	P	66.28	69.34	90.36
	R	61.15	60.22	92.22
	F-M	63.61	64.46	91.28
T12	P	58.33	68.39	95.96
	R	51.29	65.11	93.21
	F-M	54.58	66.71	94.57

For T12 (3 combined transformations), Ordinal measure results in very poor precision, recall rates (58.33% and 51.29%), compared to that of Itoh's (68.39% and 65.11%) and proposed methods (95.96% and 93.21%). Ordinal signature is less robust against region based transformations, hence it yields poor results for T12. The detection scores of proposed method is slightly less for T10-T12 transformations, because spectral features are much affected by these three transformations. Still, by integrating four robust spectral features for copy detection task, the proposed method yields improved performance compared to reference methods against combined transformations.

IV. CONCLUSION

This article highlights a novel copy detection method by utilizing audio spectral features. The spectral features are combined and clustered in order to provide compact feature description. The results prove that, the proposed copy detection method improves the detection accuracy by 27.85% when compared to the reference methods. The detection results also demonstrate the effectiveness of the proposed method against various video and audio transformations. The proposed method

is useful in applications such as digital content management and copyright protection.

Our future work will focus on how to improve robustness of proposed method in transformations such as camcording, mix with speech and very complex ones. For camcording transformations, if the proposed audio features are combined with color and motion features, then the robustness of proposed copy detection system can be improved.

ACKNOWLEDGMENT

The authors would like to thank reviewers for their valuable comments and suggestions, that improved the quality of this article.

This research work is supported by Department of Science & Technology of Government of India under research grant no. SR/WOS-A/ET-48/2010.

REFERENCES

[1] Anindya Sarkar, Vishwarkarma Singh, Pratim Ghosh, Bangalore S. Manjunath, and Ambuj Singh, "Efficient and Robust Detection of Duplicate Videos in a Large Database", IEEE Transactions on Circuits and Systems for Video Technology, vol. 20, no. 6, 2010.

[2] Chih-Yi Chiu and Hsin-MinWang, "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.

[3] Roopalakshmi.R and Ram Mohana Reddy.G, "Efficient Video Copy Detection using Simple and Effective Extraction of Color Features", in proc.of ACC-2011 in Springer-Verlag, Part IV, CCIS 193, pp. 473-480, 2011.

[4] Xian-Sheng Hua, Xian Chen and Hong-Jiang Zhang,"Robust Video Signature based on Ordinal Measure", in proc. of IEEE International Conference on Image Processing (ICIP), vol. 1, pp. 685-688, 2004.

[5] H. T. Shen, X. Zhou, Z. Huang, J. Shao, and X. Zhou,"UQLIPS: A Real-time Near-Duplicate Video Clip Detection System", in proc. of VLDB, pp. 1374-1377, 2007.

[6] David G.Lowe,"Distinctive Image Features from Scale-Invariant Key points", International Journal of Computer Vision, 91-110, 2004.

[7] Herbert Bay, Tinne Tuytelaars and Luc Van Gool, "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding, 346- 359, 2008.

[8] Yoshiaki Itoh, Masahiro Erokuumae, Kazunori Kojima, Masaaki Ishigame and Kazuyo Tanaka, "Time-space Acoustical Feature for Fast Video Copy Detection", in proc. of 2010 IEEE International Workshop on Multimedia Signal Processing (MMSP), 2010.

[9] Ahmet Saracoğlu, Ersin Esen, Tuğrul K. Ateş, Banu Oskay Acar, Zubari, Ezgi C. Ozan, Egemen özalp, A. Aydin Alatan, Tolga Çiloglu, "Content Based Copy Detection with Coarse Audio-Visual Fingerprints," 2009 Seventh International Workshop on Content-Based Multimedia Indexing (cbmi), pp.213-218, 2009.

[10] Sofia Tsekeridou,and Ioannis Pitas,"Content-Based Video Parsing and Indexing Based on AudioVisual Interaction", IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 4, April 2001.

[11] Songhao Zhu, Junchi Yan and Yuncai Liu, "Improving Semantic Scene Categorization by Exploiting Audio-Visual Features",in proc. of 2009 Fifth International Conference on Image and Graphics, 2009.

[12] Tang Jie, Liu Gang, and Guo Jun, "Improved Algorithms of Music Information Retrieval based on Audio Fingerprint", in proc. of Third International Symposium on Intelligent Information Technology Application Workshops, 2009.

[13] Tao Li,Mitsunori Ogihara and Qi Li, "A Comparative Study on Content-Based Music Genre Classification", in proc. of SIGIR-03, Toronto, Canada, 2003.

[14] R.Roopalakshmi and G.Ram Mohana Reddy, "A Novel Approach to Video Copy Detection Using Audio Fingerprints and PCA", in proc. of ANT-2011 in Elsevier Procedia Computer Science Journal, 2011. doi:10.1016/j.procs.2011.07.021

[15] Tae Hong Park, "Introduction to digital signal processing- Computer musically speaking", World scientific Press, 2010.

[16] Kris West, "Novel techniques for Audio Music Classification and Search", Doctoral Thesis, 2008.

[17] Zak Burka, "Perceptual Audio Classification Using Principal Component Analysis", M.S. Thesis, 2010.

[18] TRECVID 2010 Guidelines [Online]. Available: http://www.nlpir.nist.gov/projects/tv2010/tv2010.html

[19] Open Video Project, www.open-video.org