Recent Trends in Content-Based Video Copy Detection

R.Roopalakshmi[1], G. Ram Mohana Reddy[2]
[1,2] Information Technology Department, National Institute of Technology (NITK),
Mangalore , Karnataka ,India.
([1]roopanagendran2002@gmail.com, [2]profgrmreddy@gmail.com)

Abstract - **Video copy detection is an interesting & challenging problem, as it is becoming increasingly important with the growing rate of illegal digital media and huge piracy issues. Therefore, in the present era of Internet &Multimedia technologies,the existence of ubiquitous digital videos, has led to the requirement of robust video copy detection techniques, for content management and copyright protection. In this paper, an overview of content based video copy detection (CBVCD) is presented along with the different state-of-the-art techniques used for video feature/ signature description. This article also explores the future key directions and highlights the research challenges that need to be addressed in the CBVCD paradigm.**

Keywords – Video copy detection, global descriptors, local descriptors.

I. INTRODUCTION

The exponential growth of multimedia technologies and media streaming have increased video publishing and sharing activities tremendously. Because of this massive media consumption, there exist enormous amount of duplicate videos, which leads to huge piracy issues. Controlling the copyright of the huge number of videos uploaded everyday is a critical challenge for the owner of the popular video web servers. For example, latest survey says that users upload 65,000 new videos each day on video sharing websites like YouTube and also on an average, a viewer watches more than 70 videos online in a month [1] and this number is expected to keep growing.

In general, a copy of a video is a transformed video sequence, which is visually less similar and does not contain any new and important information, compared to the source video. There are two general approaches for detecting copies of a digital media: digital watermarking and content based video copy detection. Watermarking embeds information into the media prior to distribution. Thus all copies of the marked content contain the watermark that can be extracted, to prove ownership of the material. CBVCD is a complementary approach to watermarking. The primary idea of any CBVCD technique is, '*the media itself is the watermark',* [2], the media contains enough unique information that can be used for detecting video copies. The CBVCD approach is preferred when compared to digital watermarking, because of its following key features :

- The video signature/fingerprint generation will neither destroy nor damage video content.
- CBVCD approaches are more robust than fragile watermarking techniques.
- In CBVCD schemes, the signature extraction can be done after the distribution of digital media, which is not possible in case of watermarking, because the media gets distorted after the distribution.
- CBVCD schemes are capable of detecting copies even if the original document is not watermarked.

This paper is organized as follows: in Section II, an overview of CBVCD framework is presented; in Section III, it explores various feature description methods, including global & local descriptors of CBVCD paradigm. Finally, in Section IV, this article highlights the future research challenges & key issues in the CBVCD domain.

II. OVERVIEW OF CBVCD FRAMEWORK

The overview of CBVCD framework, shown in Fig.1, includes the following modules:

Video Database: This is the reference video database that includes all the video files.

Key Frame extraction/ Frame Sampling: This module extracts key or representative frames of a given video. To extract key frames, different methods like, shot-based key frame extraction, frame sampling, sliding window approach etc., are proposed. Shot-based key frame extraction is a well-known technique for content-based video searches. In shot-based schemes, first video is segmented into video shots and each shot is summarized using several key frames. Liu et al. [3] extracted key frames by finding the peaks of the motion energy in the video sequence. The key frame extraction methods are, limited to the comparisons of whole clips, and also they cannot detect similar subsequences. In frame sampling schemes, the total number of content-representative frames are sampled using a predefined sampling rate[4]. Video searching by a sliding window is a very popular method [5] due to its simple and effective computation. But, the fixed-length sliding window schemes cannot handle temporal transformations.

Feature Descriptor Extraction: This module extracts the feature descriptors of representative frames of the given video. The CBVCD schemes use perceptual features of the digital content to generate distinct video signatures.

Signature Database & Index: This database includes all the video signatures extracted from the reference videos

978-1-4244-5967-4/10/$26.00 ©2010 IEEE

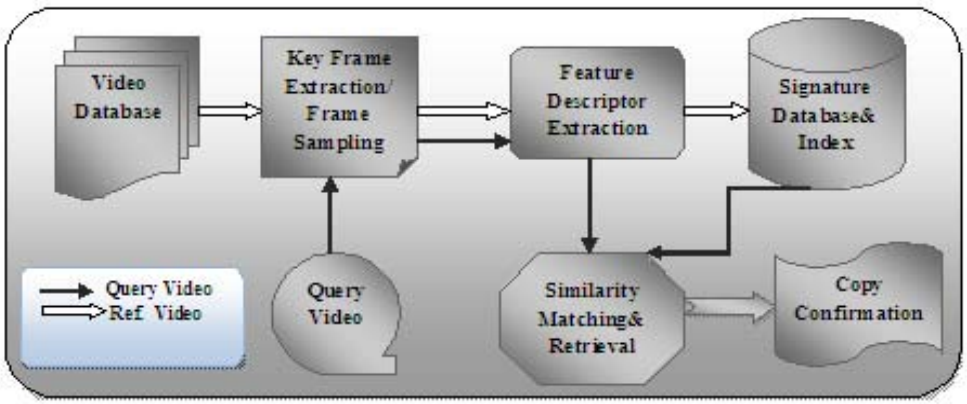

Fig.1. Overview of CBVCD Framework

and the signatures are indexed so that fast processing can be achieved.

Similarity Matching & Retrieval: This module calculates the distance between the video signatures of reference videos and query video. Based on the results, whether the query video is a copy or not is identified.

In general, CBVCD schemes include two basic tasks, implemented in the following two steps:

a. Signature Extraction – In this step, the video signatures of original video & test video are extracted.
b. Similarity Measurement – In this step, the distance between the original & query video fingerprints are compared, in order to determine if the test video is a copy of the reference video or not.

III. FEATURE DESCRIPTOR TECHNIQUES

For a successful CBVCD system, the design of an effective descriptor is the key element. A descriptor is said to be effective, if it satisfies the following properties: *robustness, discrimination ability*, and *repeatability* [6], because a robust descriptor is invariant to patterns generated by the same source, while a discriminative descriptor is sensitive to patterns belonging to different video sources. The repeatability of a descriptor specifies whether it reliably finds the same interest points for a given image under different viewing conditions. In addition to that, a descriptor must be compact and computationally efficient so that it is suitable for handling a huge amount of data streams.

Generally, feature descriptors are classified into two broad categories, Global & Local descriptors. Global descriptors summarize the global statistics of low-level features. Local descriptors derive low level features at the segment or shot level to facilitate local matching. Local features can be extracted after segmenting an image into regions and computing a set of local feature points. The basic concepts of different global & local descriptor techniques are given below:

A. Global Descriptor Techniques

Ordinal Measure

Ordinal measure [7] is a global descriptor widely used in video copy detection. To extract the ordinal measure each video frame is partitioned into non overlapping blocks and the blocks are ranked according to their average intensity values. The ranking order of a block is known as the frame's ordinal measure.

The Ordinal signature S(t) of a frame at time t is given by a vector of integers r_i ,as follows,

$$S(t) = (r_1, r_2, r_3 ..... r_n), \quad (1)$$

where r_i is the rank of block i and n is the number of blocks.

Color –shift & Centroid based Descriptors

Hoad and Zobel [8] proposed a compact video representation using Color-shift and Centroid-based signatures. These signatures are suitable for short queries and clips. The Color-shift method uses color distributions in the video frames to produce video signatures. Generally, color distributions represent the change in color in the clip over time. The Centroid-based signature represents the spatial movement of the lightest and darkest pixels in each frame over time. In order to produce the Centroid-based signature, two motion vectors are calculated for each frame: one for the darkest pixels and the other for lightest pixels. The centroid of each of

these collections of pixels is determined.

Color- Histograms

This descriptor was proposed by Naphade et al.[9]. They used YUV histograms for feature representation. First, Color histograms with 32 luminance bins and 16 bins for each chrominance channel are computed for each frame of the query clip and the target clip. Then a sliding window is used to compute a similarity measurement in the target clip.

Hauptmann et al. [10] used HSV color space for histogram calculation. In their approach, a color histogram is calculated for each key frame of the video, which is represented as:

$$H_i = (h_1, h_2, \ldots, h_m), \quad (2)$$

where H_i indicates histogram of i^{th} frame ,m indicates total number of bins, in the histogram. They have used a color histogram which is concatenated with 18 bins for Hue, 3 bins for Saturation, and 3 bins for Value, hence totally $m = 24$.

Table1, summarizes the key features of global descriptor techniques. The global descriptors are robust against spatial transformations and whole region based transformations. But their performance is poor for partial region based spatial transformations like zooming, cropping etc.,

B. Local Descriptor Techniques

SIFT (Scale Invariant Feature Transform)

The SIFT descriptor [11] is a 3D histogram of gradient locations and orientations. This descriptor selects highly distinctive image features that are invariant to several transformations like scaling, rotation, Illumination, 3D camera viewpoint, Clutter / noise, and Occlusion. First step in computing SIFT descriptor is detecting scale-space extrema using Difference-of-Gaussian(DoG) function. Once the key points are detected,then localization & orientation assignment is done and finally keypoint descriptors are constructed using the image gradients. The contribution to the location and orientation bins is weighted by the gradient magnitude and a Gaussian window overlaid over the region.The SIFT descriptor produces high dimensional description (128-D) for each local feature point ,which results in huge computations.

SURF (Speeded –Up Robust Features)

SURF descriptor [12] is a scale and rotation-invariant interest point descriptor. It describes distribution of Haar-wavelet responses within the neighborhood of interest points. Gerhard Roth et.al [13] proposed a local feature

Table1 .Salient Features of Global Descriptors

S. No.	Name of Feature Descriptor	Significant Features of the Descriptor
1.	Ordinal Measure	▪ Only 9-D vector for 3*3 block. ▪ Compact representation. ▪ Less sensitive to transforms like, brightness adjustment, histogram equalization, and frame resizing. ▪ Provides better results, when it is combined with temporal characteristics. ▪ Produces 2-D vector.
2.	Color -shift & Centroid based	▪ Compact than Ordinal measure. ▪ Color- shift descriptor is less effective for black &white content. ▪ Centroid-based descriptor gives poor performance, in case of pixel luminance degradations.
3.	Color - Histograms	▪ Compact& easy to compute. ▪ This descriptor is highly susceptible to the global variations in color.

count based approach using SURF descriptor. The SURF descriptor is quite resistant to image resizing, image deterioration and coding artifacts. But this descriptor gives poor performance, when it is dealing with videos with different frame rates.

Spatio –Temporal Features

Several CBVCD techniques are proposed [14], [15] which use spatio – temporal features. Spatio –temporal descriptors perform good localization of the features, both spatially as well as temporally, by ensuring high discriminativity and robustness. In this method, features are extracted at interesting locations by considering both spatial and temporal parameters.

PCA –SIFT (Principal Components Analysis – SIFT)

PCA –SIFT [16] descriptor is an extension of SIFT descriptor, which applies PCA to the normalized gradient patch of images, after computing Eigen space values. This descriptor is significantly smaller than the standard SIFT descriptor.But this descriptor suffers due to its shortcomings such as its implicit assumption of Gaussian

distributions and its restriction to orthogonal linear combinations.

CS -LBP (center-symmetric local binary pattern)

CS -LBP [17] descriptor is a modified version of the well-known local binary pattern (LBP) feature, which combines the strengths of the SIFT and LBP algorithms. CS-LBP is a powerful illumination invariant texture primitive, that makes use of histogram of binary patterns, computed over a region for texture description. For a neighborhood of 8 pixels, this CS-LBP produces only 2^4 binary patterns, so it is computationally simpler & compact feature descriptor, when compared with SIFT descriptor.

Gradient location-orientation histogram (GLOH)

Mikolajczyk et.al proposed a new descriptor GLOH [18] which extends SIFT by changing the location grid and it uses PCA to reduce the size. It is designed to increase the robustness and distinctiveness of SIFT descriptor.

IV. FUTURE CHALLENGES

In general, video can be viewed as collection of multi-modality of features like visual, audio, motion and textual information. Most of the currently available CBVCD techniques make use of visual content based features for detecting duplicates. Instead of that, if we use the video content information in intuitive & natural way, then obviously performance of CBVCD systems can be improved. By keeping the above factors in mind, this paper highlights few key directions for future research in context of CBVCD domain.

A. Incorporating Visual Cues

Most of the existing CBVCD systems concentrate on extracting low level features, which describe image content. But there is a semantic gap between user's high level representation of images & low level feature representation of images. In order to handle this disparity, if visual cues are incorporated into the CBVCD systems, then the performance can be enhanced.

B. Utilization of Audio Fingerprints

Generally, CBVCD systems, employ visual content of the video, but audio content of a video constitutes an indispensable information source. In most of the copy/ duplication tasks, audio content is not much manipulated, compared to the visual content. So, if we exploit audio signatures (or) combine audio & visual fingerprints for copy detection task, then detection accuracy of CBVCD systems can be enormously improved.

C. Combining Content & Context

In a given CBVCD approach, when a query video is presented, we need to search a huge database of videos. During video content analysis, if we incorporate contextual information about the video, then copy detection process can be implemented at a faster rate, which enhances the performance of the CBVCD system.

D. Detection of Near-duplicate Videos

A near-duplicate video is highly similar in content, but it appears differently, due to different external distortions like acquisitions, transformations and editions[19]. Most of the research in the CBVCD domain, concentrated only on copy detection techniques. Since video copies are subset of near-duplicate videos, and near-duplicate video detection schemes are introduced recently ,this topic remains still challenging.

E. Online Detection Process

Concurrently monitoring multiple short queries over a continuous video stream demands high accuracy. So, online copy as well as near-duplicate detection of continuously monitored video streams remains challenging issue and needs to be addressed.

F. Feature Representation using MPEG -7 Descriptors

The crucial factors, that affect the performance of CBVCD systems are, the time taken to compute feature descriptors & storage required for feature descriptors. In order to achieve fast processing of feature descriptions, we can make use of compressed domain features like MPEG -7 Descriptors[20]. Because feature extraction can be done effectively & efficiently using MPEG-7 descriptors, which is still unexplored topic in context of CBVCD domain.

G. Complicated Image Transformations

In CBVCD paradigm ,for most of the image transformations like rotation, scaling, illumination etc., a large number of solutions are proposed . But still high level of post processing actions like gamma-correction, cropping, transcending etc., poses specific challenges to the problem of content-based video copy detection.

V. CONCLUSION

This paper explores various state-of-the-art techniques used for feature description in CBVCD systems. This present work also highlights the key research challenges which need to be addressed in the context of CBVCD domain.

ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for their valuable comments and suggestions , which enhanced the quality of this article.

REFERENCES

1) Xiao Wu,Chong-Wah Hgo,Alexander G.Hauptmann ,Hung-Khoon Tan, "Real –Time Near Duplicate Elimination for Web Video Search with Content and Context," in proc. of *IEEE Transactions on Multimedia*, vol.11,no.2 ,Feb 2009**.**
2) Hampapur, A. and Bolle, R.M.,"Comparison of distance measures for video copy detection," in Proc. of *IEEE International Conference on Multimedia and Expo (ICME)*. 737–740,2001.
3) Liu, T., Zhang, H. J., and Qi, F., "A novel video key-frame extraction algorithm based on perceived motion energy model," in proc. of *IEEE Trans. Circuits Systems. Video Technology*. Vol .13,1006–1013, 2003.
4) Matthijs Douze, Hervé Jégou, and Cordelia Schmid, "An Image-Based Approach to Video Copy Detection With Spatio-Temporal Post-Filtering," in proc. of *IEEE Transactions on Multimedia*, Vol. 12, No. 4, June 2010.
5) Hua, X. S., Chen, X., and Zhang, H. J., " Robust video signature based on ordinal measure," in Proc. of *IEEE International Conference on Image Processing (ICIP)*, Volume 1, 685–688,2004.
6) Mei-Chen Yeh, Kwang-Ting Cheng, "A Compact, Effective Descriptor for Video Copy Detection," in proc. of ACM conference on Multimedia - MM'09, Beijing, China, 2009.
7) Chiu, C. Y., Chen, C. S., and Chien, L. F.," A framework for handling spatiotemporal variations in video copy detection," in proc. of *IEEE Transaction on Circuits and Systems for Video Technology, vol.* 18, 412–417, 2008.
8) Hoad, T. C. and Zobel, J.,"Detection of video sequence using compact signatures," in proc. of *ACM Transactions on Information Systems,* vol 24, pp 1–50, 2006.
9) M. Y. M. Naphade and B.-L. Yeo, "A novel scheme for fast and efficient video sequence matching using compact signatures," in Proc. *SPIE, Storage and Retrieval for Media Databases* , Vol. 3972, pp. 564–572,2000.
10) Wu, X., Hauptmann, A. G., and Ngo, C.W.,"Practical elimination of near-duplicates from Web video search," In proc.of *ACM International Conference on Multimedia (MM'07)*,pp 218–227, 2007.
11) David G. Lowe, "Distinctive image features from scale-invariant key points**,"** in proc. of *International Journal of Computer Vision*, 60, pp. 91-110,2004.
12) Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, **"**SURF: Speeded Up Robust Features," *in proc. of Computer Vision and Image Understanding*, pp 346–359, 2008.
13) Gerhard Roth, Robert Lagani`ere, Patrick Lambert, Lakhmiri, and Tarik Janati, "A Simple but Effective Approach to Video Copy Detection," in proc. of *Canadian Conference - Computer and Robot Visio*n, 2010 .
14) Willems, G., Tuytelaars, T., and Gool, L. V. ," Spatio-temporal features for robust content-based video copy detection," In proc.of *ACM International Conference on Multimedia Information Retrieval*. 283–290, 2008.
15) Chung-Chi Tsai, Chin-Song Wu, Ching-Yu Wu and Po-Chyi Su," Towards Efficient Copy Detection for Digital Videos by Using Spatial and Temporal Features," in proc. of *Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*,2009.
16) Yan Ke, Rahul Sukthankar,"PCA-SIFT: A More Distinctive Representation for Local Image Descriptors," in proc. of *Computer Vision and Pattern Recognition*, 2004.
17) M. Heikkila, M. Pietikainen, and C. Schmid, "Description of interest regions with local binary patterns,"in proc. of Pattern Recognition, vol. 42, no. 3, pp. 425–436, 2009.
18) Krystian Mikolajczyk and Cordelia Schmid,"A Performance Evaluation of Local Descriptors," in proc. of *IEEE Transactions on Pattern analysis and Machine Intelligence*, vol. 27, no. 10, October 2005.
19) Zi Huang, Heng Tao Shen, Jie Shao, Bin Cui, and Xiaofang Zhou,"Practical Online Near-Duplicate Subsequence Detection for Continuous Video Streams," in proc.of *IEEE Transactions on Multimedia,* Vol. 12, No. 5, August 2010.
20) B.S. Manjunath, P. Salembier, and T. Sikora, "Introduction to MPEG-7 - Multimedia Content Description Interface," John Wiley and Sons, 2002.