Computer and Information Science; Vol. 5, No. 5; 2012
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

Hybrid-Based Compressed Domain Video Fingerprinting Technique

Abbass S. Abbass[1], Aliaa A. A. Youssif[1] & Atef Z. Ghalwash[1]

[1] Faculty of computers and information, Helwan University, Cairo, Egypt

Correspondence: Abbass S. Abbass, Faculty of computers and information, Helwan University, Cairo, Egypt. E-mail: abbass1652@yahoo.com; aliaay@yahoo.com; aghalwash@edara.gov.eg

Received: June 4, 2012 Accepted: June 18, 2012 Online Published: July 15, 2012

doi:10.5539/cis.v5n5p25 URL: http://dx.doi.org/10.5539/cis.v5n5p25

Abstract

Video fingerprinting is a newer research area. It is also called "content-based video copy detection" or "content-based video identification" in literature. The goal is to locate videos with segments substantially identical to segments of a query video while tolerating common artifacts in video processing. Its value as a tool to curb piracy and legally monetize contents becomes more and more apparent in recent years with the wide spread of Internet videos through user generated content (UGC) sites like YouTube. Its practical applications to a certain extent overlap with those of digital watermarking, which requires adding artificial information into the contents. Fingerprints are compact content-based signature that summarizes a video signal or another media signal. Several video fingerprinting methods have been proposed for identifying video, in which fingerprints are extracted by analyzing video in both spatial and temporal dimension. However, these conventional methods have one resemblance, in which video decompression is still required for extracting the fingerprint from a compressed video. In practical, faster computational time can be achieved if fingerprint is extracted directly from the compressed domain. So far, too fewer methods are known to propose video fingerprinting in compressed domain. This paper presents a video fingerprinting technique that works directly in the compressed domain. Experimental results show that the proposed fingerprint is highly robust against most signal processing transformations.

Keywords: video fingerprinting, compressed domain, perceptual hash

1. Introduction

Text, image, audio, and video can be represented as digital data. The explosion of Internet applications leads people into the digital world, and communication via digital data becomes recurrent. However, new issues also arise and have been explored, such as data security in digital communications, copyright protection of digitized properties, and invisible communication via digital media (Nianhua, 2011). Due to the rapid development of video production technology and the decreasing cost of video acquisition tools and storage, a vast amount of video data is generated around the world every day, including feature films, television programs, personal/home/family videos, surveillance videos, game videos, etc. Digital video has opened up the potential of using video sources in ways other than the traditional serial playback. However, this requires the development of new technologies for accessing and manipulating digital video (Moxley, 2010).

Additionally, the amount of digital video data, which has the potential of becoming much greater than that of traditional analog video, necessitates the development of digital video management tools for handling massive video databases. Also, the ease with which all digital media can be flawlessly copied makes the development of appropriate rights protection and authentication tools highly desirable. There necessitates techniques for automatically managing this vast amount of information, such that users can structure them quickly, understand their content and organize them in an efficient manner. An emerging technology which is useful for the management of video, particularly with respect to rights protection, is fingerprinting (Cherubini, 2009) and (Peng, 2010) also known as perceptual hashing or replica detection. This is defined as the identification of a video segment using a representation called fingerprint (or sometimes perceptual hash), which is extracted from the video content (Saikia, 2011).

The fingerprint must uniquely identify a video segment, but does not necessarily need to represent its content. Additionally, it must remain the same when a video segment is manipulated, usually by common video processing operations such as resizing, cropping, histogram equalization, compression etc. Fingerprints can be

used for establishing whether two given segments are either identical or derived from each other, and also for establishing whether a video segment is identical with (or derived from) any segment within a given video database (Liu, 2010).

Several video fingerprinting algorithms that work at the pixel level have been proposed. Working directly with pixels is, nowadays, computationally feasible and accurate. However, those solutions do not address the magnitude of the resources needed for such a system and little analysis is provided in order to use a video fingerprinting solution in practical cases. The compressed domain processing techniques use the information extracted directly from compressed bitstream, and therefore are more advantageous than the uncompressed domain in computational means. To achieve this, the information already inherent in the video stream, which was included during the compression stage, is utilized (Maria, 2009).

A partial decompression must still be done to extract the information necessary for processing, however this overhead is small compared to full decompression of the video stream. It is shown that in MPEG video decompression, approximately 40% of the CPU time is spent in Inverse Discrete Cosine Transform (IDCT) calculations, even when using fast DCT algorithms (Young-min, 2000). This paper propose a video fingerprinting technique that work directly in the compressed domain at the stage of variable length decoding (Figure 1), so it is computationally more efficient than uncompressed domain techniques and even the methods that utilize DCT coefficients that are partially decoded.

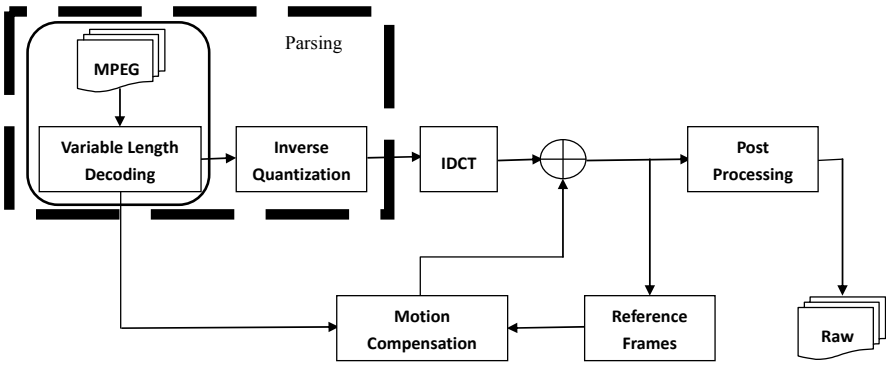

Figure 1. Full decompression versus partial decompression of the compressed video

The bold segmented rectangle is the zone where the partially decoded-based techniques work and the rounded rectangle is where the proposed technique work.

The paper is organized as follows: Related work is given in Section (2); Section (3) describes our proposed fingerprinting technique; Experimental results, comparisons and discussion are shown in Section (4); Finally, Section (5) concludes the paper and point out some directions for future work.

2. Related Works

Some video similarity detection methods use uncompressed MPEG video to directly extract the features. Content of the frames, DC values of macroblocks or motion vectors are used as features. Ardizzone (1999) use motion vectors for feature extraction. They use global motion feature or motion based segmented feature as a signature of the video. In global motion extraction step, statistical distribution of directions (i.e., an angle histogram) is calculated. The angle histogram is computed by dividing the [-180, 180] interval into subintervals. Sum of magnitudes of motion vectors in intervals constructs the angle histogram. In motion based segmentation, motion vectors are clustered and labeled. Labels are given according to the similarity of motion vectors or the histogram of motion vector magnitudes. Dominant regions are taken into account in comparison step.

Joly, Frelicot and Buisson extract local fingerprints around interest points in (Joly, 2003). These interest points are detected with the Harris detector and compared using the Nearest Neighbor method. They propose statistical similarity search in (Joly & Buisson, 2005) and (Joly, 2005). Joly et al. use this method and propose distortion-based probabilistic approximate similarity search technique in order to speed up scanning in content based video retrieval framework (Joly, Buisson, & Frelicot, 2007). Zhao et al. extract PCA-SHIFT descriptors and use it for video matching in (Zhao, 2007). They use the nearest neighbor search for matching and SVMs for learning matching patterns with their duplicates. Law et al. propose a video indexing method using temporal contextual information which is extracted from local descriptors of interest points in (Law-To, 2006) and

(Law-To & Gouet-Branet, 2006). They use this contextual information in a voting function.

Poullot et al. present a method for monitoring a real time TV channel in (Poullot, 2007). They use the method for comparing the incoming data with indexed videos in database. Innovations of the method are z-grid for building indexes, uniformity-based sorting and adapted partitioning of the components. Lienhart et al. (Yang, 2004) use color coherence vector to characterize the key frames of the video. Sanchez et al. (1999) discuss using color histograms of key frames for copy detection. They test the developed system on TV commercials and the system is sensitive to color variations. Hampapur (2000) uses edge features but he ignores the color variations. Indyk (1999) use distance between two scenes as its signature. However, it is a weak and limited signature.

So far, too fewer methods are known to propose video fingerprinting in compressed domain, one of them which DC coefficient is used to model the fingerprint (Mikhalev et al., 2008). Given the extracted DC coefficient, the video fingerprint method constructs the video frame, the modeled video frame is evaluated to obtain the key-frames of the video, which would be further analyzed for generating the fingerprints. Naphade (1999) use histogram intersection of the YUV histograms of the DC sequence of the MPEG video. It is an efficient method in terms of compression. Recently AlBaqir (2009) proposed a video fingerprinting method, in which motion vectors are used to model the fingerprint. He considers utilizing motion vector to construct approximated **motion field** since the motion vectors are commonly generated during video compression for exploiting the temporal redundancy within a video.

3. Proposed Technique

In general, MPEG normally classify the video frames into **I** (intra) frame, **P** (predicted) frame and **B** (bi-directional) frame, and each frame is divided to macroblocks (**MB)** which are 16x16 pixel motion compensation units within a frame. I frames can only have Intra blocks. P and B pictures can have different modes according to motion content. Macroblock type modes in P and B frames are given in Figure 2 and Figure 3 respectively. If **intra coding** is selected, the corresponding MB is encoded individually by exploiting the two dimensional discrete cosine transform (2D-DCT) coefficients (Equation 1). On the other hand, if **inter coding** is selected, the MB is then encoded using motion estimation-motion compensation (ME/MC) algorithms (Richardson, 2003; Heath, 2002).

$$x(n_1,n_2)=\sum_{i=0}^{M-1}\sum_{j=0}^{M-1}X_{i,j}c_ic_j\cos\left(\frac{\pi(2n_2+1)j}{2M}\right) \tag{1}$$

where,

$$c_i=\begin{cases}\sqrt{1/M} & i=0\\ \sqrt{2/M} & i>0\end{cases}$$

and $X_{i,j}$ is M×M block of pixels.

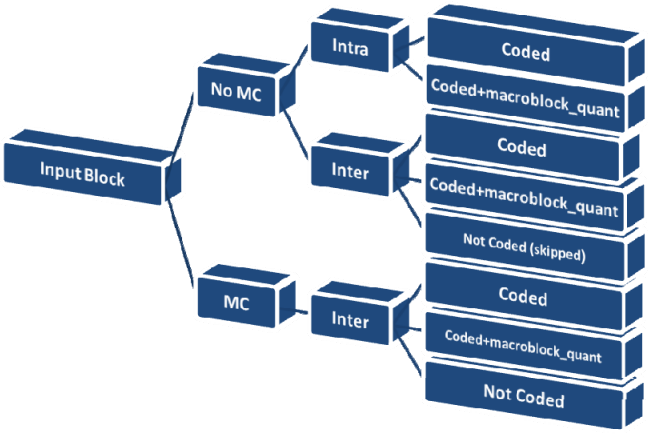

Figure 2. Macroblock type modes in P pictures

The purpose of using ME/MC is to reduce redundancy in temporal direction. To exploit temporal redundancy, MPEG algorithms compute an interframe difference called prediction error (Pei, 1999).

$$\text{prediction error}(i, j) = \frac{1}{MN} \sum_{|m| \le \frac{M}{2}} \sum_{|n| \le \frac{N}{2}} [(I_{mn}, t) - (I_{m+i,n+j}, t-1)]^2 \quad (2)$$

where M and N are macroblock sizes. For a given macroblock, the encoder first determines whether it is Motion Compensated (MC) or Non Motion Compensated (NO_MC) and a scheme is used to determine whether the current block is intra/inter coded based on the prediction error. The scheme can be quite complex, but the general idea is to code the difference between target macroblock and reference macroblock when the prediction error is small, otherwise, intra-code the macroblock. For normal scenes, prediction is performed in P and B frames. When there is a scene change, this prediction drops significantly, which lead to some macroblocks to be encoded in intra mode. Strictly speaking, if a frame is inside of a shot, then the macroblocks should be predicted well from previous or next frames. However, when the frames are on the shot boundary, the frames cannot be predicted from the related macroblocks, and a high prediction error occurs (Yeo, 1995). This causes most of the macroblocks of the P frames to be intra coded instead of motion compensated.

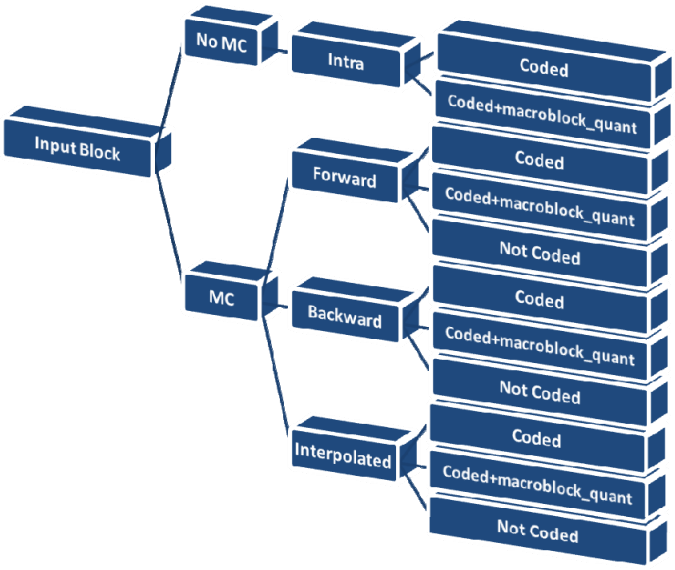

Figure 3. Macroblock type modes in B pictures

The proposed technique fuses the macroblock type's information and the motion field generated by using the motion vectors in the MPEG stream to capture the intrinsic content of the video. Only VLC decoding and number of macroblocks times addition operation is necessary to obtain the MB data (see the rounded rectangle zone in Figure 1).

The proposed technique (Figure 4) is divided into two stages namely the fingerprint extraction stage and the similarity matching stage respectively. In the first stage the compressed MPEG video clip is parsed to extract the macroblocks information and motion vector data. Then, 10 bin one dimensional histogram is constructed using the macroblocks types (Figure 5). The macroblocks types used to generate this histogram aggregate not only the normal types in the normal scenes but also the types that happened in the scene change-like scenes. So tis histogram carries important clues to the spatial and temporal content of the video. Lastly, the proposed technique merge the before mentioned feature to the motion field feature (AlBaqir, 2009) to generate two-part fingerprint. Let call the first part of the proposed fingerprint as **MBTH** (**M**acro**b**lock **T**ype Information **H**istogram) and the second part **MFH** (**M**otion **F**ield **H**istogram).

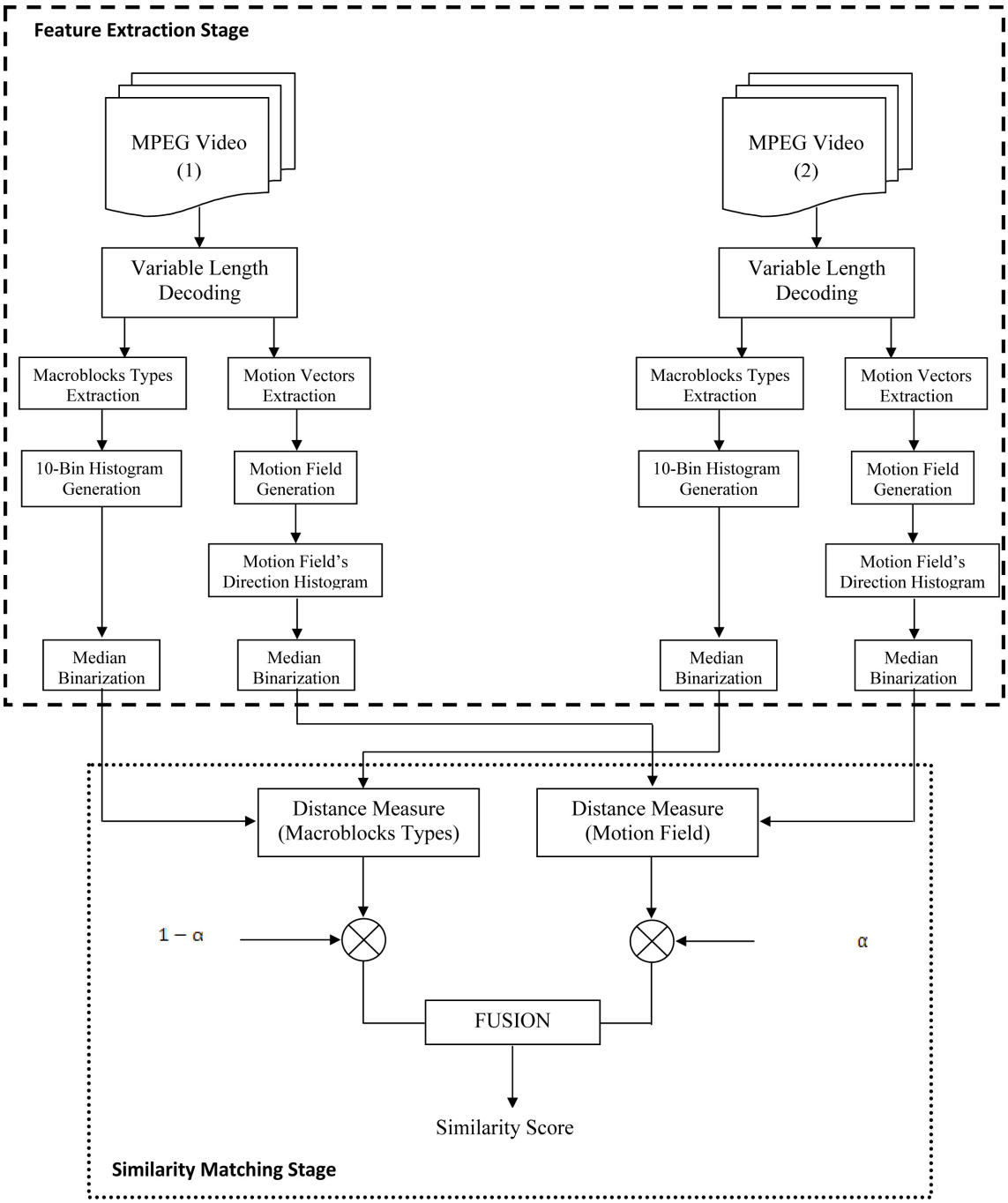

Figure 4. Proposed technique block diagram

$$MBTH^{t} = \left[X_1, X_2, X_3, \ldots X_{10}\right]^{t} \tag{3}$$

$$\sum_{i=1}^{10} X_i = \eta^{t} \tag{4}$$

where η^t is the total number of MB in frame t.

To further reduce the resultant fingerprint redundancy, the proposed technique quantizes the generated fingerprint into a binary sequence (for each part separately as Figure 4 shows) using the median filtering. This can be achieved by fixing a threshold level, and quantize every bin value within each histogram according to the threshold. By having a fingerprint in a binary sequence, a more efficient matching process can be performed

using Hamming distance between the compared fingerprints. Instead of using the histogram intersection (Equation 5) or even the Jaccard coefficient (Equation 6) to compare the resultant fingerprints.

$$W(Q,R) = \sum_{i=1}^{N} min(Q_{(i)}, R_{(i)}) \tag{5}$$

$$J(Q,R) = \frac{|Q \cap R|}{|Q \cup R|} \tag{6}$$

where Q and R are two pair of fingerprints, each containing N bins.

Figure 5. The macroblocks types used in generating the first part of the proposed fingerprint

The fusion of the two-part fingerprint in the second stage is performed as follows:

$$Similarity\ Score = \alpha\ \Psi(\mathrm{MFH}_1, \mathrm{MFH}_2) + (1\text{-}\alpha)\ \Psi(\mathrm{MBTH}_1, \mathrm{MBTH}_2) \tag{7}$$

Where α is the fusion parameter, Ψ is the distance metric, MFH_x is the MFH for video x.

4. Experimental Results

To evaluate the performance of the proposed technique, a test set of 200 videos all taken from the ocean (ReefVid) was used. Then attacks were individually mounted on the videos, so a new test set equals to 3600 videos was generated. The mounted attacks included added watermark, mosaic effect, Embossment effect, flipping, blindness effect, cropping, contrast adjustment, brightness modification, and bit rate change as Table 1 illustrates. Also, the hardware spec of the PC used for the experiments was an Intel dual-core running at 2 GHz with 3GB memory. However, all tests were ran using a single core. Finally, the technique used to compare with is the motion field technique (AlBaqir, 2009) which also operates in the compressed domain as the proposed one to be a fair comparison.

Table 1. Distortions used in the study

Index	Distortion
1	Watermark
2	Mosaic
3	Embossment
4	Horizontal Flipping
5	Vertical Flipping
6	Adding Horizontal Lines
7	Adding Vertical Lines
8	Cropping (*Big Window*)
9	Cropping (*Small Window*)
10	Contrast Adjustment(*negative image*)
11	Contrast Adjustment(*maximum contrast*)
12	Brightness Adjustment (*-50%*)
13	Brightness Adjustment (*-25%*)
14	Brightness Adjustment (*+50%*)
15	Brightness Adjustment (*+25%*)
16	Different Bit Rate(512 Kbps)
17	Different Bit Rate(800 Kbps)

Four sets of the experiments are conducted to study the following issues:

1) Determine the best value of the fusion parameter.

2) Studying the binarization issue.

3) Studying the behavior of the proposed technique against content-preserving attacks (Table 1) to investigate the robustness and uniqueness of the proposed fingerprint.

4) Comparing the proposed work against existing technique working in the compressed domain (Motion Field (AlBaqir, 2009)).

The average retrieval rate across all the 17-attacks for the proposed technique with and without using binarization is shown in Figure 6. The figure shows that the proposed work is improved when using the binarization, also using α=0.2 gives the best result in the two cases.

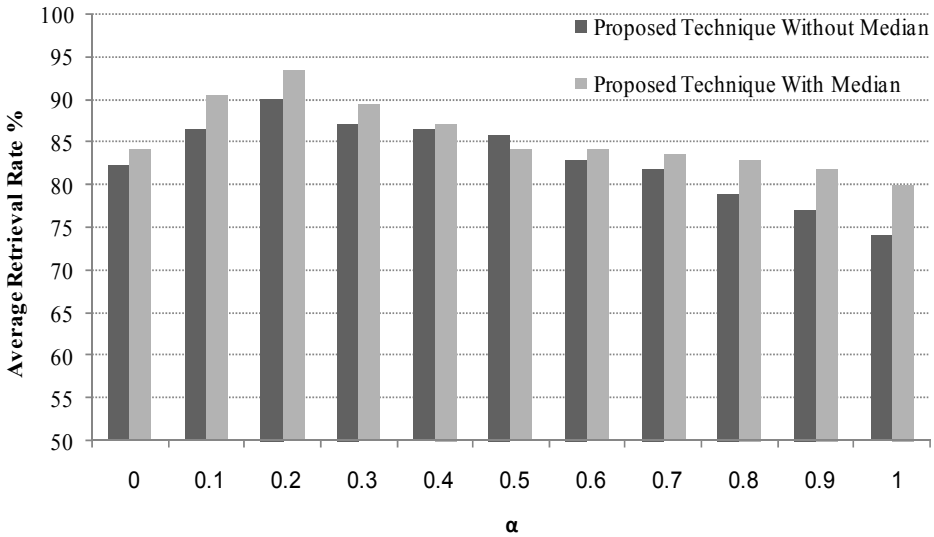

Figure 6. The proposed technique performance using the median to binarize the proposed fingerprint

Figure 7 depict the results of comparing the proposed technique against the base line technique (the motion field (AlBaqir, 2009)) in detailed manner using the before mentioned distortions (from 1 to 17 as mentioned in Table 1). It is clear that the proposed technique outperform the motion field technique. Finally Figure 8 concludes that the proposed fingerprint outperform its constituent fingerprints also the macroblocks types is more important than the motion vectors. An acceptable reasoning to the results is as follows: the motion field technique try to build the motion trajectories of the video content depending on using the available information in the compressed domain, but not all the macroblocks in the compressed domain carry motion information rather a lot of them are labeled as NO_MC or skipped macroblocks. On the other hand, the proposed technique alleviate this drawback by using the information associated with each macroblock in the compressed stream (the macroblocks types), and by the proper merging of the motion field and the macroblocks types.

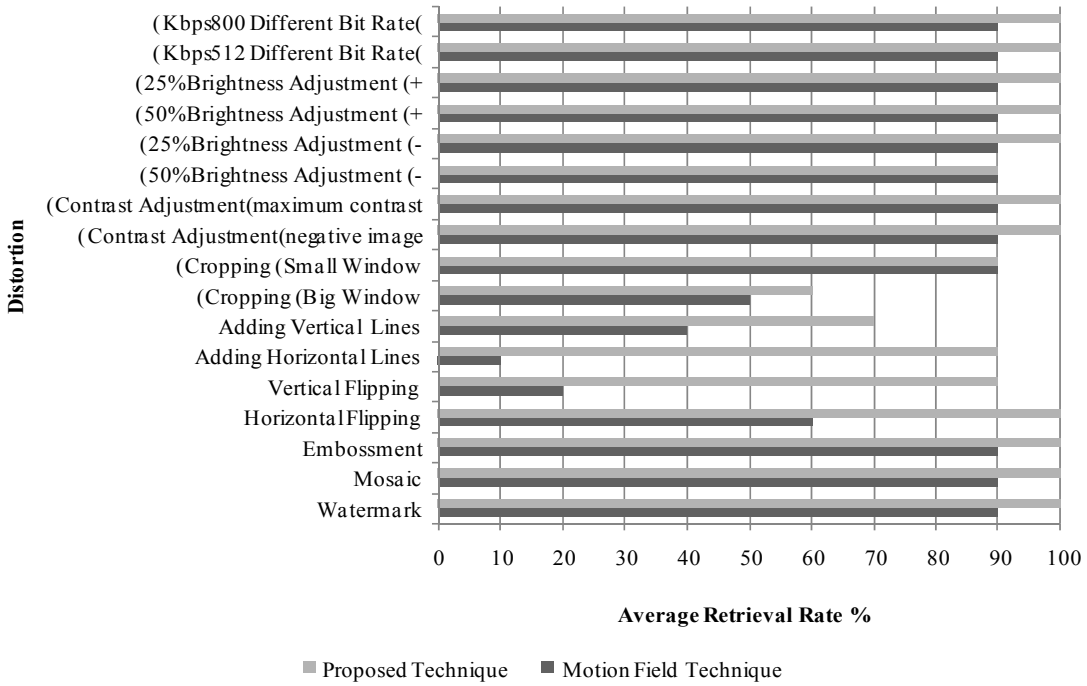

Figure 7. The performance of the proposed technique against the baseline technique across all used distortions

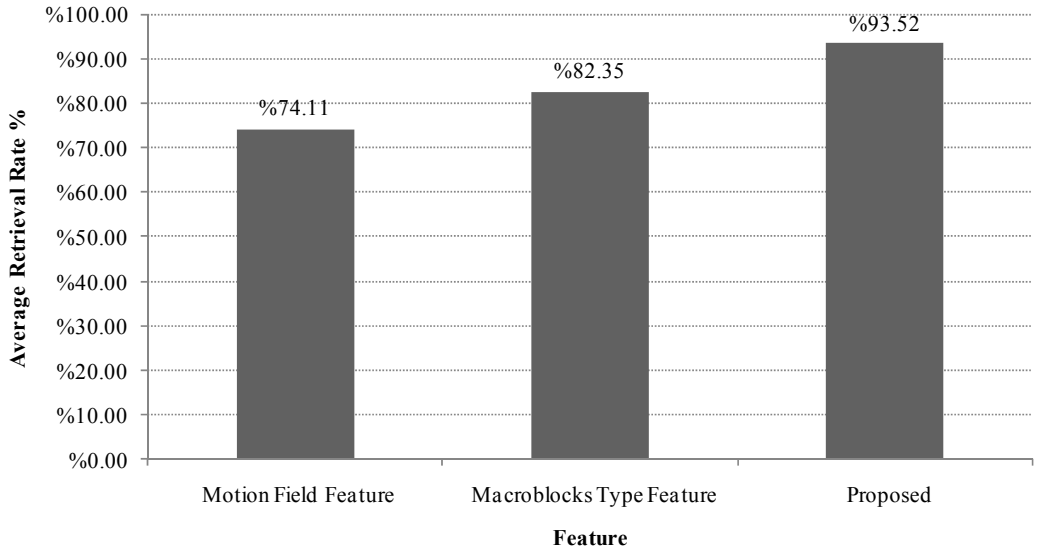

Figure 8. Comparison of the different fingerprint extraction methods

5. Conclusion

This paper proposes a video fingerprinting method in the compressed domain that utilizes the macroblock and the motion vectors information in a hybrid way. The proposed work gives promising results despite of its low computational overhead against a large spectrum of the content-based video transformations. Also, the proposed work shows that the macroblocks types is more important than the motion vectors as intrinsic content-preserving feature in the video compressed domain. One direction for future work is combining this technique with compressed domain watermarking methods to design a robust content management methodology and apply it in the broadcast monitoring area. Also, the proposed work can be adapted to work in real time environment like cellular phones associated with cloud computing metaphor.

References

AlBaqir, M. (2009). *Video fingerprinting in compressed domain.* MSc. Thesis, Delft university of technology.

Ardizzone, E., Cascia, M. L., Avanzato, A., & Bruna, A. (1999). *Video indexing using mpeg motion compensation vectors.* In ICMCS '99: Proceedings of the IEEE International Conference on Multimedia Computing and Systems, (Washington, DC, USA), p. 725, IEEE Computer Society. http://dx.doi.org/10.1109/MMCS.1999.778574

Cherubini, M., de Oliveira, R., & Oliver, N. (2009). *Understanding near-duplicate videos: A user-centric approach.* In Proc. ACM Conference on Multimedia, pp. 35-44.

Hampapur, A., & Bolle, R. (2000). Feature based indexing for media tracking. *IEEE International Conference on Multimedia, 3*, 1709-1712.

Heath, T., Howlett, T., & Keller, J. (2002). *Automatic Video Segmentation in the Compressed Domain.* IEEE Aerospace Conference.

Indyk, G., & Shivakumar, N. (1999). *Finding pirated video sequences on the internet.* Stanford Infolab Technical Report.

Joly, A., Buisson, O., & Frelicot, C. (2005). *Statistical similarity search applied to content-based video copy detection.* In ICDEW ’05: Proceedings of the 21st International Conference on Data Engineering Workshops, (Washington, DC, USA), p. 1285, IEEE Computer Society. http://dx.doi.org/10.1109/ICDE.2005.291

Joly, A., Buisson, O., & Frelicot, C. (2007). Content-based copy retrieval using distortion-based probabilistic similarity search. *IEEE Transactions on Multimedia, 9*, 293-306. http://dx.doi.org/10.1109/TMM.2006.886278

Joly, A., Frelicot, C., & Buisson, O. (2003). *Robust content-based video copy identification in a large reference database.* In Proceedings of ACM International Conference on Image and Video Retrieval (CIVR), vol. 2728, pp. 511-516.

Joly, A., Frelicot, C., & Buisson, O. (2005). *Content-based video copy detection in large databases: A local fingerprints statistical similarity search approach.* ICIP 2005, IEEE International Conference on Image Processing, vol. 1, pp. I-505-8. http://dx.doi.org/10.1109/ICIP.2005.1529798

Law-To, J., Buisson, O., Gouet-Brunet, V., & Boujemaa, N. (2006). *Robust voting algorithm based on labels of behavior for video copy detection.* In MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia, (New York, NY, USA), pp. 835-844, ACM.

Law-To, J., Gouet-Branet, V., Buisson, O., & Boujemaa, N. (2006). *Local behaviours labelling for content based video copy detection.* ICPR 2006. 18th International Conference on Pattern Recognition, vol. 3, pp. 232-235. http://dx.doi.org/10.1109/ICPR.2006.767

Liu, Y., & Yao, L. (2010). *Research of Robust Video Fingerprinting.* In Proceedings International Conference on Computer Application and System Modeling, pp. 43-46.

Maria, C., & Athanassios, N. S. (2009). *Real-time keyframe extraction towards video content identification.* 16th International Conference on Digital Signal Processing, pp.1-6.

Mikhalev, A. et al. (2008). Video fingerprint structure, database construction and search algorithms, Direct Video & Audio Content Search Engine (DIVAS) project, Deliverable number D 4.2.

Moxley, E., Mei, T., & Manjunath, B. S. (2010). Video annotation through search and graph reinforcement mining. *IEEE Transactions on Multimedia, 12*(3), 183-193. http://dx.doi.org/10.1109/TMM.2010.2041101

Naphade, M. R., Yeung, M. M., & Yeo, B. L. (1999). *Novel scheme for fast and efficient video sequence matching using compact signatures.* SPIE, vol. 3972, pp. 564-572. http://dx.doi.org/10.1117/12.373590

Nianhua, X., Li, L., Xianglin, Z., & Maybank, S. (2011). *A Survey on Visual Content-Based Video Indexing and Retrieval.* IEEE SMC, pp. 797-819.

Pei, S. C., & Chou, Y. Z. (1999). Efficient MPEG Compressed Video Analysis Using Macroblock Type Information. *IEEE Transactions on Multimedia, 1*(4), 321-333. http://dx.doi.org/10.1109/6046.807952

Peng, C., Zhipeng, W., Shuqiang, J., & Qingming, H. (2010). *Fast copy detection based on Slice Entropy Scattergraph.* IEEE International Conference on Multimedia (ICME), pp 1236-1241.

Poullot, S., Buisson, O., & Crucianu, M. (2007). *Z-grid-based probabilistic retrieval for scaling up content-based copy detection.* In CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval, (New York, NY, USA), pp. 348-355, ACM.

ReefVid: Free Reef Video Clip Database. Retrieved from http://www.reefvid.org/

Richardson, I. E. G. (2003). *H.264 and MPEG4 Video Compression - Video Coding for Next Generation Multimedia.* England: Wiley & Sons.

Saikia, N., & Bora, P. K. (2011). *Robust video hashing using the 3D-DWT.* National Conference on Communications (NCC), pp 1-5. http://dx.doi.org/10.1109/NCC.2011.5734750

Sanchez, J. M., Binefa, X., Vitria, J., & Radeva, P. (1999). *Local color analysis for scene break detection applied to TV commercials recognition.* In Visual Information Systems, vol. 1614 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg.

Yang, X., Tian, Q., & Chang, E. C. (2004). *A color fingerprint of video shot for content identification.* In Proceedings of the 12th annual ACM international conference on Multimedia Systems, pp. 276-279.

Yeo, B. L., & Liu, B. (1995). Rapid Scene Analysis on Compressed Video. *IEEE Transactions on Circuits and Systems for Video Technology, 5*(6), 533-544. http://dx.doi.org/10.1109/76.475896

Young-min, K., Sung, W. C., & Seong-whan, L. (2000). *Fast Scene Change Detection Using Direct Feature Extraction from MPEG Compressed Videos.* International Conference on Pattern Recognition (ICPR'00), vol. 3.

Zhao, W. L., Ngo, C. W., Tan, H. K., & Wu, X. (2007). Near-duplicate Keyframe identification with interest point matching and pattern learning. *IEEE Transactions on Multimedia, 9*, 1037-1048. http://dx.doi.org/10.1109/TMM.2007.898928