IDENTIFICATION VERSUS CBCD: A COMPARISON OF DIFFERENT EVALUATION TECHNIQUES

M. Corvaglia, F. Guerrini, R. Leonardi, E. Rossi

DEA-SCL, University of Brescia, Italy

ABSTRACT

Fingerprint techniques have a significant advantage in respect of watermarking: a fingerprint can be extracted in each moment of the lifetime of a multimedia content. This aspect is fundamental to solve the problem of copy detection mainly because many copies can be available in huge amount of data in circulation and because each copy can be attacked in several ways (compression, re-encoding, text-overlay, etc.). In this paper the problem of copy detection is studied and tested from two different point of views: content based and identification approaches. The results show that the proposed system is quite robust to some copy modifications and most of all show that the overall results depend on the evaluation method used for testing.

1. INTRODUCTION

The amount of digital video content present in TV Channels, Internet and Video Web Servers is ever growing thanks to the progress of multimedia and networking technologies. Due to the incredible amount of data, the monitoring of media usage becomes of great importance for various reasons.

First of all, the media should be controlled for copyright management: both the owners of web servers and the proprietors of copyright protected media should check if a given content has the right to be distributed, how and where it is used. Another application is media usage monitoring, which aims at supervising the correct placement of visual items such as advertising. Finally, the same material can be present on various sites and there should be the chance to create links by analyzing video and visual objects instead of text annotation.

These aspects can be seen in a general way as the detection and the correct identification of duplicated visual media. The solution is a tricky task because of the large amount of multimedia content to deal with and because each visual item can suffer different editing operations and modifications in terms of color, gamma, text or logo insertion geo,metric transformations, etc. that make a copy often less similar to the original video than another video.

A well known approach for the detection of duplicated visual media is given by *watermarking*, which consists in the insertion into the video stream of an imperceptible digital sign called watermark that can be later retrieved to uniquely identify the content origin and/or to establish its ownership. Watermarking presents two main disadvantages. First, the video should be pre-processed before its distribution to insert the watermark which is not a realistic condition for huge amount of video data in circulation. Second, watermarking is not robust enough to properly identify multimedia material that has been severely attacked (re-encoding, text-overlay, etc.)

Recently alternative approaches have been considered in order to overcome the evident limitations of watermarking in the considered field of interest. *Fingerprinting* approach is a passive technique and does not require any pre-processing because the content itself and intrinsic measurements are used to identify the uniqueness of a video. Given a certain video, its fingerprint consist of a significant set of features opportunely extracted and combined with the purpose to be robust across the common editing operations and sufficiently different for every original content to identify it uniquely and reliably.

In this paper a fingerprinting technique based on visual features is proposed. Its effectiveness has been evaluated under two different perspectives. In the first one the fingerprint is used to detect the transformed version of an original video; in literature this problem is called Content Based Copy Detection (CBCD). In the second case the fingerprint is used to verify its uniqueness and robustness in the identification process.

Several techniques have been proposed to solve the problem of CBCD. They can be divided into two main groups, namely those relying on global descriptors and those based on local descriptors. The first kind of descriptor is extracted from the whole frame. An example of global descriptor is the ordinal measure [1] , which consists in dividing the image into small blocks and then sorting each block depending on its average gray level: the signature is the rank of each block. Other global descriptors are the YUV color histogram of each frame or the block-based motion direction [2]. The main disadvantage of global descriptors is the lack of robustness against some attacks, for example caption insertion and geometric transformations, such as zoom, crop and letter-box. To deal with this problem, the approaches based on local descriptors compute features only on selected points of a frame, also called points of interest, which can be detected by Harris interest point detector [3] or by the frame SIFT descriptor [4].

In [5], some points of interest are tracked along the video sequence to reduce the amount of information and to take into account both spatial, temporal and dynamic behaviors of the local descriptor. The main drawback of local features based methods is their high computational cost.

In the case of identification, the first techniques were introduced for image identification [6] [7] etc. and were mainly based on different transform (wavelet, DCT, etc.) depending on application. Then many extensions for video and audio have been proposed [8].

The brief overview of the existing techniques shows that the approaches differ quite a lot according to the fact they try to solve the CBCD problem rather than the identification one. In this paper, the same approach is evaluated in both environments in order to study how the performance of a technique can vary depending on the selected approach.

This paper is organized al follows. Section 2 presents different kind of evaluation methods; Section 3 describes the proposed method; in Section 4 results are presented and discussed. Finally, in Section 5 conclusions are drawn.

2. EVALUATION TECHNIQUES

Depending on the perspective, *i.e.*, identification or CBCD, the performance evaluation technique is different.

2.1. Content based approach

In the case of CBCD, the approach is usually the exhaustive one. Given a database DB_m of N multimedia items (*e.g.*, images, videos or audio clips) I_i with $i \in [1, N]$, given a set of features F_{Ii} that characterizes each item I_i, given a query item Q extracted from a modified item of the database, given the feature F_Q extracted from the query, chosen an appropriate distance measure D between the features, the exhaustive approach consists in computing all distances D_i ($i \in [1, N]$), between F_Q and F_{Ii} and to select the most similar item I_j whose feature F_{Ij} has the minimum distance with $F_Q : D_j = \min(F_{Ii}, F_Q)$ with $i \in [1, N]$.

Two common parameters are usually considered, *i.e.*, *precision*, which is an indicator of the retrieved relevant items in the retrieved items, and *recall*, which indicates the retrieved relevant items in the overall number of relevant items in the database. In some cases these parameters are summarized in a single parameter F which is given by the weighted harmonic mean of the previous two.

Other parameters are the rate of false alarms and the probability of miss. TRECVID [9] evaluation introduces a weighted sum of these parameters at different operating points.

2.2. Identification approach

In the case of identification, the aim is to verify if a certain query item Q, described by its features F_Q, has been extracted from a multimedia item I_i, described by its features F_{Ii} with $i \in [1, N]$, which is considered to be the right one available in the database DB_m. This verification is performed by computing the distance D_{Qi} between F_Q and F_{Ii} and then comparing it with a selected threshold τ. If the condition $D_{Qi} \leq \tau$ is true, then Q is considered to be extracted from item V.

The key point of this approach is the setting of the threshold. In this sense, MPEG community for the evaluation of MPEG-7 Video Signature Tool proposed to proceed in two steps [10]: in the fist step the threshold is defined while in the second one the identification comparison is performed.

More in details, at the fist stage, called *Independence* test, given a large database of videos DB_I considered independent among them and chosen a certain distance measure between the relative features, the threshold τ is determined by the distance value which corresponds to the maximum false positive rate. MPEG proposes to set the false positive rate equal to 5 part per millions (ppm). At the subsequent stage, called *Robustness* test, a set of query items Q are available at different level and kind of modifications (database DB_R). Each query Q described by F_Q is compared with the original item I_i described by F_{Ii}. Thus they are considered related only if the distance is below the threshold τ computed by the Independence test.

3. SYSTEM OVERVIEW

The proposed system is based on the use of multiple visual features, considering their development in time along both queries Q and original items I_i. The basic idea consists in comparing the sequence of each feature extracted from the query with the same feature in the original item, through the use of a sliding window.

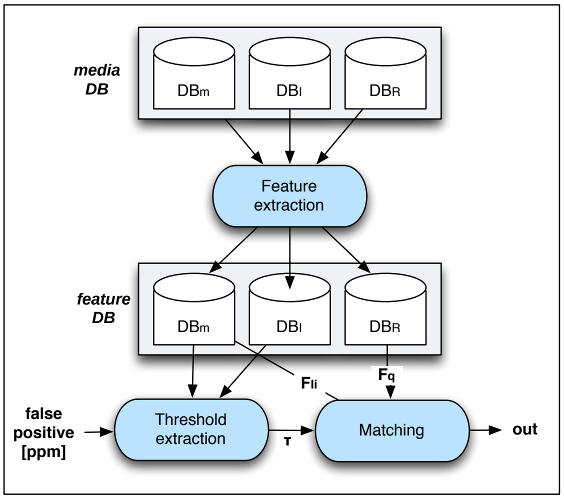

Fig. 1. Framework of the proposed system

The framework is shown in Figure 1. The set of features

F_{Ii} and F_Q respectively of each original item and query is extracted and then stored in three separate databases. The development in time of F_Q is then compared with F_{Ii} according to the evaluation technique (content based or identification) and considering a sliding window of the same duration of the query. The distance in each point of the sliding window is obtained averaging the local distances obtained applying the metric operator. When the sliding window has spanned the whole video item, it is possible to plot the distance between the query and the window of the original video item, with respect to the temporal position of the window in the video item. The global minimum of this function is considered the point where the query is positioned into the video.

The visual low-level features considered in this work are:

MPEG-7 *Dominant Color* (DC): compact representation of the representative colors in a still image or video frame [11] . More specifically, it consists of the representative colors and their percentages in the image, plus spatial coherency and color variance for each dominant color. The distance measure considered to evaluate the distance between two video frames, described by their DCs, is the Earth Mover's Distance [12].

MPEG-7 *Color Layout* (CL): compact and resolution-invariant representation that indicates the distribution of colors in a still image or video frame. This descriptor is obtained by applying the DCT transform on a 2-D array of local representative colours in Y Cb Cr color space [11]. The distance measure considered to evaluate the distance between two video frames, described by their CLs, is the MPEG-7 standard distance.

***Luminance Layout* (LL)**: representation of the distribution of luminance in an a still image or video frame. This descriptor, which a simplification of the Color Layout, has been introduced mainly to deal with monochrome videos. The distance measure considered to evaluate the distance between two video frames, described by their LLs, is the L_1 norm.

4. EXPERIMENTAL RESULTS

Experiments have been performed using a part of the data set available for MPEG-7 Video Signature Tool standardization. The main database DB_m is composed by 1900 video items of 3 minutes. The Independent database DB_I is composed by a 11300 clips of 2 seconds, which have been extracted from the videos available in DB_m and which have been divided by source in oder to guarantee the independency in threshold setting. The Robustness database DB_R contains 545 queries of 2 seconds subjected to 9 transformations (VCR recording, brightness change, camera recapturing, interlaced/progressive conversion, grayscale conversion, frame-rate reduction, resolution reduction, severe compression and text/logo overlay) with Light strength of attack; in total the number of considered queries is 4905. For each video of the three databases, features DC, CL and LL are extracted and stored in identical databases (Figure 1).

The system proposed in Section 3 has been tested in three conditions:

– Content Based (CB-min): as described in Section 2.1, each query of DB_R is compared with all the item of DB_m (exhaustive approach) and the query candidate is obtained considering the minimum distance;

– Identification (ID): as described in Section 2.2, using DB_I three thresholds have been extracted by setting the false positive rate to 5 ppm, 50 ppm and 500 ppm; then each query of DB_R is compared with the relative item of DB_m in order to decide if the query is related or not;

– Content Based with threshold (CB-thr): exhaustive approach (*CB-min*) is refined considering the minimum distance only if it is below a threshold defined in the identification approach (*ID*).

The performance results of the different features in terms of Detection Rate (%) for each modification are shown in Figure 2, 3 and 4. The classical parameters *recall* and *precision* have not been used because they are not applicable in the case if Identification test.

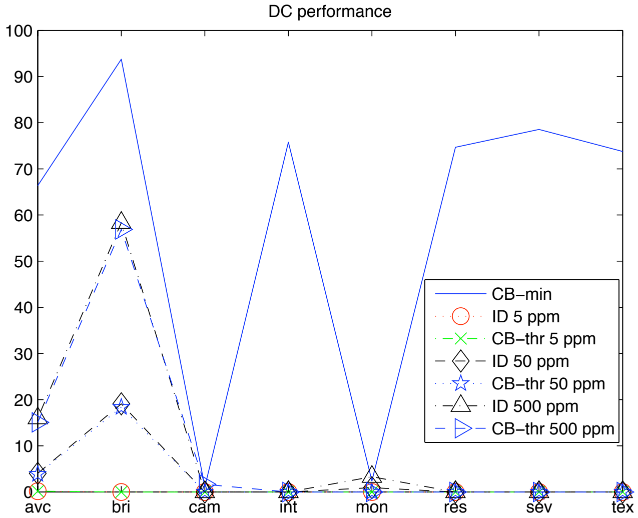

Fig. 2. Performance of DC feature varying the evaluation method

If we consider the results for *CB-min* evaluation method, we can observe that the features act similarly. The Detection Rate is higher that 70% for all modification with exception for camera recapturing (*cam*) and grayscale conversion (*mono*). We can also note that LL provides the best performance for grayscale conversion (*mono*) and that CL reaches the 90% for most of the modifications.

If we consider the results for the other two approaches (*CB-thr* and *ID*), we can clearly see that the performance are seriously compromised. As mentioned above the choice of the threshold is a crucial aspect that should be carefully dealt with. For example, the threshold obtained by the Independence test with false positive rate set to 5ppm provides per-

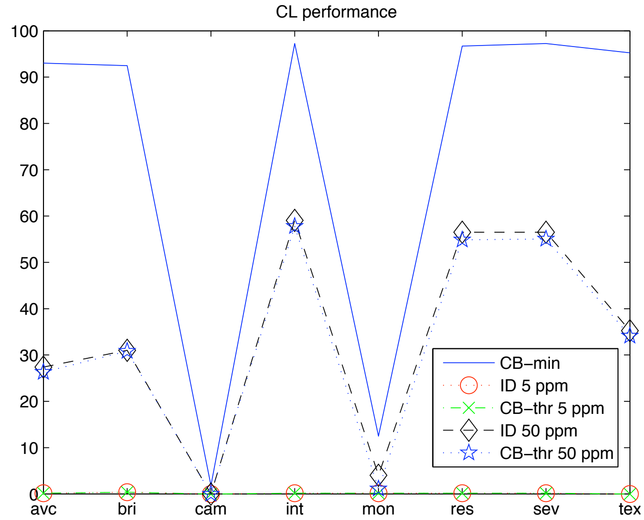

Fig. 3. Performance of CL feature varying the evaluation method

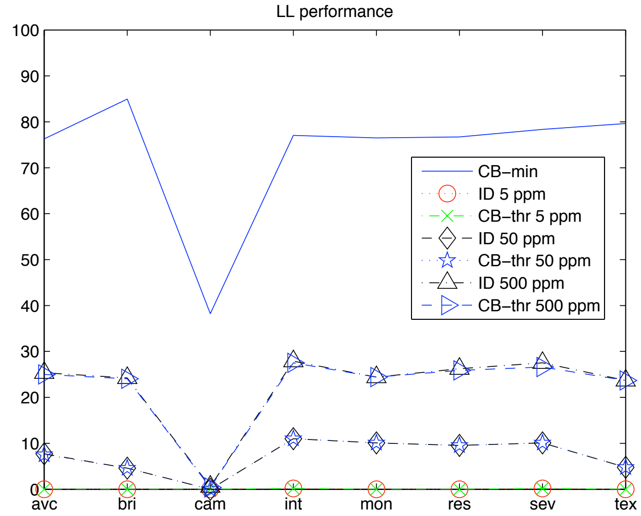

Fig. 4. Performance of LL feature varying the evaluation method

formance closed to null, while with false positive rate set to 50/500ppm the performance significantly increases. In this sense the Independence test forces the threshold to a value that is not the optimum for every technology.

5. CONCLUSIONS

In this paper a comparison of two evaluation techniques for copy detection is reported: one method is based on the classic exhaustive approach while the other one on the identification approach.

The tests have been performed using the video dataset provided by MPEG for MPEG-7 Video Signature Tool standardization. The results show that the evaluation method can seriously compromise the performance evaluation (the exhaustive approach provides significantly better results than the identification one). Moreover the choice of the threshold is crucial and the proposed system needs to be improved for the critical query modifications.

6. REFERENCES

[1] D. Bhat and S. Nayar, “Ordinal measures for image correspondence,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 4, pp. 415–423, 1998.

[2] A. Hampapur, K. Hyun, and R. Bolle, “Comparison of sequence matching techniques for video copy detection,” in *SPIE Conf. on Storage and Retrieval for Media Databases*, San Jose, CA, USA, Jan. 2002, pp. 194–201.

[3] A. Joly, O. Buisson, and C. Frelicot, “Content-based copy retrieval using distortion-based probabilistic similarity search,” *IEEE Trans. on Multimedia*, vol. 9, no. 2, pp. 293–306, Feb. 2007.

[4] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[5] J. Law-To, O. Buisson, V. Gouet-Brunet, and N. Boujemaa, “Robust voting algorithm based on labels of behavior for video copy detection,” in *ACM 14th Annual Int. Conf. on Multimedia*, 2006, pp. 835–844.

[6] R. Venkatesan, S. M. Koon, M. H. Jakubowski, and P. Moulin, “Robust image hashing,” in *ICIP’00 - IEEE Int. Conf. on Image Processing*, Vancouver, Sept. 2000.

[7] P. Brasnett and M. Bober, “Fast and robust image identification,” in *ICPR08*, 2008, pp. 1–5.

[8] B. Coskun, B., and N. D. Memon, “Spatio-temporal transform based video hashing,” *IEEE Tran. on Multimedia*, vol. 8, no. 6, pp. 1190–1208, 2006.

[9] A. Smeaton, P. Over, and W. Kraaij, “Evaluation campaigns and TRECVid,” in *ACM 8th Int. Workshop on Multimedia Information Retrieval*, 2006, pp. 321–330.

[10] MPEG Doc. No. W10154, *Updated Call for Proposals on Video Signature Tools*, Lausanne, Switzerland, Oct. 2008.

[11] B.S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7 - Multimedia Content Description Interface*, John Wiley And Sons, LTD, 2002.

[12] Y. Rubner, C. Tomasi, and L. J. Guibas, “A metric for distributions with applications to image databases,” in *IEEE Int. Conf. on Computer Vision*, Bombay, India, 1998.