TOWARD A MULTI-FEATURE APPROACH TO CONTENT-BASED COPY DETECTION

Marzia Corvaglia, Fabrizio Guerrini, Riccardo Leonardi, Pierangelo Migliorati, Eliana Rossi

DII-SCL, University of Brescia, Via Branze, 38, 25123, Brescia, Italy
E-mail: firstname.lastname@ing.unibs.it

ABSTRACT

Video Content-Based Copy Detection (CBCD) is an emergent research field which is targeted to the identification of modified copies of an original clip in a given dataset, *e.g.*, on the Internet. As opposed to digital watermarking, the content itself is used to uniquely identify the video through the extraction of features that need to be robust against a certain set of predetermined video attacks. This paper advocates the use of multiple features together with detection performance estimation to construct a flexible video signature instead of a fixed, single feature based one. To combine diverse features, a normalized linear combination is also proposed. The system performance boost is evaluated through the MPEG Video Signature Core Experiment dataset and experimental results show how the proposed signature scheme can achieve impressive improvements with respect to the single feature approach.

Index Terms— Content-Based Copy Detection, Color Features, Multi-feature System.

1. OVERVIEW

With the advent of the Internet, video content distribution has reached unprecedented peaks. Hence, finding on different web sites or even within the same database multiple copies of the same video content, perhaps transformed by some video processing, is by now very commonplace. There are many applications that aim to retrieve these copies for various reasons, from copyright protection oriented ones to the less critical video retrieval in online databases, which is a similar although not identical application.

Content-Based Copy Detection (CBCD) is a possible solution that is attracting much attention lately in the research community. The purpose of a CBCD system is to find the original video where a given query video clip, possibly modified in some way (that is edited, re-encoded, etc.) and/or immersed in a dummy video, has been extracted from, furthermore providing the start and end positions of the query in the detected original video clip. As opposed to digital watermarking, its alternative technique in this field, it is a passive approach, that is it does not require any pre-processing of the content. In this case, a video in a given database is indentified by means of its own *signature* [1] (also called fingerprint), namely some feature vector uniquely representing the video content, exactly as is the case for human signatures. CBCD is the subject of recent efforts by both the MPEG community (VST, Video Signature Tool [1]) and the TRECVid campaign [2].

The signature extracted for CBCD from the video content must obviously possess a number of suitable properties. First, it has to correctly identify the video from which the query has been taken, while limiting the false alarm rate at the same time. This in turn implies that the features composing the signature have to be robust against the range of modifications (the *attacks*) that the query is expected to possibly undergo. Additionally, the signature should also be sufficiently compact and as computationally inexpensive as possible. For these reasons, some fast feature extraction and matching techniques are generally required to be implemented for a CBCD system to be practical.

The features proposed in the CBCD context can be divided in two main groups according to the scope of the features they rely on: global, that is extracted from the whole frame, or local. The features used in this work and described in Section 3 are all examples of global features. Local features, on the other hand, try to extract the features only on selected areas of the frame. A thorough discussion on the features and relevant references can be found in [3].

The majority of the CBCD techniques proposed so far is based on a single feature used for signing the entire dataset: in the remainder, we call this approach Single Feature (SF). These techniques first set a copy detection scenario, in particular by selecting a set of attacks against which the system must be robust. Then, a reasonable feature is designed using some hypothesis on its behavior and employed as the video signature; finally, its detection performance is assessed. This is also the workflow of the MPEG-VST standardization process [1]. On the other hand, when multiple features are considered they are usually handled as different dimensions of a higher dimensional feature and the fusion is performed by finding a way to concatenate and normalize them and by employing appropriate distances (*e.g.*, see [4], [5]).

In this paper we propose an alternative point of view on the use of multiple features for the signature construction process. In fact, it is arguable that the *a priori* knowledge of the attacks could be very useful in the feature design, by letting the choice of the best feature dependent on the *a posteriori* detection performance. This implies that it is necessary to use a pool of features to adapt the system to the application framework at hand. The technique proposed in [6] embraces this philosophy too, although in a different flavor: the authors simulate a number of attacks on an original image to construct an ensemble of feature vectors (employing the DCT ordinal measure of [7]) which are used to train a classifier.

As opposed to the work cited above, we do not generate new features for every video; instead, we select a pool of already existing features and then fuse them by means of a linear combination in an optimal way to construct the signature for the original video at hand. Using this logic, the original video is attacked according to a certain scenario, generating a set of so-called original queries, and then the feature combination doing best in the given framework is selected. This has advantages not only in terms of feature extraction speed and simplicity, but also because it allows to use standard features, whose performances are already known. Moreover, using different feature spaces instead of a single one allows to improve robustness by specializing every feature for a certain set of attacks; this way we can let the application scenario decide which feature spaces are more appropriate to consider or to discard for the signature. Finally, when

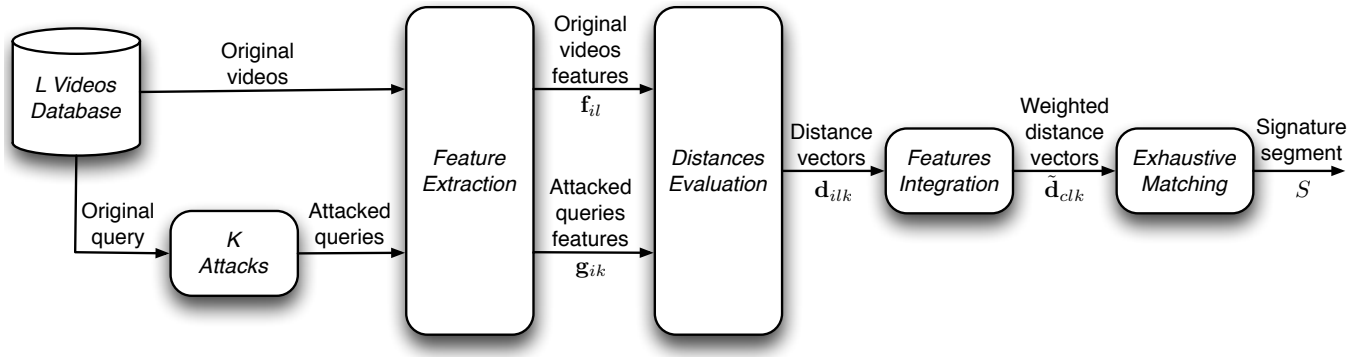

Fig. 1. Framework for the signature construction process.

the evaluation framework changes, *e.g.*, for different applications, the same features still apply since only their combination possibly changes and thus need not to be re-extracted.

The remainder of the paper is organized as follows. Section 2 provides a high level description of the proposed methodology. Section 3 describes the pool of features used in our tests. Section 4 presents the algorithm used to construct the linear combination of features. Section 5 compares the different approaches and clearly shows how the system improves even employing a straightforward features combination. Conclusive remarks are drawn in Section 6, which also indicates current and future research directions.

2. PROPOSED METHODOLOGY

The main purpose of this paper is to advocate the use of both fusion of multiple features and the *a posteriori* thinking described above to construct the signature in a copy detection system. Therefore, instead of adopting the SF approach as is the case for the MPEG-VST, we strive to use an optimal combination of features and let this combination vary from video to video and within a given video too, an approach we dub Optimal Combination (OC). Therefore, the focus of the work is not on the selection of features; instead, we point out that *any* feature proposed so far in the CBCD context, including standard ones, could be included in our multiple feature scheme to further improve detection performance.

Regarding the matching process, we apply the standard exhaustive method to compare feature vectors through a sliding window. The features are extracted from every I-frame to guarantee sufficient detection temporal accuracy. If the query feature vector is composed by q elements (that is, the query has q I-frames) and the original video feature vector has v components, a $v - q + 1$ long distance vector is obtained by evaluating and then averaging the distances between each of the query and video features, paired by closest I-frames. Though computationally expensive, this method allows to precisely measure the performance boost of our method. Of course, it is possible to further enlarge the structure of the signature, allowing some kind of hierarchical matching to speed up the matching process: a multi-purpose structure of this kind is the subject of further research.

The high level flowchart of the signature construction process is illustrated in Figure 1. A set of F features, $f_i,\ i = 1, \ldots, F$, is extracted from the I-frames of the original video database, formed by L videos, thus obtaining $F \cdot L$ feature vectors $\mathbf{f}_{il}$. Now suppose we want to construct the signature of a given video segment. That segment is first isolated in a separate clip and constitutes the original query, which then undergoes the K attacks prescribed by the application framework. Once feature extraction is performed on the I-frames of the attacked queries, an ensemble of $F \cdot K$ feature vectors $\mathbf{g}_{ik}$ is obtained. Considering now the pair composed by the l-th video and k-th attacked query, their respective F feature vectors are compared, by means of the distance measure appropriate for each feature, forming F distance vectors ($\mathbf{d}_{ilk}$).

Feature fusion, detailed in Section 4, is performed by applying a finite number C of linear combinations of the distance vectors, identified by a set of F weights w_i and represented as $\tilde{\mathbf{d}}_{clk}$ with $c = 1, \ldots, C$. Now, the performance of each combination with respect to the k-th query is evaluated by simulating the copy detection process. To that aim, the minimum of the respective weighted distance vector for each of the L videos is retrieved; the F distance values in that position are stored as this query-video pair candidate. The feature fusion and minimum search process is repeated on all L candidates; for the k-th query, the video whose candidate happens to be the said absolute minimum is the detection answer provided by the considered feature combination. Correct detection is then established if the detection answer is the original video segment used to form the original query. The linear combination of features that achieves the highest number of correct detections among the K attacked queries is finally selected as the video segment signature S.

The final formulation of the signature proposed is depicted in Figure 2, for a simple example with two features f_1 and f_2. An atomic length for the original queries of the above procedure is set (Query Minimum Size, QMS) along with TSS (Temporal Step Size) which specifies the time distance between the queries. In Figure 2, TSS is smaller than QMS, thus generating overlapping original queries Q^t; this is not necessary, although it is recommended to improve the signature temporal accuracy. Once the optimal combination of features for Q^t is found, the signature is represented by the features f_1^t, f_2^t and associated weights w_1^t, w_2^t.

When an external query is to be detected, its features are first extracted and then matched to those contained in each given video signature segment according to the combination specified therein. When the query is longer than QMS, a video signature segment of suitable length is selected by concatenating its basic elements.

3. FEATURES DESCRIPTION

The $F = 4$ features considered in this work are listed in the next paragraphs. The first two, Dominant Color and Color Layout, are part of the MPEG-7 standard Descriptors of the color feature [8] [9]. Luminance Layout is a simplification of Color Layout, obtained by

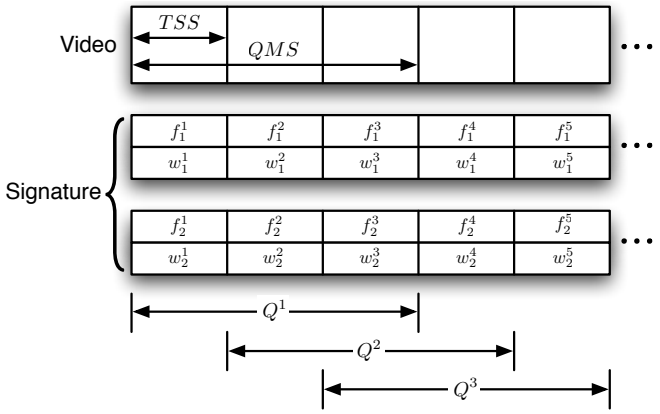

Fig. 2. Video signature conceptual scheme.

applying the latter in the grayscale domain, introduced specifically to handle monochrome videos. Last, Ordinal Measure is an additional, luminance-based feature proposed in [10].

The MPEG-7 Dominant Color (DC) describes the representative colors in a video frame. More in detail, it is composed by a number of representative colors (in this case 8) and their percentages in the image; optionally, spatial coherency and color variance could also be specified. The non-normative distance between two video frames, described by their DCs, is the Earth Mover's Distance [11] applied to the representative color distribution.

The MPEG-7 Color Layout (CL) is a compact and resolution-invariant representation of the distribution of colors in a video frame. It is obtained from the DCT transform of a 2-D array of local representative colors in the YCbCr color space. The distance proposed in MPEG-7 standard, which is a weighted Euclidean norm, is employed in this work. Luminance Layout (LL), derived from CL, is a representation of the distribution of luminance in a video frame. The simpler L_1 norm has proven to be appropriate for this feature.

Finally, Ordinal Measure (OM) is obtained by partitioning the image into blocks, then their average luminance is sorted. The feature vector is given by concatenating the rank of each block. For this feature too, the L_1 norm is used as the feature distance.

4. FEATURES FUSION

To properly mix different features in a multi-feature based signature for CBCD, there are two issues to solve: first, to define the model to combine the features; second, how to choose the best combination of features for the application at hand. Regarding the latter, we opted for hard detection, where the number of first position detections only are counted, to minimize false alarms.

The model selected to combine the feature distance vectors is to normalize them and then apply a sum constrained, discrete weights linear combination. Not only it is faster to evaluate a finite number of feature combinations instead of searching the optimal combination through some optimization algorithm for performance maximization, but it is also more reliable given that the convergence of the algorithm would be affected by the very challenging task of feeding to it the correct detection conditions.

More in detail, the feature distance vectors correspondent to a given query-video pair are normalized by their respective maximum to account for their different scale (as opposed to MinMax normalization, the minimum is not shifted to 0 to retain the significance of its value). The possible combinations $w_i,\ i = 1, \ldots, 4$ are formed by integer vectors whose components are the weights of the linear combination that sum to a given factor S. At this preliminary stage, we choose $S = F = 4$ to limit the number the combinations while including in the combinations the balanced mix of all the features. Therefore, there are $C = 35$ possible linear combinations, including for example $w = [4\,0\,0\,0]$ (pure first feature), $w = [1\,1\,1\,1]$ (balanced mix), $w = [0\,3\,1\,0]$, and so on.

For each of these, the weighted distance vectors are formed by multiplying the normalized distance vectors for the combination weights. The candidate time position for a certain query-video pair, according to the exhaustive search process outlined in Section 2, is therefore that where the features have their lowest ratio with respect to the maximum distance value in time, after being weighted by the combination. It can be noted that the different scaling the various features underwent during the normalization step have little effect in the minimum search.

5. EXPERIMENTAL RESULTS

We show the effectiveness of our method with respect to the SF approach in two steps. First, we evaluate the performance boost as the best, single feature is independently selected for every video; we refer to this approach as Optimal Feature (OF). Next, we consider the Optimal linear Combination (OC) approach, where in addition the feature integration is employed to further improve performances.

For our experiments, we used the data set provided by the MPEG community for the VST standardization process [1]. It is composed by 1900, 3 minutes long original video clips and 545 original queries; for this test we considered only the Direct (not immersed in a dummy video), 2 seconds long queries. All the queries are extracted in the same interval (60-62s). Each query is then processed by a number of attacks (Light level, as in the MPEG evaluation procedure) that are listed in the first column of Table 1. No queries pertaining to the frame rate reduction attack are available for this test due to database encoding issues. They are first used as original queries in the signature construction process by fixing $QMS = 2$s, hence constructing only a single signature element of the videos from which they were extracted (referring to Figure 2, we have a single Q^t for 545 videos). Then, we re-apply them as external queries. As already pointed out, in this work we follow a classic retrieval approach by verifying that the correct video is detected in the first position (hard decision) for any given query. The attacks are given the same relative importance; in principle, it could be adjusted by setting a diverse detection score for each attack.

The first part of Table 1 illustrates the results for the Single Feature (SF) approach. In this case, each feature is treated independently and reported in separate columns. The first 8 rows differentiate the detection percentages for each of the attacks considered, while the last is the overall detection percentage[1]. As it can be seen, LL achieves the best average detection performance because of its higher performance in the CC and MONO attacks; however, CL is evidently superior in all the other cases. This is a clear example where two features are targeted to different attacks in the application scenario, hence they would work better if combined in some way.

Before trying to combine the features, we could still boost performance by applying the *a posteriori* thinking outlined in Section 2. As already discussed, it is possible to use the best working feature for any given original video by selecting the feature working best for

[1] SC has only 417 queries, so the overall row is not exactly the average of the first 8 rows.

Attacks	*Single Feature (SF)*				*Optimal*	*Magic*	*Optimal*	*Magic*
	DC	CL	LL	OM	*Feature (OF)*	*Feature (MF)*	*Comb. (OC)*	*Comb. (MC)*
Analog VCR recording & recapturing (AVC)	66.24	92.84	76.33	50.64	95.78		96.15	
Brightness change (BC)	93.76	92.48	85.14	83.49	99.45		99.82	
Capturing on camera (CC)	1.10	1.83	38.35	14.68	44.95		53.21	
Interlaced/progressive conversion (IPC)	75.78	97.25	77.06	73.39	98.35		99.08	
Color to monochrome conversion (MONO)	2.39	12.48	76.15	69.36	88.44		93.03	
Resolution reduction (RR)	74.68	96.70	77.06	75.05	98.35		99.08	
Severe compression (SC)	74.34	97.12	77.46	70.50	98.80		99.04	
Text/logo overlay (TLO)	73.76	95.23	79.63	53.58	96.70		98.72	
OVERALL	57.25	72.52	73.28	61.06	86.55	89.84	90.86	99.50

Table 1. Various methods performance comparison in terms of correct detection percentages.

the queries extracted from said video. The Optimal Feature (OF) approach results are depicted in the second part of Table 1. The results for each attack are always better than the best feature performance of the SF method, indicating that for every video the best feature is not always the one that achieves the best average performance for that attack. The overall result improves the best result of the SF appraoch by more than 10%. For reference, an ideal approach which we call Magic Feature (MF) is also reported; this represents the technological limit of the features when taken singularly. In this ideal case, the system is free to pick a different feature not only for diverse original video clips but also in function of the attack the query underwent, that is in presence of a perfect attack estimation by the signature matching system. Obviously, considering a single attack, the OF and the MF approaches are identical; the advantages of MF are attained only when different attacks are used. The fact that the ideal approach is not very much better performant that the OF approach points out that for a given video the best feature taken singularly tends to be the always the same for every attack.

Last, instead of using of the features one at a time, Table 1 shows on the right how the performances improve in the OC approach, that is by applying the linear combination of features explained in Section 4. Remarkably, the OC approach is still superior to the ideal MF approach. In conclusion, it is shown by the overall results that using both the attacks knowledge and the linear combination of features brings the system performance above the 90% mark, which is an impressive 18% boost with respect to the single best feature performance. It is further apparent that the ideal approach dubbed Magic Combination (MC), which is analogous to the MF approach, has almost perfect score, indicating that, although the OC method achieves a great improvement in terms of detection performance, there is still room for improvement in how the features fusion is done.

6. CONCLUSIONS

In this paper a novel approach to Content-Based Copy Detection has been presented. In particular, it exploits both the fusion of multiple features and detection performance estimation in phase of signature construction to attach an adaptive signature to the original video clips. This way, it is possible to tailor the system to the application framework at hand by choosing the attacks set. The proposed signature is highly flexible and expandible through the addition of new features and can achieve significant detection improvements even using simple features, widely justifying the extra work needed to construct the video signature.

Current research is focusing on formatting the signature to allow hierarchical searching by dividing it into various levels composed of different sized segments. Also, in some cases it is possible to perform a simple attack estimation during the matching process (*e.g.*, by recognizing that the query is monochromatic), thus dividing the queries in classes that uses diverse feature combinations: this could be included in the signature structure too. Finally, other feature combination methods besides the linear one presented in this paper, such as sorted ranks based combination, are being considered. The proposed signature is then to be tested on the very challenging TRECVid database to further validate its performance.

7. REFERENCES

[1] ISO/IEC/JTC1/SC29/WG11MPEG 2009/N10345, *Description of Core Experiments in Video Signature Description*, Lausanne, Switzerland, Feb. 2009.

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

[3] A. Hampapur, K. Hyun, and Bolle R., "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.

[4] G. Qian et al., "Similarity between euclidean and cosine angle distance for nearest neighbor queries," in *ACM Symp. on Applied computing*, 2004, pp. 1232–1237.

[5] M. Bertini, A. Del Bimbo, and W. Nunziati, "Video clip matching using mpeg-7 descriptors and edit distance," in *Int. Conf. on Image and Video Retrieval (CIVR)*, 2006, pp. 133–142.

[6] J.-H. Hsiao et al., "A new approach to image copy detection based on extended feature sets," *IEEE Trans. on Image Processing*, vol. 16, no. 8, pp. 2069–2079, 2007.

[7] C. Kim, "Content-based image copy detection," *Signal Processing: Image Communication*, vol. 18, pp. 169–184, 2003.

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

[9] ISO/IEC/JTC1/SC29/WG11 MPEG 2001/N4358, *Text of ISO/IEC 15938-3/FDIS Information technology - Multimedia content description interface - Part 3 Visual*, Sydney, Australia, July 2001.

[10] 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.

[11] 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.