CBCD BASED ON COLOR FEATURES AND LANDMARK MDS-ASSISTED DISTANCE ESTIMATION

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

DEA-SCL, University of Brescia, Italy

ABSTRACT

Content-Based Copy Detection (CBCD) of digital videos is an important research field that aims at the identification of modified copies of an original clip, *e.g.*, on the Internet. In this application, the video content is uniquely identified by the content itself, by extracting some compact features that are robust to a certain set of video transformations. Given the huge amount of data present in online video databases, the computational complexity of the feature extraction and comparison is a very important issue. In this paper, a landmark based multi-dimensional scaling technique is proposed to speed up the detection procedure which is based on exhaustive search and the MPEG-7 Dominant Color Descriptor. The method is evaluated under the MPEG Video Signature Core Experiment conditions, and simulation results show impressive time savings at the cost of a slightly reduced detection performance.

Index Terms— Content-Based Copy Detection, Earth Mover's Distance, Dominant Color, Multi-Dimensional Scaling.

1. INTRODUCTION

Video content proliferation on the Internet is increasing at an impressive rate. It is common to find many copies of the same video content on different web sites, possibly modified by some video processing transformations for various reasons depending on the intended use. This fact naturally poses several challenges to copyright-oriented applications, such as copy deterrence (*e.g.*, content tracking), but also to less critical applications, *e.g.*, video copy retrieval in an online database.

To solve this problem, a possible solution is to actively process the video content *before* its distribution by embedding an imperceptible digital watermark which, if its robustness against video processing and/or tampering is granted, can be retrieved at a later time by an authorized entity to uniquely identify the (eventually modified) content origin and/or its owner.

Content-Based Copy Detection is an alternative to digital watermarking which is passive in nature and therefore does not require any pre-processing of the content. Instead of embedding external information to identify the considered video content, the content itself is used to assess whether a copy is present in a given video collection, much like a fingerprint is able to uniquely identify a human being using a fingerprint database. For this reason, the extraction of a given set of features from the content to uniquely identify itself and its modified copies is called fingerprinting.

In a standard CBCD framework, the purpose of the copy detection system is, given a query video (also called copy), to find the video (also called original clip) where the query has been taken from, even if the query has been attacked (that is it has been edited, modified, re-encoded, etc.), or the query has been immersed in a dummy video, or both. An additional information the system should provide is the start and end positions of the query in the detected original clip. Both the MPEG community (Video Signature Tool, [1]) and the TRECVid campaign [2] are interested in CBCD.

To design a CBCD system, it is necessary to select a set of features that must be able to correctly identify the original video clip from which the query has been extracted, at the same time limiting the false alarm rate (this is called the uniqueness property). Clearly, these features should be robust to the modifications potentially applied to the considered video content. Given the huge size of video data that must be processed, CBCD systems may have to rely on compactly representable features that are both easily extractable and comparable by means of a computationally inexpensive metric. Otherwise, a clever way to evaluate feature distances must be envisioned to obtain an acceptable detection time, thus making the detection process more practical.

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 descriptors is extracted from the whole frame. An example of a global descriptor is the ordinal measure [3] [4], 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, for example, the YUV color histogram of each frame [5] or the block-based motion direction [4].

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 the features only on selected points of a frame, also called points of interest, which can be detected, for example, by Harris interest point detector [6] or by the frame SIFT descriptor [7]. In [8], 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 the methods based on local features is their high computational cost.

Regardless of the kind of descriptor, robustness against processing, uniqueness and compactness are not the only challenges the system has to cope with; another main problem in CBCD applications is the time needed for feature extraction and matching due to the need to search copies among huge video databases. Many solutions have been proposed to tackle this problem. Features could be extracted only in keyframes and not on the whole original video and on the query, and some fast techniques can be used to speed up feature extraction and matching. The technique described in [9], for example, works in the compressed domain computing the full DCT coefficients directly from block DCT coefficients. The signature is the ordinal measure of some low-middle frequency full DCT coefficients.

In this work, we argue that it is possible for CBCD to use features with complex distances by using multi-dimensional scaling (MDS) techniques (see [10] [11]) to approximate the real distances in a much shorter computational time. We apply a MDS method as a part of a fully functional CBCD system, thus pointing out that features with expensive metrics could still be effectively used, *e.g.*, for incorporation in a multi-feature based system, without having to discard them because of their computational complexity. To prove our point, in our copy detection system we will employ a global, frame-based feature described in the MPEG-7 standard, that describes a sizable pool of features proposed for retrieval applications.

Using MPEG-7 terminology, a Descriptor (D) is a possible representation of a feature. One feature can be represented by many Descriptors. For example the feature "color" can be represented by Descriptors Dominant Color, Color Layout, etc. In this work we have used the Dominant Color (DC) Descriptor and the computationally expensive Earth Mover's Distance (EMD) as the metric [12]. Since we are interested in applying multi-dimensional scaling techniques to simplify feature comparison, we consider a simple approach to copy detection based on exhaustive feature search. In this paper we do not employ any further strategy to speed up the detection, such as hierarchical comparison, so that we can properly evaluate the proposed system in terms of time saving.

The paper is organized as follows. Section 2 describes the DC Descriptor and the EMD metric. Section 3 presents both the exhaustive search technique and a particular MDS algorithm used to decrease the overall computational cost. The MDS method is briefly explained in Section 4. How the system fares in terms of computational time and detection performances is reported in Section 5. Conclusive remarks are drawn in Section 6.

2. FEATURE DESCRIPTION

The feature considered in this work is the color information present in the video. In particular, the Dominant Color Descriptor is adopted, following its definition and implementation as given in the MPEG-7 Standard [13], [14]. This Descriptor specifies a set of dominant colors, that in general could be computed on a still image or on a video frame. In our case, the Descriptor is obtained for each whole I-frame of the considered video. More specifically, the Descriptor consists of the representative colors and their percentages in the region, plus spatial coherency and color variance for each dominant color. The standard also specifies how to appropriately quantize and store the Descriptors (the bit format).

In order to compute this Descriptor, the colors present in the given I-frame are first clustered following a normative algorithm by which a small number of representative colors (up to 8, that is the value we used in this work) and the relative percentages of these colors are calculated. As an option, the variances of the colors assigned to a given dominant color could also be computed. A spatial coherency value is also computed that differentiates between large color blobs versus colors that are spread all over the image.

To evaluate the distance (dissimilarity) between two frames I_1 and I_2, described by their DCs, the MPEG-7 standard proposes a non-normative distance that employs all the Descriptor components above. In this paper we have adopted the Earth Mover's Distance (EMD), described in [12], which only uses the dominant colors histogram (the LUV triplets and their associated percentages) and which has been proved to be more suitable for this problem. The EMD represents a distance measure between two statistical distributions and, in informal terms, reflects the minimal amount of work that must be performed to transform one distribution into the other

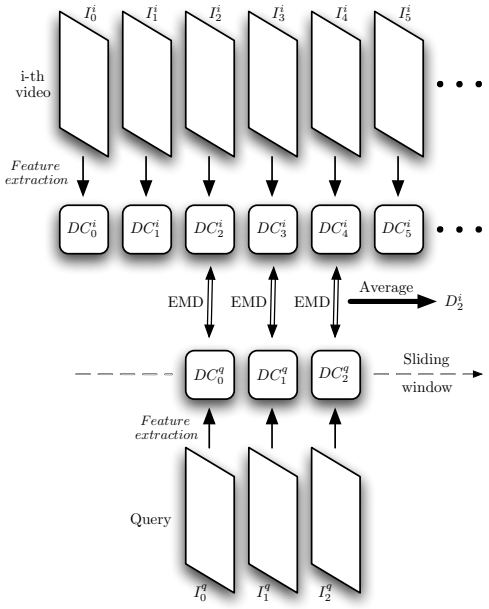

Fig. 1. Sliding window approach for exhaustive feature comparison.

by moving "distribution mass" around, once a cost metric for the latter operation is provided. This is known in literature as a special case of the transportation problem from linear optimization, for which efficient solution algorithms are available. In our case, the underlying cost function is the Euclidean distance in the LUV color domain, which is both cheap and perceptually significant. Nonetheless, the EMD is significantly more computationally expensive than simpler distances such as standard L_p metrics.

3. SYSTEM OVERVIEW: LMDS FOR EMD ESTIMATION

The most straightforward approach to design a copy detection system is the exhaustive feature comparison, performed by evaluating the distance between the Descriptors associated to the considered query and the Descriptors associated to a sliding window of the same duration in the original video clip, as depicted in Figure 1. For the i-th video of the database, the feature vector $\overrightarrow{DC}^i$ is first extracted from the I-frames $I_j^i,\ j = 0, \ldots, T^i$ (T^i is the the number of I-frames of the i-th video), as well as the query feature vector $\overrightarrow{DC}^q$. Then, a distance vector $\overrightarrow{D}^i$ is calculated for every time position; each component is obtained averaging the local distances given by the EMD metric operator. As the Group Of Pictures size value is not constant, each local distance is calculated between each I-frame of the query and the closest I-frame in the sliding window (original video) in time. Last, for every distance vector $\overrightarrow{DC}^i$, the best candidate (minimum distance) is chosen and then a single, global minimum is retained (in a more complex environment, this two-stage minimum search allows to provide, if necessary, a variable number of answers associated with a single query depending on the distance values).

As already pointed out in Section 1, the problem of this approach lies in its computational complexity, especially for large databases and complex metrics like the EMD. In this work, we propose to use multi-dimensional scaling to achieve significant time savings while

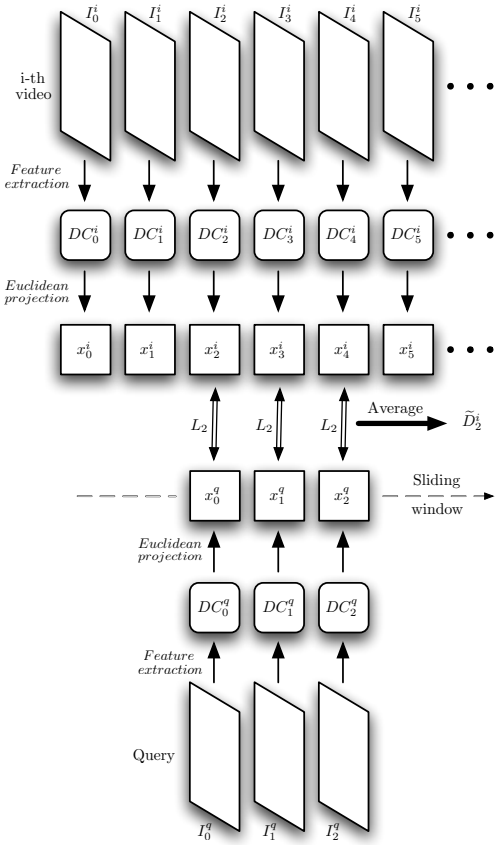

Fig. 2. LMDS method applied to the sliding window exhaustive approach.

at the same time preserving good performance in terms of copy detection. Multi-dimensional scaling is used in data analysis applications as a metric-preserving dimensionality reduction process. Given a collection of N objects and a $N \times N$ matrix containing the distances between the objects according to a specific metric, MDS algorithms try to extract a k-dimensional ($k << N$) Euclidean embedding which preserves at best the input metric.

The idea is to consider the features extracted from the video database as the input objects, and then to perform MDS to find the eigenvectors set that allows to project the features in an Euclidean space. The query features are then projected on the same space and hence the sliding window search can be performed using the standard Euclidean distance instead of the much heavier EMD. MDS is performed according to the technique presented in [15], which is briefly recapped in Section 4. Consequently, the EMD step in Figure 1 is replaced by a feature projection in the k-dimensional Euclidean space and a standard L_2 metric computation, as depicted in Figure 2, and the output is an estimated distance vector with components $\widetilde{D}_j^i$ which undergoes the same minimum search process as before.

4. LMDS ALGORITHM

As opposed to classical MDS algorithms, which are still computationally expensive for large collections of objects, we adopted the algorithm based on landmark points described in [15]. First, a small number n of objects are randomly selected and classical MDS is performed on this subset. Then, the coordinates of all the other objects are obtained by triangulating their position using the distances to the landmark points only. A focal point is the selection of the landmark points, and many strategies could be envisioned. Random selection is anyway cheap and yet effective and is employed here. More sophisticated approaches of landmark selection are under consideration for future improvements.

More in detail, given the $n \times n$ matrix Δ_n of squared EMD distances between the random landmark feature points (DCs taken from random I-frames in the video database), a mean-centered inner-product matrix B_n is computed as $B_n = -\frac{1}{2} H_n \Delta_n H_n$, where $|H_n|_{ij} = \delta_{ij} - \frac{1}{n}$. Then, to obtain a k-dimensional embedding, the k largest positive eigenvalues of B_n are taken, which we indicate with $\lambda_i,\ i = 1, \ldots, k$. Negative eigenvalues account for the non-Euclidean part of the distance matrix Δ_n. The embedding matrix L_k rows are given by $\sqrt{\lambda_i} \cdot \vec{v_i}^T$, where $\vec{v_i}^T$ are the associated eigenvectors. Last, the pseudo-inverse transpose matrix $L_k^{\#}$ of L_k can be readily evaluated; in fact, its rows are given by $\vec{v_i}^T / \sqrt{\lambda_i}$. To embed a DC feature point a taken from an I-frame I_a, that is to say to find its coordinate vector $\vec{x}_a$ in the k-dimensional Euclidean space, one need only to compute $\vec{x}_a = -\frac{1}{2} L_k^{\#} (\vec{\delta}_a - \vec{\delta}_u)$ (the "Euclidean projection" process in Figure 2), where $\vec{\delta}_a$ is the vector containing the squared EMD distances between the features of a and those of the n landmark features and $\vec{\delta}_u$ is the mean column of Δ_n. It is important to note that the computation of the projection matrix $L_k^{\#}$ has to be performed only once for a given database.

5. EXPERIMENTAL RESULTS

In this section we present some results of the experimental evaluation, according to the criteria specified in [1]. Regarding the video data set, MPEG community provided 1900 original videos and 545 queries (here we are considering Direct queries, that is those not immersed in a dummy video). The attacks on the queries are reported in the first column of Table 1. Each query has been attacked at different levels of modification, where applicable (Light, Medium, Heavy). In this work we only considered Light levels as in the MPEG evaluation procedure. However, for this experiment we did not employ the identification approach proposed in [1], which aims at setting a global threshold to differentiate between false and true detection when a query is compared with a given video clip. Instead, we verified that the video with the smallest distance with respect to the given query is indeed the true one as in classical retrieval scenarios.

The parameters of our method are the number n of landmark features and the number k of eigenvalues λ_i to retain. For the latter, we could simply use all the positive eigenvalues, obtaining the maximum admissible value for k, or limit the maximum spectral radius (*i.e.*, the maximum eigenvalue divided by the minimum eigenvalue) of the eigenvectors sub-matrix, thus excluding some of the smaller eigenvalues (associated with the less significant Euclidean coordinates). In Table 2 the values of the various landmarks number n and consequent subspace dimensions k used in this work are reported.

Detection performance of both the EMD metric and its approximated version using LMDS are illustrated in Table 1. The former represents the technological limit of the DC feature, since the exhaustive search is the best method, although the most computational expensive and thus impractical. Note that in the case of camera capture attack LMDS achieves better results because a single weak miss becomes a true positive due to poor overall estimation for this particular modification. As it can be observed, using $n = 20$ induces a sharp decrease in performance because the subspace has too small

Attacks	EMD	LMDS												*Max spectral radius*
		∞				10^5				10^4				
		20	50	150	250	20	50	150	250	20	50	150	250	*N. of landmarks (n)*
Analog VCR recording	66.4	0.2	55.7	0	0	45.1	55.7	62.3	63.5	45.1	55.7	63.0	64.8	
Brightness change	93.8	0.4	83.5	0.6	0.9	79.1	83.5	85.9	85.9	79.1	83.5	85.7	86.4	
Capture on camera	1.1	0.4	1.3	0.2	0.2	1.1	1.3	1.3	0.9	1.1	1.3	1.1	1.3	
I/P conversion	75.8	0	65.7	0.4	0.2	55.4	65.7	69.0	71.0	55.4	65.7	70.1	71.9	
Monochrome conversion	2.4	0.4	1.5	0.2	0	1.3	1.5	1.7	1.1	1.3	1.5	1.8	1.8	
Resolution reduction	74.7	0.2	66.2	0.6	0.2	56.7	66.2	70.3	71.2	56.7	66.2	70.1	72.3	
Severe compression	74.3	0.5	61.4	0.2	0.2	53.5	61.4	65.5	68.1	53.5	61.4	67.4	69.5	
Text/logo overlay	73.8	0.2	63.5	0.2	0.2	54.5	63.5	66.2	68.6	54.5	63.5	69.2	69.9	

Table 1. Detection performance comparison between EMD method and LMDS method with different k (expressed in percentages).

Max spectral radius	*N. of landmarks (n)*			
	20	50	150	250
∞	15	28	80	124
10^5	14	28	78	122
10^4	14	28	75	117

Table 2. Euclidean subspace dimension k depending on n and the maximum spectral radius.

Max spectral radius	*N. of landmarks (n)*			
	20	50	150	250
∞	34.5	24.2	10.1	6.3
10^5	37.4	24.2	10.6	6.5
10^4	37.4	24.2	11.0	6.9

Table 3. Time savings ratio (EMD method time divided by LMDS method time) depending on n and the maximum spectral radius.

dimensionality. Using larger values of n without limiting the spectral radius, on the other hand, reduces the performance because small eigenvalues result in noise amplification for the associated coordinates since the distance matrix is not Euclidean. Other than that, higher n are beneficial, though obviously more expensive. It can be seen that using a small number of dimensions still guarantees performance close to those of the real metric, hence the Dominant Color feature could be effectively used, *e.g.*, in a multi-feature CBCD system.

As for time savings, they are shown in Table 3 under the same conditions used for Table 2. Using LMDS allows to save an impressive amount of time even with a limited video data set. These figures could be even more significant for larger databases and are expected to be confirmed even with search methods smarter than exhaustive comparison.

6. CONCLUSIONS

In this paper we have shown how the MDS technique based on landmark points can be effectively employed to reduce the computational complexity of the Dominant Color feature comparison using an expensive distance like the EMD. We tested our system in a CBCD framework, under the MPEG-VST Core Experiment conditions, and we obtained significant speed-up in the detection process at the price of a minor performance degradation. We plan to use this method, possibly improved in the landmark selection side, to incorporate this and other suitable features in a multi-feature CBCD system.

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] 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.

[4] 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.

[5] M. Naphade, M. Yeung, and B. Yeo, "A novel scheme for fast and efficient video sequence matching using compact signatures," in *SPIE Conf. on Storage and Retrieval for Media Databases*, San Jose, CA, USA, Jan. 2000, pp. 564–572.

[6] 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.

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

[8] 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.

[9] Z. Xu et al., "Fast and robust video copy detection scheme using full DCT coefficients," in *IEEE Int. Conf. on Multimedia & Expo*, June 2009, pp. 434–437.

[10] T. Cox and M. A. Cox, *Multidimensional Scaling*, Chapman and Hall, London, 1994.

[11] J. Kruskal and M. Wish, *Multidimensional Scaling*, Sage Publications, Beverly Hills, CA, 1978.

[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.

[13] 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.

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

[15] V. de Silva and J. B. Tenenbaum, "Sparse multidimensional scaling using landmark points," Tech. Rep., Stanford University, 2004.