A Compact, Effective Descriptor for Video Copy Detection

Mei-Chen Yeh
Department of Electrical and Computer Engineering
University of California at Santa Barbara, CA, USA
meichen@umail.ucsb.edu

Kwang-Ting Cheng
Department of Electrical and Computer Engineering
University of California at Santa Barbara, CA, USA
timcheng@ece.ucsb.edu

ABSTRACT
Large scale video copy detection tasks require a compact and computational-efficient descriptor that is robust to various transformations that are typically applied to generate copies. In this paper, we propose a new frame-level descriptor for such a task. The descriptor encodes the internal structure of a video frame by computing the pair-wise correlations between geometrically pre-indexed blocks. It is conceptually simple, small in size, and fast to compute. Experiments using the MUSCLE VCD benchmark show its superior performance compared to existing approaches.

Categories and Subject Descriptors
I.4.7 [**Image Processing and Computer Vision**]: Feature Measurement – *feature representation*. H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering.*

General Terms
Algorithms.

Keywords
Video copy detection, frame descriptor, graph representation.

1. INTRODUCTION
With increasing bandwidth available to average users and the exploding popularity of social media, digital video availability has grown exponentially through the use of online distribution technologies such as web-TV, video blogs, and video sharing websites. To manage video contents and to protect intellectual properties, Content-Based Copy Detection (CBCD) techniques provide an alternative approach to watermarking for identifying video sequences from the same source [5, 7, 9, 11, 16]. For example, the identification of an advertisement is useful for categorizing broadcasting streams and for determining whether a video was broadcast at a suitable time and was of the correct duration.

As pointed out in [7, 9], one major challenge in CBCD arises from the fact that a copy is not necessarily an identical or a near replication, but rather a transformed video sequence. Therefore, copies can be visually dissimilar. The design of an effective descriptor is the key to a successful video copy detection system. An effective descriptor must have the following two properties: *robustness* and *discriminability*. A robust descriptor is invariant to patterns generated by the same source, while a discriminative descriptor is sensitive to patterns belonging to different video sources. Furthermore, to enable applications in large video distribution websites such as YouTube.com, a descriptor must be compact and computationally efficient to be suitable for handling a huge amount of data streams.

In this paper, we propose a new, effective frame descriptor for content-based video copy detection. This descriptor is based on the statistical properties of pair-wise correlations among partitioned grids. It encodes the intrinsic structure of a visual pattern. The 16-d compact descriptor is fast to compute and suitable for parallel implementation. Experiments using the MUSCLE VCD benchmark [8] demonstrate its superior performance compared to existing solutions.

In the remainder of the paper, we first describe related work in this field. Section 3 presents the main contribution of the paper—the spatial correlation descriptor. We then present a fast video copy detection framework in Section 4 and, finally, demonstrate its performance and conclude the paper with a short discussion summarizing our findings.

2. RELATED WORK
There are numerous descriptors for near-duplicate image or video detection available in the literature [2, 4, 7, 11, 12, 14, 16, 17]. Global statistics, such as color histograms, are widely used to efficiently work with a large corpus [2, 14, 17]. These global descriptors are, in general, efficient to compute, compact in storage, but insufficiently accurate in terms of their retrieval quality. Alternatively, local statistics, such as interest points calculated with local descriptors, were proposed in [7, 11, 12]. This description type is relatively invariant and, thus, robust to image transformations such as occlusions and cropping. However, local descriptors require more storage space and matching between them is computationally more complex. In the video domain, both global and local descriptors have been extended to incorporate temporal information [4, 16].

Law-To *et al.* presented a comparative study for video copy detection and concluded that, for small transformations, temporal ordinal measurements [4] are effective, while methods based on local features demonstrate more promising results in terms of robustness [9]. However, Thomee *et al.* conducted a large-scale evaluation of image copy detection systems and reached a somewhat different conclusion. Their chosen method that used interest points performed poorly due to its inability to find similar sets of points between copies [15]. They concluded that either a simple median method or the retina method performs the best. To design a practical copy detection system which meets the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MM'09, October 19–24, 2009, Beijing, China.
Copyright 2009 ACM 978-1-60558-608-3/09/10...$10.00.

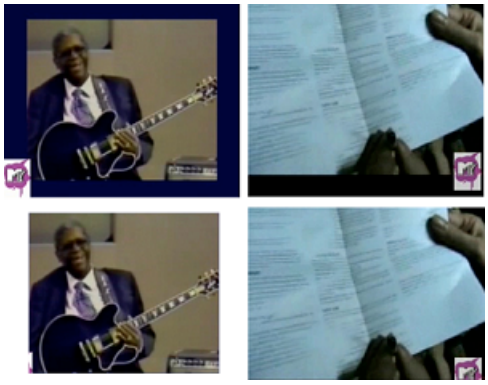

Figure 1. Examples of border removal. Top: original frames, bottom: same frames after border removal. These are sample images from the MUSCLE VCD dataset.

scalability requirements, a compact, frame-level descriptor that retains the most relevant information, instead of just sets of interest point descriptors, is desirable [12]. Furthermore, frame-level descriptors are readily integrated into fast detection frameworks such as the one presented in [12, 17].

3. DESCRIPTION SCHEME

3.1 Simple Preprocessing

Adding borders on video frames is one of the most common transformations made to a copy [9], as shown in the top row of Fig. 1. We first remove the borders using a simple, heuristic method. To determine the border color, the lightness component of the HSL color space is utilized:

$$L = \frac{\max(R,G,B) + \min(R,G,B)}{2}. \quad (1)$$

In particular, we identify the border color by inspecting the mean and the variance of the lightness values of the first few lines. A pixel is considered to lie on the border if its lightness value is less than the border color. Using this criterion, the image is scanned row-by-row starting from the top. A particular row is labeled as part of the border if it contains more than 60% border pixels. The same scanning procedure is repeated for the other sections (left, right and bottom). The bottom row of Fig. 1 shows the resulting frames of this procedure.

3.2 Spatial Correlation Descriptor

Our descriptor is based on computing the properties of a graph built from the partitioned blocks of a frame. Let's denote $G = (V, E, \Omega)$ an undirected weighted graph where V is the set of vertices, indexed by $i \in \{1,...,N\}; E \subset V \times V$ is the set of weighted edges; and Ω is the adjacency matrix:

$$\Omega(i,j) = \begin{cases} w(i,j) & if\ (i,j) \in E \\ 0 & otherwise, \end{cases} \quad (2)$$

where $w(i, j) = w(j, i)$ is the weight of the edge between vertex i and vertex j.

We first construct a graph by imposing a K by K grid on the input frame, as shown in Fig. 2 (a) for which $K = 2$. Each partitioned

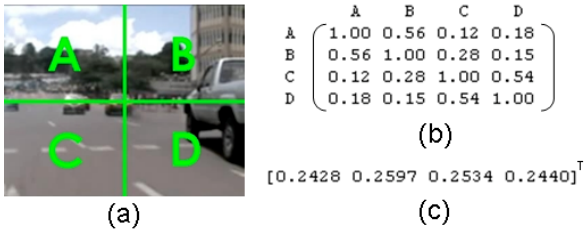

Figure 2. Illustrations of our description scheme. (a) Frame partitions; (b) The correlation matrix; (c) The descriptor.

block corresponds to a vertex in G, thus, $N = K^2$. The connection between two nodes is determined by the *content proximity* between two blocks. We observe that two copies may not share common visual properties such as colors, textures, and edges; however, they often maintain a similar inter-block relationship. For simplicity, we use the content proximity between block i (denoted as X_i) and block j (denoted as X_j) as the edge weight and define it as:

$$w(i,j) = \exp(-\frac{1}{A} D(X_i, X_j)), \quad (3)$$

where $D(X_i, X_j)$ is the sum of square differences (SSD) between block pixels[1] and A is a scaling parameter. Figure 2 (b) shows the correlation matrix of the left image.

The number of elements in the correlation matrix is quartic with respect to the partition factor K. For example, a K by K grid would generate K^2 nodes and K^2 $(K^2\text{-}1)/2$ features[2]. The dimension can be further reduced based on the spectral properties of the graph [6]. One practical method for describing the structure of a graph is to compute the stationary distribution of a random walk [6]:

$$\pi(i) = \frac{\sum_j w(i,j)}{vol\, G}, \quad (4)$$

where each element is the degree of a node divided by the volume of the graph $vol\ G = \sum_i \sum_j w(i, j)$. The stationary distribution of dimension N from the correlation matrix yields a compact, yet robust frame descriptor. Thus, the comparison of two graphs is shifted to the comparison of two corresponding distributions. In our implementation, we used the χ^2 statistics for measuring the dissimilarity between two descriptors.

Another aspect of the stationary distribution-based descriptors can be analyzed using the Markov chain models [3], where each block is treated as a state and the correlation is interpreted as the transition probability between blocks. We can transform the correlation matrix to the transition probability matrix based on the following:

$$P(i,j) = \frac{w(i,j)}{\sum_j w(i,j)}. \quad (5)$$

[1] We computed the averaged value using the CIE L*a*b* space.

[2] Elements on the diagonal are redundant because the correlation of two identical blocks is always 1.

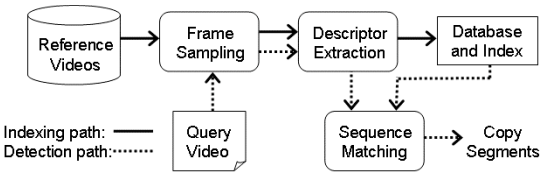

Figure 3. The video copy detection framework.

The stationary distribution π satisfies $\pi P = \pi$ and can be computed by finding the eigenvector of P^T whose corresponding eigenvalue is 1. Note that image representation based on stationary distribution has shown some success in video concept detection [10]. However, the work in [10] focuses on the extension of histogram-like representations where Markov chains are used to characterize histogram bins. In this paper, we propose a completely different graph construction.

In our experiment, we divide a frame into 4x4 blocks, resulting in a 16-d compact descriptor. This proposed descriptor, which encodes the pair-wise correlation between blocks within a frame, is referred to as the *spatial correlation* descriptor.

4. CBCD FRAMEWORK

The spatial correlation descriptor can be used for large-scale video mining [12], video retrieval [1, 2], and copy detection [5]. In this paper, we demonstrate the effectiveness of this descriptor for video copy detection. Figure 3 shows our system framework. Since a video can be naturally represented as a sequence of frames, temporal constraints should be employed in the design of metrics that compare the similarities between two videos [1, 2, 5]. In our framework, we used sequence matching techniques, or the so-called *edit distance*, for measuring video similarities.

First, a video is partitioned into a sequence of frames. For simplicity, we sampled one frame per second in our experiments. Next, each frame is summarized by our spatial correlation descriptor. Now, given two video sequences $X = [x_1, x_2, \ldots, x_m]$ and $Y = [y_1, y_2, \ldots, y_n]$, and a set of pre-defined operations (e.g. insertion, deletion, and substitution), the edit distance between X and Y is the minimal cost of applying a sequence of operations that transforms X into Y. The distance can be computed using a dynamic programming approach. We used the χ^2 statistics for comparing two spatial correlation descriptors, x_i and y_j, to determine the operation costs, as suggested in [2]. Finally, we retrieve those segments whose distance is less than a predefined threshold.

5. EXPERIMENTS

5.1 Dataset

We conducted experiments using the MUSCLE VCD benchmark [8]. This publicly available benchmark provides ground truth data for evaluating a system's detection accuracy based on two tasks: finding copies (ST1) and finding extracts (ST2). The first task evaluates a system's ability to find copies of whole videos in the database, while the second task is to detect regions of copies in the query. Both tasks are challenging because the transformations applied to this benchmark (e.g. change of color/brightness, blur, recording with an angle, inserting logos/subtitles, etc.) were very diverse.

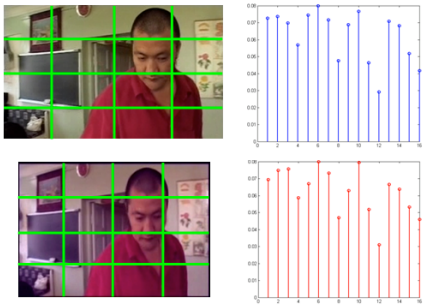

Figure 4. The robustness of the spatial correlation descriptor. Left: corresponding sampled frames in the query and the copy video; right: their descriptors. Despite the large difference in visual properties between two frames, their descriptors are quite similar.

This database consists of 101 videos with a combined length of 80 hours. These videos come from different sources—web video clips, TV archives, movies—and cover various program types including documentaries, movies, sporting events, TV shows, and cartoons. Also, the videos in this dataset have different bit-rates, resolutions, and video formats.

5.2 Detection Results

To evaluate the effectiveness of our proposed descriptor in detecting video copies, Table 1 shows the direct comparison of our results to other methods using the same performance measures. Although the spatial correlation descriptor is a global feature, it achieves the state-of-the-art performance. Figure 4 displays the spatial correlation descriptors computed from two corresponding frames sampled from the MUSCLE VCD dataset. Despite the large differences caused by color changes and vertical deformation between two frames, their descriptors are similar. Figure 5 shows the only video that failed in the ST1 task. It is apparent that flipping a frame changes the structure of the graph that pre-indexes its nodes with an order. A vertical shift causes the other failed videos in the ST2 task. However, the proposed descriptor is in general very robust to signal-based attacks.

The spatial correlation descriptor has similar merits of the ordinal intensity signature—both methods use the relationship among partitioned blocks. However, the ordinal feature solely explores the *order* of those blocks' averaged intensities while ours captures more completed and meaningful internal patterns. The idea of matching internal similarities has also been utilized in [13]. But the descriptor in [13] is computed *locally*. That is, a frame can have hundreds or thousands of those local descriptors. Furthermore, the correlation is calculated between a certain patch and its surrounding patches, while ours measures the correlation between all pairs of those patches.

Note that the descriptor can be easily extended to incorporate temporal information. The correlation matrix then records the correlation among cubes, instead of blocks, by using additional frames that are temporally adjacent to the frame under consideration.

Table 1: Accuracy on the MUSCLE VCD benchmark.

Method	ST1 score	ST2 segment score
CIVR07 Teams	0.46 ~ 0.86	0.17 ~ 0.86
Poullot *et al.* [12]	0.93	0.86
Ours	0.93	0.86

6. CONCLUSION

Frame descriptors are very crucial to video copy detection performance. In this paper, we present an effective, compact descriptor, which is conceptually simple and computationally efficient. The descriptor is constructed by encoding pair-wise correlations within a frame. Although visual properties change due to transformations, this descriptor uses the internal structure of a video frame, which makes it robust to signal-based attacks such as contrast enhancement, color changes, blurring, as well as to certain geometric attacks such as frame scaling.

There are a few directions we may explore to further enhance the descriptor. For example, the current design of our descriptor is not invariant to rotation. Furthermore, we are somewhat surprised that this descriptor is still robust despite a small amount of cropping. One interesting direction would be to investigate the use of multi-scale grids that might provide the descriptor with more robustness against severe geometric distortion.

7. REFERENCES

[1] D. A. Adjeroh, M. -C. Lee, and I. King. A distance measure for video sequence similarity matching. In *Proceedings of the International Workshop on Multi-Media Database Management Systems*, pages 72-79, 1998.

[2] M. Bertini, A. D. Bimbo, and W. Nunziati. Video clip matching using MPEG-7 descriptors and edit distance. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 133-142, 2006.

[3] L. Breiman. *Probability*. Society for Industrial and Applied Mathematics, 1992.

[4] L. Chen and F. W. M. Stentiford. Video sequence matching based on temporal ordinal measurement. *Pattern Recognition Letters*, 29(13):1824-1831, 2008.

[5] C. -Y. Chiu, C. -H. Li, H. -A. Wang, C. -S. Chen, and L. F. Chien. A Time warping based approach for video copy detection. In *Proceedings of the IEEE International Conference on Pattern Recognition*, pages 228-231, 2006.

[6] F. R. K. Chung. Spectral Graph Theory. American Mathematical Society, May 1997.

[7] A. Joly, O. Buisson, and C. Frelicot. Content-based copy retrieval using distortion-based probabilistic similarity search. *IEEE Transactions on Multimedia*, 9(2):293-306, 2007.

[8] J. Law-To, A. Joly, and N. Boujemaa. Muscle-VCD-2007: a live benchmark for video copy detection, 2007. <http://www-rocq.inria.fr/imedia/civr-bench/>.

Figure 5. The only example in the ST1 task for which our method failed. Left: query, right: reference.

[9] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, and F. Stentiford. Video copy detection: a comparative study. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 371-378, 2007.

[10] J. Li, W. Wu, T. Wang, and Y. Zhang. One step beyond histograms: image representation using Markov stationary features. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1-8, 2008.

[11] E. Maani, S. A. Tsaftaris, and A. K. Katsaggelos. Local feature extraction for video copy detection in a database. In *Proceedings of the IEEE International Conference on Image Processing*, pages 1716-1719, 2008.

[12] S. Poullot, M. Crucianu, and O. Buisson. Scalable mining of large video databases using copy detection. In *Proceedings of the ACM International Conference on Multimedia*, pages 61-70, 2008.

[13] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1-8, 2007.

[14] S. H. Srinivasan and N. Sawant. Finding near-duplicate images on the web using fingerprints. In *Proceedings of the ACM International Conference on Multimedia*, pages 881-884, 2008.

[15] B. Thomee, M. J. Huiskes, E. Bakker, and M. S. Lew. Large scale image copy detection evaluation. In *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pages 59-66, 2008.

[16] X. Wu, Y. Zhang, Y. Wu, J. Guo, and J. Li. Invariant visual patterns for video copy detection. In *Proceedings of the IEEE International Conference on Pattern Recognition*, pages 1-4, 2008.

[17] M. Yeh and K. -T. Cheng. Video copy detection by fast sequence matching. In P*roceedings of the ACM International Conference on Image and Video Retrieval*, 2009.