LOCAL FEATURE EXTRACTION FOR VIDEO COPY DETECTION IN A DATABASE

Ehsan Maani, Sotirios Tsaftaris, Aggelos K. Katsaggelos

Northwestern University

ABSTRACT

In this paper a new content-based copy identification method for video sequences is presented that is robust to a number of image transformations and particulary robust to compression artifacts. A scale and rotation invariant local image descriptor for corner points in detected key frames is proposed based on a generalized radon transform. In addition, a distance similarity metric is used that fuses intensity and geometry information to compare key frames extracted using a scene detection algorithm. Furthermore, to achieve low querying computational complexity a DP approach is employed. Experimental results demonstrate the effectiveness of the proposed approach.

1. INTRODUCTION

With the advent of media sharing web portals, peer to peer networks, and online media stores, digital rights managements has become an integral requirement to protect revenue and avoid copyright infringement litigation. Solutions on digital fingerprinting have gained a particular momentum. Watermarking based methods rely on the embedding of a signal independent (or dependent) signature in the signal that could be found in an exact or (attacked) copy of the original signal [1]. However, these methods assume the insertion of the watermark in all possible versions of the signal. On the other hand digital fingerprinting solutions rely on the extraction of a content-based digital signature from each signal and thus closely related to content-based retrieval methods (CBR)[2, 3]. To identify copies, signatures are extracted and are searched in an indexed database containing the signatures of all stored signals.

With video as a target signal, audio acoustic, joint audio-visual, or a image sequence based [2] fingerprinting approach could be adopted. In this paper an image sequence based approach is adopted.

Overall the requirements for a signature based video fingerprinting system for copyright control are:

- Small signature footprint (small file size).
- Fast signature generation for extraction and querying.
- Robustness to geometric attacks, such as rotation, scaling, translation, and cropping.
- Robustness to signal based attacks, such as gamma correction, contrast enhancement, partial occlusion, and low bit-rate compression.

In this paper a new content-based video detection (CBCD) method is presented that is robust to a number of geometric attacks and particulary robust to compression artifacts. A scale and rotation invariant local image descriptor for corner points in an image based on a generalized radon transform is proposed. In addition, a distance similarity metric is used that fuses intensity and geometric information to compare key frames extracted using a scene detection algorithm. Furthermore, to achieve low querying computational complexity a DP approach is employed.

This paper is organized as follows: In section 2, the system overview is given. Section 3 discusses the signature generation scheme emphasizing on the proposed local descriptor. Section 4 addresses the querying aspect. Section 5 presents the experimental results and discusses the performance of the system. Finally, section 6 concludes this paper and offers future extensions.

2. LOCAL FEATURE EXTRACTION SYSTEM

In general, local feature based approaches are more robust to geometric attacks but have higher complexity compares to global image features as the ones used in [5, 6]. In this work we consider a method based on local feature extraction. In this approach, extraction of the feature (fingerprints) involves three major steps:

- key frames are detected based on the mean of the frame differences (also called intensity of motion) [2].
- In each key frame, interest points (regions) are identified utilizing an improved version of the Harris detector.
- a *description* of the region of interest is computed for each interest point and stored in the database. This step is discussed in the next section in details.

During the querying process, a sequence $\mathcal{S}$ is defined. Each time a key frame is detected, its fingerprint is extracted as explained above. Then, the fingerprint of the key frame is compared against all fingerprints in the database using a DP based algorithm as described in section 4. If a match is found in the database the sequence $\mathcal{S}$ is flagged as “copied”.

3. LOCAL IMAGE DESCRIPTORS

Local photometric descriptors obtained for regions of interest have proven to be very successful in many applications such as texture recognition, image/video retrieval, video mining, and video copy detection. These local descriptors emphasize different image properties such as pixel intensities, color, texture, and edges. These descriptors are distinctive and robust to partial occlusion, cropping or translation. Furthermore, many of them are also invariant under image transformations such as scaling or rotation [7].

Depending on the application, one needs to choose among many different techniques developed for describing local image regions. These techniques include distribution based descriptors, spatial frequency technique,and differential based descriptors [7]. A simple and suited descriptor for CBCD applications is a differential based descriptor. This descriptor is formed with a set of image derivatives (local jets) computed up to a given order to approximate a point neighborhood. Nevertheless, there are two predominant drawbacks associated with derivative based descriptors. Firstly, they are not as distinctive since the derivative is only taken along two specific directions(x and y axis). Therefore, the actual change of the signal along other directions is undetermined. Secondly, the derivatives are sensitive to compression noise which can be quite large especially along the edges.

3.1. Angular intensity variation descriptor

The main contribution of this work is the new local descriptor introduced for increased robustness to compression noise. Let $f(x, y)$ denote the gray-scale image, and $\boldsymbol{p} = (x, y)$ denote an interest point(i.e., a corner or junction) in the center of the interest region with radius R. Then, the angular intensity variation (AIV) function $S(\theta)$ around the point $\boldsymbol{p}$ is defined by

$$S(\theta) = \frac{1}{R}\int_0^R f(x + r\cos(\theta), y + r\sin(\theta))dr, \quad (1)$$

where θ is a real number between 0 and 2π measured with respect to local image orientation (local gradient) for rotational invariance. The local orientation is obtained by convolution of a Gaussian gradient with the image. Note that $S(\theta)$ contains all the information on sharpness of the edges in the region of interest as well as their relative angles. In other words, $S(\theta)$ characterizes the structure of the region of interest which is invariant to image transformations. Nonetheless, for the video copy detection application, invariance to rotation is not mainly a necessity. Consequently, in this work we simply measure the angle θ with respect to the x-axis.

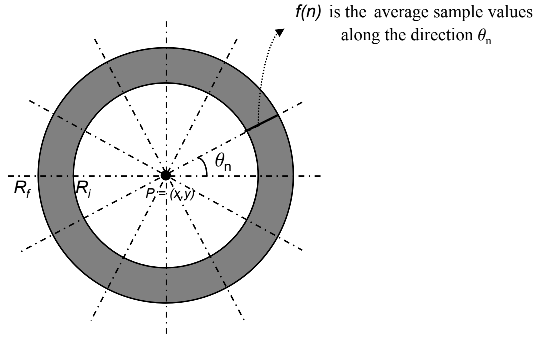


Fig. 1. Sampling around a circle centered on an interest point.

In order to account for the discrete nature of the images and also to reduce the size of the fingerprints, we employ a discrete version of equation (1)

$$\tilde{S}[n] = S(\theta_n) = \frac{1}{R_f - R_i}\sum_{r=R_i}^{R_f} f(x+r\cos(\theta_n), y+r\sin(\theta_n)), \quad (2)$$

where $\theta_n = 2\pi\frac{n}{N}$, $n = 0, .., N-1$. Here R_f is the region radius and R_i is the initial radius. Figure 1 illustrates calculation of the signature function $\tilde{S}[n]$. The value of the image function f is interpolated whenever the location of the points on the circle is not an integer.

Once the signature $\tilde{S}[n]$ for a region of interest is determined, its N-point discrete cosine transform (DCT) is evaluated. Next, for some $M < N$, a number of $N - M - 1$ higher frequency components of the DCT transform are discarded, leaving $M + 1$ of the transform coefficients. This can be justified because most of the high frequency information in the signature function is unreliable due to noise and interpolation error. The DC component is also discarded since it is simply the average of the sample values and has no useful information about the structure of the interest region. This leaves M components which form the *sub-fingerprint* $\boldsymbol{b}(\boldsymbol{p})$ for the current interest region $\boldsymbol{p}$. Since any ratio between two transform coefficients is invariant to contrast variations, i.e., $f'(x, y) = af(x, y)$, we chose to use the normalized sub-fingerprint vector $\boldsymbol{b}/||\boldsymbol{b}||$.

3.2. The center of mass technique

Traditionally, to evaluate the similarity measure between an image M and M', the signature of every keypoint $\boldsymbol{p}$ in M is compared against the signature of a number of keypoints in M'. The keypoint $\boldsymbol{p}'$ is then considered to match $\boldsymbol{p}$ if their signature distance is minimized. One weakness of this approach is that it is possible that the signature of a point on right side of the image matches the signature of a point on the left side even though they are unrelated. Therefore, is is necessary to also capture the geometry of the interest points themselves.

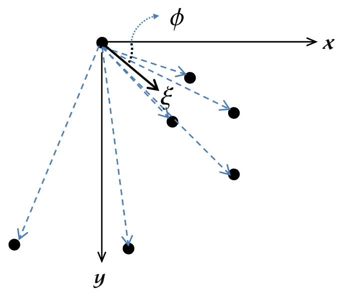


Fig. 2. Center of mass technique for capturing the geometry of the points.

Let $\boldsymbol{p}_i = (x_i, y_i)$ denote the i^{th} keypoint in the image. Then, for each point $\boldsymbol{p}_i$ we calculate a weighted average of the separation vectors $\boldsymbol{\xi}_{ij} = \boldsymbol{p}_i - \boldsymbol{p}_j$ according to

$$\bar{\boldsymbol{\xi}}_i = \frac{1}{K-1} \sum_{\substack{j=1 \\ j \neq i}}^{K} w(|\boldsymbol{\xi}_{ij}|) \boldsymbol{\xi}_{ij}, \tag{3}$$

where $w(.)$ is a monotonically decreasing function on $\Re^+$ and K is the total number of interest points in the image. Once the vector $\bar{\boldsymbol{\xi}}_i$ has been determined, its angle ϕ_i is also calculated as illustrated in Fig. 2. This angle indicates the relative position of the current keypoint $\boldsymbol{p}_i$ with respect to other keypoints. Therefore, this angle is recorded along with the corresponding sub-fingerprint $\boldsymbol{b}_i$. This process is repeated for all keypoints to determine the fingerprint of the frame.

4. MATCHING TO THE DATABASES

Let M and M' denote two key frames from the original and the query respectively. The distance between these two key frames is measured according to

$$\begin{aligned} D(M, M') = \sum_i \min_{i-d \le i' \le i+d} \rho(\boldsymbol{b}_i^M - \boldsymbol{b}_{i'}^{M'}) \\ + \kappa[1 - Cos(\phi_i^M - \phi_{i'}^{M'})], \end{aligned} \tag{4}$$

where $\boldsymbol{b}_i^M$ and ϕ_i^M denote the i^{th} local fingerprint vector and its associated angle. The function $\rho(.)$ represents a vector norm and κ is a fixed parameter. Note that the first term in equation (4) measures the similarity of the local features, while the second term measures the local *geometric similarity* of the points. Since fingerprints are always stored in a raster scan order, each local fingerprint i in M is only compared against $2d + 1$ local fingerprints in M', where d is a user-defined parameter. In this work a simple Euclidean norm is utilized for evaluation of the the vector norms in Eq. (4).

In order to determine whether a query sequence matches any of the sequences in the database, we have developed an algorithm similar to Needleman-Wunsch [8] algorithm to meet the requirements of this application. In this algorithm, for each keyframe in the query, we list all key frames in the database whose similarity distance measure is below a threshold. Then, using DP, we try to find a continuous set of key frames from a sequence among the matched frames. Figure 3 demonstrates this concept. In this figure all the images in the database that matched a key frame of the query are listed below the key frame. In this example, the DP found a *continuous* subsequence of the sequences A, C and D in the matched list. The number of consecutive frame matches from a sequence determines its matching score. Furthermore, connectors (dashed lines) are inserted between consecutive matches (solid lines) to account for a missed, extra or misplaced key frame in the query. However, they do not count towards the sequence matching score. In the example of Fig. 3, sequences A, B, C and D have matching scores of 3, 0, 4, and 2 respectively. For each query, the sequence in the database with the highest score is identified as a possible match. If the score is larger than a specific percentage (usually 50-60) of the query's length, the query is considered to have a match in the database.

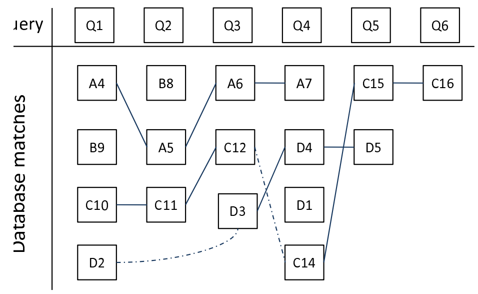
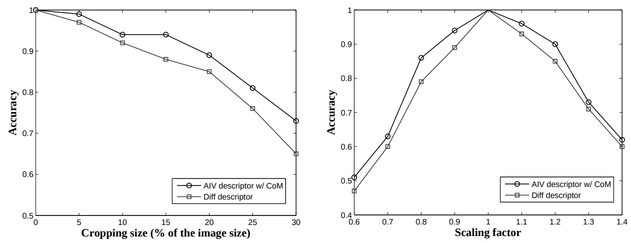


Fig. 3. Sequence matching algorithm based on DP.

5. EXPERIMENTAL RESULTS

In this section we evaluate the robustness of the proposed system to some common attacks. For this purpose, a database of over 200 hours of various video content was created. These video clips have different frame resolutions from CIF to HD 720p. Furthermore, *positive* and *negative* queries are set up for the retrieval test. A positive query is a segment from a clip known to be part of the database, while a negative query is a segment from a clip not in the database. For our experiments, we set up 500 positive and 100 negative queries with a length of 2 minutes each. The following attacks were considered: 1) low bit-rate compression 2) spatial cropping 3) spatial scaling (frame resizing). 100 video clips from the positive queries are selected for each case and re-edited to meet the requirements of each experiment. The performance of the system is evaluated by its *accuracy*, that is, the fraction of its classifications that are correct. The performance of our system is compares to a derivative based local descriptor system [2]. The descriptor contains image derivatives up to the third order (9 derivatives). Our system is set up to have 1 key frame



(a) (b)

Fig. 4. (a) Comparison of the detection accuracy of cropped video queries (b) Detection accuracy of the two systems under scaling.

for each second of video and 15 key points per key frame. A Gaussian function with standard deviation equal to $\frac{1}{8}$ of the image diagonal is considered for $w(.)$ in Eq. (2). The two parameters R_i and R_f in Eq. (2) are set to 2 and 5 pixels respectively.

Table 1 shows the accuracy of the two systems for different query qualities. The queries were generated by recompressing the original video clip in the database with different MPEG-4 quantization parameters (QP). In our first experiment (AIV-1 and Diff-1) we only considered 6 seconds of the video query for identification. However, the length of the queries were increased to 12 seconds in the second (AIV-2 and Diff-2) experiment. In AIV-3 we used the same parameters as in AIV-2 but the center of mass technique is not used to investigate its effect.

Table 1. Identification accuracy for different compressed queries.

QP	AIV-1	AIV-2	AIV-3	DIFF-1	DIFF-2
4	0.96	1.0	0.89	0.85	0.92
7	0.95	0.98	0.86	0.81	0.84
10	0.93	0.98	0.83	0.74	0.79
13	0.88	0.97	0.81	0.70	0.78
16	0.79	0.89	0.78	0.64	0.69

Figure 5 demonstrate the performance of the two systems under cropping and scaling attacks. In this experiments we used 12 seconds of the video queries for detection. The gain in the AIV system is due to the more discriminative nature of the descriptor and also the center of mass technique which reduces number of false positives. As it can be seen from the figure, both systems perform poorly when the query is scaled out of 0.8-1.2 range. This is in fact due to poor repeatability of the Harris points under scaling transformation. One resolution to this issue is to extract the fingerprints at multiple scales and store them in the database for detection.

6. CONCLUSION

We presented a new content-based copy identification method for video sequences that is robust to a number of image transformations and particularly robust to compression artifacts. A scale and rotation invariant local image descriptor for corner points in detected key frames was proposed. In addition, a distance similarity metric is used that fuses intensity and geometry information to compare key frames extracted using a scene detection algorithm. Furthermore, to achieve low querying computational complexity a DP approach is employed. The experimental results in a database consisting of more than 200 hours of video demonstrates remarkable accuracy in detecting attacked copies. At its present form our method is not invariant to rotation. However, a change in the descriptor to measure the angles from the gradient orientation will render the signature rotation invariant.

7. REFERENCES

[1] D. Simitopoulos, S. A. Tsaftaris, N. V. Boulgouris, A. Briassouli, and M. G. Strintzis, "Fast watermarking of mpeg-1/2 streams using compressed-domain perceptual embedding and a generalized correlator detector," *EURASIP J. Appl. Signal Process.*, vol. 2004, no. 1, pp. 1088–1106, 2004.

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

[3] M. Petkovic and W. Jonker, *Content-Based Video Retrieval: A Database Perspective (Multimedia Systems and Applications)*, Springer, 2003.

[4] Z. Li, L. Gao, and A. K. Katsaggelos, "Locally embedded linear subspaces for efficient video indexing and retrieval," *Multimedia and Expo, 2006 IEEE International Conference on*, pp. 1765–1768, 9-12 July 2006.

[5] A. Hampapur, K. Hyun, and R. M. Bolle, "Comparison of sequence matching techniques for video copy detection," in *Proc. SPIE*, Dec. 2001, vol. 4676, pp. 194–201.

[6] C. Mikolajczyk, K.; Schmid, "A performance evaluation of local descriptors," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[7] N. and W. , "A general method applicable to the search for similarities in the amino acid sequence of two proteins.," *J Mol Biol*, vol. 48, no. 3, pp. 443–53, 1970.