A ROBUST AND LIGHTWEIGHT FEATURE SYSTEM FOR VIDEO FINGERPRINTING

Tzu-Jui Liu[1], Hye Jung Han[1], Xin Xin[2], Zhu Li[2], Aggelos K. Katsaggelos[1]

Dept. of EECS, Northwestern University, Evanston, IL, USA

ABSTRACT

In this paper, a new content-based feature identification method for video sequences is presented. It is robust to a number of image transformations and relatively lightweight compare to most state of the art methods. A scale and rotation invariant descriptor for a set of interest points in detected key frames is proposed based on modified minimal spanning tree algorithm. In addition, a predicative coding scheme is used to achieve minimal size of the descriptor for transmission. Furthermore, the pairwise distance between the frequency responses of the curvature vector from the descriptors is calculated and compared to efficiently match query with a large database. Experimental results demonstrate the effectiveness of our approach.

Index Terms— Robust video hashing, content-based fingerprinting, multimedia fingerprinting, video copy detection

1. INTRODUCTION

With the recent trend in legislations that target copyright infringement [1] from websites that host user generated contents and peer to peer sharing services, the ability to efficiently and automatically detect videos that may contain questionable content become an integral requirement for websites to operate according to the law. Solutions on digital fingerprinting have gained a significant momentum. A common approach, watermarking based methods, relies on the embedding of signature independent/dependent of the signal that could be found in copy of the original signal [2]. However, the embedded signature can often be detected easily and removed without affecting the original content. A more complicated and computational intensive approach, content-based methods, does not have such weakness since the modified video must have contents that differ from the original video in order to avoid being detected by such method. To identify copies using content-based retrieval (CBR) methods [3, 4], signatures are extracted from the query and are searched in an indexed database containing the signatures of all stored signals.

The benefit of CBR methods for video fingerprinting is the robustness against geometric attacks and signal based attacks since many of proposed algorithms are based on well-established image retrieval techniques such as SIFT or SURF [5][6]. However, since these image retrieval techniques are designed for complex tasks that also need to compensate for the change in viewpoints and shear transformations, the size and the complexity of the descriptors are often too large to be calculated in real-time when applied to video retrieval. Small signature footprint and fast signature extraction and querying are crucial for a video fingerprinting system to be applicable in real-world situation where system resource and bandwidth for transmission are limited.

In this paper, a new lightweight and robust feature system for content-based copy detection (CBCD) is presented. A scale and rotation invariant image descriptor for a set of interest points in an image based on a modified minimal spanning tree algorithm is proposed. In addition, a predicative coding scheme is used to achieve minimal size of the descriptor for transmission. Furthermore, a predicative coding scheme in combination with tree representation algorithm is used to achieve minimal size of the descriptor for transmission. Furthermore, the pairwise distance between the frequency responses of the curvature vector from the descriptors is calculated and compared to efficiently match query with a large database.

This paper is organized as follows: In section 2, the system overview is given. Section 3 discusses the signature generation scheme with an emphasis on the proposed dominant curve 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. OVERVIEW OF THE FEATURE SYSTEM

In general, local feature based approaches are more robust to geometric attacks, but have high complexity compared to global image features. In this work, we consider a method that is based on local feature extraction technique which is then combined with essential information to form a global feature that is both lightweight and robust against transformation. In this approach, extraction of the features forming the fingerprints involves three major steps:

- Interest points are detected using Hessian-based detector [5].

- A minimal spanning tree is formed using the distance and Hessian values as weight and then reduced to the dominant curve that captures the essential information
- A compact one dimensional vector of values is formed from the curve using predicative coding for transmission

During the querying process, the curve S is interpolated and its frequency response is estimated for a pairwise comparison against all fingerprints in the database as describe in section 4. If the matching score is below a certain threshold, the curve S is flagged as copied.

3. DOMINANT CURVE IMAGE DESCRIPTORS

Local image descriptors obtained for regions of interest have proven to be very successful in many applications such as texture/object recognition, image/video retrieval, video mining, and video copy detection. These descriptors emphasize different image properties such as pixel intensities, color, and wavelet responses. These descriptors are distinctive and robust to partial occlusion, cropping, or translation. Furthermore, many of them are also invariant under affine transformations such as scaling, rotation, shear or reflection [5].

When applying these techniques to video content-based detection, the virtue of having such robust descriptors on every frame becomes a computational and transmission burden. Many affine transformations such as shear and reflection or complex cases such as partial occlusion or cropping seldom occur on video since these transformations can dramatically degrade the content. The common video transformations include strong re-encoding, scaling, brightness/contrast change and noise insertion. One key similarity among these attacks is that the relative positions of interest points in a given frame are often preserved. Exploring the geometric relationship among the interest points is the solution to a lightweight and easy to compute fingerprint system for video CBCD.

3.1. Hessian-based detector

The first step of forming the proposed dominant curve descriptor is finding a stable set of interest point that is robust against the common transformation. The widely used Harris corner detector is not scale-invariant and therefore unsuitable for this system [6]. On the other hand, the Hessian-based detector used in [5] is scale invariant and ideal for the purpose of our research. As presented in [5], given a point of *x* within a frame *I*, the Hessian determinant at scale *s* is defined as

$$\mathrm{H}(\mathbf{x},\mathbf{s}) = \begin{bmatrix} L_{xx}(\boldsymbol{x},\boldsymbol{s}) & L_{xy}(\boldsymbol{x},\boldsymbol{s}) \\ L_{xy}(\boldsymbol{x},\boldsymbol{s}) & L_{yy}(\boldsymbol{x},\boldsymbol{s}) \end{bmatrix}, \quad (1)$$

where $L_{xx}(\boldsymbol{x},\boldsymbol{s})$ is the convolution of the Gaussian second order derivative with the image I in point x, and same applies for $L_{xy}(\boldsymbol{x},\boldsymbol{s})$ and $L_{yy}(\boldsymbol{x},\boldsymbol{s})$.

As presented in [5], in order for the Hessian-based detector to be scale invariant, the Hessian filter is applied at various scales, and the non-maxima suppression algorithm is applied to obtain the most stable Hessian response across the scale-space. The value acquired through the detection method for a given interest point is positively related to its resilience against transformation attacks. In this particular feature system, the top 200 interest points are selected according to the associated Hessian value. Tests are conducted to examine how much of the 200 point set is preserved through the common transformation which includes blurring, contrast/brightness change, strong re-encoding, scaling, and noise/pattern insertion. The result shows that on average around 80% of the set is preserved, which indicates that Hessian-based detector used in [5] is very suitable for selecting stable interest points.

3.2. MST representation of the interest points

Traditionally, to evaluate the similarity between two images I and I', the signature of every interest point p in I is compared against the signature of a number of interest points in I'. The interest point p' is then considered to match p if their signature distance is minimized [4][5]. Some recent works also introduce a geometric verification stage to further improve their accuracy [7]. However, as mentioned in the previous section, methods presented in [4][5] and [7] are cumbersome to compute and require high bandwidth for transmission. Therefore, instead of using geometric information as verification, a novel MST representation of interest points that capture both visual and geometric information is incorporated in this feature system.

Given a connected and undirected graph, a spanning tree of that graph is a subgraph that has a form of a tree that connects all the vertices together. A minimum spanning tree (MST) is then a spanning tree with weight less than or equal to the weight of every other spanning tree. Given a set of stable interest points P_i with positions designated as (x_i, y_i) and the associated Hessian value h_i (and the maximum Hessian value among the set as h_{max}), a complete and undirected graph G is formed with each interest point as vertices V. The weight of edge E between vertex $V_i$and V_j is defined as

$$W(V_i, V_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} + H(V_i, V_j), \quad (2)$$

where,

$$H(V_i, V_j) = \frac{1}{2}\left(\frac{h_{max}}{h_i} + \frac{h_{max}}{h_j}\right)^2, \quad (3)$$

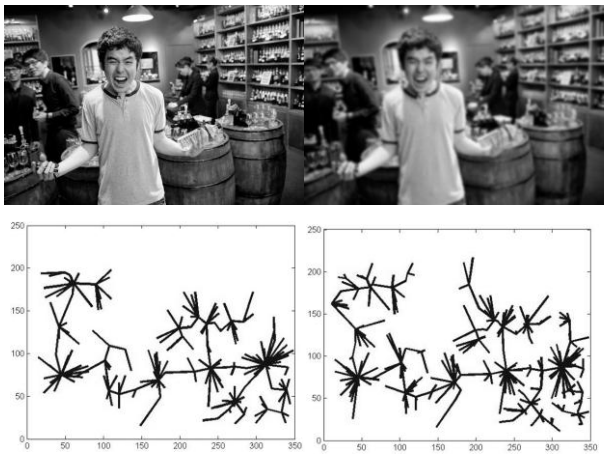

Fig. 1.(left) Original image and its associated *M*; (right) Blurred image and *M'*.

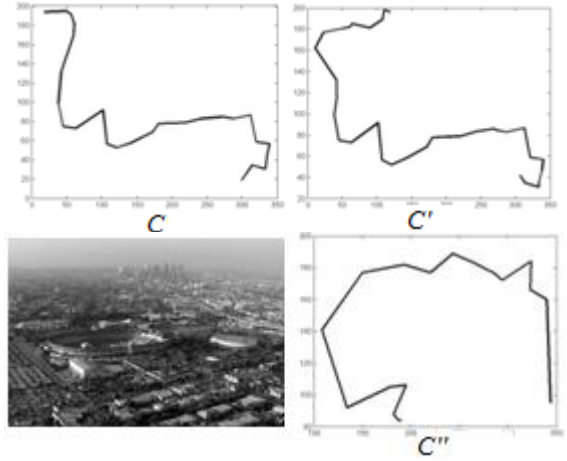

Fig. 2.(Top Left) Original image and its associated *C*; (Top right) Blurred image and *C'* from Fig. 1. (Bottom) New image and its associated *C''*.

A combination of the Euclidean distance between two vertices and the normalized summation of Hessian values is used as edge weight. With the weight defined for all edges, a MST *M* is then constructed on the complete undirected graph using Kruskal's algorithm [8]. The proposed MST representation *M* needs to be resilient against common transformation attacks, therefore, an edge should form between two stable and close by vertices. In cases where the weaker interest points (with lower Hessian value) might be left out of *G', M'* can still preserve similar interest points although they do not contain the identical information as *M*. As shown in Fig 1, the *M'* after the blurring transformation on the original image still preserve the overall structure of *M*.

3.3. Dominant curve descriptor and the encoding scheme

One key observation on the MST representation of interest points is that there is still much trivial information included in the tree diagram. For example, the two MSTs in Figure 1 show strikingly similar results in terms of the main orientation, distribution of white spaces and some characteristic shapes included in their sub-trees. These similarities need to be further amplified by trimming down the less significant branches of the tree. When trimming the tree, it is imperative to maintain the structural and geometry information of the MST. Hence, the third stage of the proposed feature building system involves forming a dominant curve *C* that captures the most important information from the MST. Our method has implemented a depth-first search algorithm to search a dominant curve in a tree. It is a blind search that initially extends a child node until it reaches the end of each branch in the MST tree. During the search process, it preserves the path that has a length greater than a preset threshold *l* with the highest average in Hessian value for each branch. It also successfully maintains the topological information of the MST tree. This approach achieves a higher efficiency in space complexity and is better suited for a heuristic method that looks for like-looking branches than other methods such as breadth-first search method. As shown in Figure 2, the proposed dominant curve *C* successfully captured the distinct characteristic of *M*. One can easily distinguish the similarity of the different curves by observing the overall shape of *C*.

Being able to use a compact coding technique such as predicative coding scheme is one of the benefits of having a curve structure instead of a set of scattered points. The encoding scheme of the dominant curve descriptor is presented as follows. First, an initial point is defined, which in this case can be either of the end points (x_0, y_0) on *C*. Secondly, the full position vector of the initial is recorded and followed by the offset $(\Delta x_{i,j}, \Delta y_{i,j})$ of the position vectors between current point*i* and the next point *j* for all *n* interest points in *C*.

$$CV = \langle x_0, y_0, \Delta x_{0,1}, \Delta y_{0,1} \ldots \Delta x_{n-1,n}, \Delta y_{n-1,n} \rangle, \quad (4)$$

Finally, the entire vector of positions CV is converted into binary string with minimal possible bits using Huffman coding scheme. The coding scheme guarantees a minimal size of descriptor for transmission, which is essential since many video uploading devices are mobile phones where the bandwidth of their networks are more restricted than that of traditional broadband network.

4. MATCHING CURVE IN FREQUENCY DOMAIN

The dominant curve descriptor is robust against some common video attacks that do not affect the positions of the interest points such as blurring, re-encoding, color adjustment or pattern insertion. The affine transformation such as rotation, scaling and shifting attacks need to be addressed at the matching stage of the proposed feature system. In order to achieve scale-invariant, rotation-invariant and shift invariant, the matching of query is composed of three main steps:

- Scale-invariant: an *n* point normalized curve is constructed by interpolating the dominant curve *C*.
- Rotation-invariant: a *K* curve is built by calculating the curvature of the normalized *C*.
- Shift-invariant: the Fourier-Transformed *K* is estimated for pair-wise comparison to the database

Since any given two curves might be composed with different number of interest points, normalizing with interpolation achieves scale-invariant and decreases the effect of cropping attack. Matching the two reconstructed dominant curves is a problem similar to challenges faced in object/pattern recognition, and the curvature approaches have gained a wide spread popularity for their robustness against rotation, scaling and translation transformation [9]. The curvature of the reconstructed dominant curve is calculated with

$$K = \frac{x'y'' - y'x''}{(x'^2 + y'^2)^{\frac{3}{2}}}, \quad (5)$$

Finally, using Fast-Fourier Transformation, shift-invariant property is achieved since any amount of shift in the time domain is transformed to a phase change in the frequency domain. A simple pair-wise comparison of the resulting vectors from the above steps determines the similarity between the two frames. The decision threshold is determined with generic experimental results from various cases.

5. EXPERIMENTAL RESULTS

5.1 Dataset and Simulation Method

In this section, the robustness of the proposed system against several challenges is evaluated. For this purpose, a database of over 15 hours of various randomly selected videos was created. These video clips have different frame resolutions from 360p to 720pHD. Furthermore, using screen capturing tools currently available on the market, key screen shots are captured at a rate of one frame per a second. Also, some artificially created filters for imitating common video attacks are applied to setup positive and negative queries for the retrieval test.

5.2 Experiment Setup

In database preparation, hessian filter was applied to all the frames extracted from videos and the top 200 interest points were selected from the hessian response as discussed in section 3.1. Weight is assigned to each edge based on the Euclidean distance between two vertices and normalized Hessian values to create MST as described in section 3.2. Then, dominant curve descriptors expressed in fast fourier transform were built for every video frame according to the algorithm described in section 3.3.

A positive query is a set of frames from a clip known to be part of the database, while a negative query is a set of frames from a clip not in the database. For our experiments, we set up 25 positive and 5 negative queries with variable number of frames in each video. The following attacks were considered:1) change in brightness; 2) change in contrast; 3) frame cropping 4) image overlay and 5) text insertion. 25 video clips from the positive queries and 5 video clips from the negative queries are selected for each case and re-edited to meet the requirements of each experiment.

Every video input was processed using the algorithms described in section 3. Each element in the resulting fourier transformed vector was compared to the database. For each pairwise comparison, weight is assigned and accumulated if there is a match, which means an error less than 1%. According to the weight assigned for each frames in videos, the name of matching video was retrieved. Also, if the number of matching frame is less than our experimentally determined threshold, the video returns a "no match."

5.3 Evaluation Methods

The performance of the system is evaluated by its *accuracy*, that is, the fraction of its classifications that are correct which is represented in following way: Accuracy$= (TP + TN)/(TN + TP + FN + FP)$. TP denotes the number of true positives, FP is the number of false positives, TN is the number of true negatives, and FN is the number of false negatives.

5.4 Results

Test Criteria	tp	fn	tn	fp	Accuracy
Increased Brightness	22	3	3	2	83%
Decreased Brightness	21	4	3	2	80%
Increased Contrast	18	7	3	2	70%
Decreased Contrast	21	4	1	4	73%
Crop (20%)	18	7	3	2	70%
Text Insertion	13	12	2	3	50%
Image overlay	2	23	1	4	10%
Image overlay-2	4	21	1	4	17%

Table1. Video matching result and its accuracy

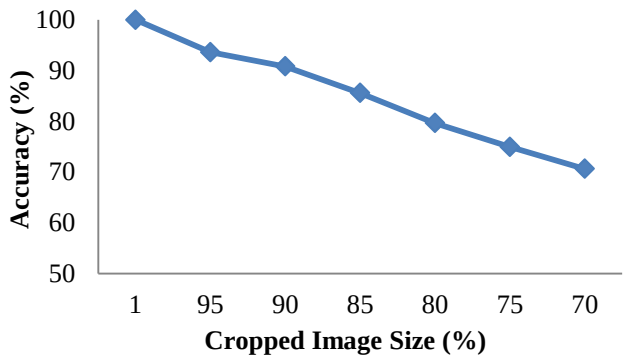

Figure 3. Relationship between the accuracy and image cropping.

Table 1 shows the accuracy of our proposed method. Among the 8 cases of common attacks applied on the video inputs, changes in the brightness did not affect the matching and the retrieval results greatly and hence, our method returned a relatively reliable result. Change in contrast is another important criterion to take account of. Although the accuracy was not as good as that of change in brightness, it still presents an accuracy above 70% which can be considered reliable. On the other hand, our method demonstrated vulnerability in the cases of image cropping, image overlaying, and text insertion. This problem occurs due to the method's nature of taking interest points from each frame. One possible solution to this problem is to find and eliminate some related interest points that repetitively appear in the video since such features are highly likely to be the source of error in video fingerprinting such as noises or text.

Figure 3 demonstrates the performance of our proposed method in relation to image cropping. The accuracy was determined based on the number of cropped frames that match with the original video frames. Clearly, the accuracy is linearly and inversely proportional to the cropped image size. Since our approach incorporates the interest points which are largely based on the content of the image, the loss of important information from cropping results in a decrease in accuracy of retrieval results.

6. CONCLUSION

We presented a new content-based copy identification method for video sequences that is lightweight, easy to compute and robust to common image transformations. A scale, rotation and shift invariant image descriptor for interest points in detected key frames was proposed. In addition, a distinct dominant curve descriptor is used that fuses interest point stability and geometry information to compare key frames extracted using a Fourier-transformed curvature algorithm. The experimental results in a database consisting of more than 15 hours of video demonstrate the reliability of our method in detecting attacked copies. Future work includes a large-scale benchmark testing with real-world data that provides more insights on the current settings for fine-tuning. In addition, many post-processing indexing techniques can be incorporated to further enhance the matching efficiency in large-scale databases. Furthermore, some preprocessing schemes can be introduced to the proposed system that removes subtitles, PiP, flipping and other complex image transformations to increase its robustness.

7. REFERENCES

[1] Lamar S. Smith, "Stop Online Piracy Act," *House Judiciary Committee*, United States House of Representatives, Washington D.C., October 26, 2011.

[2] 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.

[3] Mani Malek Esmaeili, Mehrdad Fatourechi, and Rabab Kreidieh Ward,"A Robust and Fast Video Copy Detection System Using Content-Based Fingerprinting," *IEEE Transactions on information forensics and security.*, vol. 6, no. 1, pp. 213-226, March 2011.

[4] David G. Lowe, "Object recognition from local scale-invariant features," *Proceedings of the International Conference on Computer Vision*, vol. 2, pp. 1150-1157, 1999.

[5] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool,"Speed-Up Robust Features," *Computer Vision and Image Understanding.*, vol. 110, no. 3, pp. 246-359, 2008.

[6] C. Harris and M. Stephens, "A combined corner and edge detector," *Proceedings of the Alvey Vision Conference*, pages 147-151, 1988.

[7] B. Girod, V. Chandrasekhar, D. Chen, N. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. Tsai and R. Vedantham, "Mobile Visual Search," *IEEE Signal Processing Magazine*, vol. 76, July 2011.

[8] Joseph. B. Kruskal, "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem," *Proceedings of the American Mathematical Society*, vol. 7, no. 1, pp. 48–50, 1956.

[9] Farzin Mokhtarian, Sadegh Abbasi, "Shape similarity retrieval under affine transforms," *Pattern Recognition*, vol. 35, pp. 31–41, 2002.

[10] Sunil Lee and C.D. Yoo, "Robust video fingerprinting for content-based video identification," *Circuits and Systems for Video Technology*, IEEE Transactions on, vol. 18, no. 7, pp. 984 -988, July 2008.

[11] Mani Malekesmaeili, Mehrdad Fatourechi, "Video Copy Detection Using Temporally Informative Representative Images" *International Conference on Machine Learning and Applications Vol.1* p.69-74