ROBUST VIDEO FINGERPRINTING BASED ON AFFINE COVARIANT REGIONS

Sunil Lee and Chang D. Yoo

Div. of EE, School of EECS, KAIST,
373-1 Guseong Dong, Yuseong Gu, Daejeon 305-701, Republic of Korea
sunillee@kaist.ac.kr, cdyoo@ee.kaist.ac.kr

ABSTRACT

This paper proposes a robust video fingerprinting method based on affine covariant regions. In video fingerprinting, a video clip is identified using short feature vectors referred to as fingerprints. In the proposed method, local fingerprints based on the centroid of gradient orientations are extracted from affine covariant regions detected in each frame. For the region detection, the maximally stable extremal region (MSER) detector which is considered to have high repeatability and low complexity is used. For reliable matching of the local fingerprints, only spatio-temporally consistent matches are taken into account. The experimental results show that the proposed method is robust against both geometric and non-geometric transformations.

Index Terms— Video fingerprinting, Content-based video identification, MSER, Affine covariant region, Local feature.

1. INTRODUCTION

In video fingerprinting, an unknown video clip is identified using its *fingerprints* which are short feature vectors that can uniquely characterize one video clip from another [1]. In general, a video fingerprint should be robust against various content-preserving distortions (robustness) while being discriminative so that perceptually different video clips can be distinguished (pairwise independence). Promising applications of video fingerprinting are filtering for file sharing services on the Internet, broadcast monitoring, automated indexing of a large-scale video library, etc [1]–[3].

Recently, many video fingerprinting methods have been proposed [1], [3]–[5]. The methods in the literatures used one of the following video fingerprints—the centroid of gradient orientations (CGO) [1], the differential block luminance [3], the radial hash (RASH) [4], and the spatio-temporal transform coefficients [5]. These fingerprints based on the global structure of video frames are robust against various common video processing steps including lossy compression, frame rate change, adjustment of contrast and gamma, etc. However, they are vulnerable to geometric transformations such as horizontal or vertical shift (translation), rotation, cropping, etc. One promising approach to achieve the robustness against the geometric transformations is to use *local fingerprints* which can characterize the local structures of video frames. Such approach has been used in the image retrieval literatures [6], [7], and is recently applied to video fingerprinting [8]. Joly *et al.* detected interest points from video frames and extracted the local fingerprints based on the differential decomposition of the region around the detected point [8]. Their method is quite fast and its performance is extensively evaluated using a large-scale video database (DB). However, the robustness of their method is not satisfactory for some distortions, e.g. scaling (resizing).

In this paper, a robust video fingerprinting method based on local fingerprints from affine covariant regions is proposed. In the proposed method, the CGO-based local fingerprints are extracted from the affine covariant regions in each frame. For the detection of the affine covariant regions, the maximally stable extremal region (MSER) detector [9] is used due to its reasonably high repeatability at low computational complexity [10]. For reliable matching of the local fingerprints, the parameters of geometric transformation which relates a query video to video clips in the DB are estimated, and only those matches that are spatio-temporally consistent are taken into account. The experimental results show that the proposed method is robust against various geometric and non-geometric transformations including scaling, rotation, cropping, additive noise, and adjustment of contrast and gamma.

The rest of the paper is organized as follows. Section 2 and 3 present the methods for the extraction and the reliable matching of the local fingerprints, respectively. Section 4 evaluates the performance of the proposed video fingerprinting method. Finally, Section 5 concludes the paper.

2. EXTRACTION OF LOCAL FINGERPRINTS

Fig. 1 shows the overview of the extraction of local fingerprints. First, an input video clip is resampled at a certain frame rate, and key-frames are detected. Next, affine covariant regions are detected in the chosen key-frames. Then, the detected regions are geometrically normalized, and the CGO-based local fingerprints are extracted from them. The detec-

This work was supported by grant No. R01-2007-000-20949-0 from the Basic Research Program of the Korea Science and Engineering Foundation.

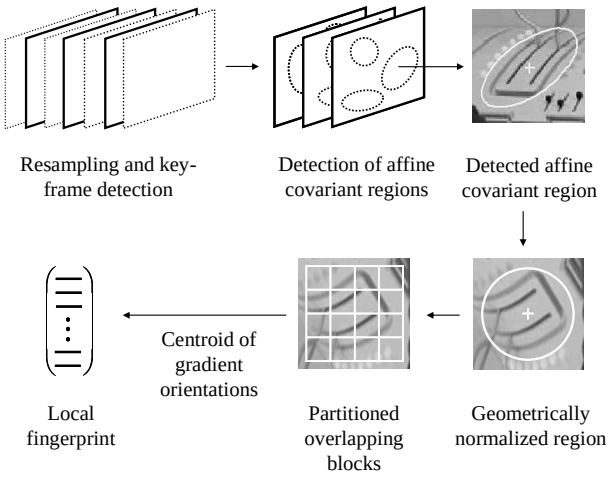

Fig. 1. Overview of extraction of local fingerprints.

tion of the affine covariant regions and the extraction of the CGO-based local fingerprints are further explained in the following sections.

2.1. MSER Detector

In the proposed method, affine covariant regions in each frame are detected by the MSER detector [9]. The MSERs are connected components of an image (each individual frame in the proposed method) where local binarization is stable over a large range of thresholds. The MSER has a number of desirable properties, e.g. invariance to affine transformation of pixel intensities, covariance to adjacency preserving (continuous) transformation, etc. The performance evaluation by Mikolajczyk *et al.* [10] also showed that the MSER detector performs better on a wide range of test sequences while requiring much less computational complexity than the other affine covariant region detectors such as Harris or Hessian-affine region detector [11].

2.2. CGO-Based Local Fingerprint

In the proposed method, the CGOs [1] are extracted and used as local fingerprints of the detected affine covariant regions. First, the detected regions are geometrically normalized and resized to the fixed size as shown in Fig. 1. Then, the normalized region is partitioned into $N \times M$ blocks which overlap each other, and the CGOs are calculated from each block.

Let $L(x, y)$ be the luminance value at location (x, y) of a geometrically normalized region. The gradient vector $[L_x\ L_y]^T$ at coordinate (x, y) is obtained by

$$L_x(x,y) = L(x,y) * G_x(x,y), \quad (1)$$
$$L_y(x,y) = L(x,y) * G_y(x,y) \quad (2)$$

where $G_x = \partial G/\partial x$ and $G_y = \partial G/\partial y$ are the partial derivatives of a Gaussian kernel G with a certain standard deviation σ. The magnitude r and the orientation θ of the gradient vector are calculated as

$$r(x,y) = \sqrt{L_x^2(x,y) + L_y^2(x,y)}, \quad (3)$$
$$\theta(x,y) = \mathrm{atan2}(L_y(x,y), L_x(x,y)) \quad (4)$$

where $\mathrm{atan2}(\cdot,\cdot)$ is the four quadrant arctangent function, thus $\theta \in [-\pi, \pi]$. The magnitude of each gradient vector is weighted by the Gaussian weighting function with standard deviation equal to one half the width of the normalized region. The purpose of the weighting is to reduce the effect of the error in the region detection in terms of both position and shape. The CGO of the block $B_{n,m}$ in the nth row and the mth column is obtained by

$$\theta_c(n,m) = \frac{\sum_{(x,y)\in B_{n,m}} r'(x,y)\theta(x,y)}{\sum_{(x,y)\in B_{n,m}} r'(x,y)} \quad (5)$$

where r' is the weighted gradient magnitude, $1 \le n \le N$, and $1 \le m \le M$. Since the weighted sum of gradient orientation is normalized by the sum of gradient magnitude, the value of the CGO also ranges from $-\pi$ to π. Finally, the local fingerprint $\mathbf{f} = \left[\theta_c(1,1)\ \theta_c(1,2) \cdots\ \theta_c(N,M)\right]^T$ is obtained as a vector of the CGOs.

3. MATCHING OF LOCAL FINGERPRINTS

Given the local fingerprints extracted from a query video, the candidate fingerprints which are close to the query fingerprints are retrieved by performing a range search on the DB which contains fingerprints and associated meta-data of a large library of video clips. Since the exhaustive search of a large-scale DB is infeasible, an efficient DB structure such as the k-d-tree [12] should be employed.

As a result of the DB search, a set of pairs of matched regions in the query video and the video clip in the DB is obtained. Let (x_q, y_q, t_q) and (x_n, y_n, t_n) be the spatio-temporal positions of the matched regions in the query video and the video clip in the DB, respectively. The goal of the fingerprint matching is to reliably estimate the time offset $t_o = t_n - t_q$. Since each frame of a video clip in the DB has a unique time code, once the time offset is obtained, the title of the query video and its temporal position in the original clip can be instantaneously identified from the meta-data in the DB. When only the temporal consistency is considered, the problem of estimating the time offset is equivalent to obtaining the model $t_n = t_q + t_o$, given a set of 2D coordinates $\{(t_q, t_n)\}$ which is a result of the DB search. To further improve the performance of the fingerprint matching, the spatial consistency should be also taken into account. In the proposed method, it is assumed that the spatial coordinates of two matched regions are related by (non-isotropic) scaling, rotation, and translation as follows

$$\begin{pmatrix} x_n \\ y_n \end{pmatrix} = \begin{pmatrix} s_x \cos\alpha & -s_y \sin\alpha \\ s_x \sin\alpha & s_y \cos\alpha \end{pmatrix} \begin{pmatrix} x_q \\ y_q \end{pmatrix} + \begin{pmatrix} d_x \\ d_y \end{pmatrix} \quad (6)$$

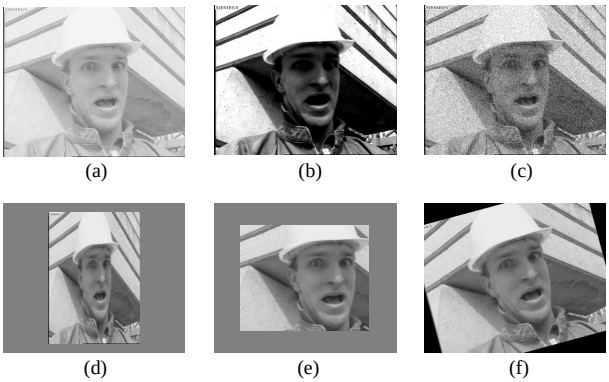

Fig. 2. Examples of the distorted frames: (a) Gamma correction with a factor of 0.5, (b) Contrast adjustment with a factor of 1.5, (c) Additive Gaussian noise with standard deviation 25, (d) Non-isotropic scaling with $(s_x, s_y) = (0.5, 0.88)$, (e) Cropping of 50%, and (f) Rotation at 15 degrees.

where (s_x, s_y), α, and (d_x, d_y) are the parameters of scaling, rotation, and translation, respectively. The time offset t_o and the parameters of the geometric transformation in (6) can be estimated using the random sample consensus (RANSAC) [13] algorithm. Based on the estimated parameters, only spatio-temporally consistent matches of the regions are taken into account for the identification of the query video.

4. PERFORMANCE EVALUATION

The performance of the proposed video fingerprinting method is evaluated using the fingerprint DB generated from 60 videos belonging to various genres, such as commercial, movie, music video, sports, news, documentary, etc. From the DB, 591 excerpts of 10 seconds long are randomly chosen and subjected to the following geometric and non-geometric distortions:

- Gamma correction with a factor of $0.5 \sim 1.5$.
- Contrast adjustment with a factor of $0.5 \sim 1.5$.
- Additive white Gaussian noise with standard deviation from 1 to 25.
- Frame rotation at angle from 1 to 15 degrees.
- Frame cropping: : $50 \sim 90\%$ of the central portion of the frame are retained while the boundaries are removed.
- Isotropic/non-isotropic resizing with factors of (s_x, s_y)= $(0.5, 0.88)$, $(0.8, 0.8)$, $(1.2, 1.2)$, and $(1, 1.76)$.

The distorted video clips are used as query videos in the experiments. Fig. 2 shows examples of the distorted frames.

In the experiment, the detected affine covariant regions are geometrically normalized and resized to 49×49. To calculate the gradient vectors, the partial derivatives of 5×5 Gaussian kernel with standard deviation $\sigma = 1$ is used. The normalized region is partitioned into 16($= 4 \times 4$) overlapping blocks, and the 16-dimensional CGO-based local fingerprint is extracted. As a DB structure, the k-d-tree is used, and two regions are declared as similar (matched) when the squared Euclidean distance between the local fingerprints from those regions is below a certain threshold, typically 0.1.

Fig. 3 shows the identification rate of the proposed video fingerprinting method for various geometric and non-geometric distortions. The symbols TC and STC in the legend denote that the corresponding results are obtained by considering the temporal consistency (TC) and the spatio-temporal consistency (STC), respectively. The identification is declared as a success when the query video is found in the DB with a temporal precision of 0 frame (TC-0 and STC-0) or $\pm$ 1 frame (TC-1 and STC-1). As shown in the figure, the proposed method achieves the identification rate higher than 95 % for all kinds of distortions when $\pm$ 1 frame error is allowed. The results also show that the proposed method performs reasonably well even when only the temporal consistency is considered, however, its performance is always improved by considering both spatial and temporal consistency. We also note that the proposed method is robust not only against the isotropic scaling but also against the non-isotropic scaling, since the affine covariant region retains its local structure under affine transformations. The overall results show that the proposed method achieves the robustness against various geometric transformations while preserving the robustness against common, non-geometric transformations.

5. CONCLUSION

In this paper, a robust video fingerprinting method based on affine covariant regions is proposed. In the proposed method, the CGO-based local fingerprints are extracted from the affine covariant regions detected by the MSER detector. The extracted local fingerprints are reliably matched to those in the DB by considering the spatio-temporal consistency among the fingerprints. The experimental results show that the proposed method is robust against various geometric and non-geometric transformations such as cropping, rotation, etc. The future work is to propose a local fingerprint with improved robustness and pairwise independence.

6. REFERENCES

[1] Sunil Lee and Chang D. Yoo, "Video Fingerprinting Based on Centroids of Gradients," In *Proc. ICASSP 2006*, Toulouse, France, vol. 2, pp. 401-404, May 2006.

[2] T. Kalker, J. A. Haitsma,and J. Oostveen, "Issues with

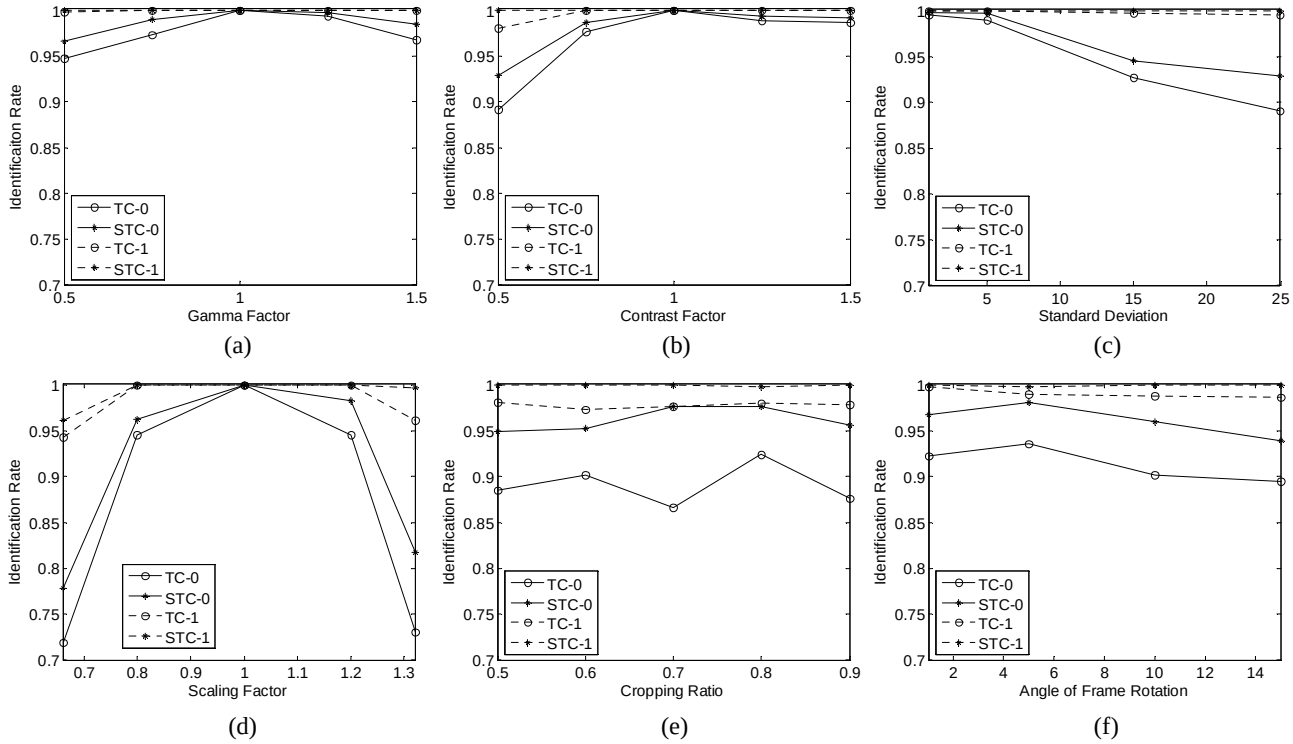

Fig. 3. Identification rate for various distortions: (a) Gamma correction, (b) Contrast adjustment, (c) Additive noise, (d) Scaling, (e) Cropping, and (f) Rotation.

digital watermarking and perceptual hashing," in *Proc. SPIE 4518, Multimedia Systems and Applications IV*, Nov. 2001.

[3] J. Oostveen, T. Kalker, and J. A. Haitsma, "Feature Extraction and a Database Strategy for Video Fingerprinting", in *Proc. International Conference on Recent Advances in Visual Information Systems*, pp. 117-128, 2002.

[4] C. D. Roover, C. D. Vleeschouwer, F. Lefebvre, and B. Macq, "Robust video hashing based on radial projection of key frames," *IEEE Trans. Signal Processing*, vol. 53, no. 10, pp. 4020-4037, Oct. 2005.

[5] B. Coskun, B. Sankur, and N. Memon, "Spatio-temporal transform based video hashing," *IEEE Trans. Multimedia*, vol. 8, no. 6, pp. 1190-1208, Dec. 2006.

[6] C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 530-535, 1997.

[7] V. Monga and B. L. Evans, "Perceptual image hashing via feature points: Performance evaluation and tradeoffs," *IEEE Trans. Image Processing*, vol. 15, no. 11, pp. 3453-3466, Nov. 2006.

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

[9] J. Matas, O. Schum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," in *Proc. Brit. Mach. Vision Conf.*, pp. 384-393, 2002.

[10] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 1/2, pp. 43-72, 2005.

[11] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *Int. J. Comput. Vis.*, pp. 63-68, 2004.

[12] Robinson J. T., "The k-d-b-tree: A Search Structure for Large Multidimensional Dynamic Indexing," In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, pp. 10-18, 1981.

[13] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381-395, 1981.