ROBUST VIDEO FINGERPRINTING BASED ON 2D-OPCA OF 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 2-Dimensional Oriented Principal Component Analysis (2D-OPCA) of affine covariant regions. The goal of video fingerprinting is to identify a video clip using perceptual features called fingerprints. In the proposed method, to achieve the robustness against geometric transformations, fingerprints are extracted from local regions covariant with a class of affine transformations. The detected affine covariant regions are normalized geometrically and photometrically, and local fingerprints are extracted by applying a novel discriminant analysis algorithm, 2D-OPCA to the normalized regions. For the reliable matching of 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, 2D-OPCA, MSER, Affine covariant region, Local fingerprint.

1. INTRODUCTION

The goal of video fingerprinting is to identify an unknown video clip using its *fingerprints* which are perceptual features that uniquely characterize one video clip from another [1]. The fingerprints 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 large-scale video archives, content authentication, etc [1].

The conventional video fingerprinting methods, e.g. [1] and [2], 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., since video fingerprints used in the conventional methods are based on the global structure of video frames. 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 [3], and is recently applied to video fingerprinting [4].

In this paper, a robust video fingerprinting method based on the *2-Dimensional Oriented Principal Component Analysis* (2D-OPCA) of affine covariant regions is proposed. An affine covariant region [5] is a connected region covariant with a class of affine transformations. The detected affine covariant regions are normalized geometrically and photometrically, and local fingerprints are extracted from the normalized region using a novel discriminant analysis algorithm, 2D-OPCA. Contrary to the conventional OPCA [6], the proposed 2D-OPCA can be directly applied to 2D local regions without transforming them into 1D vectors, thus can implicitly avoid undersample or singularity problem [7]. 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 spatio-temporally consistent matches are taken into account.

The rest of this paper is organized as follows. Section 2 describes the proposed 2D-OPCA algorithm. Section 3 and 4 present the methods for the extraction and the reliable matching of the local fingerprints, respectively. Section 5 evaluates the performance of the proposed method. Finally, Section 6 concludes the paper.

2. 2D-OPCA ALGORITHM

The OPCA [6] is a generalization of the standard PCA algorithm. In this paper, the 2D-OPCA, which can be conducted directly on 2D matrices, is proposed. Consider two $h \times w$ random matrices X and Z whose elements have zero mean. Let $\mathcal{L}$ and $\mathcal{R}$ be the spaces spanned by $\{\mathbf{u}_i\}_{i=1}^{h'}$ and $\{\mathbf{v}_i\}_{i=1}^{w'}$ where $\mathbf{u}_i \in \mathbb{R}^h$ and $\mathbf{v}_i \in \mathbb{R}^w$, respectively. Then, the projection of X onto the $(h' \times w')$-dimensional space $\mathcal{L} \otimes \mathcal{R}$, which is the tensor product of two spaces $\mathcal{L}$ and $\mathcal{R}$, is given as

$$Y = L^T X R \in \mathbb{R}^{h' \times w'} \tag{1}$$

where $L = [\mathbf{u}_1, \cdots, \mathbf{u}_{h'}] \in \mathbb{R}^{h \times h'}$ and $R = [\mathbf{v}_1, \cdots, \mathbf{v}_{w'}] \in \mathbb{R}^{w \times w'}$. The goal of the 2D-OPCA is to find the projection matrices L and R which maximizes the projection energy of X while simultaneously minimizing the projection energy of Z.

Unlike the classical OPCA, there is no closed-form solutions for L and R. Instead, we tackled the problem of the 2D-OPCA by using the following 2-layer approach. In the first layer, only the projection by L is considered as follows

$$Y_1 = L^T X \in \mathbb{R}^{h' \times w}. \tag{2}$$

When the Frobenius norm is used, the optimal projection matrix L should maximize $J_1(\cdot)$ given by

$$J_1(L) = \frac{E[||L^T X||_F^2]}{E[||L^T Z||_F^2]} = \frac{\text{trace}\{L^T C_X L\}}{\text{trace}\{L^T C_Z L\}} \tag{3}$$

where $C_X = E[XX^T]$ and $C_Z = E[ZZ^T]$. The solution to (3) can be obtained by solving the following generalized eigenvalue problem

$$C_X \mathbf{u} = \lambda C_Z \mathbf{u}. \tag{4}$$

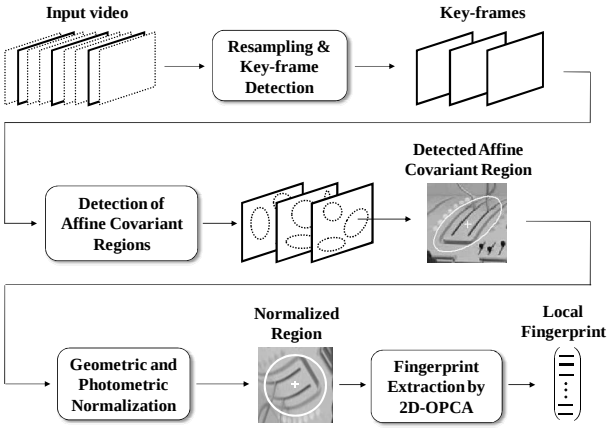

Fig. 1. Overview of extraction of local fingerprints.

The column vectors of L are obtained as the eigenvectors of $C_X^{-1}C_Z$ or $C_Z^{-1}C_X$ corresponding to the h' largest eigenvalues.

In the second layer, the projection of Y_1 by R is considered:

$$Y_2^T = R^T Y_1^T \in \mathbb{R}^{h' \times w'} \tag{5}$$

The optimal projection matrix R must maximize $J_2(\cdot)$ given by

$$J_2(R) = \frac{\text{trace}\{R^T C_X^L R\}}{\text{trace}\{R^T C_Z^L R\}} \tag{6}$$

where $C_X^L = E[X^T L L^T X]$ and $C_Z^L = E[Z^T L L^T Z]$. The column vectors of R are obtained as the eigenvectors of $(C_X^L)^{-1}C_Z^L$ or $(C_Z^L)^{-1}C_X^L$ corresponding to the w' largest eigenvalues.

3. EXTRACTION OF LOCAL FINGERPRINTS

Fig. 1 shows the overview of the extraction of local fingerprints. First, an input video is resampled at a fixed frame rate, and key-frames are detected. In the proposed method, a frame is declared as a key-frame when its global intensity of motion corresponds to local extrema as in [8]. Next, affine covariant regions are detected in each key-frame using the maximally stable extremal region (MSER) detector [9]. Then, each detected MSER is normalized geometrically and photometrically to obtain the normalized local region X. Finally, the dimension of X is reduced using the projection obtained by the 2D-OPCA as in (1), and the resulting principal oriented component Y is obtained as a local fingerprint of the region.

3.1. Detection and Normalization of Affine Covariant Regions

In the proposed method, affine covariant regions in each key-frame are detected using the MSER detector [9]. The MSERs are connected components of an image where local binarization is stable over a large range of thresholds. The MSER has a number of desirable properties, e.g. invariance to common photometric changes, covariance to adjacency preserving transformation, etc. The performance evaluation by Mikolajczyk *et al.* [5] showed that the MSER detector performs better for various test sequences while requiring much less computational complexity than the other detectors. To recover the geometric transformations applied to the regions, the detected regions are geometrically normalized prior to the fingerprint extraction. In the proposed method, the regions are normalized using the square root of their second moment matrix of the intensity gradient [10] and the dominant orientation [11]. The geometrically normalized regions are also photometrically normalized by equalizing their histogram.

3.2. Extraction of Local Fingerprints Using 2D-OPCA

In the proposed method, the projection matrices $L \in \mathbb{R}^{h \times h'}$ and $R \in \mathbb{R}^{w \times w'}$ are obtained as follows. Let $\{X_i\}_{i=1}^n$ be a set of $h \times w$ matrices $X_i \in \mathbb{R}^{h \times w}$, where each X_i represents a normalized local region from an original (undistorted) video. Suppose that for each X_i, a set of m distorted versions $\{\tilde{X}_i^j\}_{j=1}^m$ is available, and define the corresponding difference (distortion, noise) matrices $Z_i^j = \tilde{X}_i^j - X_i$. In the proposed method, the original and the following $m = 9$ sets of local regions are used as training data to obtain the projection matrices: A) additive white Gaussian noise with standard deviation 25, B) brightness $\uparrow$ 30%, C) brightness $\downarrow$ 30%, D) Gaussian blurring with radius 2 pixels, E) contrast with factor 0.5, F) contrast with factor 1.5, G) lossy compression (DivX at 256 kbps), H) gamma correction with $\gamma = 0.5$, and I) gamma correction with $\gamma = 1.5$. Then, the 2D-OPCA algorithm is applied to $\{X_i\}$ and $\{Z_i^j\}$. The matrices C_X and C_Z in (3) are obtained by

$$C_X = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T, \tag{7}$$

$$C_Z = \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}(Z_i^j - \bar{Z})(Z_i^j - \bar{Z})^T \tag{8}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\bar{Z} = \frac{1}{nm}\sum_{j=1}^m\sum_{i=1}^n Z_i^j$. Given the projection matrix L obtained in the first layer, the matrices C_X^L and C_Z^L in (6) are obtained by

$$C_X^L = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^T L L^T (X_i - \bar{X}), \tag{9}$$

$$C_Z^L = \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}(Z_i^j - \bar{Z})^T L L^T (Z_i^j - \bar{Z}). \tag{10}$$

The dimension of a normalized region X can be effectively reduced using the projection matrices L and R as in (1), and Y is used as a local fingerprint of the region. The obtained projection matrices L and R maximize the projection energy of the original signal X while simultaneously minimizing the projection energy of the distortion Z. Thus the resulting fingerprint Y is expected to preserve the structure of the original signal while being robust against the distortions in the training data. The extraction of local fingerprint using the 2D-OPCA is also computationally efficient since what is required for the extraction is only the simple matrix multiplication.

4. MATCHING OF LOCAL FINGERPRINTS

The distance between two local fingerprints Y_1 and Y_2 are measured using the Frobenius norm given by

$$D(Y_1, Y_2) = \frac{1}{h'w'}||Y_1 - Y_2||_F^2 = \frac{1}{h'w'}\sum_{i=1}^{h'}\sum_{j=1}^{w'}\{Y_1(i,j) - Y_2(i,j)\}^2. \tag{11}$$

Given local fingerprints extracted from a query video, the candidate fingerprints close to the query fingerprints are retrieved by

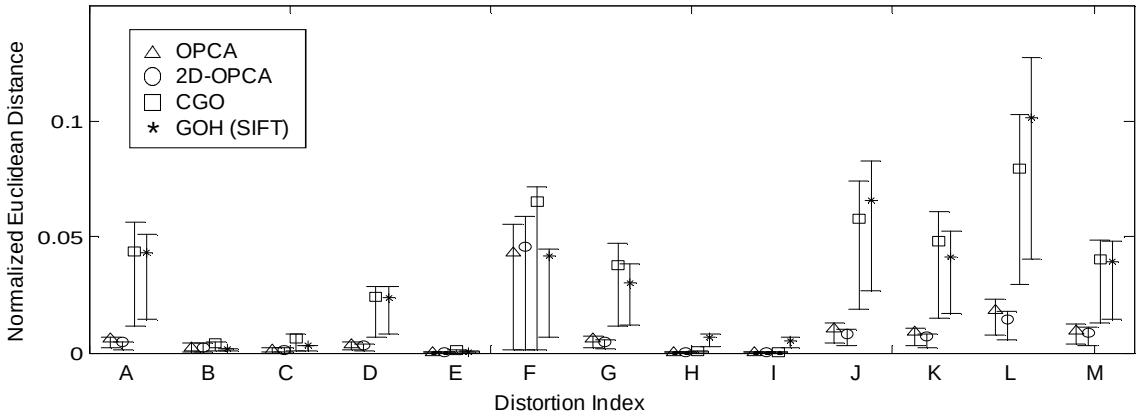

Fig. 2. OPCA, 2D-OPCA, CGO, and GOH mean Euclidean distances, and upper and lower quartiles for various distortions. Among the considered distortions, 9 (A∼I) distortions are included in the training data, while 4 (J∼M) distortions are not.

performing a range search on the DB which contains fingerprints and associated metadata of a large library of video clips. 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_o, y_o, t_o) 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 $t_d = t_o - t_q$, since each frame of a video clip in the DB has a unique time code. When only the temporal consistency is considered, the problem of obtaining the time offset is equivalent to estimating the model $t_o = t_q + t_d$, given a set of 2D coordinates $\{(t_q, t_o)\}$ 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 an affine transformation including scaling, rotation, and translation as follows

$$\begin{pmatrix} x_o \\ y_o \\ t_o \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} & 0 \\ h_{21} & h_{22} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_q \\ y_q \\ t_q \end{pmatrix} + \begin{pmatrix} x_d \\ y_d \\ t_d \end{pmatrix}. \quad (12)$$

The time offset t_o and the parameters of the geometric transformation in (12) can be estimated using the random sample consensus (RANSAC) [12] 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.

5. PERFORMANCE EVALUATION

In the experiments, the detected affine covariant regions are normalized to 49×49 matrices ($h = w = 49$). The average number of key-frames detected per one second is 1 (T-1 and ST-1) or 2 (T-2 and ST-2), respectively. As explained in Section 3.2, a set of original local regions and its 9 distorted versions are used as training data for the 2D-OPCA. Each set contains 17,816 regions detected from various video clips. In the experiments, $h' = w' = 4$ principal oriented components are used as a fingerprint. The k-d-tree [13] is used as the DB indexing structure, and two regions are declared as similar when the squared Euclidean distance between the local fingerprints from those regions is below a certain threshold, typically 0.1.

The performance of the proposed fingerprint is measured and compared with that of the OPCA-based fingerprint and the fingerprints based on the centroid of gradient orientations (CGO) [1] and the gradient orientation histogram (GOH) of the scale invariant feature transform (SIFT) [11]. For all the considered fingerprints, the average Euclidean distance between the fingerprints extracted from original and distorted video clips is calculated. The performance is evaluated for 13 distortions which include 4 distortions that are not included in the training data: J) resizing to CIF + lossy compression (DivX at 256kbps) + histogram equalization, K) resizing to CIF + color variation (red $\uparrow$ 20%, green $\downarrow$ 10%, blue $\uparrow$ 5%) + frame rate change from 24 to 20 fps + contrast $\uparrow$ 30% + lossy compression (DivX at 256kbps), L) resizing to CIF + lossy compression (DivX at 256kbps) + Gaussian blurring with radius 2 pixels, and M) sharpening. Each of original and distorted sets includes 26,379 local regions from various video clips, and they are used as the testing data. The experimental results are shown in Fig. 2. For each fingerprint, all the Euclidean distances are scaled such that the mean distance between two different local fingerprints is scaled to unity. The results show that the distance of the proposed 2D-OPCA-based local fingerprint is lower than that of the others, which means that the proposed fingerprint outperforms the OPCA-based, the CGO-based, and the GOH-based fingerprints not only for the training distortions but also for the distortions which are not considered during the training.

The performance of the proposed video fingerprinting method is also evaluated in terms of the identification rate using the fingerprint DB generated from 13 movies belonging to various genres. The resolution and the frame rate of the movies are 640×272 and 24 fps, respectively, and the total length of the movies is approximately 29 hours. From the DB, 591 10-seconds-long excerpts are randomly chosen and subjected to various geometric and non-geometric distortions. Fig. 3 shows the identification rate of the proposed video fingerprinting method for the considered distortions. The symbols T and ST in the legend denote that the corresponding results are obtained by considering the temporal consistency (T) and the spatio-temporal consistency (ST), respectively. The identification is declared as a success when the query video is found in the DB with a temporal precision of $\pm$ 1 frame. As shown in the figures, the proposed method achieves the identification rate higher than 95 % for all the distortions when the spatio-temporal consistency is considered. 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-

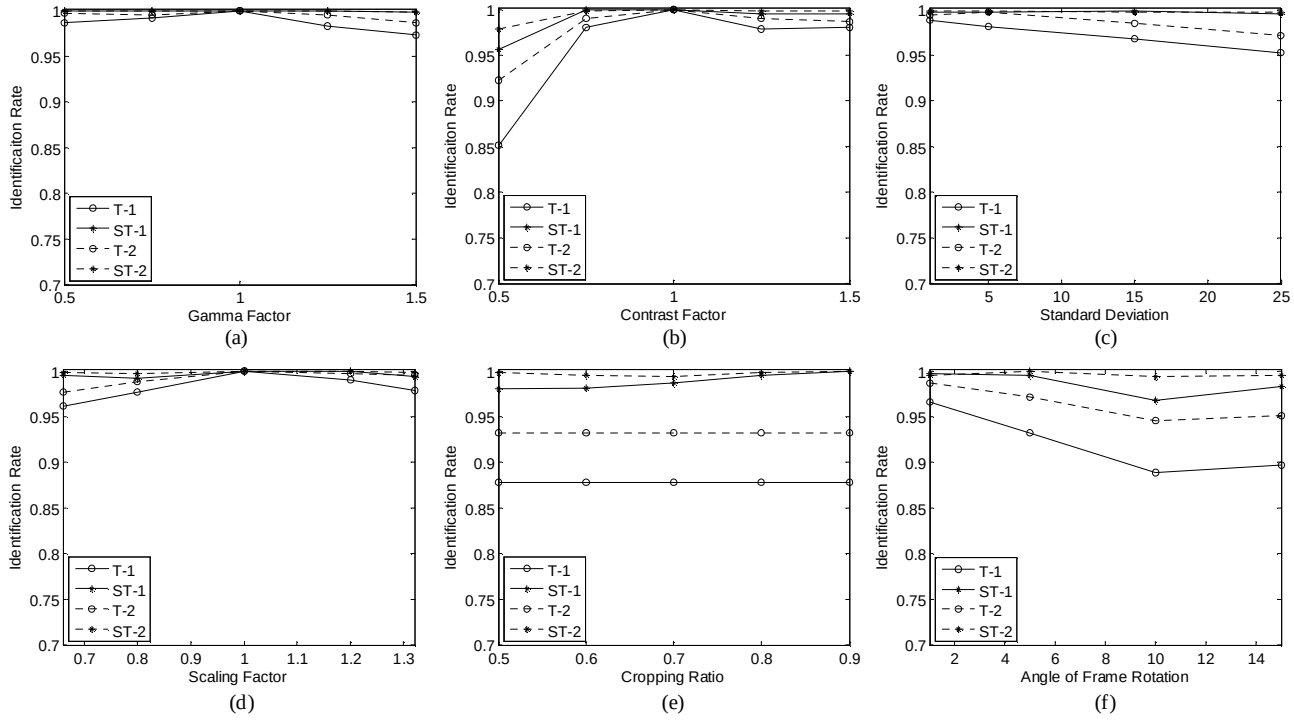

Fig. 3. Identification rate for various distortions: (a) Gamma correction with a factor of $0.5 \sim 1.5$, (b) Contrast adjustment with a factor of $0.5 \sim 1.5$, (c) AWGN with standard deviation from 1 to 25, (d) Isotropic/non-isotropic scaling, (e) Cropping, and (f) Rotation at angle from 1 to 15 degrees.

isotropic scaling, since the affine covariant region retains its local structure under affine transformations.

6. CONCLUSION

In this paper, a robust video fingerprinting method based on 2D-OPCA of affine covariant regions is proposed. In the proposed method, the 2D-OPCA-based local fingerprints are extracted from the affine covariant regions detected by the MSER detector. The extracted local fingerprints are reliably matched to the fingerprints in the DB by considering the spatio-temporal consistency. The experimental results show that the proposed method is robust against various geometric and non-geometric transformations.

7. REFERENCES

[1] Sunil Lee and Chang D. Yoo, "Robust Video Fingerprinting for Content-Based Video Identification," *IEEE Trans. Circuits and Systems for Video Technology*, Accepted for pulication.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] K. Diamantaras and S. Kung, *Principal Component Neural Networks*, John Wiley, 1996.

[7] W. J. Krzanowski, P. Jonathan, W. V. McCarthy, and M. R. Thomas, "Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data," *Applied Statistics*, 44:101-115, 1995.

[8] S. Eickeler and S. Muller, "Content-based video indexing of TV broadcast news using hidden markov models," in *Proc. ICASSP 1999*, Phoenix, AZ, USA, pp. 2997-3000, Mar. 1999.

[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] A. Baumberg, "Reliable feature matching across widely separated view," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, South Carolina, USA, pp. 774-781, 2000.

[11] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. ICCV*, pp. 1150-1157, 1999.

[12] 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.

[13] J. T. Robinson, "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.