Robust Video Fingerprinting for Content-Based Video Identification

Sunil Lee, *Member, IEEE*, and Chang D. Yoo, *Member, IEEE*

***Abstract*—Video fingerprints are feature vectors that uniquely characterize one video clip from another. The goal of video fingerprinting is to identify a given video query in a database (DB) by measuring the distance between the query fingerprint and the fingerprints in the DB. The performance of a video fingerprinting system, which is usually measured in terms of pairwise independence and robustness, is directly related to the fingerprint that the system uses. In this paper, a novel video fingerprinting method based on the centroid of gradient orientations is proposed. The centroid of gradient orientations is chosen due to its pairwise independence and robustness against common video processing steps that include lossy compression, resizing, frame rate change, etc. A threshold used to reliably determine a fingerprint match is theoretically derived by modeling the proposed fingerprint as a stationary ergodic process, and the validity of the model is experimentally verified. The performance of the proposed fingerprint is experimentally evaluated and compared with that of other widely-used features. The experimental results show that the proposed fingerprint outperforms the considered features in the context of video fingerprinting.**

***Index Terms*—Content-based video identification, perceptual video hashing, video fingerprinting.**

I. Introduction

IN THE LAST decade, the amount of video contents digitally produced, stored, distributed, and broadcasted has grown enormously. The proliferation of digital videos has made accessibility of video contents much easier and cheaper while being the source of many problems, e.g., the illegal distribution of copyrighted movies via file sharing services on the Internet. The problems associated with digital videos require an efficient method for protecting, managing, and indexing video contents. Among various solutions to these problems, fingerprinting, which is also known as perceptual hashing or content-based media identification, is receiving increased attention [1]. Fingerprints are perceptual features or short summaries of a multimedia object, and the goal of fingerprinting is to provide fast and reliable methods for content identification [1], [2]. Specifically, video fingerprints are feature vectors that uniquely characterize one video clip from another [3], and the goal of video fingerprinting is to identify a given video query in a database (DB) by measuring the distance between the query fingerprint and the fingerprints in the DB. Promising applications of video fingerprinting are filtering for file-sharing services, broadcast monitoring, automated indexing of large-scale video archives, etc.

Fig. 1. Overall structure of the proposed video fingerprinting method.

Video fingerprints should be carefully chosen since they directly affect the performance of the entire video fingerprinting system. In general, the video fingerprints need to satisfy the following properties [1]–[3].

- **Robustness** (invariance under perceptual similarity): Fingerprints extracted from a video clip subjected to content-preserving distortions should be similar to the fingerprints extracted from the original video clip.
- **Pairwise independence** (collision free): If two video clips are perceptually different, the fingerprints extracted from them should be considerably different.
- **Database search efficiency**: For applications with a large-scale DB, fingerprints should be conducive to efficient DB search.

Fig. 1 shows the overall structure of the proposed video fingerprinting method which consists of three parts: 1) fingerprint extraction; 2) DB search; and 3) fingerprint matching. In the fingerprint extraction, video fingerprints based on the *centroid of gradient orientations* are extracted from an unknown video clip to be identified. In the DB search, a range search is performed to find the candidate fingerprints for matching. The DB includes fingerprints from a large library of video clips and the corresponding metadata such as the video title. To retrieve candidates quickly, an efficient indexing structure such as k-d-tree [4] needs to be employed. However, since the focus of this paper is on the fingerprint extraction and matching, DB search algorithms are not explained in detail. Finally, in the fingerprint matching, the query fingerprints are exhaustively searched among the candidates found in the DB search, and the metadata associated with the candidate closest to the query fingerprints is declared as the fingerprinting result. A threshold used to reliably determine a fingerprint match is theoretically derived by modelling the proposed fingerprint as a stationary ergodic process, and the validity of the model is experimentally verified.

The rest of the paper is organized as follows. Sections II and III describe the fingerprint extraction and the fingerprint matching parts of the proposed video fingerprinting method, respectively. Section IV evaluates the performance of the proposed fingerprinting method. Finally, Section V concludes the paper.

Manuscript received February 15, 2007; revised August 27, 2007. This work was supported by the Brain Korea 21 Project, the School of Information Technology, KAIST. This paper was recommended by Associate Editor Q. Sun.

The authors are with the Division of Electrical Engineering, School of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea (e-mail: sunillee@kaist.ac.kr, cdyoo@ee.kaist.ac.kr).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2008.920739

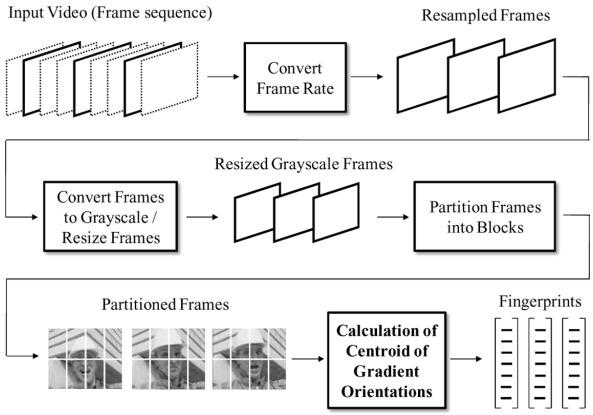

Fig. 2. Overall procedure of the proposed video fingerprint extraction.

II. Fingerprint Extraction

A. Overall Procedure of Fingerprint Extraction

Fig. 2 shows the overall procedure of the proposed video fingerprint extraction. In the first step, an input video is resampled at a fixed frame rate S frames per second (fps) to cope with frame rate change. In the second step, each resampled frame is converted to grayscale to make the proposed fingerprinting method robust against the color variation and applicable not only to color video clips but also to classic black-and-white films. In the third step, each grayscale frame is resized so that its width and height are normalized to the fixed values X and Y, respectively. This step makes the proposed fingerprinting method robust against resizing of an arbitrary factor. In the fourth step, each resized frame is partitioned into a grid of N rows and M columns, resulting in $N \times M$ blocks. Finally, the centroid of gradient orientations is calculated for each of these blocks, and an (NM)-D fingerprint vector is obtained for each frame.

B. Centroid of Gradient Orientations

Let $f[x, y, k]$ be the luminance value at location (x, y) in the kth frame. The gradient of f at coordinates (x, y) is defined as the vector

$$\nabla \mathbf{f} = [G_x \quad G_y] = \left[\frac{\partial f}{\partial x} \quad \frac{\partial f}{\partial y}\right]. \tag{1}$$

The gradient vector points in the direction of maximum rate of change of f at coordinates (x, y) [5]. In the proposed method, the partial derivatives G_x and G_y are approximated as follows:

$$G_x = f[x+1, y, k] - f[x-1, y, k] \tag{2}$$

$$G_y = f[x, y+1, k] - f[x, y-1, k]. \tag{3}$$

The gradient vector $\nabla \mathbf{f}$ can also be represented as its magnitude $r[x, y, k]$ and orientation $\theta[x, y, k]$ which are given by

$$r[x, y, k] = \sqrt{G_x^2 + G_y^2} \tag{4}$$

$$\theta[x, y, k] = \tan^{-1}\left(\frac{G_y}{G_x}\right). \tag{5}$$

In the proposed fingerprinting method, the following value called the centroid of gradient orientations is obtained from each block:

$$c[n, m, k] = \frac{\sum_{(x,y)\in B_{n,m,k}} r[x, y, k]\theta[x, y, k]}{\sum_{(x,y)\in B_{n,m,k}} r[x, y, k]} \tag{6}$$

where $B_{n,m,k}$ is the block in the nth row and the mth column of the kth frame and $c[n, m, k]$ is the centroid obtained from the block $B_{n,m,k}$ $(1 \le n \le N, 1 \le m \le M)$. Due to the normalization by the sum of gradient magnitudes, the centroid has a value between $-(\pi/2)$ and $(\pi/2)$. The (NM)-D fingerprint vector $\mathbf{c}_k$ of the kth frame is obtained by

$$\mathbf{c}_k = [\, c[1,1,k] \quad c[1,2,k] \quad \ldots \quad c[N,M,k] \,]. \tag{7}$$

The gradients from which the proposed fingerprint is obtained are closely related to the distribution of edges which provide relevant information about visual content of video frames, e.g., object boundaries [6]. Since the gradients are based not on the pixel values but on the pixel differences, the proposed fingerprint is automatically robust against global change in pixel intensities such as brightness, color, and contrast. Although nonlinear operations such as gamma correction are known to cause a large change in relative magnitudes for some gradients, the proposed fingerprint is still robust against nonlinear operations since they are less likely to affect the gradient orientations [7].

The gradient-based features have been used as a descriptor which represents local image regions [7] and also as a video fingerprint [8]. Lowe used the histogram of gradient orientations as a local descriptor which characterizes a region around the detected interest points [7]. The comparative test in [9] shows that Lowe's local descriptor based on gradients outperforms other local descriptors. However, the high dimensionality of Lowe's descriptor renders the histogram of gradient orientations unsuitable for video fingerprinting. Hampapur and Bolle used the centroid of gradient magnitudes as a video fingerprint along with dominant color [8]. Since they extract the fingerprints only from chosen key-frames, the high-D fingerprint had to be used to maintain pairwise independence. However when the fingerprints are extracted from every resampled frame as in the proposed method, the fingerprint with lower dimension must be used. The proposed video fingerprint based on the centroid of gradient orientations achieves good robustness and pairwise independence at reasonably low dimension. The performance of the proposed fingerprint and that of the gradient-based features explained above are compared in Section IV-D, and the comparison results show that the proposed fingerprint outperforms other gradient-based features in the context of video fingerprinting.

III. Fingerprint Matching

In the DB search, given K fingerprints from the query video clip, the candidate fingerprints for the matching are found by performing a range search on the DB. However, a single fingerprint with low dimension is not sufficient for a reliable matching.

To alleviate this problem, in the proposed method, a *fingerprint sequence* is generated by concatenating the fingerprints extracted from K consecutive frames. For example, suppose that $\mathbf{c}_{v,k'}$, the k'th fingerprint of a video clip v in the DB, is retrieved as a nearest-neighbor of $\mathbf{c}_k$ of the query video. Then, the (NMK)-D candidate fingerprint sequence $\mathbf{c}'$ is generated by

$$\mathbf{c}' = [\, \mathbf{c}_{v,(k'-k+1)} \quad \dots \quad \mathbf{c}_{v,k'} \quad \dots \quad \mathbf{c}_{v,(k'+K-k)} \,]. \tag{8}$$

For all the candidate fingerprints retrieved in the DB search, the corresponding fingerprint sequences are generated as in (8), and they are matched to the query fingerprint sequence $\mathbf{c}$ given by

$$\mathbf{c} = [\, \mathbf{c}_1 \quad \mathbf{c}_2 \quad \dots \quad \mathbf{c}_K \,]. \tag{9}$$

We note that the range search in the DB search part is based on an individual fingerprint, while the fingerprint sequence is used only in the fingerprint matching part. Since the dimension of the fingerprint is low, e.g., 8–12, DB search can be efficiently performed and does not suffer from the curse of dimensionality.

In the fingerprint matching, two video clips are declared similar if the distance between their fingerprint sequences is below a certain threshold T. In determining T, the false alarm rate P_{FA} and the false rejection rate P_{FR} are considered. The false alarm rate P_{FA} is the probability to declare different videos as similar, while the false rejection rate P_{FR} is the probability to declare the videos from the same video as dissimilar. For a good match, one would like to simultaneously minimize both P_{FA} and P_{FR}. However, it is not possible since as P_{FA} decreases, P_{FR} tends to increase, and conversely as P_{FR} decreases, P_{FA} increases [10]. Furthermore, P_{FR} is difficult to analyze in practice since there are plenty of video processing steps of which the exact characteristics are unknown. Thus, it is common to determine a threshold T such that P_{FR} is minimized subject to a fixed P_{FA}[2], [3]. This approach is equivalent to the well-known Neyman–Pearson criterion [10].

A. Fingerprint Modelling

The problem of fingerprint matching is approached by assuming the proposed fingerprint sequence as a realization of a stationary ergodic process. We note that similar analysis has been performed for watermark detection [11], and matching of audio [2] and video [3] fingerprints. First, the centroids $\{c[n,m,k] \mid 1 \le n \le N, 1 \le m \le M, 1 \le k \le K\}$ of a fingerprint sequence are further normalized by its mean μ_c and the standard deviation σ_c as follows:

$$p[n,m,k] = \frac{c[n,m,k] - \mu_c}{\sigma_c}. \tag{10}$$

where $1 \le n \le N$, $1 \le m \le M$, and $1 \le k \le K$. The normalized fingerprint sequence $\mathbf{p}$ is a random process with zero mean and unit variance. Let R and Q be the autocorrelations of $\mathbf{p}$ which are given by

$$R[\tau_1,\tau_2,\tau_3] = E[p[n,m,k]p[n+\tau_1,m+\tau_2,k+\tau_3]] \tag{11}$$

$$Q[\tau_1,\tau_2,\tau_3] = E[p^2[n,m,k]p^2[n+\tau_1,m+\tau_2,k+\tau_3]] \tag{12}$$

where $0 \le \tau_1 \le N-1$, $0 \le \tau_2 \le M-1$, and $0 \le \tau_3 \le K-1$. Based on the ergodic assumption, the autocorrelations R and Q can be estimated from the time-averaged autocorrelation of actual fingerprint sequences, and they are used to derive the probability of false alarm given a certain threshold.

B. Determination of Threshold T

Fast and mathematically tractable fingerprint matching can be achieved by using the squared Euclidean distance as follows:

$$D(\mathbf{p},\mathbf{q}) = \frac{1}{NMK}\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{k=1}^{K}(p[n,m,k]-q[n,m,k])^2 \tag{13}$$

where $\mathbf{p}$ and $\mathbf{q}$ are the fingerprint sequences which are extracted from different video clips. By the central limit theorem, the distance D has a normal distribution if (NMK) is sufficiently large and the contributions in the sums are sufficiently independent [11]. Let μ_D and σ_D be the mean and the standard deviation of the distance D, respectively. Based on the normal assumption, the distance D follows the normal distribution $N(\mu_D, \sigma_D^2)$, and then the probability of false alarm P_{FA} can be obtained as follows:

$$\begin{aligned} P_{\mathrm{FA}} &= \int_{-\infty}^{T} \frac{1}{\sqrt{2\pi}\sigma_D} \exp\left[\frac{-(x-\mu_D)^2}{2\sigma_D^2}\right] dx \\ &= \frac{1}{2}\mathrm{erfc}\left(\frac{\mu_D - T}{\sqrt{2}\sigma_D}\right). \end{aligned} \tag{14}$$

The remaining problem is to obtain the mean μ_D and the variance σ_D^2 of the distance D. Assuming that the two fingerprint sequences $\mathbf{p}$ and $\mathbf{q}$ are independent, the mean μ_D of the distance D is given as

$$\begin{aligned} \mu_D &= E[D] \\ &= 2. \end{aligned} \tag{15}$$

The variance σ_D^2 of the distance D is obtained as

$$\begin{aligned} \sigma_D^2 &= E[D^2] - (E[D])^2 \\ &= E[D^2] - 4 \\ &= \frac{2}{N^2M^2K^2}\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{n'=1}^{N}\sum_{m'=1}^{M}\sum_{k'=1}^{K} \\ &\quad \times \Big\{ Q(|n-n'|,|m-m'|,|k-k'|) \\ &\quad + 2R^2(|n-n'|,|m-m'|,|k-k'|) \Big\} - 2 \end{aligned} \tag{16}$$

where R and Q are the autocorrelations of $\mathbf{p}$ as defined in (11) and (12), respectively.[1] As explained in Section III-A, R and Q in (16) can be estimated from the time-averaged autocorrelation of actual fingerprint sequences for given N, M, and K. Now, for a certain value of P_{FA}, the threshold T can be determined from (14). For example, we can expect the false alarm rate to be as low as 4.6365×10^{-7} when $N = 2$, $M = 4$, $K = 100$, and $T = 0.4$.

[1]The detailed derivation of (15) and (16) is available at http://mmp.kaist.ac.kr/~sunillee/vf_tcsvt.html.

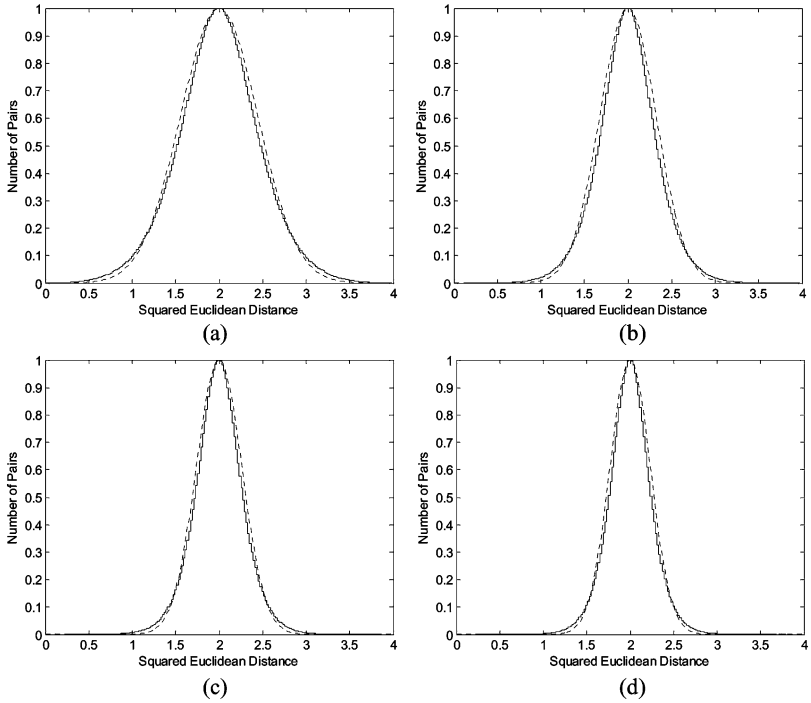

Fig. 3. Comparison of theoretically derived normal distribution (dotted line) and empirically obtained distribution (histogram) of the distance (solid line) when the fingerprint dimension per frame is (a) 4, (b) 8, (c) 12, and (d) 16.

IV. Performance Evaluation

The performance of the proposed video fingerprinting method is evaluated using the fingerprint DB generated from 300 movies belonging to various genres. The total length of the movies in the DB is approximately 590 hours. Unless stated explicitly, the parameters used for the experiments are $S = 10$, $X = 320$, $Y = 240$, $M = 4$, $N = 2$, and $K = 100$ which corresponds to 10 seconds. As a performance measure, the receiver operating characteristics (ROC) curve [12], which plots false rejection rate versus false alarm rate at various operating points (thresholds), is mainly used.

A. Pairwise Independence

The model derived in Section III shows that fingerprints from different video clips are considerably different, and this leads to the assumption that the proposed fingerprint is pairwisely independent. The validity of the model is evaluated as follows. First, the fingerprint DB with different dimensions are generated from the aforementioned movies. The fingerprint dimensions of the generated DB are 4 ($N = M = 2$), 8 ($N = 2$, $M = 4$), 12 ($N = 3$, $M = 4$), and 16 ($N = M = 4$). The other parameters S, X, Y, and K are set to the default values. Next, 554,197,443 ($> 10^8$) pairs of fingerprint sequences from perceptually different 10-seconds-long video excerpts are generated from each DB. Then, the squared Euclidean distance D between fingerprint sequences in each pair is calculated, and its distribution (histogram) is compared with the normal distribution $N\left(\mu_D, \sigma_D^2\right)$ whose mean and standard deviation are derived using the parameters of each fingerprint DB. Fig. 3 compares the theoretically derived distribution of the distances and the histogram of the distances measured from the pairs. The results in Fig. 3 show that the proposed fingerprint follows the stochastic model assumption and the normal approximation fairly well for all the considered dimensions. This leads to the belief that the proposed fingerprint is pairwisely independent, and the threshold T obtained from (14) can be used in practice with reasonable accuracy.

TABLE I
Probability of False Rejection (P_{FR}) for Different Kinds of Video Processing Steps With Threshold $T = 0.4$

Processing	P_{FR}
Lossy compression (DivX 256kbps [14])	0.0098
Resizing to CIF	0.0031
Frame rate change from 24 to 15 fps	0.0219
Gaussian blurring with radius 1 pixel	0.0026
Global change green color (+20%)	0.0019
Global change in brightness (+30%)	0.0054
Global change in gamma correction (+30%)	0.0014
AWGN (Standard deviation: 1, 5, **15**, 25)	0.0971
Rotation (**1**, 2, 3 degrees) + Inside-box cropping	0.0463
Frame cropping (70, **80**, 90%)	0.0509
Random frame drop (10, 30, 50, **70**, 90%)	0.0170
Resizing to CIF + DivX 256kbps + Frame rate change from 24 to 15 fps	0.0388

B. Robustness

To evaluate the robustness of the proposed video fingerprinting method, various sets of distorted video clips are generated. Due to the limit of storage space and processing time, only 50 movies are chosen from the DB and used for the evaluation. The distortions applied to the original video clips are summarized in Table I.

Fig. 4 shows the ROC curves for various distortions, and Table I summarizes the measured false rejection rate (P_{FR}) for the considered distortions with threshold $T = 0.4$. Note that the underlined parameters, e.g., 70% in random frame drop, are those used to obtain the false rejection rate in the table. As shown in the figures and the table, the proposed fingerprint is

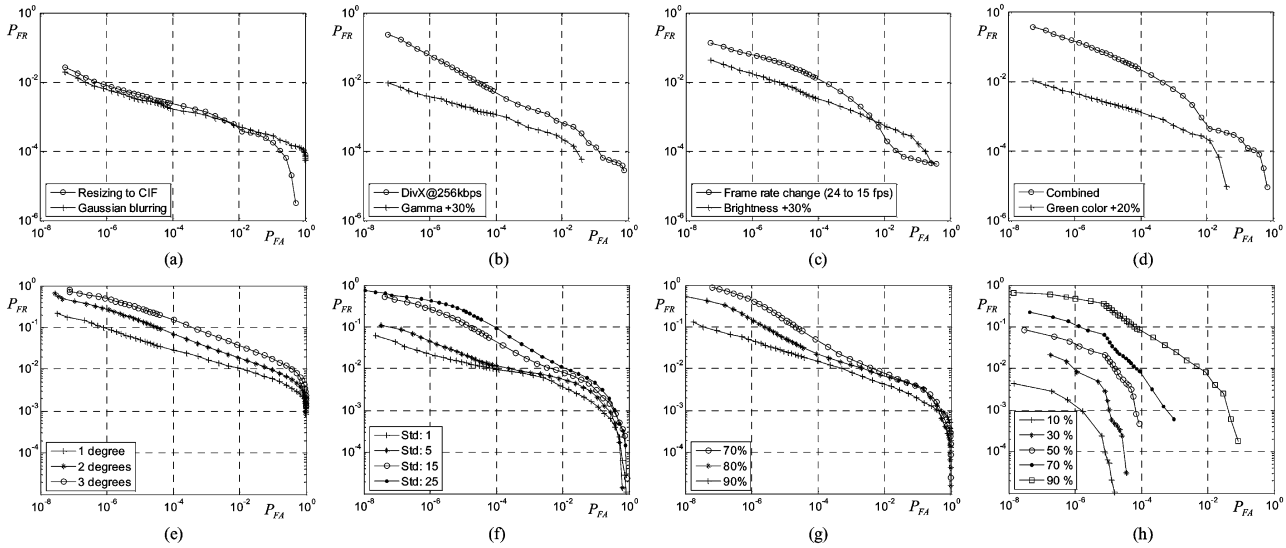

Fig. 4. ROC curves for various distortions: (a) Resizing to CIF and Gaussian blurring with radius 1 pixel. (b) Lossy compression (DivX 256kbps) and global change in gamma $(+30\%)$. (c) Frame rate change from 24 to 15 fps and global change in brightness $(+30\%)$. (d) Global change in green color $(+20\%)$ and combined distortion (resizing to CIF + Frame change from 24 to 15 fps + DivX 256kbps). (e) Rotation at angles of 1, 2, and 3 degrees followed by inside-box cropping, (f) AWGN with standard deviation 1, 5, 15, and 25. (g) Frame cropping (70, 80, and 90%). (h) Random frame drop with drop rate from 10 to 90%.

highly robust against nongeometric distortions including lossy compression, global change in color, brightness, and gamma, resizing, Gaussian blurring, additive noise, and combined distortion. The proposed fingerprint is also robust against the temporal distortions such as frame rate change and random frame drop, even when 70% of frames are lost.

The performance of the proposed fingerprint degrades when video clips are distorted by geometric transformations such as frame rotation and cropping. The vulnerability against general geometric transformations is a common problem of the video fingerprinting methods which use global features of a frame as a fingerprint. However, as shown in Fig. 4(e) and (g), the proposed fingerprint is robust against minor geometric transformations, e.g., frame rotation up to 1 degree and frame cropping which retains more than 80% of central portion of a frame. This result shows that the proposed fingerprint can match an original video clip and its geometrically distorted version as long as the geometric transformation does not severely degrade the perceptual similarity between them.

C. Effects of Parameters on Performance

Fig. 5 shows the effect of the parameters (frame size, fingerprint dimension, query length, and frame rate) on the performance. As shown in Fig. 5(a), the performance was similar for all the considered frame size, however, the performance was slightly better when the frame size was QVGA (320×240), especially in terms of the false alarm rate. Fig. 5(b) shows that the performance is improved as the fingerprint dimension increases, however, the amount of the improvement decreases as the dimension increases and becomes marginal when the dimension exceeds 12. Since the increase of the fingerprint dimension degrades the DB search efficiency, the appropriate dimension has to be chosen. The experimental results show that the dimension between 8 and 12 would be a reasonable choice. Fig. 5(c) shows that the performance is improved as the query length increases. However, since the query length is limited in practice, it should

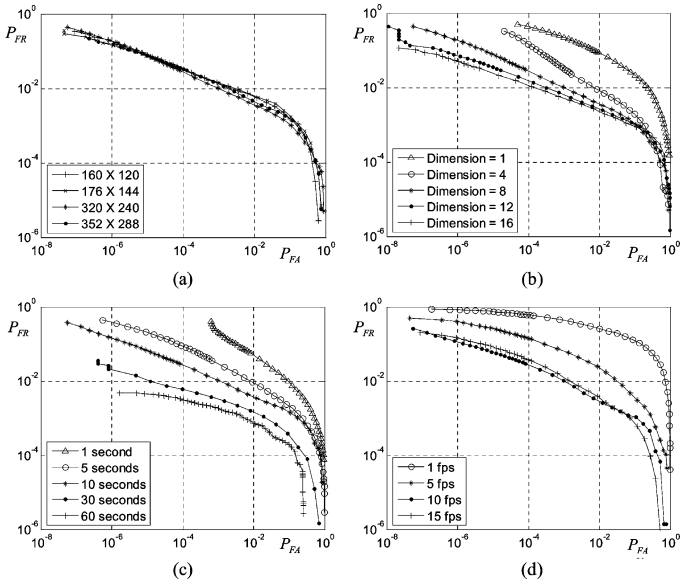

Fig. 5. Effects of parameters on the performance of the proposed video fingerprinting method. (a) Width and height from 160×120 to 352×288. (b) Fingerprint dimension from 1 to 16. (c) Query length from 1 to 60 seconds. (d) Frame rate from 1 to 15 fps.

be carefully determined considering the requirements of the applications. The ROC curves in Fig. 5(a)–(c) are obtained using the video clips distorted by the combined distortion as in Fig. 4(d). We note that the effects of the parameters on the performance are similar for other distortions. Fig. 5(d) shows the effect of the frame rate S which is closely related to the random start distortion introduced by the misalignment of the resampled frames. As shown in the figure, the robustness against the random start is improved as the frame rate increases and starts to saturate when the frame rate exceeds 10 fps. This results suggest that the frame rate around 10 fps would be a reasonable choice for S.

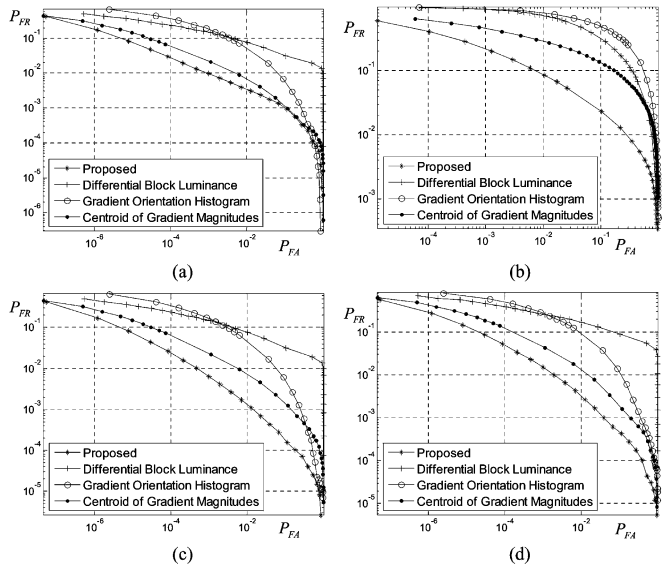

Fig. 6. ROC curves of proposed fingerprint, differential block luminance [13], gradient orientation histogram [7], and centroid of gradient magnitudes [8] for: (a) distortion set 1, (b) distortion set 2, (c), distortion set 3, and (d) distortion set 4.

D. Comparison of Proposed Method With Other Features

The performance of the proposed fingerprint is compared with that of other widely-used features, differential block luminance [13], gradient orientation histogram [7], and centroid of gradient magnitudes [8]. For a fair comparison, an input video clip is resampled, converted to grayscale, and resized as in the proposed method prior to the feature extraction, and the dimensions of all the features are set to the same value, 8 per frame. The differential block luminance is obtained by first partitioning a frame into 2×5 blocks, and then by taking the difference of the mean luminances of blocks adjacent in both spatial and temporal domain as in [13]. Although Oostveen *et al.* take signs of differences and form binary fingerprints, the values of the differences are directly used as fingerprints in this comparative test. The gradient orientation histogram is widely used as a local descriptor in the literature [7]. However, in this comparative test, the histogram of an entire frame is obtained and used as a fingerprint. The number of bins in the gradient orientation histogram is also determined as 8 for a fair comparison. The centroid of gradient magnitudes [8] is obtained by first partitioning a frame into 2×2 blocks, and then by calculating the centroid of gradient magnitudes for each block. Since the centroid of gradient magnitude is given as (x, y) location in each block, 8-D fingerprint is obtained for each frame. As a distance measure, squared Euclidean distance metric is used for all the features.

The comparative test is performed using 50 movies chosen from the DB. First, the four sets of distorted video clips are generated by applying the following sets of distortions.

- **Set 1**: Resizing to CIF, Frame rate change from 24 to 15 fps, and Lossy compression (DivX 256 kbps) [14].
- **Set 2**: Luminance histogram equalization, Resizing to CIF, and Lossy compression (DivX 256 kbps).
- **Set 3**: Brightness $+15\%$, Frame rate change from 24 to 15 fps, Resizing to QVGA, and Lossy compression (DivX 256 kbps).
- **Set 4**: Color variation (Red $+20\%$, Green -10%, Blue $+5\%$), Frame rate change from 24 to 20 fps, Contrast $+30\%$, Resizing to CIF, and Lossy compression (DivX 256 kbps).

Each distortion set is a combination of various distortions common in practical applications. Fig. 6 shows the ROC curves of the considered features and the proposed fingerprint for the four distortion sets. As shown in the figure, the proposed fingerprint achieves the lowest false rejection rate for a given false alarm rate (vice versa). This means that the proposed fingerprint outperforms the considered features in the context of video fingerprinting.

V. Conclusion and Future Work

In this paper, a novel video fingerprinting method based on the centroid of gradient orientations is proposed. The proposed video fingerprinting method is not only pairwisely independent but also robust against common video processing steps including lossy compression, resizing, frame rate change, global change in brightness, color, gamma, etc. The problem of reliable fingerprint matching is approached by assuming the fingerprint as a realization of a stationary ergodic process. The matching threshold is theoretically derived for a given false alarm rate using the assumed stochastic model, and its validity is experimentally verified. The experimental results show that the proposed fingerprint outperforms other features in the context of video fingerprinting. The future work is to propose a secure video fingerprinting method robust against general geometric transformations, e.g., rotation, shift, cropping, etc.

References

[1] T. Kalker, J. A. Haitsma, and J. Oostveen, "Issues with digital watermarking and perceptual hashing," *Proc. SPIE 4518, Multimedia Systems and Applications IV*, pp. 189–197, Nov. 2001.

[2] J. S. Seo, M. Jin, S. Lee, D. Jang, S. Lee, and C. D. Yoo, "Audio fingerprinting based on normalized spectral subband moments," *IEEE Signal Process. Lett.*, vol. 13, no. 4, pp. 209–212, Apr. 2006.

[3] S. Lee and C. D. Yoo, "Video fingerprinting based on centroids of gradient orientations," in *Proc. ICASSP*, Toulouse, France, May 2006, vol. 2, pp. 401–404.

[4] J. T. Robinson, "The k-d-b-tree: A search structure for large multidimensional dynamic indexing," in *Proc. ACM SIGMOD Int. Conf. Management Data*, 1981, pp. 10–18.

[5] C. G. Rafael and E. W. Richard, *Digital Image Processing*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 2002.

[6] D. A. Forsyth and J. Ponce, *Computer Vision—A Modern Approach*. Englewood Cliffs, NJ: Prentice Hall, 2003.

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

[8] H. Arun and M. B. Rudolf, VideoGREP: Video Copy Detection using Inverted File Indices IBM Research, Yorktown Heights, NY, 2001, Tech. Rep..

[9] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *Proc. CVPR*, vol. 2, pp. 257–263, 2003.

[10] L. M. James and L. C. David, *Decision and Estimation Theory*. New York: McGraw-Hill, 1978, pp. 27–38.

[11] J. P. Linnartz, T. Kalker, G. Depovere, and R. Beuker, "A reliability model for the detection of electronic watermarks in digital images," in *Proc. Symp. Commun. Vehicular Technol.*, 1997, pp. 202–209.

[12] C. E. Metz, "Basic principles of ROC analysis," *Seminars Nucl. Med.*, vol. 8, no. 4, pp. 283–298, 1978.

[13] J. Oostveen, T. Kalker, and J. Haitsma, "Feature extraction and a database strategy for video fingerprinting," in *Proc. Int. Conf. Recent Adv. Vis. Inf. Syst.*, 2002, pp. 117–128.

[14] DivX Codec DivX, San Diego, CA [Online]. Available: http://www.divx.com