Robust Video Fingerprinting for Content-Based Video Identification

Sunil Lee * *Student 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 by measuring the distance between the query fingerprint and the fingerprints in the database. 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 modelling 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.**

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.

Manuscript received February, 2007, revised August 2007.

Sunil Lee and Chang D. Yoo are with the Div. of Electrical Engineering, School of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology, 373-1 Guseong Dong, Yuseong Gu, Daejeon 305-701, Korea (Phone: + 82-42-869-5470, Fax: + 82-42-862-0559, E-mail: sunillee@kaist.ac.kr, cdyoo@ee.kaist.ac.kr).

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.

Recently, many video fingerprinting methods have been proposed [4]–[15]. The methods in [4]–[6] use histogram-based fingerprints. For example, Cheung and Zakhor estimated the video similarity by first summarizing each video with a small set of its sampled frames and then by measuring the distance between color histograms obtained from the corresponding frames [4]. There also exist video fingerprinting methods which use a spatial feature other than a histogram [7], [8]. Roover *et al.* proposed the radial projection of the image pixels denoted as radial hashing (RASH) [7], while Hampapur and Bolle used dominant color and the centroid of gradient magnitudes [8]. Some methods use an ordinal measure [9] of a feature as a fingerprint [10]–[12]. For example, Kim and Vasudev used the ordinal measure of the block mean luminance for video copy detection [11]. There are video fingerprinting methods which also exploit temporal characteristics of videos [11], [13]–[15]. For example, Oostveen *et al.* used the differential block luminance where the differentiation is taken in both spatial and temporal directions [14]. Coskun *et al.* proposed a spatio-temporal transform based video fingerprinting method [15].

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 [16] or locality sensitive hashing (LSH) [17] needs to be employed. However, since the focus of this paper

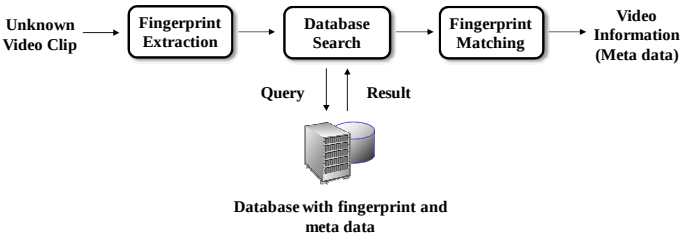

Fig. 1. Overall structure of the proposed video fingerprinting method.

is on the fingerprint extraction and matching, DB search algorithms are not explained in detail. A survey of various DB search methods can be found in [18]. 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. In the proposed method, the metadata includes the title of the query video clip and its temporal position in the original video. Note that if the obtained minimum distance is above a certain threshold, it is declared that the query video clip does not exist in the DB. 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. Section 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.

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 which is one of the most common video processing steps. In the second step, each resampled frame is converted to grayscale, so that only luma components (pixel intensity in grayscale) of an input video are used for the fingerprint extraction. This step makes 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)-dimensional 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

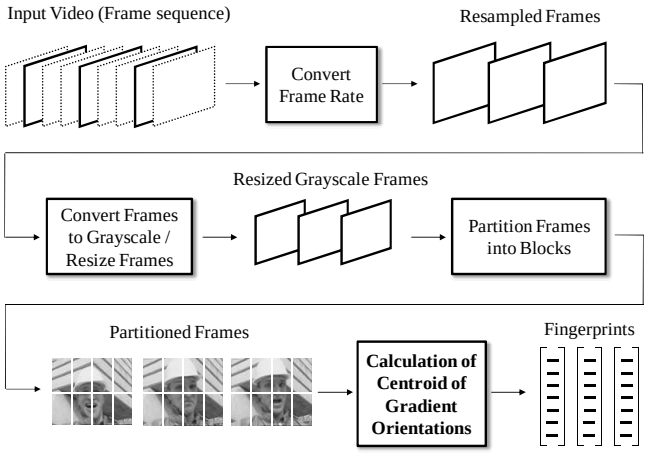

Fig. 2. Overall procedure of the proposed video fingerprint extraction.

as the vector

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

The gradient vector points in the direction of maximum rate of change of f at coordinates (x, y) [19]. There are many operators which approximate the partial derivatives G_x and G_y, e.g. Roberts cross-gradient operators, Sobel operators, etc. [19]. In the proposed method, G_x and G_y are simply approximated as follows

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

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

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

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

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

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 $-\frac{\pi}{2}$ and $\frac{\pi}{2}$. The (NM)-dimensional fingerprint vector $\mathbf{c}_k$ of the kth frame is obtained by

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

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 [20]. Thus the proposed fingerprint is expected to be robust against the content-preserving distortions while being discriminative so that perceptually

different video clips can be distinguished. 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 non-linear 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 non-linear operations since they are less likely to affect the gradient orientations [21].

The gradient-based features have been used as a descriptor which represents local image regions [21] and also as a video fingerprint [8]. Lowe's scale invariant feature transform (SIFT) [21] is a combination of an interest point detector and a local descriptor based on gradients. He used the histogram of gradient orientations as a local descriptor which characterizes a region around the detected interest points. The comparative test in [22] shows that Lowe's local descriptor based on gradients outperforms other local descriptors. However, the high dimensionality of Lowe's descriptor (128 for each region) renders the histogram of gradient orientations unsuitable for video fingerprinting. Ke and Sukthankar tried to reduce the dimension of the SIFT descriptor by applying traditional principal component analysis (PCA) [23], however, the dimension was still too high for the purpose of 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-dimensional fingerprint (225 per each frame) 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-F, 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)-dimensional candidate fingerprint sequence $\mathbf{c}'$ is generated by

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

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

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

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. The problem of the fingerprint matching can be formulated as a hypothesis testing using the following hypotheses H_0 and H_1 based on the fingerprinting function $F(\cdot)$ and distance measure $D(\cdot,\cdot)$:

- H_0: Two video clips v_1 and v_2 are from the same video if the distance $D\left(F(v_1), F(v_2)\right)$ is below a certain threshold T.
- H_1: Two video clips v_1 and v_2 are from the different video if the distance $D\left(F(v_1), F(v_2)\right)$ is above 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 [24]. 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 [24].

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 [25], 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

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

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

where $0 \leq \tau_1 \leq N-1$, $0 \leq \tau_2 \leq M-1$, and $0 \leq \tau_3 \leq K-1$. 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:

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

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 [25]. 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:

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

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

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

where R and Q are the autocorrelations of $\mathbf{p}$ as defined in (11) and (12), respectively. The detailed derivation of (15) and (16) is available at http://mmp.kaist.ac.kr/~sunillee/vf_tcsvt.html. 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.

IV. Performance Evaluation

A. Experimental Setup

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. These values are considered as the default value of each parameter. Note that the parameters of the proposed fingerprinting method can be adjusted according to the requirements of the application. The effect of the parameters on the performance is discussed in Section IV-E. As a performance measure, the receiver operating characteristics (ROC) curve [26], which plots false rejection rate versus false alarm rate at various operating points (thresholds), is mainly used.

B. Computational Complexity

Since the proposed fingerprint is obtained by taking the weighted sum of the gradient magnitude and orientation calculated for every pixel of a resized frame, the computational complexity required for the extraction of a single fingerprint vector is given as $O(XY)$. Note that the complexity is not related to the fingerprint dimension (NM) while being directly proportional to the size of the resized frame (XY). We also note that since the resizing parameters X and Y are fixed constants, e.g. X=320 and Y=240, the complexity of the fingerprint extraction is almost constant. The computational complexity of the fingerprint matching is given as $O(NMKK_c)$, where K_c is the number of candidate fingerprint sequences obtained in the DB search. The number of the candidates K_c varies according to the fingerprint dimension, the DB search strategy, and the DB size.

The complexity of the proposed method is experimentally evaluated by measuring the search time of 1,000 10-seconds-long video queries on the PC with 3 GHz CPU and 1 GB main memory. The fingerprint dimension was set to 12 (N=3, M=4), the DB size was 590 hours as introduced in Section IV-A, and the k-d-tree [16] was used as an indexing structure. The result shows that it takes 2.13 seconds on average for the identification. Even in the worst case, it takes only 6.29 seconds, and 96% of the queries are identified in 3.30 seconds. Note that the decoding time is not included in the search time since it heavily depends on the video resolution and the used encoder. However, the overall search time including the decoding time does not exceed the query length in the experiments, which means that the proposed video fingerprinting method can be performed in real-time.

C. 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

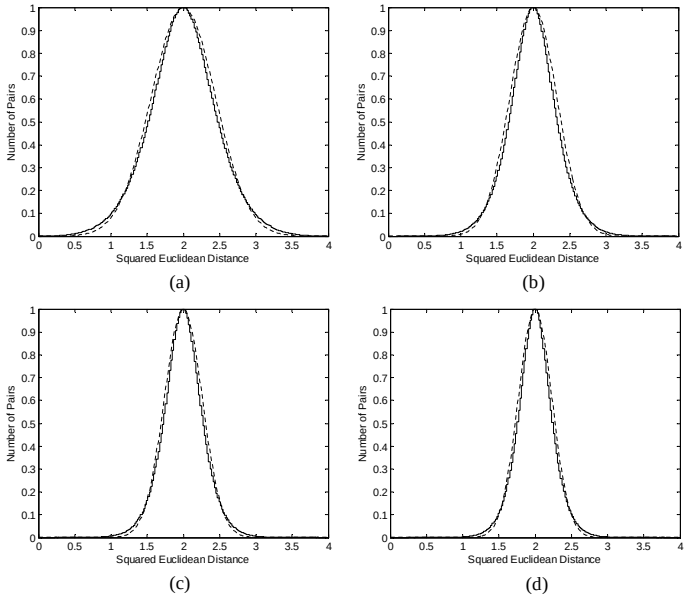

Fig. 3. Comparison of theoretically derived normal distribution (dotted line) and empirically obtained distribution (histogram) of the distance (solid line) when the fingerprint dimension per frame is (a) 4, (b) 8, (c) 12, and (d) 16.

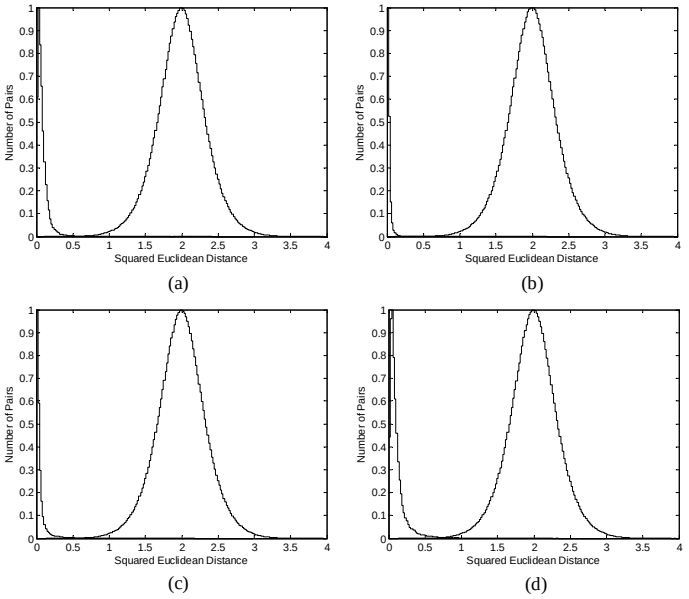

Fig. 4. Comparison of distribution of intra and inter distance: (a) Lossy compression (DivX [27] 256kbps), (b) Resizing to CIF (352×288), (c) Frame rate change from 24 to 15 fps, and (d) Combined distortion: Resizing to CIF, Frame change from 24 to 15 fps, and Lossy compression (DivX 256kbps).

(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 $N(\mu_D, \sigma_D^2)$ 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.

To consider the tradeoff between pairwise independence and robustness, the following two types of distances are defined:

- Inter distance: The distance between fingerprint sequences from perceptually different video clips.
- Intra distance: The distance between fingerprint sequences from an original video clip and its distorted version.

Fig. 4 shows the distribution (histogram) of the inter and the intra distances of the proposed fingerprint for 4 kinds of distortions – lossy compression, resizing, frame rate change, and their combination. To evaluate the intra distance, 700,803 pairs of fingerprint sequences from original 10-seconds-long video clips and their distorted versions are used. As shown in the figure, the intra distance is much smaller than the inter distance for all the considered distortions. The histograms of the inter and the intra distances scarcely overlap each other, which means that both the false alarm rate and the false rejection rate of the proposed fingerprinting method can be made very low at a certain threshold.

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)	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

D. 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. 5 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 highly robust against lossy compression, global change in

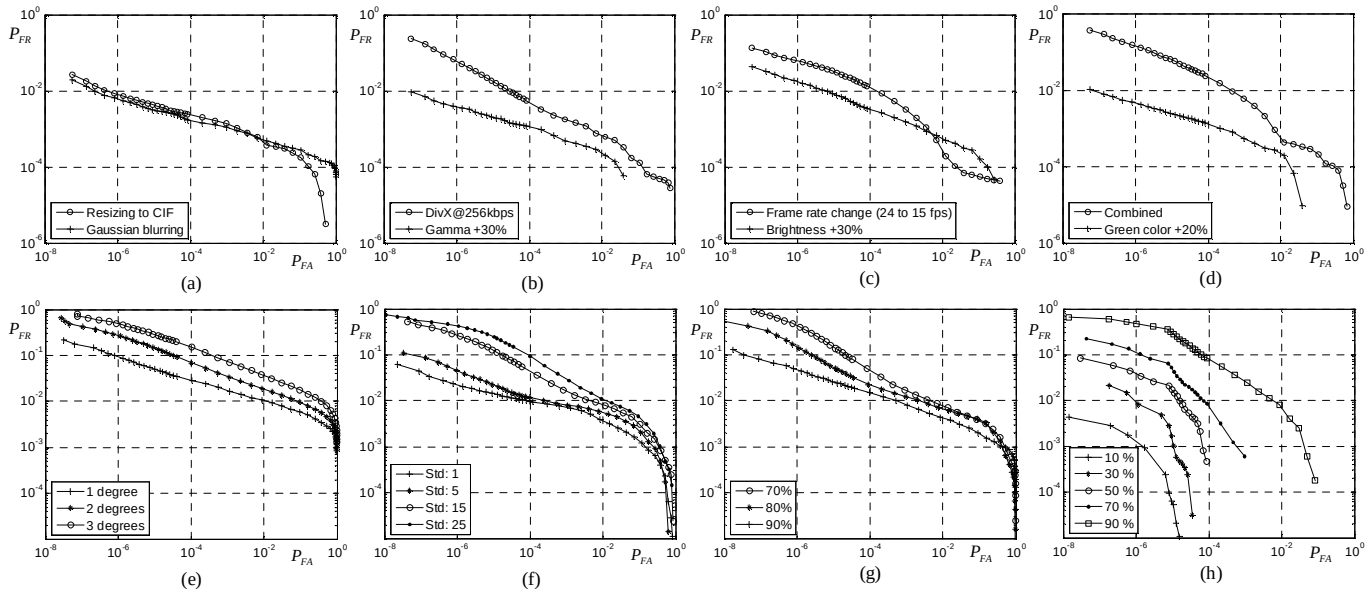

Fig. 5. ROC curves for various distortions: (a) Resizing to CIF and Gaussian blurring with radius 1 pixel, (b) Lossy compression (DivX 256kbps) and global change in gamma (+30%), (c) Frame rate change from 24 to 15 fps and global change in brightness (+30%), (d) Global change in green color (+20%) and 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%), and (h) Random frame drop with drop rate from 10 to 90%.

color, brightness, and gamma, resizing, frame rate change, and Gaussian blurring. Even under combined distortion which simulates the situation where video clips are transmitted via low-bandwidth channel or encoded for the playback in hand-held devices with limited computing power, the false rejection rate does not exceed 4% at the threshold $T = 0.4$. These results show that the proposed fingerprint is robust against distortions which do not obliterate the geometric structure of video frames. The proposed fingerprint is also robust against the temporal distortion such as 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. 5(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.

In general, the gradient orientation is known to be susceptible to random noise. However, as shown in Fig. 5(f), the proposed fingerprint is reasonably robust against additive noise unless the noise introduces severe visual artifacts, e.g. when its standard deviation is smaller than 15. We note that the robustness against additive noise could be achieved since the centroid of gradient orientations is obtained as a weighted sum of orientations over a sufficiently large area, so that the overall effect of additive noise is reduced. Furthermore, the orientation of a gradient vector is less likely affected by additive noise when the magnitude of the vector is large. This property also improves the robustness of the proposed fingerprint against additive noise since the orientations with the large magnitudes are more likely to determine the centroid.

E. Effects of Parameters on Performance

Fig. 6(a) shows how the performance of the proposed method changes when the frame size (=$X \times Y$) varies from QQVGA (160×120) to CIF. As shown in the figure, even though the performance was similar for all the considered frame size, the performance was slightly better when the frame size was QVGA (320×240), especially in terms of the false alarm rate. Fig. 6(b) shows the effect of the fingerprint dimension. As shown in the figure, 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. 6(c) shows the change of the performance according to the query length. As shown in the figure, the performance is improved as the query length increases. However, since the query length is limited in practice, it should be carefully determined considering the requirements of the applications. The ROC curves in Fig. 6(a), (b), and (c) are obtained using the video clips distorted by the combined distortion as in Fig. 5(d). We note that the effects of the parameters on the performance are similar for other distortions. Fig. 6(d) shows the effect of the frame rate S. When the frame rate is low, both the DB size and the computational complexity are reduced. However, when

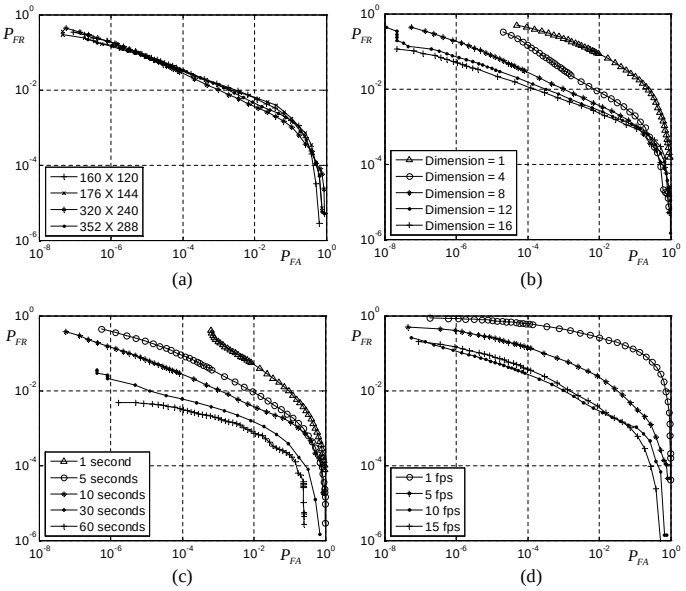

Fig. 6. Effects of parameters on the performance of the proposed video fingerprinting method: (a) Width and height from 160×120 to 352×288, (b) Fingerprint dimension from 1 to 16, (c) Query length from 1 to 60 seconds, and (d) Frame rate from 1 to 15 fps.

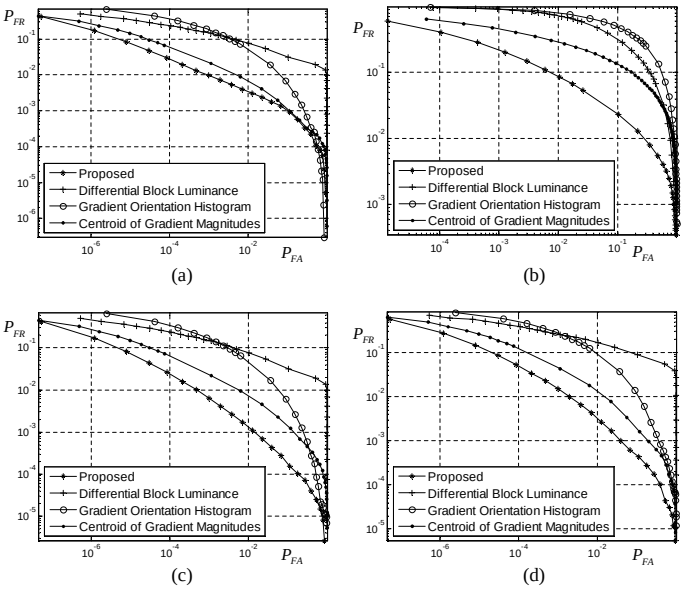

Fig. 7. ROC curves of proposed fingerprint, differential block luminance [14], gradient orientation histogram [21], and centroid of gradient magnitudes [8] for (a) distortion set 1, (b) distortion set 2, (c), distortion set 3, and (d) distortion set 4.

the frame rate is too low, the fingerprinting system suffers from the random start distortion which is introduced by the misalignment of the resampled frames. The ROC curves in Fig. 6(d) are obtained by applying the worst random start distortion to the video clips for each frame rate. 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.

F. 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 [14], gradient orientation histogram [21], 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 5×2 blocks, and then by taking the difference of the mean luminances of blocks adjacent in both spatial and temporal domain as in [14]. 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 [21]. 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-dimensional 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 256kbps)
- **Set 2**: Luminance histogram equalization, Resizing to CIF, and Lossy compression (DivX 256kbps)
- **Set 3**: Brightness +15%, Frame rate change from 24 to 15 fps, Resizing to QVGA, and Lossy compression (DivX 256kbps)
- **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 256kbps)

Each distortion set is a combination of various distortions common in practical applications. Fig. 7 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.

The differential block luminance is a simple yet effective feature that represents the mean luminance difference between both spatially and temporally adjacent blocks. Therefore it is expected to be robust against global change in pixel intensities. However, the introduction of differentiation in the temporal direction makes the fingerprint unreliable when the video clip includes still or slowly varying scenes. Oostveen *et al.* simply alleviate this problem by intentionally leaving the DC component of mean luminance when taking the difference

in the temporal direction. However, this approach degrades the performance of the fingerprint for global change in pixel intensities which the differential block luminance is expected to be robust to. The differential block luminance is also vulnerable to non-linear operations such as gamma correction which changes the relative intensities of adjacent pixels. The comparison results in Fig. 7 show that the differential block luminance performs worse than the proposed fingerprint and the centroid of gradient magnitude, while its performance is comparable to or slightly better than that of the gradient orientation histogram. The proposed fingerprint also outperforms the centroid of gradient magnitude for all the considered sets of distortions since the gradient magnitude is more likely affected by non-linear distortions which cause a large change in relative magnitudes for some gradients. The experimental results also show that the gradient orientation histogram, which is known to be the best local descriptor [22], does not perform well when it is used to characterize an entire frame with limited number of bins. We guess that the gradient orientation histogram is well suited to representing the structure of local regions rather than to capturing the global characteristics of an entire frame. Note that we do not argue that the proposed fingerprint absolutely outperforms the gradient orientation histogram. However, the proposed fingerprint performs better than the gradient orientation histogram in terms of both robustness and pairwise independence when the dimension of both fingerprints are kept low to meet the requirements 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. To achieve this goal, a robust local fingerprint and an efficient binarization method need to be proposed.

References

[1] T. Kalker, J. A. Haitsma,and J. Oostveen, “Issues with digital watermarking and perceptual hashing,” in *Proc. SPIE 4518, Multimedia Systems and Applications IV*, Nov. 2001.

[2] Jin S. Seo, Minho Jin, Sunil Lee, Dalwon Jang, Seungjae Lee, and Chang D. Yoo, “Audio Fingerprinting Based on Normalized Spectral Subband Moments,” *IEEE Signal Processing Letters*, vol. 13, no. 4, pp. 209-212, Apr. 2006.

[3] 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.

[4] S. C. Cheung and Avideh Zakhor, “Efficient video similarity measurement with video signature,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 13, no. 1, pp. 59-74, Jan. 2003.

[5] M. R. Naphade, M. M. Yeung, and B. L. Yeo, “A novel scheme for fast and efficient video sequence matching using compact signatures,” *SPIE: Storage and Retrieval for Media Database*, pp. 564-572, 2000.

[6] K. Kashino, T. Kurozumi, and H. Murase, “A quick search method for audio and video signal based on histogram pruning,” *IEEE Trans. Multimedia*, vol. 5, no. 3, pp. 348-357, Sep. 2003.

[7] 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.

[8] Arun Hampapur and Rudolf M. Bolle, “VideoGREP: Video Copy Detection using Inverted File Indices,” Technical Report, IBM Research, 2001.

[9] D. N. Bhat and S. K. Nayar, “Ordinal measures for image correspondence,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 4, pp. 415-429, Apr. 1998.

[10] R. Mohan, “Video sequence matching,” in *Proc. ICASSP*, vol. 5, pp. 3697-3700, Jan. 1998.

[11] Changick Kim and B. Vasudev, “Spatiotemporal sequence matching for efficient video copy detection,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 127-132, Jan. 2005.

[12] Xian-Sheng HUA, Xian CHEN, Hong-Jiang ZHANG, “Robust Video Signature Based on Ordinal Measure,”, in *Proc. International Conference on Image Processing (ICIP 2004*), vol. 1, pp. 685-688, October 24-27, Singapore, 2004.

[13] T. C. Hoad and J. Zobel, “Fast video matching with signature alignment,” in *Proc. Int. Workshop Multimedia Inf. Retrieval*, pp. 262-268, 2003.

[14] Job Oostveen, Ton Kalker, and Jaap 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.

[15] 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.

[16] 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.

[17] M. Datar, P. Indyk, N. Immorlica, and V. Mirrokni, “Locality-Sensitive Hashing Scheme Based on p-Stable Distributions,” in *Proc. Symposium on Computational Geometry*, 2004.

[18] C. Bohm, S. Berchtold, and D. Keim, “Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases,” *ACM Comput. Surv.*, vol. 33, no. 3, pp. 322-373, 2001.

[19] Rafael C. Gonzalez and Richard E. Woods, *Digital Image Processing, 2nd Edition*, Prentice Hall, 2002.

[20] D. A. Forsyth and J. Ponce, *Computer Vision - A Modern Approach*, Prentice Hall, 2003.

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

[22] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” in *Proc. CVPR*, vol. 2, pp. 257-263, 2003.

[23] Yan Ke and Rahul Sukthankar, “PCA-SIFT: A more distinctive representation for local image descriptors,” in *Proc. CVPR 2004*, vol. 2, pp. 506-513, 2004.

[24] James L. Melsa and David L. Cohn, *Decision and Estimation Theory*, McGraw-Hill Book Company, pp. 27-38, 1978.

[25] J. P. Linnartz, T. Kalker, G. Depovere, and R. Beuker, “A reliability model for the detection of electronic watermarks in digital images,” in *Symposium on Communications and Vehicular Technology*, 1997.

[26] C. E. Metz, “Basic principles of ROC analysis,” *Seminars in Nuclear Medicine*, vol. 8, no. 4, pp. 283-298, 1978.

[27] DivX Codec [Online]. Available: http://www.divx.com.