VIDEO FINGERPRINTING BASED ON CENTROIDS OF GRADIENT ORIENTATIONS

Sunil Lee and Chang D. Yoo

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

ABSTRACT

Fingerprints are feature vectors that can uniquely characterize the video signal. The goal of a video fingerprinting system is to judge whether two videos have the same contents by measuring distance between fingerprints extracted from the videos. In this paper, a novel video fingerprinting method based on the centroids of gradient orientations is proposed. The centroid of gradient orientations is chosen due to its reliability and robustness against common video processing steps. A threshold used to reliably determine a fingerprint match is theoretically derived, and its validity is experimentally verified. The experimental results show that the proposed fingerprint is not only pairwise independent but also robust against common video processing steps.

1. INTRODUCTION

Today an enormous amount of video contents is digitally produced, stored, and distributed. The proliferation of digital videos has made accessibility of video contents that much easier and cheaper while being the source of many problems such as illegal distribution of copyrighted movies via file sharing services on the Internet. Protection, management, and indexing of the video contents have become essential with the increasing popularity of digital videos. Among various solutions to these problems, fingerprinting is receiving increased attention. Fingerprints are perceptual features or short summaries of a multimedia object [1], and the goal of fingerprinting is to provide fast and reliable methods for content identification [2]. Specifically, the goal of a video fingerprinting system is to judge whether two videos have the same contents even under quality-preserving distortions, e.g. resizing, frame rate change, lossy compression such as MPEG, DivX, and WMV, etc.

Features for fingerprinting 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].

- **Robustness**: The fingerprints extracted from a degraded video should be similar to the fingerprints of the original video.
- **Pairwise independence**: Two videos, that are perceptually different, must have different fingerprints.
- **Database search efficiency**: Fingerprints must be suitable for fast database (DB) search.

Many features have been proposed for the video fingerprinting, e.g. color (luminance) histogram [3], mean luminance and its variants [4][5][6], dominant color [7], etc. In this paper, a novel video fingerprinting method based on the *centroid of gradient orientations (CGO)* is proposed. The gradient orientation is the direction in which the directional derivative has the largest value. Lowe [8] used the histogram of gradient orientations as local descriptors for object recognition, and comparative test in [9] showed that it performs best among various local descriptors. However its high dimensionality renders the histogram of gradient orientations unsuitable for video fingerprinting. Instead, we propose the CGO as a video fingerprint. Fingerprint matching is performed using the squared Euclidean distance. By modelling the proposed fingerprint as a stationary process, a threshold for the reliable matching is obtained. The experimental results show that the CGO satisfies the main requirements of fingerprints and outperforms other features in the context of video identification.

The rest of the paper is organized as follows. Section 2 describes the proposed video fingerprinting method in detail. Section 3 provides the results of the various performance evaluation on the proposed video fingerprinting method. Finally, Section 4 concludes the paper.

2. PROPOSED VIDEO FINGERPRINTING METHOD

An overview of the proposed fingerprinting method is shown in Fig. 1. First, an input video is resampled at a fixed frame rate F frames per second (fps) to cope with the frame rate change (typically $F = 10$ fps). Next, each resampled frame is converted to the grayscale and its width and height are normalized to the fixed values R_x and R_y (typically $R_x = 320$ and $R_y = 240$), respectively. These steps make the proposed fingerprints robust against variations in color characteristics and resizing. Then, each resized frame is partitioned into

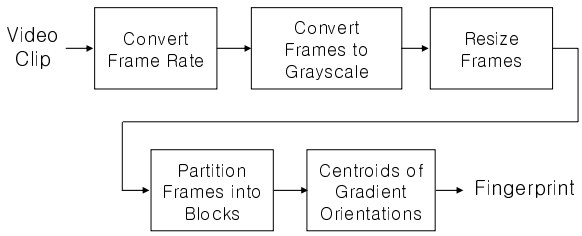

Fig. 1. Overview of the proposed video fingerprint extraction

$M = M_x \times M_y$ blocks (typically $M_x = 4$ and $M_y = 2$), and the centroid of gradient orientations is calculated for each block. Finally, M-dimensional vector of the centroids is obtained and used as a fingerprint for the frame. For fingerprint matching, a fingerprint sequence that consists of fingerprints extracted from K consecutive frames is used (typically $K = 100$ that corresponds to 10 seconds when $F = 10$ fps). The details of the proposed method are explained in the next subsections.

2.1. Fingerprint based on centroid of gradient orientations

For each luminance value $s(x, y)$ of a block S in a video frame, the gradient magnitude $m(x, y)$ and orientation $\theta(x, y)$ are calculated as follows:

$$m(x,y) = \sqrt{G_x^2 + G_y^2} \quad (1)$$

$$\theta(x,y) = \tan^{-1}(G_y/G_x) \quad (2)$$

where the partial derivatives G_x and G_y are approximated by $G_x = s(x+1, y) - s(x-1, y)$ and $G_y = s(x, y+1) - s(x, y-1)$. Then, the centroid of gradient orientations of the block S is calculated as follows:

$$c = \frac{\sum_{x=2}^{X-1}\sum_{y=2}^{Y-1}\theta(x,y)m(x,y)}{\sum_{x=2}^{X-1}\sum_{y=2}^{Y-1}m(x,y)} \quad (3)$$

where $X = R_x/M_x$ and $Y = R_y/M_y$. The value of the CGO ranges from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$ regardless of the location of the block. A vector of M CGOs is used as a fingerprint for a frame, and a vector of MK CGOs extracted from K consecutive frames is used as a fingerprint sequence for fingerprint matching.

2.2. Fingerprint matching

In the fingerprint matching, two videos are declared similar if the distance between their fingerprints is below a certain threshold T. For the selection of 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. In practice, P_{FR} is difficult to analyze since there are plenty of video processing steps of which the exact characteristics are unknown. Thus it is common to deal with only P_{FA} for choosing the threshold T [1].

2.2.1. Fingerprint modelling

The problem of fingerprint matching is approached by assuming the proposed fingerprint as a realization of a stationary process. We note that similar analysis has been performed for watermark detection in [10], and matching of audio fingerprints in [1]. Let $c[n]$ be the CGO of a fingerprint sequence ($1 \le n \le N = MK$). We further normalize $c[n]$ by its mean m_c and variance σ_c^2 as follows:

$$p[n] = \frac{c[n] - m_c}{\sigma_c}. \quad (4)$$

So that p is a random process with zero mean and unit variance. By simplifying the stochastic model of the CGO as the first-order autocorrelation, the following expressions are obtained:

$$\begin{aligned} R[k] = E\Big[p[n]p[n+k]\Big] &= a^{|k|}, \\ Q[k] = E\Big[p^2[n]p^2[n+k]\Big] &= 1 + (\mu_4 - 1)b^{|k|} \end{aligned} \quad (5)$$

where $\mu_k = E[p^k[n]]$, and a and b represent a measure of the correlation of CGO. By the normalization, $\mu_1 = 0$ and $\mu_2 = 1$. Experimental results obtained from the actual video data show that the proposed fingerprint follows the first-order model reasonably well, and the values of a, b, and μ_4 are typically 0.800, 0.685, and 4.349, respectively.

2.2.2. Reliability analysis

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

$$D = \frac{1}{N}\sum_{n=1}^{N}(p[n] - q[n])^2 \quad (6)$$

where p and q are the CGOs of the different fingerprint sequences. By the central limit theorem, the distance D has a normal distribution if N is sufficiently large and the contributions in the sums are sufficiently independent [10]. The expectations of D and D^2 are given as

$$\begin{aligned} E[D] &= \frac{1}{N}E\Big[\sum_{n=1}^{N}(p[n] - q[n])^2\Big] \\ &= 2\mu_2 + 0 = 2, \end{aligned} \quad (7)$$

$$E[D^2] = \frac{1}{N^2} E\left[\Big(\sum_{n=1}^{N}(p[n]-q[n])^2\Big)^2\right]$$
$$= \frac{4}{N^2}\sum_{k=1}^{N-1}(N-k)[1+(\mu_4-1)b^k+2a^{2k}]$$
$$+2+(2\mu_4+4)/N. \quad (8)$$

Using the typical values of a, b, and μ_4, the standard deviation σ_D of the distance D is obtained as 0.2596. Through the normal approximation of the distance $N(2,\sigma_D^2)$, the probability of false alarm P_{FA} is given as follows:

$$P_{FA} = \int_{-\infty}^{T} \frac{1}{\sqrt{2\pi}\sigma_D} \exp\Big[\frac{-(x-2)^2}{2\sigma_D^2}\Big]dx. \quad (9)$$

For a certain value of P_{FA}, the threshold T for D can be determined. In the experiments we use $T = 0.4$. Then we arrive at a very low false alarm rate of 3.512×10^{-10}.

3. PERFORMANCE EVALUATION

3.1. Performance of the proposed method

The performance of the proposed video fingerprinting method is evaluated using the fingerprint DB generated from 60 videos belonging to various genres, such as commercial, movie, music video, sports, news, documentary, etc. The length and the resolution of the videos in the DB range from 2 to 4 minutes, and from 320×240 to 720×400, respectively. The frame rate is 29.97 fps for all videos. The parameters used for the simulations are $F = 10$, $R_x = 320$, $R_y = 240$, $M_x = 4$, $M_y = 2$, and $K = 100$.

Pairwise independence of the proposed video fingerprint is evaluated using 101,768 randomly selected pairs of fingerprint sequences. Fig. 2 shows the histogram of the squared Euclidean distance between the chosen pairs. The histogram shows that the proposed fingerprint follows the stochastic model in Section 2.2 reasonably well. The mean and the standard deviation of the measured distance were 1.9663 and 0.2936, respectively, and both of them are close to the theoretically derived values.

To evaluate the robustness of the proposed video fingerprint, the original videos are subjected to various video processing steps. Mean, standard deviation, and false rejection rate P_{FR} of the measured distance between the original and the processed video fingerprints are summarized in Table 1 for 329 randomly selected fingerprint sequences whose total length corresponds to about 55 minutes. The measured distance was below the threshold $T = 0.4$ for all video processing steps except the histogram equalization which causes 9 false rejections out of 329 fingerprint sequences. This result is not surprising since the histogram equalization sometimes severely degrades the perceptual quality of the video, and 9 fingerprint sequences falsely rejected in the experiments were those extracted from severely distorted parts of the video. The overall results show that the proposed video fingerprint is highly robust against common video processing steps.

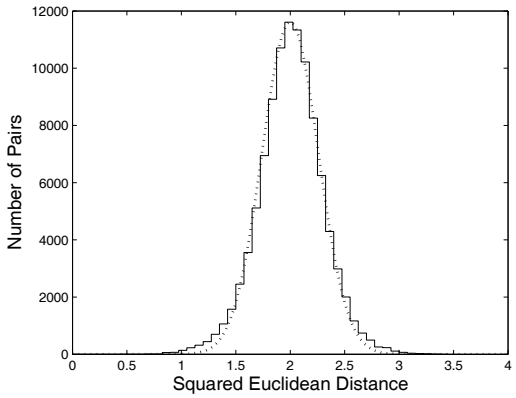

Fig. 2. Histogram of the squared Euclidean distance between the fingerprint sequences extracted from different videos.

Table 1. Mean, standard deviation (Std) and false rejection rate (with threshold $T = 0.4$) of the measured distance for different kinds of video processing steps.

Processing	**Mean**	**Std**	P_{FR}
DivX@1Mbps	0.0286	0.0320	0.0
DivX@500kbps	0.0378	0.0634	0.0
Brightness +15%	0.0288	0.0443	0.0
Red channel +20%	0.0245	0.0297	0.0
Gaussian blurring (1 pixel)	0.0349	0.0542	0.0
Histogram equalization	0.1156	0.1111	0.0274
Resizing to CIF (352x288)	0.0345	0.0398	0.0
Frame rate change (29.97 $\rightarrow$ 24 fps)	0.0408	0.0436	0.0

3.2. Comparison of the proposed method with other features

The robustness of the proposed video fingerprint is compared with that of luminance histogram, block mean luminance, and difference of block mean luminance. Difference of block mean luminance is obtained by taking difference of mean luminances of blocks adjacent in spatial and temporal domain as in [5]. Oostveen *et al.* take signs of differences and form binary fingerprints, however the values of the differences are used as fingerprints in this comparative test. The dimensions of the evaluated fingerprints are set to 8 per frame except difference of block mean luminance whose dimension is set to 9 per frame.

The comparative test is performed using the DB generated from 60 videos. Manhattan distance for luminance histogram

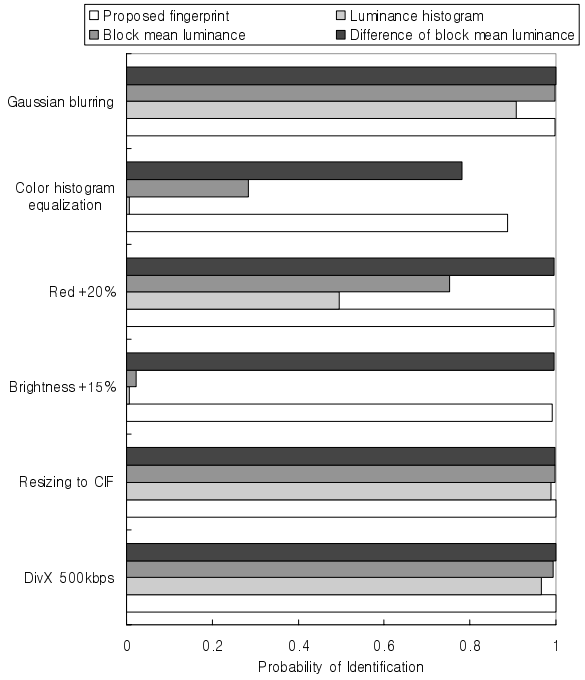

Fig. 3. Comparison of robustness of the features against video processing steps.

and squared Euclidean distance for other features are used as distance measures. The length of the fingerprint sequence for the matching is set to $K = 10$ which corresponds to 1 second. For each processed video, 4 kinds of fingerprints are extracted, and the DB position with the minimum distance is found by exhaustively searching the DB. If the DB position with the minimum distance exactly corresponds to the temporal positions of the input fingerprint sequence in the processed video, it is assumed that the input fingerprint sequence is correctly identified. Fig. 3 shows the probability of correct identification for 4 fingerprints. The results show that the proposed video fingerprint is more robust than other features for the video identification.

4. CONCLUSION

In this paper, a novel video fingerprinting method based on the centroids of gradient orientations is proposed. The performance of the proposed fingerprint is evaluated with videos from various genres using the matching threshold theoretically derived by modelling the CGO as a stationary process. The experimental results show that the proposed fingerprint is discriminative and highly robust against common video processing steps. In comparative test, the proposed fingerprint outperforms other widely-used features. It would be the future work to propose an efficient DB search algorithm suitable for the proposed fingerprint.

5. ACKNOWLEDGMENTS

This work was supported by grant No. R01-2003-000-10829-0 from the Basic Research Program of the Korea Science and Engineering Foundation and by University IT Research Center Project.

6. REFERENCES

[1] Jin S. Seo, Minho Jin, Sunil Lee, Dalwon Jang, Seungjae Lee, Chang D. Yoo, "Audio Fingerprinting Based on Normalized Spectral Subband Centroids," In *Proc. ICASSP 2005*, Philadelphia, USA, vol. 3, pp. 213-216, Mar. 2005.

[2] 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.

[3] Sen-ching Samson 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.

[4] Changick Kim, "Spatio-Temporal Sequence Matching for efficient video copy detection," SPIE Storage and Retrieval for Media Databases 2004, Jan. 2004.

[5] 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.

[6] 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.

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

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

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

[10] 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.