Video Fingerprinting by Common Features in a Scene

Jongweon Kim[*], Sungjun Han[**], Yongbae Kim[**]
[*]Dept. of Contents and Copyright, [**]Creative Content Labs
Sangmyung University
Seoul, Korea
Email: jwkim@smu.ac.kr, sungjun@cclabs.kr,
ybkim@cclabs.kr

Jungjae Lee[***]
[***]Dept. of Entertainment Business
Soongsil Cyber University
Seoul, Korea
Email:jjlee@mail.kcu.ac

***Abstract*—Video fingerprinting is an important aspect in the copyright protection field, as digital environment enables the copyright infringement to get easier and easier. There are many video contents on the Internet. Copyright owners want to identify contents on the net and to block infringed contents. The video content is invaluable because the owners invested a huge amount of money when they made the movie. In this paper, we propose an efficient algorithm to identify video contents even if we only have a video frame. The algorithm divides a video content into scenes and then extracts common features from a scene. The feature database contains only a set of features per scene. That means the proposed algorithm optimizes the feature database and the time it takes to compare the features. Also, this algorithm can identify a video using only a frame of the video.**

Keywords-video identification; common feature; scene; scale invariant feature transform .

I. INTRODUCTION

Video fingerprinting technology is an important aspect related to the video search and copyright protection. In the copyright protection field, content fingerprinting is a technical measure to block the illegal distribution of copyrighted contents on the net. Generally, there are three kinds of content filtering, namely keyword-based, hash-based, feature-based filtering. The feature-based filtering is the most powerful technology to identify the contents under several distortion attacks.

Digital Rights Management (DRM) is the most secure measure [1] to protect the content because it uses encryption technology. In February of 2007, Steve Jobs, who was former CEO of Apple Inc., announced the introduction of DRM-free service [2]; next, DRM started to disappear from contents market. Digital watermarking technology [3][4] is an alternative to the DRM-free service, but the watermark information should be embedded into the content before distribution. Indeed, the digital watermarking is not perfect to protect contents and it is sensitive to malicious attacks. These are drawbacks of the digital watermarking technology.

Watermarks offer some advantages over fingerprinting. A unique watermark can be added to the content at any stage in the distribution process and multiple independent watermarks can be inserted into the same video content. This can be particularly useful in tracing the history of video copies. Detecting watermarks in a video can indicate the source of an unauthorized copy.

While video fingerprinting systems must search a potentially large database of reference fingerprints, a watermark detection system only has to do the computation to detect the watermark. This computation can be significant and, when multiple watermark keys must be tested then, watermarking can fail to scale to User Generated Content (UGC) site volumes.

Fingerprinting technology has a number of advantages over the conventional DRMs and digital watermarking technologies. Fingerprinting technology does not need to embed any information before distribution and just stores the features into database in contrast with digital watermarking. The key problem of DRM is interoperability among DRMs and the fingerprinting enables the authorized users to use the content without any barriers. Fig. 1 shows the use case of the fingerprinting technology for the copyright protection.

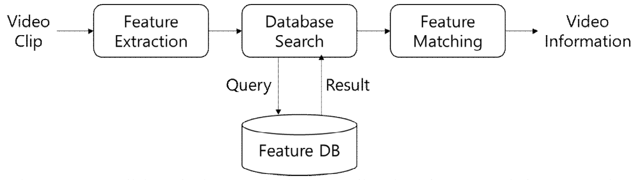


Figure 1. Traditional Fingerprinting Application for Copyright Protection

Normally, digital contents are compared based on hash values that are directly derived from the digital components of a content. However, such methods are incomplete as they can only determine absolute equality or non-equality of video data files or parts. More often than not, differences in a video codec and digital processing artifacts may cause small differences in the digital components without changing the video perceptually. Thus, when employing hash methods, a comparison for absolute equality may fail even when two video clips are perceptually identical. Moreover, hash based filtering is also of little value when one wishes to identify video clips that are similar (but not identical) to a given reference clip. The limitation of the equality and inequality decision by hash value is that the hash-based technique is not available for the similar searching [5].

On the other hand, video fingerprinting technique enables identification of videos with a different resolution compared with the original (smaller or larger) as well as identifying videos that have been slightly modified (blurring, rotation, acceleration or deceleration, cropping, insertions of new elements in the video), and videos where the audio track has been modified [5].

For the video fingerprinting, Mani Malek Esmaeili et al. suggested a fast video fingerprinting technology [6], Bo Wu et al. proposed a robust video fingerprinting using sparse represented features [7] and Mu Li et al. proposed a compact video fingerprinting using structural graphical model [8]. Although their algorithms show good performance to identify a video, the algorithms require several frames or scenes.

There are many fingerprinting algorithms in the image processing field. Nowadays, the Scale Invariant Feature Transform (SIFT) [9] algorithm is powerful to extract features from images. Although SIFT is the most powerful feature extraction algorithm, SIFT is inefficient for extracting features from video because it is a complicated algorithm and it occupies many computational resources.

In this paper, we propose a new method to block contents using fingerprinting. The rest of this paper is organized as follows. Section II describes the theory of SIFT and the burdens of the algorithm. Section III describes the proposed method and its procedure. Section IV addresses the experiment and results. Section V concludes the paper.

II. Scale Invariant Feature Transform

The SIFT can extract image features that are invariant to scale and rotation. The SIFT algorithm is comprised of four main stages: scale space extrema detection, keypoint localization, orientation computation and keypoint descriptor extraction.

The first stage to detect scale space extrema is the process to detect the invariant interest point using Difference of Gaussians (DoG) to identify the potential keypoints which are extrema. DoG is an approximation of Laplacian of Gaussians (LoG) and has low computational complexity. The Gaussian blurred images at six different scales are produced from the input image and DoGs are computed from neighbors to extract local extrema in scale space. In the second stage for keypoint localization, candidates of keypoint are localized by detecting extrema in the DoG images that are locally extremal in space and scale. The unstable kepoints (usually edges) in space are removed by thresholding for the ratio of eigenvalues of the Hessian matrix (unstable edge keypoints have high ratios, and stable corner keypoints have low ratios), low contrast keypoints are removed and the remaining keypoints are localized by interpolating across the DoG images. The third stage for orientation computation is the process to assign a principal orientation of keypoint. The directions of pixels around keypoint are computed and the histogram of the directions is used to select the orientation of keypoints. If there is another orientation over 80% of maximum histogram, the stage assigns additional keypoint. This means there can be one or more keypoints at the same point. The final stage computes the orientation of the gradients around a keypoint. This is the stage to make a highly distinctive descriptor for each keypoint. For the orientation invariance, the descriptor coordinates and gradient orientations are rotated relative to the orientation of keypoint.

For every keypoint, a set of orientation histograms is created on 4x4 pixel neighborhoods with 8 bins each (using magnitudes and orientation of samples in 16 x 16 region around the keypoint). The resulting feature descriptor will be a vector of 128 elements that is then normalized to unit length to handle illumination differences. Descriptor size can be varied, however best results are reported with 128D SIFT descriptors. SIFT descriptors are invariant to rotation, scale, contrast and partially invariant to other transformations. The SIFT descriptor size is controlled by its width, i.e. the array of orientation histograms (n x n) and number of orientation bins in each histogram (r). The size of resulting SIFT descriptor is rn^2. The value of n affects the window size around the keypoint as we use 4 x 4 region to capture pattern information, e.g. for n = 3, we will use a window, size of 12 x 12 around the keypoint. Various sizes were analyzed in [10] and it was reported that 128D SIFT is superior in terms of matching precision, i.e. n = 4 and r = 8. Most other works have used standard 128D SIFT features while very few have tried smaller SIFT descriptors for small scale works, e.g. 36D SIFT features from 3 x 3 subregions, each with 4 orientation bins, with few target images are used in [11].

Smaller sized descriptors use less memory and result in faster classification but precision rates may be affected. No research article has investigated the classification performance of SIFT descriptors of size other than 128.

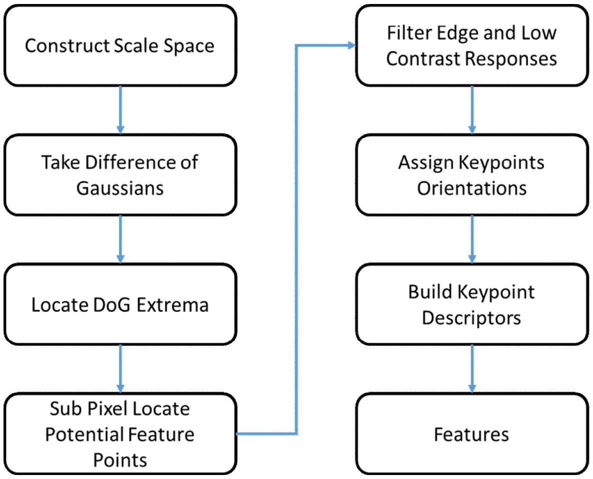

Construct Scale Space
Take Difference of Gaussians
Locate DoG Extrema
Sub Pixel Locate Potential Feature Points
Filter Edge and Low Contrast Responses
Assign Keypoints Orientations
Build Keypoint Descriptors
Features

Figure 2. Procedure of SIFT

Video clips have 30 frames per second. If SIFT is applied to video content for the video fingerprinting, the algorithm should process to extract features from each frame. This means the computational amount is very high and it is not useful for video identification.

III. Proposed Method

Our goal of the paper is how to identify the video from a frame. For this purpose, we developed a simple scene detection algorithm and a video fingerprinting technology using SIFT which has the low complexity for the image

ISBN: 978-1-61208-533-3

identification. First of all, we have to improve the computational complexity before using SIFT as a video fingerprinting technology. There are some candidates to reduce the computational amount of the feature extraction. One candidate is a temporal feature extraction from video clips. Although the temporal feature has low computational complexity, this feature cannot distinguish the video clips frame by frame. The other candidate is the binary feature extraction which is proposed by Lee et al [12]. The binary fingerprints are obtained by filtering and quantizing intermediate features extracted from an input video clip. The filters and their associated quantizers for the fingerprint extraction are selected from a class of candidate filters and quantizers using the Symmetric Pairwise Boosting (SPB) algorithm.

Our approach is to reduce the computational amount for video clips when extracting the features. Even if the proposed algorithm reduces the computational operation, it can also identify the video clips by only a frame.

A. *Scene Segmentation*

A video program such as motion pictures, TV movies, etc., has a story structure and organization. As illustrated in Fig. 3, three levels define this syntactic and semantic story structure: narrative sequence, scene and camera shot. A camera shot is a set of continuous frames representing a continuous action in time or space. It represents the fundamental unit of video production, reflecting a basic fragment of story units. A scene is a dramatic unit composed of a single or several shots. It usually takes place in a continuous time period, in the same setting, and involves the same characters. At a higher level, we have the narrative sequence, which is a dramatic unit composed of several scenes all linked together by their emotional and narrative momentum [13].

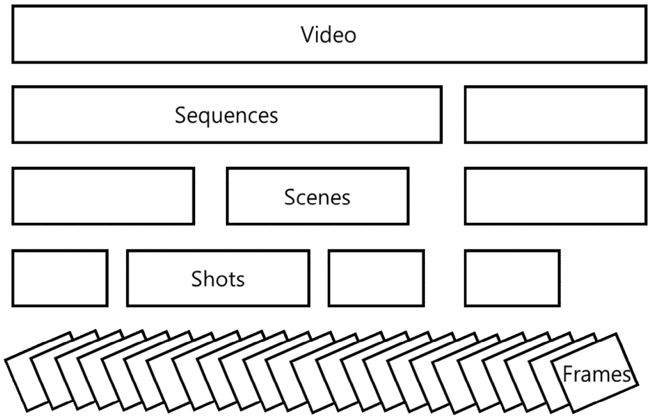


Figure 3. Video structure

In the proposed algorithm, the scene segmentation is achieved by using average of difference between frames. In a scene, the difference value of the consecutive frames is lower than that of the consecutive frames between scenes. This algorithm has simple architecture and computational advantage. Especially, this method is efficient to segment similar frames as a scene.

B. *Features for Video Fingerprinting*

There are many common features between frames in a scene. The implementation process of the feature database is as follows:

Step 1: Segment the scene from video clips.
Step 2: Extract features from each frame.
Step 3: Choose common features from features of frames.
Step 4: Store common features into database.

In step 3, all features of a scene are arranged by same feature values and descriptors. By the frequency of the same feature in a scene, the features are sorted and then selected as the common feature.

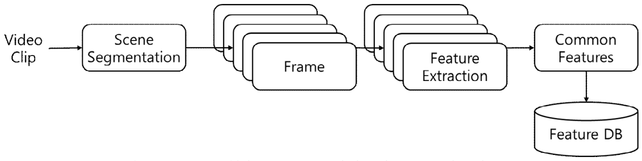


Figure 4. Build process of the feature database

Fig. 4 shows the build process of the feature database.

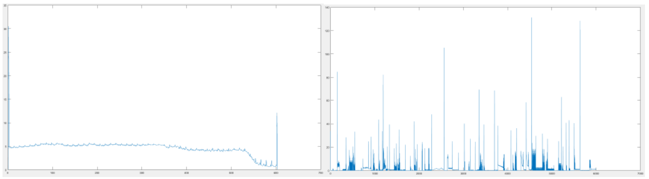
Once the feature database is implemented, the identification process is as follows:

Step 1: Choose a frame from video clips.
Step 2: Extract features from the selected frame.
Step 3: Compare the features with the database.
Step 4: Identify the video.

The main idea of the proposed algorithm is using the common features among the video scene. Any one of the fingerprinting algorithms, such as SIFT, SURF, etc., can be used to extract common features.

IV. EXPERIMENTS AND RESULTS

To evaluate the proposed method for the video fingerprinting, we have taken 4 video clips. At the first step, the video clips have been segmented by scenes and we extracted the features from each frame of the scene. The features from each frame of the scene are refined as common features between frames.



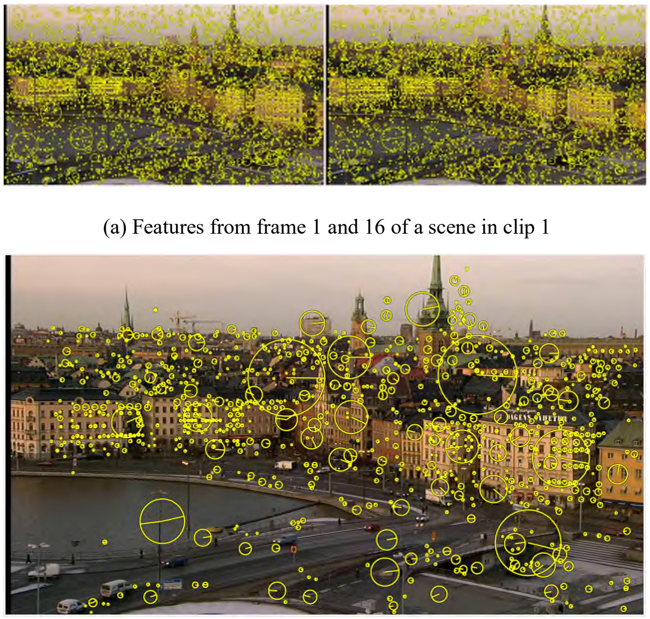
(a) Sample video clip 1 (b) Sample video clip 2

Figure 5. Scene segmentation graphs

Fig. 5 depicts the scene segmentation results of the test video clips. The pulses in the graph are boundaries of the scene. Graph (a) shows the video clip 1 has only a scene, which is slightly changed between the frames in the scene. Video clip 2 has many scenes, as shown in graph (b).

If the interval between the peaks is long, there is a long scene and if the interval is short, there is a short scene. The graph for scene segmentation is calculated from average value of the difference between a frame and the neighbor frame.

After scene segmentation, the features of every frame in the scene are extracted and then we choose the common features of all frames. Fig. 6 shows the extracted features from frames and the common features. (a) shows the features from the first frame of a scene in the test video clip 1 and the 16th frame of same scene. (b) shows the common features among all frames of the scene in the clip 1. The yellow circles are features which are extracted by SIFT. The different circle size means the feature is extracted in the different scale domain and the line in the circle represents the orientation of the feature.



(a) Features from frame 1 and 16 of a scene in clip 1

(b) Common features of a scene in clip 1

Figure 6. Features from each frame and common features

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a video fingerprinting method to identify video clips using only a frame. The video fingerprinting technology can block the illegal distribution of the infringed video contents on the internet. Our approach used spatial features of each frames and reduced the size of the feature database and amount of features in a scene. For achieving this purpose, we segmented the video clips into scenes, extracted the features of each frame in a scene and chose the common features in a scene. As a result, the number of the common features is less than the average number of the features of each frame. This results in less computation complexity when the video fingerprinting is applied to filtering of infringed contents. Moreover, the approach can identify the video clip even if there is only a frame of the video clip.

In the future work, we are going to improve the identification speed and to develop a common feature extraction algorithm. We are also planning to study a fast algorithm for comparison search in feature database.

ACKNOWLEDGMENT

This work was supported by the Technological Innovation R&D Program (S2380813) funded by the Small and Medium Business Administration (SMBA, Korea)"

REFERENCES

[1] Digital Rights Management, Wikipedia, visited Oct. 28th 2016. https://en.wikipedia.org/wiki/Digital_rights_management

[2] Steve Jobs, Thoughts on Music, Apple Wet Site, Feb., 2007. http://www.apple.com/kr/hotnews/thoughtsonmusic/

[3] I. J. Cox, J. Kilian, T. Leighton and T. Shamoon. Secure Spread Spectrum Watermarking for Multimedia. IEEE Transactions on Image Processing, vol. 6, pp.1673-1687, 1997.M

[4] J. H. Nah, J. W. Kim and J. S. Kim, Video Forensic Marking Algorithm using Peak Position Modulation, Appl. Math. Inf. Sci., vol.7, no.6, pp.2391-2396, 2013.

[5] Digital video fingerprinting, Wikipedia, visited Oct., 25, 2016, https://en.wikipedia.org/wiki/Digital_video_fingerprinting

[6] M. M. Esmaeili, M. Fatourechi, and R. K. Ward, A Robust and Fast Video Copy Detection System Using Content-Based Fingerprinting, IEEE Transactions on Information Forensics and Security, vol. 6, no. 1, pp.213-226, 2011

[7] B. Wu, S. Krishnan, N. Zhang and L. Su, Compact and Robust Video Fingerprinting using Sparse Represented Features, Multimedia and Expo (ICME), 2016 IEEE International Conference on, pp.1-6, 2016

[8] M. Li and V. Monga, Compact Video Fingerprinting via Structural Graphical Models, IEEE Transactions on Information Forensics and Security, vol.8, no.11, pp.1709-1721, 2013

[9] N. Y. Khan, B. McCane and G, Wyvill, " SIFT and SURF Performance Evaluation Against Various Image Deformations on Benchmark Dataset", International Conference on Digital Image Computing: Techniques and Applications, pp.501-506, 2011.

[10] D. Lowe, Distinctive Image features from scale invariant keypoints, International journal of Computer Vision, 60, pp.91-110, 2004.

[11] W. Daniel, R. Gerhard, M. Alessandro, D. Tom and S. Dieter, Pose Tracking from Natural Features on Mobile Phones, Proc. International Symposium on Mixed and Augmented Reality, pp.125-134, 2008.

[12] S. Lee, C. D. Yoo and T. Kalker, "Robust Video Fingerprinting Based on Symmetric Pairwise Boosting," IEEE Transactions on Circuit and Systems for Video Technology, vol.19, no.9, pp.1379-1388, Sep. 2009, doi: 10.1109/TCSVT.2009.2022801

[13] W. Mahdi, L. Chen and M. Ardebilian, Automatic video scene segmentation based on spatial-temporal clues and rhythm, 2014. https://arxiv.org/abs/1412.4470v1