J. Vis. Commun. Image R. 32 (2015) 120–129

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Graph-based video fingerprinting using double optimal projection ☆

Xiushan Nie [a,*], Ju Liu [b], Qian Wang [a], Wenjun Zeng [c]

[a] School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014, China
[b] School of Information Science and Engineering, Shandong University, Jinan 250100, China
[c] Department of Computer Science, University of Missouri, Columbia, MO 65211, United States

ARTICLE INFO

Article history:
Received 14 April 2014
Accepted 3 August 2015
Available online 10 August 2015

Keywords:
Video copy detection
Content protection
Video fingerprinting
Dimensionality reduction
Double optimal projection
Graph model
Intra-cluster
Inter-cluster

ABSTRACT

A double optimal projection method that involves projections for intra-cluster and inter-cluster dimensionality reduction are proposed for video fingerprinting. The video is initially set as a graph with frames as its vertices in a high-dimensional space. A similarity measure that can compute the weights of the edges is then proposed. Subsequently, the video frames are partitioned into different clusters based on the graph model. Double optimal projection is used to explore the optimal mapping points in a low-dimensional space to reduce the video dimensions. The statistics and geometrical fingerprints are generated to determine whether a query video is copied from one of the videos in the database. During matching, the video can be roughly matched by utilizing the statistics fingerprint. Further matching is thereafter performed in the corresponding group using geometrical fingerprints. Experimental results show the good performance of the proposed video fingerprinting method in robustness and discrimination.

1. Introduction

Recent advancements in network and multimedia technologies have facilitated the distribution and sharing of digital videos over the Internet. To protect intellectual property as well as to promote multimedia services, the proper distribution and use of video content must be ensured. Watermarking can be employed to mitigate this security issue, and it is suitable if the embedder has controlled over the content creation stage. However, this requirement may be difficult to satisfy in practical applications, particularly in content filtering on User Generated Content (UGC) sites. In contrast, Robust video fingerprinting, which is also known as content-based video identification or perceptual video hashing, does not require access to the content at the time of creation. It can be used to identify existing multimedia content and ensure the video content protection.

In a video fingerprinting system, original video fingerprints are extracted by using the appropriate methods and are then stored in a database. For query videos that may be subjected to content-preserving attacks, such as noises, geometrical transformations, and compression, the fingerprints are extracted and then compared with the fingerprints stored in the database. The query videos are then identified as copies if the fingerprints match. Otherwise, the query is considered as a new video. Therefore, robustness and discrimination are two primary considerations in designing a fingerprinting system. i.e., the fingerprint should resist to content-preserving operations, and it also should be distinguishable when two videos have perceptually different contents.

Many video fingerprinting methods have been developed in recent years. The spatial-based (e.g., luminance of blocks [1] and centroid of gradient [2]) and transform-domain-based methods (e.g., radon transformation [3] and three-dimensional Discrete Cosine Transformation (3D-DCT) [4]) are two types of video fingerprinting in the early days. Some researchers also improved the early methods, such as the method in [5] which used a new learned basis set algorithm based on 3D-DCT method to generate video fingerprint. Recently, more and more methods combined different features to form video fingerprints. Ngo et al. [6] used a temporal network and partial alignment to detect video copies. Esmaeili and Ward [7] proposed a video fingerprinting based on temporally informative representative images (TIRIs), and from then on many researchers began to exploit TIRIs to extract video fingerprint, such as Rubini et al. [8] and Divay and Wiselin [9] who calculated horizontal and vertical coefficients of DCT from TIRIs as the video fingerprint. Sun et al. [10] proposed a temporally visual weighting (TVW) method based on visual attention to produce video fingerprints.

Although these methods have shown success in video copy detection, there is a problem that is seldom exploited. Most

☆ This paper has been recommended for acceptance by M.T. Sun.
* Corresponding author.

http://dx.doi.org/10.1016/j.jvcir.2015.08.001
1047-3203/

state-of-the-art video fingerprinting technologies generate features in the original video feature space, the dimensionality of which can range from several hundreds to thousands. Therefor, in real-world image/video retrieval systems, a serious problem in applying statistical techniques to image/video retrieval is the "curse of dimensionality" [11]. That is, when the dimension of feature space (such as color, shape and textures) is high, the indexing efficiency will fall rapidly. Therefore, dimensionality reduction should be considered in image/video analysis. In addition, most natural signals have a smooth and regular structure, i.e., piecewise smoothness, that permits substantial dimension reduction with minimal or no loss of content information[12], therefore, dimensionality reduction is feasible. Previous works [13–15] have performed a preliminary exploration of video copy detection using dimensionality reduction. These works used multidimensional scaling and locally linear embedding to generate video fingerprints. However, these methods cannot consider both the local and global structures of a video. The global and local structures are both important to generate video fingerprints. We propose a video fingerprinting scheme which consists of intra-cluster and inter-cluster dimension reduction to optimally preserve global and local information together to address this limitation in this study, and we call this global and local-preserving projection as double optimal projection (DOP).

Video frame clustering is the first step of DOP. Determining the method by which to divide video frames into different clusters is important in DOP because frames in a cluster that are similar to each other should be considered as the local structure, whereas relationship among different clusters should be considered as the global structure. In this study, a K-means-like clustering algorithm is used to partition video frames. Initial cluster center and distance metric are two important components in the K-means-like clustering algorithm, given that relatively optimal cluster centers can better partition the original data. We propose a graph model to select initial cluster centers and define a good distance metric for the proposed clustering algorithm.

The contributions of the proposed work are summarized below.

- *Graph model*: When applying dimensionality reduction on video frames, relationships among frames may be changed in low-dimensional space. Therefore, we should partition frames to preserve the local and discriminative structure. The graph model is an efficient method for video frames partition, because it can select the relatively optimal center of each cluster using mature graph theory aside from providing an efficient distance metric for the clustering algorithm. Moreover, the graph Laplacian theory can applied in videos based on the graph model when performing DOP.
- *DOP*: DOP preserves both the local and global structures of video, in which different weights are proposed to construct inter- and intra-cluster projections. The video frames are projected from their original high-dimensional feature space to a low-dimensional space where the video fingerprint is computed.
- *Fast matching*: In the proposed fingerprinting method, the fingerprint consists of statistics and geometrical fingerprints. To determine whether a video query is copied from a video in a database, the statistics fingerprint is initially matched to divide fingerprints in the database into different groups. Further matching is then performed in the selected group using the geometrical fingerprint.

The remainder of the paper is structured as follows: The proposed method is described in Section 2. Section 3 shows the experimental results and analysis, and the conclusions are summarized in Section 4.

2. Proposed method

The framework of the proposed fingerprinting method is shown in Fig. 1. Before fingerprint extraction, video signals are preprocessed. Copies of the same video clips with different frame sizes, numbers, and frame rates usually exist in the video database. Given that the proposed scheme targets a constant number of binary bits as the video fingerprint, we standardize the input video clips to keep the size of the fingerprint sequence constant. In the experiments, we normalized the video clip as $320 \times 240 \times 500$ (frame width $\times$ frame height $\times$ number of frames). As a result, the fingerprinting method is robust to changes in frame size and frame rate. Graph model-based clustering and DOP were then performed after preprocessing. DOP was performed to obtain compact representations of the original data. The fingerprint sequence was generated from the low-dimensional space, while querying the video database, a fingerprint sequence was generated and then matched with the stored sequences from the database.

2.1. Video frames clustering

Both the global and local information are important for video analysis. However, no general dimensionality reduction method that preserves both the global and local information exists. To address this problem, DOP is used in our study. As mentioned in Section 1, clustering is an important step prior to conducting DOP. K-means clustering is a well-known algorithm used in many applications. However, the parameter K is difficult to estimate and should thus be given in advance. Given that the initial centroid in each cluster has a significant effect on the clustering result, it should be selected carefully. To address these two issues, we propose a graph model-based K-means clustering method in this study, which we call K-means-like clustering to distinguish it from the classic K-means-clustering.

In the proposed method, the video is first converted to a graph. Unlike K-means clustering, which selects random vertices, we select several specific vertices of the graph as the initial cluster centers. The parameter K is set as the number of initial cluster centers. In addition, we use the spatio-temporal information of a video to define the weights of the graph. These weights are taken as the distance metric, which not only denotes the comprehensive relations among frames, but also reduces the complexity of the proposed clustering algorithm.

A video can be represented as a graph $G_w = (V, E, \mathbf{W})$ in a high-dimensional space with frames as its vertex set V. The vertices are connected by edges in edge set E. In addition, each edge is assigned a positive weight $w_{ij} \in \mathbf{W}$. Defining the weight w_{ij} of the corresponding edge, which measures the distance between pairs of frames i and j, is a key step in building the graph. In this study, a smaller w_{ij} indicates greater similarity of frames i and j. The weight w_{ij} is proportional to temporal similarity and is inversely proportional to spatial similarity. The weight w_{ij} is taken as the distance metric in the K-means-like clustering algorithm, and the distance matrix $\mathbf{W} = \{w_{ij}\}$ is calculated before clustering.

Let $sim(i,j)$ and $|f_j - f_i|$ denote the spatial and temporal similarity between the ith and jth frames, respectively. To satisfy the aforementioned relation, w_{ij} is computed as:

$$w_{ij} = \exp\left\{\frac{-k \cdot sim(i,j)}{|f_j - f_i|}\right\}, \tag{1}$$

where k is used to emphasize the importance of spatial similarity, and $|f_j - f_i|$ is the time difference between frames i and j.

Natural image signals are highly structured. The pixels of such signals exhibit strong dependencies, especially when they are