Int. J. Electron. Commun. (AEÜ) 69 (2015) 82–89

Contents lists available at ScienceDirect

International Journal of Electronics and Communications (AEÜ)

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

Robust video identification approach based on local non-negative matrix factorization

Zhe-Ming Lu [a], Bo Li [a], Qing-Ge Ji [b,*], Zhi-Feng Tan [b], Yong Zhang [c]

[a] School of Aeronautics and Astronautics, Zhejiang University, Hangzhou 310027, China
[b] School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510006, China
[c] ATR National Defense Technology Key Laboratory, College of Information Engineering, Shenzhen University, Shenzhen 518060, China

ARTICLE INFO

Article history:
Received 11 March 2014
Accepted 25 July 2014

Keywords:
Video identification
Non-negative matrix factorization
Local non-negative matrix factorization
Shot detection
Content preserved distortion

ABSTRACT

With the popularization of media-capture devices and the development of the Internet's basic facilities, video has become the most popular media information in recent years. The massive capacity of video imposes the demand of automatic video identification techniques which are very important to various applications such as content based video retrieval and copy detection. Therefore, as a challenging problem, video identification has drawn more and more attention in the past decade. The problem addressed here is to identify a given video clip in a given set of video sequences. In this paper, a robust video identification algorithm based on local non-negative matrix factorization (LNMF) is presented. First, some concepts about LNMF are described and the way of finding the factorized matrix is given. Then, its convergence is proven. In addition, a LNMF based shot detection method is proposed for constructing a video identification framework completely based on LNMF. Finally, a LNMF based identification approach using Hausdorff distance is introduced and a two-stage search process is proposed. Experimental results show the robustness of the proposed approach to many kinds of content-preserved distortions and its superiority to other algorithms.

1. Introduction

With the prevalence of media-capture devices and the perfection of network infrastructure, the volume of multimedia information has showed a great increase in recent years. By virtue of the properties of intuitive, easy capturing and content-rich, video has become the most popular media information which can be seen from the growing popularity of many kinds of video sharing web sites like YouTube. As a result, the massive capacity of video imposes the demand of identification techniques. The task of video identification is to find the video sequences derived from the same source. One typical application is content based video retrieval. For instance, a part of the video clip may be released on the Internet as content abstraction. The person who is attracted by its content can find the whole version using the short clip as reference with the aid of video identification. Another typical application is video copy detection, which is designed to judge whether there exists a common segment in two different video sequences. Since there are various visual transformations on video sequences, video copy segments possessing identical visual content may not share the same appearance. It is important to research how to uncover the underlying common patterns among visually similar segments and construct robust features against various usual transformations.

Video identification has drawn much attention in the past decade. The global features such as intensity or color histograms were adopted by most of the early approaches [1]. Motion direction [2] and trajectory [3] were also utilized for facilitating video identification with the consideration of the dynamic nature of video. Moreover, some researchers exploited the combined features for video identification. For example, Hoad et al. [4] combined the shot length, color information and centroid motion to generate the signature, and some information such as the subtitle and audio was employed for identification in [5]. In recent years, many algorithms based on local spatial–temporal features have been proposed [6]. In addition, robust hash algorithms [7,8] were also applied for video identification. Different from the aforementioned schemes, robust hash has to consider the security aspect of feature extraction to resist content forgery attacks. As it is reported by most literatures, resisting some content preserved distortions, such as translation and rotation, is still one of the most challenging problems in video identification. Our goal in this paper is to develop a video identification approach that is robust against such distortions.

* Corresponding author. Tel.: +86 20 84110614; fax: +86 20 84110614.
E-mail addresses: issjqg@mail.sysu.edu.cn, zhemingl@yahoo.com (Q.-G. Ji).

http://dx.doi.org/10.1016/j.aeue.2014.07.021
1434-8411/

Non-negative matrix factorization (NMF) is a relatively new matrix factorization algorithm proposed in recent years [9]. Unlike the traditional factorization algorithm (such as SVD and QR decomposition), non-negative constraints are imposed by NMF. As all the entries in the factorized matrices are non-negative, NMF has more intuitive meaning than other methods. NMF has attracted broad attention by the researchers in the field of matrix theory and signal processing, and NMF has been successfully applied to many fields such as face recognition [10,11], text mining [12,13], audio signal analysis [14]. Besides, some scholars imposed other constraints on NMF according to the characteristics of the application settings they researched on, which gave rise to various extensions of NMF. It can be seen from the following parts that NMF owns the ability of dimensionality reduction and can be used for extracting the most representative content of an image set.

However, NMF is seldom applied to video signal processing. As we all know, there is a great amount of redundant information in the video sequence. Since the properties of NMF are suitable for video identification, a video identification approach based on an extension of NMF is proposed in this paper. The rest of this paper is organized as follows. In Section 2, the background knowledge of NMF used in face recognition is introduced. In Section 3, the proposed approach is described in detail. In Section 4, experimental results of the proposed approach and comparisons with other algorithms are shown. In Section 5, conclusions are drawn and the future work is suggested.

2. Preliminaries

Since the proposed approach is related to NMF, we first introduce some concepts about NMF and how it is applied in face recognition in this section. The problem of face recognition can be simply stated as follows: Given a set of labeled face images (the learning set) and an unlabeled set of face images from the same group of people (the test set), the task is to identify which person each unlabeled face image belongs to. A very simple solution to this problem is using a nearest neighbor classifier directly in the image space [15]. Under this approach, an unlabeled image is classified by assigning to it the label of the closest point, where distances are simply measured in the image space. Simple as it is, its disadvantage is obvious. However, such approach is computationally expensive and requires a large number of storage units. Besides, it is not robust when the images in the learning set and the images in the test set are collected under different lighting conditions [16]. As the correlation approach costs lots of time and space, a natural idea is to adopt dimensionality reduction methods. Generally, such methods can be described as follows: Let a set of N training images be given as an $n \times N$-sized matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, where $\mathbf{x}_i \in \mathbf{R}^n$ $(1 \le i \le N)$ is viewed as a vector in the n-dimensional space. Let us consider a linear transformation that maps the original n-dimensional space into an m-dimensional feature space, where $m < n$. The new feature vector $\mathbf{h}_i \in \mathbf{R}^m$ $(1 \le i \le N)$ satisfies the following equation:

$$x_i = Bh_i \quad i = 1, 2, \ldots, N \tag{1}$$

where $\mathbf{B} \in \mathbf{R}^{n \times m}$ is a matrix with orthonormal columns. And the following equations hold:

$$X = BH \tag{2}$$

$$h_i = B^T x_i \quad i = 1, 2, \ldots, N \tag{3}$$

where $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N]$. Eq. (3) is derived from the property of the matrix with orthonormal columns, i.e., $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Let $\mathbf{b}_k$ denote the kth column vector in $\mathbf{B}$ and h_{ki} denote the kth entry in $\mathbf{h}_i$, we can get the following equation based on Eq. (2).

$$x_i = \sum_{k=1}^{m} h_{ki} b_k \tag{4}$$

Every $\mathbf{b}_k$ has the same dimension as each original image and they are called the basis images. From Eq. (4) we can see that each image $\mathbf{x}_i$ can be represented as the linear combination of the basis images whose coefficients are stored in $\mathbf{h}_i$. Then $\mathbf{h}_i$, instead of $\mathbf{x}_i$, is used for later training. In the classification stage, the unlabeled image is also projected onto the m-dimensional feature space via the same matrix $\mathbf{B}$, i.e., the projection is done through Eq. (3). Different methods may give different ways to construct $\mathbf{B}$, and in most of the methods, each "=" in the above equations should be replaced by "≈" as the matrix $\mathbf{X}$ is only approximately factorized.

2.1. Principal component analysis

A common technique used for dimensionality reduction in face recognition is principal component analysis (PCA). It is the technique that chooses a dimensionality reduction linear projection that maximizes the scatter of all projected samples. Formally, the total scatter matrix $\mathbf{S}_T$ of the original training images is defined as:

$$S_T = \sum_{i=1}^{m} (x_i - \mu)(x_i - \mu)^T \tag{5}$$

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{6}$$

Applying the linear projection, the total scatter of the projected feature vectors $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N$ can be calculated as $\mathbf{B}^{\mathrm{T}}\mathbf{S}_T\mathbf{B}$. In PCA, the projection $\mathbf{B}_{\mathrm{opt}}$ is chosen to maximize the determinant of the total scatter matrix, i.e.,

$$B_{opt} = \arg\max_B |B^T S_T B| = [b_{opt1}, b_{opt2}, \ldots, b_{optm}] \tag{7}$$

where $\{\mathbf{b}_{\mathrm{opt}i} \mid i = 1,2,\ldots,m\}$ is the set of n-dimensional eigenvectors of $\mathbf{S}_T$ corresponding to the m largest eigenvalues. Here, the columns of $\mathbf{B}_{\mathrm{opt}}$ are orthonormal and the rows of $\mathbf{H}$ are mutually orthogonal. Obviously, the PCA factorization approach imposes no other constraints except for orthogonality, thus arbitrary sign of the entries in $\mathbf{B}$ and $\mathbf{H}$ is allowed. Consequently, many basis images or eigenfaces lack intuitive meaning, and a linear combination of them involves in complex cancellations between positive and negative numbers.

2.2. Non-negative matrix factorization

In NMF, non-negative constraints are imposed instead of the orthogonality that is imposed in PCA. As a result, the entries of $\mathbf{B}$ and $\mathbf{H}$ are all non-negative, thus, only non-subtractive combinations are allowed. It is compatible to the intuitive notion that the whole is formed by accumulating all the combining parts, and it is the way how the NMF learns a part-based representation [9].

Let $\mathbf{Y}$ denote the product of $\mathbf{B}$ and $\mathbf{H}$, then NMF uses the divergence of $\mathbf{X}$ from $\mathbf{Y}$ as the measure of the cost of factorizing $\mathbf{X}$ that is defined as:

$$D(X||Y) = \sum_{i,j} \left(x_{ij} \log \frac{x_{ij}}{y_{ij}} - x_{ij} + y_{ij} \right) \tag{8}$$