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:

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

Here, when $\sum_{i,j}x_{ij}=\sum_{i,j}y_{ij}=1$, $D(\mathbf{X}||\mathbf{Y})$ reduces to Kullback–Leibler divergence. Thus, the NMF factorization can be defined as:

$$\min_{B,H} \quad D(\mathbf{X}||\mathbf{BH}) \quad \text{s.t} \quad \mathbf{B},\mathbf{H}\geq 0;\quad \sum_i b_{ij}=1,\forall j \tag{9}$$

where all entries in **B** and **H** are non-negative. The above optimization can be done by multiplicative update rules [17], and **X** can be approximately factorized into two non-negative matrices **B** and **H**.

2.3. Local non-negative matrix factorization

As Li *et al.* stated in [18] that the NMF model defined by the constrained minimization of Eq. (8) imposes no constraints on the spatial locality and hence such factorization can hardly reveal local features in the data **X**. So they imposed additional constraints on **B** and **H** to strengthen the ability in learning local features and achieved higher recognition accuracy than NMF in their experiments. Such approach is called local non-negative matrix factorization (LNMF). The constraints on **B** and **H** will be discussed in detail in Section 3. An example given in Fig. 1 illustrates the ability of learning local features between LNMF and NMF. We can see from Fig. 1 that each basis image generated by the LNMF procedure reveals a much smaller part of the human face than the NMF scheme, and hence provides truly parts-based representation.

Since LNMF can be used for dimensionality reduction and has the ability to effectively extract the essential information from an image set (learning parts of the human face), it owns the properties that video identification needs. For video sequence, LNMF can not only remove much of the redundant spatial–temporal information, but also result in basis images that can serve as a simplified description of a sequence of consecutive frames. In this paper, we propose a video identification approach based on LNMF, which can be described in detail as follows.

3. The proposed approach

3.1. Feature extraction

Before showing how the matrix **X** is factorized, we first show how it is constructed. Each $\mathbf{x}_i$ is the feature vector extracted from the ith image, and it is generated as follows. First, the i-image is represented using the HSV color model. Then the V component of the image is extracted for histogram calculation with M bins. Finally, the calculation result is stored to form a M-dimensional feature vector $\mathbf{x}_i$. The reason that we choose such feature extraction scheme in our approach lies in two aspects. First, we adopt the histogram information other than the image pixels since the former is more robust to some content preserved distortions such as translation and rotation. Second, as the LNMF procedure needs lots of computation time, only the most significant component of the image is selected for feature extraction so as to save computation cost.

3.2. Proposed constraints on NMF

As it is stated in Section 2.3, LNMF is extended from NMF by adding some constraints related to the application setting. In this paper, we choose two additional constraints to impose on the NMF as follows.

Constraint 1. In order to minimize the number of basis components required to represent **X**, we should prevent a basis component from being further decomposed into more components. Given the existing constraints $\sum_i b_{ij}=1$ $(1\leq j\leq m)$, we hope that $\sum_i b_{ij}^2$ $(1\leq j\leq m)$ should be as small as possible so that $\mathbf{b}_j$ contains as many non-zero entries as possible. Let $\mathbf{U}=[u_{ij}]=\mathbf{B}^{\mathrm{T}}\mathbf{B}$, then Constraint 1 can be expressed as $\sum_i u_{ii}=\min$.

Constraint 2. Different basis images should be as orthogonal as possible, so as to minimize the redundancy between different basis images. Therefore, Constraint 2 can be expressed as $\sum_{i\neq j}u_{ij}=\min$.

By incorporating the above two constraints, we have $\sum_{i,j}u_{ij}=\min$, which results in a new constrained divergence as the object function for LNMF:

$$D(\mathbf{X}||\mathbf{BH})=\sum_{i,j}\left(x_{ij}\log\frac{x_{ij}}{y_{ij}}-x_{ij}+y_{ij}\right)+c\sum_{i,j}u_{ij} \tag{10}$$

where c is a constant. Thus the aim of the LNMF procedure is to find the best **B** and **H** which can minimize the constrained divergence (10).

A local solution to this constrained minimization can be found by using the following three-step update rules:

$$h_{kl}=h_{kl}\sum_i x_{il}\frac{b_{ik}}{\sum_k b_{ik}h_{kl}} \tag{11}$$

$$b_{kl}=\frac{b_{kl}\sum_j x_{kj}\left(h_{lj}/\sum_t b_{kt}h_{tj}\right)}{\sum_t h_{lt}} \tag{12}$$

$$b_{kl}=\frac{b_{kl}}{\sum_t b_{tl}} \tag{13}$$

The searching process is iterative, and in each iteration, **H** and **B** are updated sequentially through (11)–(13) with the other fixed.

To prove the correctness of above searching process, we adopt an auxiliary function similar to that used in the Expectation-Maximization algorithm [19].

Definition 1. $G(p,p')$ is an auxiliary function for a common function $L(p)$ if the following conditions are satisfied.

$$G(p,p')\geq L(p) \quad \text{and} \quad G(p,p)=L(p) \tag{14}$$

The auxiliary function is a useful concept due to the following lemma.

Lemma 1. *If G is an auxiliary function, then L is non-increasing under the update*

$$p^{t+1}=arg\ \min_p G(p,p^t) \tag{15}$$

Proof. According to (14) and (15), we can easily get $L(p^{t+1})\leq G(p^{t+1},p^t)\leq G(p^t,p^t)=L(p^t)$.

Based on the above definition and lemma, we can use the auxiliary function for minimizing the constrained divergence (10) as follows:

Updating H: **H** is updated by minimizing $L(\mathbf{H})=D(\mathbf{X}||\mathbf{BH})$ with **B** fixed. An auxiliary function is constructed for $L(\mathbf{H})$ as

$$G(\mathbf{H},\mathbf{H}')=\sum_{i,j}x_{ij}\log x_{ij}-\sum_{i,j,k}x_{ij}\frac{b_{ik}h'_{kj}}{\sum_t b_{it}h'_{tj}}\left[\log(b_{ik}h_{kj})-\log\frac{b_{ik}h'_{kj}}{\sum_t b_{it}h'_{tj}}\right]$$
$$+\sum_{i,j}y_{ij}-\sum_{i,j}x_{ij}+c\sum_{i,j}u_{ij} \tag{16}$$

It is easy to verify $G(\mathbf{H},\mathbf{H})=L(\mathbf{H})$ when substitute $\mathbf{H}'=\mathbf{H}$ into Eq. (16). Now we turn to prove $G(\mathbf{H},\mathbf{H}')\geq L(\mathbf{H})$. As we all know, the Jensen's inequality holds:

$$f\left(\sum_k \lambda_k x_k\right)\leq\sum_k\lambda_k f(x_k) \tag{17}$$

if $f(x)$ is a convex function, $\lambda_k \geq 0$, and $\sum_k \lambda_k = 1$.

Since $\log(x)$ is a concave function, then $-\log(x)$ is a convex function. Thus we substitute $f(x) = -\log(x)$, $x_k = x_k/\lambda_k$ into Eq. (17), and we have

$$-\log\left(\sum_k x_k\right) \leq -\sum_k \lambda_k \log\left(\frac{x_k}{\lambda_k}\right) \tag{18}$$

Then we have the following result by substituting $x_k = b_{ik}h_{kj}$, $\lambda_k = \left(b_{ik}h'_{kj}/\sum_t b_{it}h'_{tj}\right)$:

$$-\log\left(\sum_k b_{ik}h_{kj}\right) \leq -\sum_k \frac{b_{ik}h'_{jk}}{\sum_t b_{it}h'_{tj}}\left(\log b_{ik}h_{kj} - \log\frac{b_{ik}h'_{jk}}{\sum_t b_{it}h'_{tj}}\right) \tag{19}$$

That is, $G(\mathbf{H},\mathbf{H}') \geq L(\mathbf{H})$ holds.

Thus, to minimize $L(\mathbf{H})$ with respect to $\mathbf{H}$, we can update $\mathbf{H}$ using

$$H^{t+1} = arg \min_H G(H, H^t) \tag{20}$$

Such an $\mathbf{H}$ can be found by setting $(\partial G(\mathbf{H}, \mathbf{H}')/\partial h_{kl}) = 0$ for all k, l. Then we have:

$$\frac{\partial G(\mathbf{H}, \mathbf{H}')}{\partial h_{kl}} = -\sum_i x_{il}\frac{b_{ik}h'_{kl}}{\sum_k b_{ik}h'_{kl}}\frac{1}{h_{kl}} + \sum_i b_{ik} = 0 \tag{21}$$

Thus we find the following update rule for h_{kl} with the existing constraint $\sum_i b_{ik} = 1$

$$h_{kl} = h'_{kl}\sum_i x_{il}\frac{b_{ik}}{\sum_k b_{ik}h'_{kl}} \tag{22}$$

Updating B: $\mathbf{B}$ is updated by minimizing $L(\mathbf{B}) = D(\mathbf{X}||\mathbf{BH})$ with $\mathbf{H}$ fixed. The auxiliary function for $L(\mathbf{B})$ is constructed as follows

$$G(\mathbf{B}, \mathbf{B}') = \sum_{i,j} x_{ij}\log x_{ij} - \sum_{i,j,k} x_{ij}\frac{b'_{ik}h_{kj}}{\sum_t b'_{it}h_{tj}}\left[\log(b_{ik}h_{kj}) - \log\frac{b'_{ik}h_{kj}}{\sum_t b'_{it}h_{tj}}\right]$$
$$+\sum_{i,j} y_{ij} - \sum_{i,j} x_{ij} + c\sum_{i,j} u_{ij} \tag{23}$$

Similarly, we can prove $G(\mathbf{B},\mathbf{B}) = L(\mathbf{B})$ and $G(\mathbf{B},\mathbf{B}') \geq L(\mathbf{B})$. By setting $(\partial G(\boldsymbol{B}, \boldsymbol{B}')/\partial b_{kl}) = 0$, we find

$$b_{kl} = \frac{b'_{kl}\sum_j x_{kj}\left(h_{lj}/\sum_t b'_{kt}h_{tj}\right)}{\sum_t h_{lt} + 2c\sum_t b_{kt}} \tag{24}$$

Since $b_{i,j} \in [0,1]$ and $\mathbf{B}$ is an approximately orthogonal basis, and $x_{i,j} \in [0, N_H]$, where N_H is the total number of pixels in the image, there must be $h_{ij} \geq x_{i,j} \in [0, N_H]$. Therefore, we can set c to be not too large (e.g., $c = 0.5$) so that $\sum_t h_{lt} \gg 2c\sum_t b_{kt}$ and thus the denominator of Eq. (24) is approximately equal to $\sum_t h_{lt}$. Then we have Eq. (12).

We can see from the above discussion that $c\sum_{i,j} u_{ij}$ has no effect on Eqs. (11) and (12) under the setting of our approach. So we maintain the constraint of the column normalization by adding the update rule that satisfies the constraints $\sum_k b_{kl} = 1$, for all l. Then we have Eq. (13).

From the above analysis, we prove that the three-step update rules (11)–(13) result in a sequence of non-increasing values of $D(\mathbf{X}||\mathbf{BH})$, and hence converges to a local minimum. As they are multiplicative update rules, it follows that every updated matrix $\mathbf{B}^i$ and $\mathbf{H}^i$ is non-negative as long as $\mathbf{B}^0$ and $\mathbf{H}^0$ are initialized as non-negative. In addition, a small constant e is added to every denominator in (11)–(13) to prevent it from being zero. And this iterated process is stopped when the iteration number reaches the predefined constant *Iter_max*.

3.3. Shot detection

Before the video identification process, the shot detection is a necessary step. Although the shot detection field has been researched for many years and there are a lot of methods, we choose to develop a shot detection approach based on LNMF in our work for achieving a video identification system completely based on LNMF. And some results produced in the shot detection step can even be used in the following identification process which can save much computation cost. The proposed shot detection method can be described as follows.

First, the video is scaled to the same resolution 360×240. As diverse sequences usually have different spatial size, we normalize them into the same size in order to make them comparable. Then a two dimensional Gaussian filter is implemented on the scaled frame to eliminate the spatial noise. The pre-processing step not only enhances the robustness of the approach to some noise, but also reduces the computational burden as the size of the sequences decreases.

After that, the first N_D frames of the video sequence are chosen sequentially (In our experiments, N_D is set to 2000 if $N_V \geq 2000$, where N_V is the total number of frames in the video, otherwise, N_D is set to N_V) and the feature extraction process described in Section 3.1 is performed to construct the original feature matrix $\mathbf{X}_V$ with size $M \times N_D$.

Once the pre-processing step is finished and $\mathbf{X}_V$ is constructed, the factorization described in Section 3.2 can be carried out with the number of basis images m set to m_V (In the experiment, $m_V = 10$). Then the basis image matrix $\mathbf{B}_V$ can be obtained. As every basis vector (basis image) is approximately orthogonal to the other, then $\mathbf{B}_V{}^{\mathrm{T}}\mathbf{B}_V \approx \mathbf{I}$ holds, and Eq. (3) can be still used for projection. Let $\mathbf{v}_i$ denote the projected vector of $\mathbf{x}_i$, then it can be defined as:

$$\mathbf{v}_i = \mathbf{B}_V^T\mathbf{x}_i \quad i = 1, 2, \ldots, N_V \tag{25}$$

Similar as it is stated in [20], the projected feature space removes the noise or the trivial variations in the video sequence and the clustering of visually similar frames in this refined feature space will yield better results than in the raw feature space.

To measure the similarity between two frames f_i and f_j, the cosine similarity measure (CSM) is applied to the projected vector $\mathbf{v}_i$ and $\mathbf{v}_j$ [11]:

$$\Phi(f_i, f_j) = \cos(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{||\mathbf{v}_i|| \cdot ||\mathbf{v}_j||} \tag{26}$$

The value of Eq. (26) ranges from 0 to 1, where 1 stands for the identical projected vector. The more different the vectors are, the closer to 0 is the value.

With the definition of the projection and similarity measure, we should detect the shots by a dynamic clustering method. This algorithm works as follows.

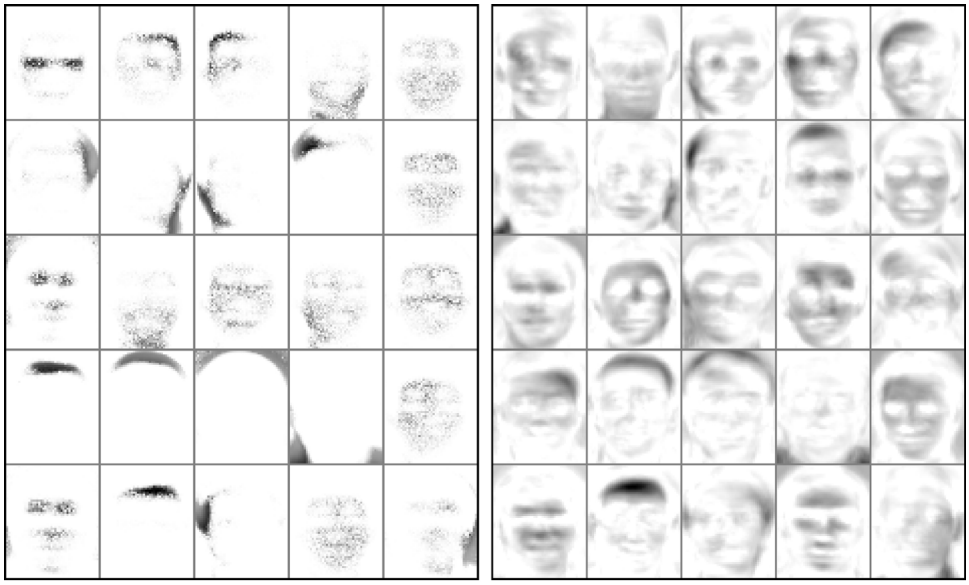
Initialization: Let c_i denote the ith cluster and $\mathbf{m}_i$ denote the mean of all the vectors in c_i. When f_1 is to be processed, $\mathbf{v}_1$ becomes a seed for a new cluster c_1 and $\mathbf{m}_1 = \mathbf{v}_1$ is set. And c_1 is set as the current cluster.

Recursion: When frame f_l $(2 \leq l \leq N_V)$ is to be processed, we are interested in testing if f_l is to be included into the current cluster c_j. Then, we test if the following condition holds

$$\cos(\mathbf{m}_j, \mathbf{v}_l) < Th\_c \tag{27}$$

where *Th_c* is a predefined threshold. If the inequality (27) is satisfied, we create a new cluster c_{j+1} represented by $\mathbf{m}_{j+1} = \mathbf{v}_l$, and c_{j+1} is set as the current cluster. Otherwise, $\mathbf{m}_j$ is updated with the inclusion of $\mathbf{v}_l$.

Since the video sequence is clustered, the shots can be detected based on these clusters. As the sparse clusters usually show the



(a) Basis images by LNMF (b) Basis images by NMF

Fig. 1. Comparison of the ability in learning local features between LNMF and NMF.

transition between shots, the dense ones are identified as shots accordingly.

3.4. Video identification

Since the pre-processing step is done, the features are extracted and the shots are detected, the signature for each shot can be generated through the process described as follows.

For each shot in the video sequence, the feature extraction process is performed on every frame of the shot and a feature matrix $\mathbf{X}_S$ with size $M \times M_S$ is constructed, where M_S is the total number of frames in the shot. Then X_S is factorized by LNMF with the number of basis images m set to set to m_S. (In the experiment, $m_S = 4$.) After that, m_S basis images are obtained for the shot and they serve as the signature of that shot.

The similarity of two shots is measured by the distance between their signatures. However, the synchronization between basis images can be easily broken under some content preserved distortions. Therefore, if the distance between basis images is measured by correspondence-based metrics, the signature distance between the original shot and its distorted version may be quite high. The Hausdorff distance, as an alternative metric to measure the similarity between two unorganized point sets, is superior to other metrics in robustness for content-preserved distortions. Hence, the modified Hausdorff distance (MHD) [21] is adopted for signature comparison in this paper.

Let us denote the signatures of two shots as: $\mathbf{S}_1 = [\mathbf{sb}_{11}, \mathbf{sb}_{12}, \ldots, \mathbf{sb}_{1mS}]$, $\mathbf{S}_2 = [\mathbf{sb}_{21}, \mathbf{sb}_{22}, \ldots, \mathbf{sb}_{2mS}]$. Then the distance between two shots $D(\mathbf{S}_1, \mathbf{S}_2)$ can be computed as follows:

$$h_{MHD}(\mathbf{S}_1, \mathbf{S}_2) = \frac{1}{m_S} \sum_{\mathbf{sb}_{1i} \in \mathbf{S}_1} \min_{\mathbf{sb}_{2j} \in \mathbf{S}_2} d(\mathbf{sb}_{1i}, \mathbf{sb}_{2j}) \tag{28}$$

$$h_{MHD}(\mathbf{S}_2, \mathbf{S}_1) = \frac{1}{m_S} \sum_{\mathbf{sb}_{2j} \in \mathbf{S}_2} \min_{\mathbf{sb}_{1i} \in \mathbf{S}_1} d(\mathbf{sb}_{1i}, \mathbf{sb}_{2j}) \tag{29}$$

$$D(\mathbf{S}_1, \mathbf{S}_2) = \max[h_{MHD}(\mathbf{S}_1, \mathbf{S}_2), h_{MHD}(\mathbf{S}_2, \mathbf{S}_1)] \tag{30}$$

where $d(\bullet, \bullet)$ denotes the distance of two basis images. In our work, histogram intersection is used for measuring such distance. As each basis image is normalized, $d(\bullet, \bullet)$ can be defined as follows:

$$d(hist_1, hist_2) = 1 - \sum_{i=1}^{M} \min(hist_1[i], hist_2[i]) \tag{31}$$

where $hist_1[i]$ and $hist_2[i]$ are the numbers contained in the ith bin of the histograms respectively.

If a query shot is to be processed, the search for the best matching shot can be performed using the following two stages:

Stage 1: After all shots in the database have been compared with the query shot based on the above measure approach, a list of shots with N_S smallest distances is stored. If the least distance in the list is smaller than the predefined threshold value Th_s, then a best matching is found. Otherwise, the value is considered to be unable to provide reliable information about the similarity of two shots and Stage 2 is processed so as to find a more reliable one.

Stage 2: Let A denote the query shot, B denote the shot to be compared with in the list, and L_A and L_B denote the lengths of A and B, respectively. A more precise similarity value is gained by frame comparison. As the video basis image matrix $\mathbf{B}_V$ is obtained in the shot detection step, each frame in A and B can be projected through Eq. (25). These projected vectors are used for similarity calculation as follows.

Let us denote the two matrices which contain the projected vectors of A and B as: $\mathbf{V}_A = [\mathbf{v}_{11}, \mathbf{v}_{12}, \ldots, \mathbf{v}_{1LA}]$, $\mathbf{V}_B = [\mathbf{v}_{21}, \mathbf{v}_{22}, \ldots, \mathbf{v}_{2LB}]$. As L_A is not necessarily equal to L_B and in some case A is a portion of B, we adopt a directed Hausdorff distance to measure the distance $D'(\mathbf{V}_A, \mathbf{V}_B)$ between $\mathbf{V}_A$ and $\mathbf{V}_B$:

$$D'(\mathbf{V}_A, \mathbf{V}_B) = \frac{1}{L_A} \sum_{\mathbf{v}_{1i} \in \mathbf{V}_A} \min_{\mathbf{v}_{2j} \in \mathbf{V}_B} d'(\mathbf{v}_{1i}, \mathbf{v}_{2j}) \tag{32}$$

where $d'(\bullet, \bullet)$ denotes the Euclidean distance between two vectors. After all the N_S shots in the list have been compared with A based on the above approach, the shot with the least distance is chosen. If its distance is smaller than a predefined threshold Th_f, then a best matching is found. Otherwise, A is not thought to exist in the database. The proposed video identification approach can be easily

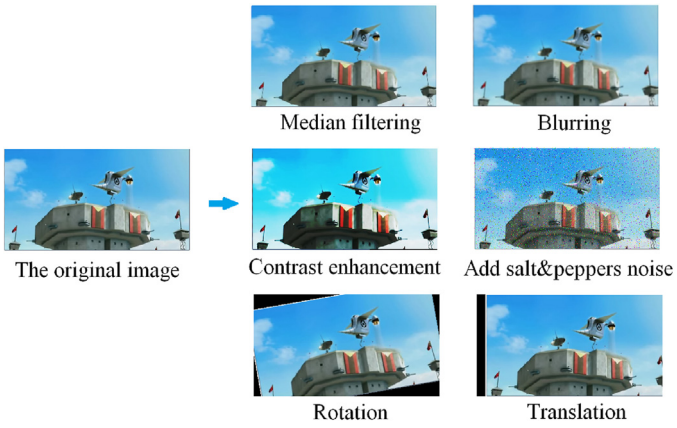


Fig. 2. The images affected by content-preserved distortions.

Table 1
The parameters in the experiment.

Parameter	Description	Value
M	The feature vector dimension	32
e	A small constant added to the denominator	0.0001
c	The constant used in LNMF	0.5
m_V	The number of basis images for factorizing $\mathbf{X}_V$	10
m_S	The number of basis images for factorizing $\mathbf{X}_S$	4
Iter_max	The max number of iterations	20
Var_1	The horizontal variance of the Gaussian filter	5
Var_2	The vertical variance of the Gaussian filter	5
Th_c	The threshold used in shot detection	0.9
N_S	The number of shots stored in Stage 1	10
Th_s	The threshold used in Stage 1	0.2
Th_f	The threshold used in Stage 2	900

extended to the query with several shots and the details are omitted here.

4. Experimental results

To evaluate the performance of the proposed approach, a similar scenario to the one presented in [22] is used. That is, the problem addressed in this paper is the identification of a given video clip in a given set of videos. Our database consists of 100 distinct video sequences with durations varying from 3 to 10 min including different contents such as news, sports matches, cartoons, commercials, and standard test sequences. The performance of the shot detection approach is not included here because it has little effect on the results of identification as long as every sequence is processed through the same shot detection method. Hence, each video sequence is divided into several segments of length 30 s and each segment is considered as a shot artificially [22].

For robustness evaluations, a series of content-preserved distortions are implemented on test sequences, including 3×3 median filter, blur, contrast enhance, 4% salt and pepper noise addition, 10° rotation, 18-pixel horizontal translation, and 50% frame dropping. The images affected by some of these distortions are illustrated in Fig. 2.

In our experiment, the parameters described in Section 3 are given in Table 1, where Var_1 and Var_2 are the variances of the Gaussian filter used in the pre-processing step. These parameters are selected mainly based on experiments. Since the V component of each frame image has 256 grayscales, thus we adopt $M = 32$ bins considering the trade-off between the complexity and feature effectiveness. Since the number of frames used in a video clip is relatively large (about 1000–2000) and the performance requirement of shot detection is relatively loose, the number of basis images for factorizing $\mathbf{X}_V$ is set to be 10, nearly reducing the feature dimension

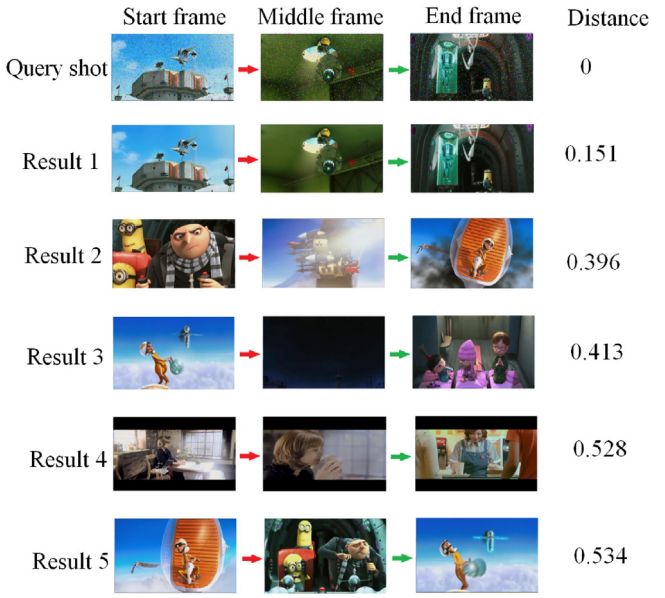


Fig. 3. The query example by our approach for the query shot with salt and pepper noise.

by 99%. While since the number of frames in each shot is relatively small (about 20–100) but the performance of the final result is greatly affected by m_S, the number of basis images for factorizing $\mathbf{X}_S$ is set to be 4, nearly reducing the feature dimension by 80%. To test the robustness of the proposed approach, for each type of the content preserved distortions, 50 distinct shots are randomly chosen from the database and the distortion is performed on each of them to construct the query set. Fig. 3 shows an example of query result generated by our approach with the first five returned shots, each shot with three representative frames. We can see that most of the returned shots are from the same video sequences as the query shot because they have the similar color histograms.

To demonstrate the advantages and the disadvantages of our proposed algorithm, two existing schemes are compared. Recent advances have shown that video tomography (VT) and bag-of-visual-words (BoVW) can be successfully used for the purpose of video fingerprinting. In [22], the VT technique is adopted to capture the spatiotemporal changes in videos to obtain a measure of local and global motion in videos, which can uniquely characterize and identify videos. In [23], a video identification scheme that takes advantage of both VT and BoVW was proposed, where the video signature is created by first extracting inclined tomography images from the video content, and by subsequently applying the BoVW approach to the inclined tomography images obtained. The same query sets generated above are used here again to ensure more precise comparison result. Here, in order to show the superiority of LNMF over NMF, we also provide the result using NMF for comparison. Fig. 4 shows the comparison result of the four approaches.

From Fig. 4, we can see the good performance of the proposed method using LNMF under various distortions, and the NMF based feature is worse than the LNMF based feature. The accuracy of the identification against different distortions is more than 92% on average and it is even completely unaffected by some types of them. The accuracy will further increases if N_S, the number of shots returned in Stage 1, increases. Of course, the run time will increase consequently as more computation is needed in Stage 2. Trade-off should be made to balance the accuracy against the computation cost. It also can be seen from Fig. 4 that the proposed approach using LNMF outperforms the VT and VT-BoVW schemes in most of

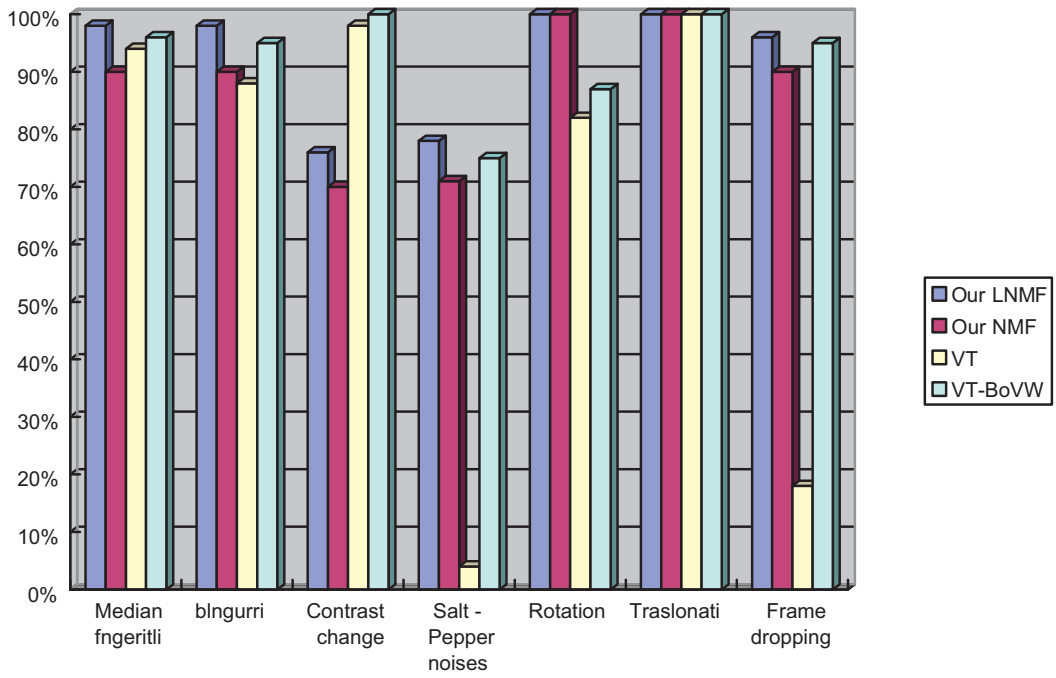


Fig. 4. Performance comparison between the proposed scheme with LNMF, the proposed scheme with NMF, the VT approach in [22], and the VT-BoVW approach in [23].

the cases especially salt & pepper noise addition, rotation and frame dropping. As the features extracted in VT algorithm are based on edge counting, the above two distortions dramatically affect them and hence impair the reliability of the shot signatures. On the other hand, the VT and VT-BoVW schemes show better robustness to contrast enhancement than the proposed algorithm as the color histograms change greatly under such distortion. Therefore, more types of features and more methods should be considered to obtain better robustness in the future.

5. Conclusions

In this paper, a robust video identification algorithm based on the local non-negative matrix factorization is presented. Some details about LNMF are described and the updated rules for it are given. Besides, the convergence is proven. Then a video identification framework completely based on LNMF is given including the shot detection approach and the identification approach. Then several query sets with various kinds of content preserved distortions are constructed for robustness evaluation. The experimental results show the good performance of the proposed approach under such distortions and its superiority to the other method. On the other hand, the results also reveal some weak points of the proposed approach and give the hints for improvement. For future development, more types of features and more approaches should be considered to further enhance the robustness of the LNMF based approach.

Acknowledgements

This work was partially supported by Zhejiang Provincial Natural Science Foundation of China (No. R1110006) and the Special Funds on Strategic New Industry Development of Shenzhen under grants JCYJ20120817163934173, ZDSY20120613125016389, and JCYJ20130329105534856.

References

[1] Tan YP, Kularni SR, Ramadge PJ. A framework for measuring video similarity and its application to video query by example. In: Proceedings of international conference on image processing. 1999. p. 106–10.
[2] Hampapur A, Bolle R. Comparison of sequence matching techniques for video copy detection. In: Proceedings of international conference on storage and retrieval for media databases. 2002. p. 194–201.
[3] Li Z, Katsaggelos AK, Gandhi B. Fast video shot retrieval based on trace geometry matching. IEE Proc Vision Image Signal Process 2005;152(3):367–72.
[4] Hoad TC, Zobel J. Detection of video sequences using compact signatures. ACM Trans Inform Syst 2006;24(1):1–50.
[5] Hauptmann AG, Jin R, Ng TD. Multi-modal information retrieval from broadcast video using OCR and speech recognition. In: Proceedings of ACM/IEEE-CS joint conference on digital libraries. 2002. p. 160–1.
[6] Leon G [M.Sc. Thesis] Content identification using video tomography. College of Engineering and Computer Science, Florida Atlantic University; 2008.
[7] Coskun B, Sankur B, Menon N. Spatio-temporal transform-based video hashing. IEEE Trans Multimedia 2006;8(6):1190–208.
[8] Roover CD, Vleeschouwer CD, Lefebvre F, Macq B. Robust video hashing based on radial projections of key frames. IEEE Trans Signal Process 2005;53(10):4020–37.
[9] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999;401:788–91.
[10] Jiao F, Gao W, Chen X. A face recognition method based on local feature analysis. In: Proceedings of Asian conference on computer vision. 2002. p. 1–5.
[11] Buciu I, Pitas I. Application of non-negative and local non negative matrix factorization to facial expression recognition. In: Proceedings of IEEE international conference on pattern recognition. 2004. p. 288–91.
[12] Nielsen FA, Balslev D, Hansen LK. Mining the posterior cingulate: segregation between memory and pain components. NeuroImage 2005;27(3):520–2.
[13] Berry MW, Browne M. Email surveillance using non-negative matrix factorization. Comput Math Organ Theory 2005;11(3):249–64.
[14] Fevotte C, Bertin N, Durrieu JL. Nonnegative matrix factorization with the itakura-satio divergence: with application to music analysis. Neural Comput 2009;21(3):1–32.
[15] Brunelli R, Poggio T. Face recognition: features vs. templates. IEEE Trans Pattern Anal Mach Intell 1993;15(10):1042–53.
[16] Belhumeur PN, Hespanha JP, Kriegman DJ. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 1997;19(7):711–20.
[17] Lee DD, Seung HS. Algorithms for non-negative matrix factorization. In: Proceedings of neural information processing systems. 2000. p. 556–62.
[18] Li SZ, Hou XW, Zhang HJ. Proceedings of international conference on computer vision and pattern recognition. In: Learning spatially localized, parts-based representation. 2001. p. 207–12.

[19] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 1977;39(1):1–38.
[20] Cernekova Z, Kotropoulos C, Pitas I. Video shot segmentation using singular value decomposition. In: Proceedings of IEEE conference on acoustic, speech and signal processing. 2003. p. 181–4.
[21] Dubuisson MP, Jain AK. A modified Hausdorff distance for object matching. Pattern Recogn 1994;1(9):566–8.
[22] Leon G, Kalva H, Furht B. Video identification using video tomography. In: Proceedings of IEEE international conference on multimedia and expo. 2009. p. 1030–3.
[23] Min HS, Kim SM, Neve WD, Ro YM. Video copy detection using included video tomography and bag-of-visual-words. In: Proceedings of IEEE international conference on multimedia and expo. 2012. p. 562–7.