

# Video-based face recognition using tensor and clustering

Jidong Zhao<sup>1\*</sup>, Wanjie Zhang<sup>2</sup>, Jingjing Li<sup>1</sup>, Ke Lu<sup>1</sup>

<sup>1</sup>University of Electronic Science and Technology of China, Chengdu, 610731, China

<sup>2</sup>Beijing Up-tech Harmony Co., LTD, Beijing, China

Received 1 March 2014, www.cmnt.lv

---

## Abstract

Video-based face recognition has become one hot topic in the field of pattern recognition recently. How to fully utilize the spatial and temporal information in video to overcome the difficulties existing in the video-based face recognition, such as low resolution of face images in video, large variations of face scale, radical changes of illumination and pose as well as occasionally occlusion of different parts of faces, is the focus. In this paper, we propose a novel manifold-based face recognition algorithm using tensor and clustering (TCVLPP), which can discover more space-time semantic information hidden in video face sequence, simultaneously make the best of the intrinsic nonlinear structure information to extract discriminative manifold features. We also compare our approach with other algorithms on our own video databases. The experimental results show that TCVLPP can get a higher recognition accuracy rate for video-based face recognition.

*Keywords:* video-based face, tensor, manifold learning

---

## 1 Introduction

Face recognition [1] is always the focus of Machine learning and Pattern recognition. Many techniques have been developed over the past few decades to solve the problems under different assumptions and constraints. The representative approaches for reducing dimensions include Principal Component Analysis (PCA) [2] Linear Discriminant Analysis (LDA) [3] and Locality Preserving Projections (LPP) [4, 5]. Among them, LPP is a manifold-based algorithm to find an embedding which preserves the local information and obtain a face subspace, so it can reveal the intrinsic face structure. LPP has been successfully used in static face recognition, which inspires us to apply it to video-based face recognition to uncover more intrinsic information of faces then improve the performance.

Recently, with the wide use of video monitoring, the great popularity of video chat and the crazy development of Video Website, video has been one of the most important media carriers. Consequently, video-based face recognition [6-7] has received an extensive attention in the latest five years. Different from static face image, video face sequence is spatiotemporal continuous and certain contacts exist between images of one video stream. If we apply the video images with traditional methods, it inevitably will discount the inherent temporal coherence information. Since the spatiotemporal information plays a significant role in face recognition, how to fully exploit redundancy information in the video sequence is the key issue for video-based recognition. There are already many video-based algorithms proposed, such as Adaptive Hidden Markov Model (HMM) [8] and Gaussian Mixture

Model (GMM) [9], they all try to treat the video sequence as a whole. This inspires that it is critical to uncover the whole information of the video face for video face recognition.

Recently, some kinds of trials by introducing manifold learning into video-based face recognition have been proposed to take full use of the space-time information inside video. Among them, Zhao et al. [17] developed a tensor feedback algorithm for video-face recognition, which treats a video as a tensor by taking the space-time information into consideration, then uses feedback technique to improve the recognition performance. Lu et al. [18] introduced locality preserving projection into video face by constructing a novel data model via k-means to give each frame in video a weight, so that a video can be reconstructed with the weighted frames. These works have demonstrated that manifold learning and space-time information are both very useful in video-based face recognition.

In this paper, we propose a novel manifold-based approach (TCVLPP) using tensor and clustering for video-based face recognition. We proposed a novel data-model method to capture the spatial-time information within video, then utilize the tensor LPP [12] to discover the manifold structure. Experimental on our own collected video-face database demonstrated that the proposed TCVLPP can obtain more semantic contents from the spatiotemporal information in the video sequence and extract discriminative manifold features by finding a low-dimensional embedding of face data, therefore can achieve a higher accuracy rate, compared with other algorithms.

The rest of this paper is organized as follows. Section II gives a brief description of Tensor Locality Preserving

---

\* Corresponding author's e-mail: jdzhao2014@163.com

Projections. We introduce our novel face recognition algorithm (TCVLPP) for video in section III. The experimental results are shown in section IV. Finally, we give conclusions and future work in section V.

**2 Related work**

Locality Preserving Projections (LPP) is a linear approximation of the nonlinear Laplacian Eigenmap [10] which aims to preserve the local structure of the data and can be explained by spectral graph theory [11]. Next, we will introduce the detail of tensor-based LPP [12], which is a 2D version of the original LPP [4 5].

Given  $m$  data points  $X = \{x_1, x_2, \dots, x_m\}$  sampled from the face sub-manifold  $M \in R^c \otimes R^d$ , one can build a nearest neighbor graph  $G$  to model the local geometrical structure of  $M$ . Let  $W$  be the weight matrix of  $G$ . One popular way to compute  $W$  is Heat Kernel, which is defined as follows:

$$W_{ij} = \begin{cases} e^{-\frac{\|x_i - x_j\|^2}{t}}, & \text{KNNs} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Let  $U$  and  $V$  be the transformation matrices. A reasonable transformation respecting the graph structure can be obtained by solving the following objective functions:

$$\min_{U,V} \sum_{ij} \|U^T x_i V - U^T x_j V\|^2 W_{ij} \quad (2)$$

The objective function incurs a heavy penalty if neighbouring points  $x_i$  and  $x_j$  are mapped far apart. Therefore, minimizing it is an attempt to ensure that if  $x_i$  and  $x_j$  are ‘‘close’’ then  $U^T x_i V$  and  $U^T x_j V$  are ‘‘close’’ as well. Let  $y_i = U^T x_i V$  be the low-dimensional feature. The goal of tensor LPP is to learn a more discriminative low-dimensional representation of the high-dimensional original data, which contains a lot of noise and would hinder the recognition performance. Also the high-dimensional data would need a lot of computation cost, which makes it harder to apply the real world applications. Comparing with the original LPP, tensor LPP treats an image as a tensor so that the original structure of the image has been preserved, which inspires us to apply to video-based face recognition to obtain more structure information in video.

For the objective function (2), we can transform it by letting  $D$  be a diagonal matrix  $D_{ii} = \sum_j W_{ij}$ . Then objective Equation (2) can be rewritten as follows:

$$\begin{aligned} & \frac{1}{2} \sum_{ij} \|U^T x_i V - U^T x_j V\|^2 W_{ij} = \\ & \frac{1}{2} \sum_{ij} \text{tr}((y_i - y_j)(y_i - y_j)^T) W_{ij} = \\ & \frac{1}{2} \sum_{ij} \text{tr}(y_i y_i^T + y_j y_j^T - y_i y_j^T - y_j y_i^T) W_{ij} = \\ & \text{tr}(\sum_i D_{ii} y_i y_i^T - \sum_{ij} W_{ij} y_i y_j^T) = \\ & \text{tr}(\sum_i D_{ii} U^T x_i V V^T x_i^T U - \sum_{ij} W_{ij} U^T x_i V V^T x_j^T U) \end{aligned} \quad (3)$$

Finally, the optimization problem of objective Equation (3) can be achieved by minimizing the following two sub-problems:

$$\begin{cases} \min_{U,V} \frac{\text{tr}(U^T (D_U - W_U) U)}{\text{tr}(U^T D_U U)} \\ \min_{U,V} \frac{\text{tr}(V^T (D_V - W_V) V)}{\text{tr}(V^T D_V V)} \end{cases}, \quad (4)$$

where the variables in (4) are defined as following:

$$D_U = \sum_i D_{ii} x_i^T U U^T x_i, \quad W_U = \sum_{ij} W_{ij} x_i^T U U^T x_j$$

$$D_V = \sum_i D_{ii} x_i^T V V^T x_i, \quad W_V = \sum_{ij} W_{ij} x_i^T V V^T x_j$$

We first fix  $U$ , then  $V$  can be computed by solving the following generalized eigenvector problem:

$$(D_V - W_V)v = \lambda D_U v \quad (5)$$

Once  $V$  is obtained,  $U$  can be updated by solving the following generalized eigenvector problem:

$$(D_U - W_U)u = \lambda D_V u \quad (6)$$

The optimal  $U$  and  $V$  can be obtained by iteratively computing the generalized eigenvectors.

**3 TCVLPP algorithm design**

An image retrieval system is a computer system for browsing, searching and retrieving images from a large database of digital images. Most traditional and common methods of image retrieval utilize the way of static image. When applying it to video-based face recognition, we just extract some images from each face video to construct the database. These images of each video are just in chaos, so it may return the ideal suited objects from all the different people and ignore the integrality of each video. On this view, different approaches should be utilized for video-based recognition. At the present stage, there has raised many methods to treat the video as a whole.

Besides, real data of natural and social sciences is often very high-dimensional. However, the underlying structure can in many cases be characterized by a small number of parameters. Reducing the dimensionality of such data is beneficial for visualizing the intrinsic structure and it is

also an important pre-processing step in many statistical pattern recognition problems. Among the classical techniques, PCA aims at preserving the global Euclidean structure, LDA can be used to find a linear subspace which is optimal for discrimination, and LPP preserves the local neighbourhood structure on the data manifold.

According to this knowledge, we try to compare PCA, LDA and LPP in video-based face environment. As known to all, a video contains front and side faces in various illuminations and directions, so the experiment is divided into two situations, including only front face and mixed front-side face.

### 3.1 MANIFOLD-BASED EXPERIMENTS

First, we compare them in the all front face database (Figure 1). The results are shown in Table 1. Second, we compare them in a database with mixed front-side face in various illuminations and directions (Figure 2). Table 2 shows the recognition.

From the results above, it can be seen that these three algorithms all perform well in Table 1, and with the number of training data growing, the recognition rate increases. In Table 2, all three algorithms play worse than the previous. But LPP reduces 15 percentages while LDA and PCA about 20 to 25 percentages. From this, we find LPP can play much better when face under changes in viewpoint and illumination. In the real world, there are face direction and illumination changing in videos, that's

why we choose LPP as our basic algorithm. In the next section, we will introduce our proposed algorithm for video-based face recognition.



FIGURE 1 Samples of frontal face in our own collected database



FIGURE 2 Samples of mixed face: frontal face and side face in our own collected database

TABLE 1 Results of three methods

Methods	3 train	4 train	5 train	6 train
PCA	73.0%	80.5%	86.3%	88.43%
LDA	74.23%	81.09%	86.9%	89.02%
LPP	78.11%	83.91%	87.7%	89.99%

TABLE 2 Results of three methods

Methods	3 train	4 train	5 train	6 train
PCA	46.9%	50.1%	55.9%	61.6%
LDA	54.32%	57.58%	61.23%	65.34%
LPP	62.47%	65.22%	68.8%	69.57%

### 3.2 TCVLPP ALGORITHM

For video-based face recognition, we aim to discover the time-space information. Before this, we have done some search, like VLPP [13]. In that paper, we use the simple average progressing data modeling method and select some images from each video to form the data set, then we average the vectors to get the representative vector of each video. Results display VLPP has a more precise outcome than LPP for on video-based face recognition. Other works [14-17] also showed that methods on the basis of LPP can acquire more semantic information inside the video face.

From another point of view, the spatiotemporal information existed in the video sequence plays a great

important role. That also proves our approach is very available for video-based recognition.

CVLPP [18] is based on VLPP using clustering instead of average process. But we find the results not very satisfied. So we consider TCVLPP combining CVLPP and Tensor.

TCVLPP first discovers the space-time information via constructing data model using k-means clustering. Then tensor LPP is applied to find the intrinsic manifold structure. At last Euclidean distance is used for discrimination. Detail steps are listed following.

#### Step 1. Data Modeling.

In our algorithm,  $n$  is the frame number of every video, we use  $k$ -means method to cluster the images in the same video, then we get  $k$  sets  $\{z_1, z_2, \dots, z_k\}$  with their matrix

means  $\{\mu_1, \mu_2, \dots, \mu_k\}$ . According to the number of each cluster, we choose a set of weight parameters  $\{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_k\}$ ,  $0 < \varepsilon_i < 1$ , where  $\varepsilon_i$  is the weight of  $\mu_i$ . In the end, we achieve a matrix  $y_i$  as the representation of the  $i^{\text{th}}$  video as  $y_i = \sum_{i=1}^k \mu_i \varepsilon_i$ .

### Step 2. Tensor LPP.

#### 1) Constructing the Nearest-Neighbor Relation.

After clustering process, one video can be seen as a matrix. At this state, we can construct the nearest graph by compare each matrix. If  $M_i$  and  $M_j$  are adjacent or labeled, we put  $S_{ij}=1$ , else  $S_{ij}=0$ .

#### 2) Computing the Projections.

Compute the eigenvectors and eigenvalues for the generalized eigenvector problem:

$$\begin{cases} (D_U - W_U)v = \lambda D_U v \\ (D_V - W_V)u = \lambda D_V u \end{cases}$$

## 4 Experimental results

In this section, the experiment will be carried out to show the efficiency and effectiveness of our proposed novel video-based face recognition algorithm. Here we first compare LPP-based methods: TCVLPP, CVLPP and VLPP, then some current ones. The datasets we use are the Honda/UCSD Video Database (Figure 3) and our own collection database.



FIGURE 3 Samples of Honda/UCSD video face database

### 4.1 DATABASE DESCRIPTION

For Honda/UCSD video face database, we select 18 different persons (Figure 3). First we divide each training video into 10 groups and select 5 images from each group as one training data set. So there every person has 10 training classes, also we choose 4 face images from 18 testing videos as the testing data. For the static method, we choose 3 images from one person as training set and the left as test set.

Our own video database concludes 15 different persons, and each has 8 videos. When gathering the video data, we intend to arrange to have the front and profile face posture, and want to establish the different illumination environment, simultaneously pay attention to the changing of facial expression. The main goal is to simulate the real video environment in the shortest time as possible. In the practical application, the video frequency time is long, and generally the profile, the illuminations and the expression are changing over time. Therefore, we collected such targeted video database for comparative experiment is very suitable.

### 4.2 COMPARISON WITH LPP-BASED METHODS

For video-based face recognition, we aim to discover the time-space information. Before this, we have done some search (VLPP, CVLPP). From these two algorithms, we have gained that rational data modeling can help improve recognition accuracy for video-based face recognition. Here we compare our TCVLPP with CVLPP and VLPP on our own collection database.

In the experiment, we randomly select 2,4,6,8 as the training samples and get 4 groups of results with 60 test images. First, we use clustering to construct the video data. After this, tensor LPP is applied to find the subspace of the two-dimension images. The results are displayed following Figure 4.

From the results, we see that three algorithms all perform well, that is to say, how to construct data model to discover the special space-time information hidden in video is the key point method for improving video face recognition. Along with the increase of training class, it achieves a better of three algorithms.

Furthermore, TCVLPP has a better result than CVLPP and VLPP. That shows tensor can preserve more structure of face image. But TCVLPP just use tensor after clustering, and it more or less ignores some time information. Next we will work on more efficient method.

### 4.3 COMPARISON WITH SOME CURRENT METHODS

During the past several years, face recognition in video has received significant attention. There have been many approaches to solve the changes in illumination and/or pose and/or facial occlusion and/or low resolution of acquired image. Here we just choose 2 current algorithms on UCSD video face database. The detail of two algorithms are described as follows.

Torres et al. proposed a Self-Eigenfaces Algorithm [19], which performed on each set of views of the same person giving different sets of eigenfaces, one for each of the persons who want to be recognized. The test images to be recognized are projected and reconstructed using each one of the sets of the different eigenfaces. In our experiment, we select 5 images from each video to

construct self-eigenface, then test images are used to recognition.

Li et al. developed a supervised algorithm to capture the label information [20]. The set of features is then compressed with a signature, which is composed of numbers of points and their corresponding weights. During matching, the distance between two signatures is computed by Earth Movers Distance (EMD). In our experiment, we first use KLDA to do dimension reduction, then k-means are applied to compute the weights and ratios

of each video. At last, test images' weight is its feature vector and its ratio is 1, then compute the distance by EMD.

The results are shown in Figure 5 and the recognition rate is computed as the mean rate of 60 test images. From the results, TCVLPP can perform much better, just about 10 percentages higher. One point is that LPP can preserve more intrinsic manifold structure and another point tells that our data modeling is also available for video-based face recognition.

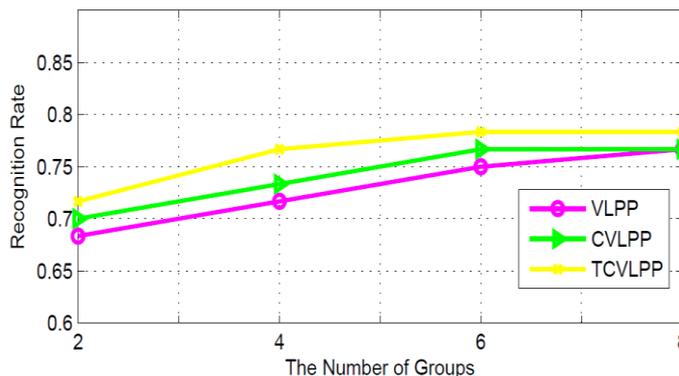


FIGURE 4 Results of VLPP, CVLPP and TCVLPP

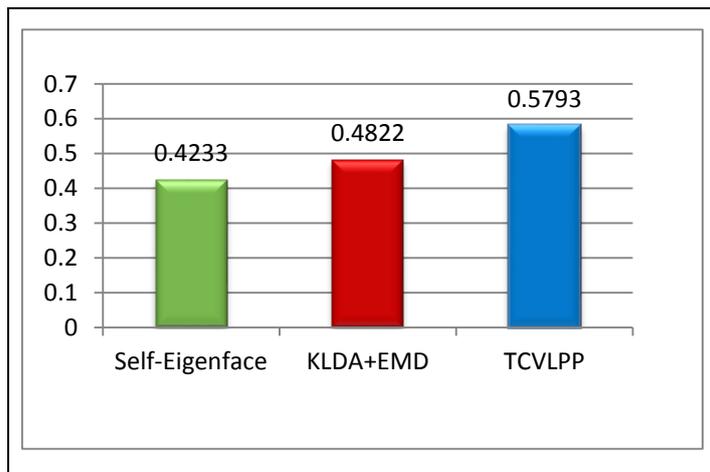


FIGURE 5 Recognition results of three algorithms

**5 Conclusion and future work**

In this article, we propose a novel manifold-based face recognition algorithm for video using tensor and clustering (TCVLPP). In the algorithm, data model hopes to obtain the spatiotemporal connection between faces of the same video, and manifold learning is applied to preserve local relationships within the data set for uncovering its essential manifold structure. The experiment has proved its effectiveness.

There are still other problems remaining unclear. For instance, it remains unclear how to determine the parameter *k* in the *k*-nearest neighbor search. In our

algorithm, we will do more tests to ensure that whether suitable data modeling is helpful for video-based face recognition, then to search for an efficient data modeling method.

**Acknowledgements**

Project supported by the Fundamental Research Funds This work was supported by National Science Foundation of China under Grant 61371183, National High Technology Research and Development Program of China under Grant 2012AA041403.

## References

- [1] Chellappa R, Wilson C, Sirohey S 1995 Human and machine recognition of faces: A survey *Proceedings of the IEEE* **83**(5) 705-40
- [2] Turk M, Pentland A 1991 Face recognition using eigenfaces *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 586-91
- [3] Belhumeur P N, Hespanha J P, Kriegman D J 1997 Eigenfaces vs Fisherfaces: recognition using class specific linear projection *IEEE Transaction on Pattern Analysis and Machine Intelligence* 711-20
- [4] He X, Niyogi P 2003 Locality Preserving Projections *Advances in Neural Information Processing Systems 16 (NIPS) Vancouver Canada*
- [5] He X 2005 Face Recognition Using Laplacianfaces *IEEE transactions on pattern analysis and machine intelligence* **27**(3)
- [6] Yan Y, YuJin Z 2009 State of the Art on Video-Based Face Recognition *Chinese Journal of Computer Ters* **32**(5)
- [7] Wang H, Wang Y, Cao Y 2009 Video-based face recognition: A survey *Proceedings of World Academy of Science, Engineering and Technology* 293-302
- [8] Liu X M, Chen T 2003 Video-based face recognition using adaptive hidden Markov models *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. Madison* 340-5
- [9] Arandjelovi O, Cipolla R 2004 Face recognition from face motion manifolds using robust kernel resistor average distance *Proceedings of the IEEE Conference on Compute Vision and Patter Recognition workshop* 88-93
- [10] Belkin M, Niyogi P 2002 Laplacian eigenmaps and spectral techniques for embedding and clustering *Proceedings of advance in Neural Information Processing Systems 14 Cambridge, MA:MIT Press* 585-91
- [11] Cvetkovi D M, Doob M, Sachs H, Torgasev A 1987 Recent Results in the Theory of Graph Spectra *Annals of Disrete mathematics series, North-Holland*
- [12] He X, Cai D, Niyogi P 2005 Tensor subspace analysis *In Proceedings of advance in Neural Information Processing Systems*
- [13] Lu Ke, Ding Z, Zhao J, Wu Y 2010 Video-based face recognition *3rd International Congress on Image and Signal Processing CISP*
- [14] Lu Ke, Ding Z, Zhao J 2011 A Manifold Learning Algorithm for Video-based Face Recognition *Journal of Information & Computational Science* **7**(9)
- [15] Lu Ke, Ding Z, Zhao J 2010 Video-based Face Recognition using Relevance Feedback *Journal of Information & Computational Science* **6**(12)
- [16] Ke Lu, Ding Z, Zhao J 2012 A novel Face Recognition Algorithm for Video *International Journal of Advancements in Computing Technology* **4**(13) 315-22
- [17] Zhao J, Ding Z 2013 Video-based Face Recognition using Tensor and Feedback *International Journal of Advancements in Computing Technology* **5**(7) 1009-16
- [18] Lu Ke, Ding Z, Zhao J, Wu Y 2011 A Manifold Learning Algorithm for Video-based Face Recognition *Journal of Information & Computational Science*
- [19] Torres L, Vila J 2002 Automatic face recognition for video indexing applications *Pattern Recognition* **35**(3) 615-25
- [20] Li J W, Wang Y H, Tan TN 2005 Video-based face recognition using earth mover's distance *Proceedings of the International Conference on Audio- and Video-based person authentication New York* 229-23
- [21] Lee K C, Ho J, Yang M H, et al 2003 Video-based face recognition using probabilistic appearance manifolds *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*
- [22] Mian A 2008 Unsupervised learning from local features for video-based face recognition *IEEE International Conference on Automatic Face & Gesture Recognition* 1-6

## Authors

	<p><b>Jidong Zhao, born in 1976.</b></p> <p><b>Current positon, grades:</b> associate professor in School of Computer Science and Engineering, University of Electronic Science and Technology of China.</p> <p><b>University studies:</b> PhD degrees in Computer Application Technology from the University of Electronic Science and Technology of China, in 2009.</p> <p><b>Scientific interest:</b> pattern recognition and computer vision.</p>
	<p><b>Wanjie Zhang, born in 1981.</b></p> <p><b>Current positon, grades:</b> engineer in Beijing Up-tech Harmony Co.</p> <p><b>University studies:</b> B.S. degree in automation from North China University of Technology in 2003.</p> <p><b>Scientific interest:</b> pattern recognition and computer vision.</p>
	<p><b>Jingjing Li, born in 1988.</b></p> <p><b>Current positon, grades:</b> a master candidate in School of Computer Science and Technology, University of Electronic Science and Technology of China.</p> <p><b>University studies:</b> M.Sc. degree in Information Security from University of Electronic Science and Technology of China in 2013.</p> <p><b>Scientific interest:</b> video-based data analysis, and visual object tracking.</p>
	<p><b>Ke Lu, born in 1974.</b></p> <p><b>Current positon, grades:</b> professor in School of Computer Science and Engineering, University of Electronic Science and Technology of China.</p> <p><b>University studies:</b> PhD degrees in Computer Application Technology from the University of Electronic Science and Technology of China, in 2006.</p> <p><b>Scientific interest:</b> include pattern recognition, multimedia and computer vision.</p>