Spatio-temporal Embedding for Statistical Face Recognition from Video

Wei Liu[1], Zhifeng Li[1], and Xiaoou Tang[1,2]

[1] Department of Information Engineering,
The Chinese University of Hong Kong,
Hong Kong, China
{wliu5, zli0}@ie.cuhk.edu.hk
[2] Microsoft Research Asia, Beijing, China
xitang@microsoft.com

Abstract. This paper addresses the problem of how to learn an appropriate feature representation from video to benefit video-based face recognition. By simultaneously exploiting the spatial and temporal information, the problem is posed as learning Spatio-Temporal Embedding (STE) from raw video. STE of a video sequence is defined as its condensed version capturing the essence of space-time characteristics of the video. Relying on the co-occurrence statistics and supervised signatures provided by training videos, STE preserves the intrinsic temporal structures hidden in video volume, meanwhile encodes the discriminative cues into the spatial domain. To conduct STE, we propose two novel techniques, Bayesian keyframe learning and nonparametric discriminant embedding (NDE), for temporal and spatial learning, respectively. In terms of learned STEs, we derive a statistical formulation to the recognition problem with a probabilistic fusion model. On a large face video database containing more than 200 training and testing sequences, our approach consistently outperforms state-of-the-art methods, achieving a perfect recognition accuracy.

1 Introduction

As still image-based recognition accuracy is still too low in some practical applications comparing to other high accuracy biometric technologies, video-based face recognition has been proposed recently [15][12][6][16][7][9][13]. One major advantage of video-based techniques is that more information is available in a video sequence than in a single image. Naturally, the recognition accuracy could be improved if the abundant information can be properly exploited.

It has been demonstrated that modeling temporal dynamics is very useful to the video-based problems. Hence, recent video-based face recognition research employs them to improve recognition performance. Using the statistical coherence over time, Zhou *et al.* [16] model the joint probability distribution of identity and head motion using sequential importance sampling, which leads to a generic framework for both tracking and recognition. Liu *et al.* [9] analyze video sequences over time by HMMs, each of which learns the temporal dynamics within

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part II, LNCS 3952, pp. 374–388, 2006.

a video sequence. Comparing likelihood scores provided by the HMMs, the identity of a testing video sequence is yielded with the highest score. Because learning temporal dynamics during the recognition stage is very time-consuming, above statistical models can not suffice for the real-time requirement of automatic face recognition systems. Lee *et al.* [7] approximate a nonlinear appearance manifold which stands for one subject (person) as a collection of linear submanifolds, and encode the dynamics between them into the transition probability. The manifold learning algorithms in [7] are subject-specific and lack the discriminating power, so they do not adapt well to the face recognition scenario which is a supervised classification problem.

Opposite to face hallucination techniques [1][8][10] which try to infer the lost image content of a facial image, the video-based face recognition scenario is confronted with the abundance of consecutive frames in face videos. Hence it is crucial to efficiently exploit the spatial and temporal information.

In this paper we present a novel spatio-temporal representation for each video sequence that we call "Spatio-Temporal Embedding" (STE). The STE of a video sequence is its miniature, condensed version containing the intrinsic spatio-temporal structures inherent in the space-time video volume. Based on STEs, we develop a statistical face recognition framework from video by integrating several novel techniques including Bayesian keyframe learning for learning

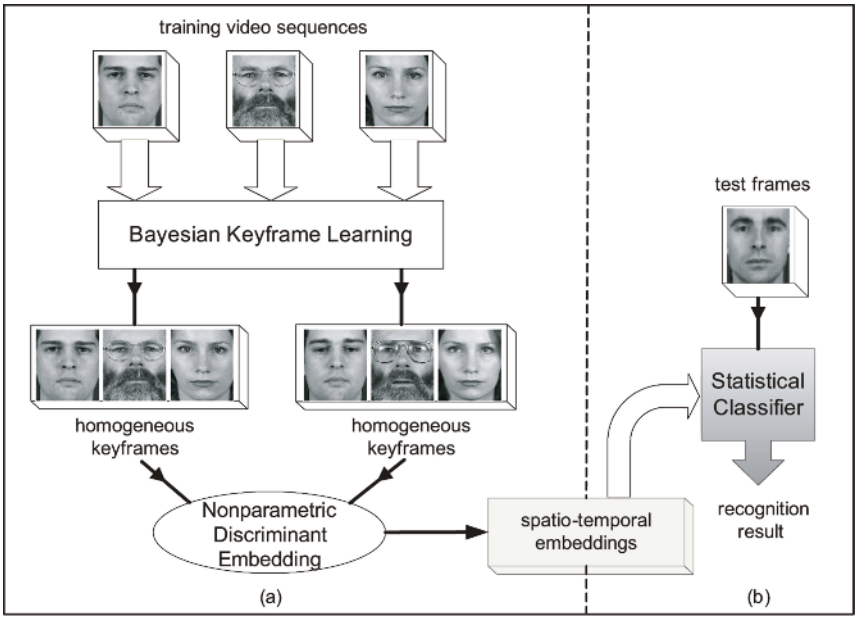

Fig. 1. The framework of our video-based face recognition approach. (a) Training stage: learn keyframes from video sequences and then arrange them into K groups of homogeneous keyframes which will be input to NDE; (b) testing stage: construct a statistical classifier in terms of learned spatio-temporal embeddings.

temporal embedding, Nonparametric Discriminant Embedding (NDE) for learning spatial embedding, and statistical classification solution. This framework takes full advantage of the effective amount of potential information in videos and at the same time overcomes the processing speed and data size problems. The detailed diagram of the proposed framework is plotted in Fig. 1.

The rest of this paper is organized as follows. In Section 2, we propose to learn temporal embedding "keyframes" which are robust to data perturbation. In Section 3, we develop NDE to further learn spatial embedding of keyframes. A statistical classifier is designed in Section 4. Experimental results on the largest standard video face database, the XM2VTS database [11], are reported in Section 5. Finally, we draw conclusion in Section 6.

2 Temporal Embedding

Recent literature proposes to extract the most representative frames called "exemplars" or "keyframes" from the raw videos. Keyframes extracted from a video sequence just span the temporal embedding of the video. However, previous approaches for extracting keyframes only consider the temporal characteristics of individual video, the extracted keyframes thus tend to differ in describing the temporal structures. In this section, we present our approach for automatically learning the homogeneous keyframes used to support discriminant analysis.

2.1 Previous Work

Krueger and Zhou [6] apply radial basis functions to select representative images as exemplars from training face videos, and this facilitates both tracking and recognition tasks. Our previous work [13] uses information in audio signals of video to locate maximum audio amplitudes of temporal segments to find the corresponding video keyframes. We [13] have demonstrated that audio-guided keyframes well represent a video sequence, and reach a satisfactory video-to-video matching level using subspace approaches.

For videos of varying frame contents, a simple matching of two video sequences frame-by-frame will not help much to video-to-video matching, since we may be matching a frame in one video with a frame of different expression in another video. This may even deteriorate the face recognition performance. The key to the performance improvement is that face frames in each sequence are in a consistent order of temporal dynamics, so that neutral face matches with neutral face and smile face matches with smile face. The consistent order implies synchronized pose, orientation, and expression variations of face images in each video sequence. Therefore, in order to make use of keyframes to boost recognition performance, keyframes across different video sequences should be extracted in a synchronized way. We call this "frame synchronization" as we will guarantee that keyframes extracted from different videos are temporally synchronized from each other.

In literature [6], the best keyframes (exemplars) are sought such that the expected distance between them and frames in the raw video sequence is minimized.

Due to the mentioned synchronization criterion for keyframe extraction, the algorithm proposed in [6] fails to generate "good" keyframes because it only works on individual video. Our work [13] succeeds in learning synchronized keyframes by utilizing audio signals in videos. Specifically, when recording video data in the XM2VTS database, each person is asked to recite two sentences "0,1,2,...,9" and "5,0,6,9,2,8,1,3,7,4" which span a video sequence of 20 seconds. Since the audio signals are in the same order over time and approximately reflect the temporal structures of videos, the audio-guided method guarantees frame synchronization. Nevertheless, in some applications, it may be difficult to get audio signals contained in videos. In addition, the method is vulnerable to data perturbation. For example, if a person reads the digit sequence "0,1,2,...,9" in a random order, skips one digit, or repeats one digit, then the audio-guided method will fail the frame synchronization and a wrong frame match may appear.

Beyond the audio limitation, we should design a novel keyframe learning approach which could comply with frame synchronization with only the image information of video data adopted.

2.2 Synchronized Frame Clustering

A prelude to learning keyframes is clustering on video frames. Previous clustering on videos only focuses on spatial (e.g. appearance) correlations and skip temporal correlations that also play an important role in clustering. For exemplars provided by XM2VTS, when one reads one particular digit, the associated frames should be mapped to the same cluster that corresponds the digit. Due to concerns of spatial and temporal continuity inherent in video data, we propose a synchronized clustering method which incrementally outputs aligned clusters across all video sequences based on K-means clustering [4].

Let $V = \{\mathbf{x}_1, \cdots, \mathbf{x}_t, \cdots, \mathbf{x}_N\} (\mathbf{x}_t \in \Re^d)$ represent a set of video frame samples belonging to the same video sequence. In this work we assume temporal coherence on the order in which data points arrive one-by-one. Let V be a stream of data, its temporal ordering specified by the corresponding subscript. For the training video set $\{V^{(i)}\}_{i=1}^{M}$, we cluster each one $V^{(i)}$ into K clusters $\{\mathcal{C}_k^{(i)}\}_k$ at one time, and then merge these formed clusters $\{\mathcal{C}_k^{(i)}\}_i$ into a larger one $\mathcal{C}_k = \bigcup_i \mathcal{C}_k^{(i)}$.

For clustering the sequence $V^{(i)}$, we promote the classical K-means algorithm [4] using the following spatio-temporal objective function to assign k^* to $\mathbf{x}_t^{(i)}$

$$k^* = \arg \min_{k \in \{1,\cdots,K\}} \frac{(\mathbf{x}_t^{(i)} - \bar{\mathbf{x}}_k)^T \mathrm{Q} (\mathbf{x}_t^{(i)} - \bar{\mathbf{x}}_k)}{\lambda_1} + \frac{(t - time(\mathbf{x}_t^{(i)}, \mathcal{C}_k^{(i)}))^2}{\lambda_2}, \qquad (1)$$

where $\bar{\mathbf{x}}_k$ is the average of frames in cluster $\mathcal{C}_k$; Q is an adaptive distance metric which will be updated after each sequence clustering; function $time(\mathbf{x}_t^{(i)}, \mathcal{C}_k^{(i)})$ computes the temporally nearest order in the cluster $\mathcal{C}_k^{(i)}$ for frame $\mathbf{x}_t^{(i)}$; scaling parameters λ_1, λ_2 control the trade-off between spatial and temporal similarity.

Table 1. Synchronized frame clustering algorithm

Step 1: *PCA.* To reduce the high dimensionality of images, we project the training frame ensemble $\{\mathbf{x}_t^{(i)}\}_{t,i}$ into the PCA subspace. For brevity, still use $\mathbf{x}$ to denote the frames in the PCA subspace in the following steps.

Step 2: *Initialization.* For the first video sequence $V^{(1)}$, randomly select K frames as the initial cluster center $\bar{\mathbf{x}}_1, \cdots, \bar{\mathbf{x}}_K$, then perform K-means clustering on $V^{(1)}$ by eq. (1). K small clusters $\{\mathcal{C}_k^{(1)}\}_{k=1}^K$ are obtained, with which large clusters $\{\mathcal{C}_k\}_{k=1}^K$ are generated as $\mathcal{C}_k = \mathcal{C}_k^{(1)}$ $(k = 1, \cdots, K)$. So the initial metric Q can be computed using eq. (2) in terms of initial $\{\mathcal{C}_k\}_{k=1}^K$.

Step 3: *Synchronized clustering.*

For $i = 2, \cdots, M$

use current cluster centers $\{\bar{\mathbf{x}}_k\}_{k=1}^K$ to conduct K-means clustering on sequence $V^{(i)}$ resulting in new small clusters $\{\mathcal{C}_k^{(i)}\}_{k=1}^K$;

update $\mathcal{C}_k \longleftarrow \mathcal{C}_k \bigcup \mathcal{C}_k^{(i)}$ $(k = 1, \cdots, K)$;

update $\bar{\mathbf{x}}_k \longleftarrow \sum_{\mathbf{x} \in \mathcal{C}_k} \mathbf{x}/|\mathcal{C}_k|$ $(k = 1, \cdots, K)$;

update Q with updated $\{\bar{\mathbf{x}}_k\}$ and $\{\mathcal{C}_k\}$;

End.

Outputs: Synchronized clusters $\{\mathcal{C}_k^{(i)}\}_{k=1,\cdots,K}^{i=1,\cdots,M}$ across all video sequences $\{V^{(i)}\}_{i=1}^M$.

Motivated by Relevant Component Analysis (RCA) [2] which gives a good metric with contextual information among samples explicitly encoded, a stepwise update for the metric is done absorbing the current context of clustering

$$\mathrm{Q} \longleftarrow \left(\sum_{k=1}^{K} \frac{1}{|\mathcal{C}_k|} \sum_{\mathbf{x} \in \mathcal{C}_k} (\mathbf{x} - \bar{\mathbf{x}}_k)(\mathbf{x} - \bar{\mathbf{x}}_k)^T + r\mathrm{I} \right)^{-1} . \tag{2}$$

r is an regularization constant, which is often necessary when $\mathcal{C}_k$ contains a small number of frames. As long as $\{\mathcal{C}_k\}$ expand further, Q will become more accurate and reliable, even need not regularization.

To overcome the problems incurred by disordered video data, we propose a synchronized frame clustering algorithm plotted in Tab. 1. As we take centers of large clusters $\mathcal{C}_k$ as K means for clustering individual sequence, small clusters $\mathcal{C}_k^{(i)}$ across different video sequences within the same large cluster $\mathcal{C}_k$ tend to be homogeneous. What' more, historical clustering results provide the reference order for following sequence clustering. Consequently, our clustering algorithm guarantees frame synchronization and outputs synchronized and aligned clusters $\{\mathcal{C}_k^{(i)}\}_{k,i}$ across all training video sequences $\{V^{(i)}\}_i$.

2.3 Bayesian Keyframe Learning

Since the audio-guided method is sensitive to disordered video sequences, we thus propose an automatic keyframe learning method and make it robust to disordered video sequences. The learning method enables us to select not only

synchronized but also distinctive keyframes spanning the temporal embeddings of videos. The intuition of this method is that excellent keyframe extraction should be pursued jointly in temporal and spatial domains.

After synchronized frame clustering, each video sequence has at most K clusters $\mathcal{C}_k^{(i)}(k = 1, \cdots, K)$ with the consistent order. Our purpose is to select the keyframes, i.e most representative exemplars, from each particular cluster $\mathcal{C}_k^{(i)}$ in each sequence. Given the synchronized property of clusters $\mathcal{C}_k = \{\mathcal{C}_k^{(i)}\}_i$, we model the co-occurrence statistics among all video frames in the constructed cluster $\mathcal{C}_k$ as the joint probability distribution

$$p(\mathbf{x}, \mathcal{C}_k) \propto \exp\left(-\frac{\|\Lambda_k^{-1/2}\mathrm{U}_k^T(\mathbf{x} - \bar{\mathbf{x}}_k)\|^2}{2}\right), \tag{3}$$

in which the eigensystem $(\mathrm{U}_k, \Lambda_k)$ is solved by performing PCA on frames in cluster $\mathcal{C}_k$, and keeping the leading eigenvectors retaining 98% of the energy.

Given each sequence $V^{(i)}(i = 1, \cdots, M)$, candidates $\mathbf{e}$ in small cluster $\mathcal{C}_k^{(i)}$ with maximum likelihood to $\mathcal{C}_k$ are selected as keyframes. In practice, the frames with the m greatest conditional probabilities $p(\mathbf{e}|V^{(i)}, \mathcal{C}_k)$ to each cluster $\mathcal{C}_k$ are selected as top-m keyframes. By Bayesian law, we choose the optimal exemplar $\mathbf{e}^*$ such that (the deviation parameter δ_k can be evaluated using data in $\mathcal{C}_k$)

$$\begin{aligned}
\mathbf{e}^* &= \arg\max_{\mathbf{e}}\ p(\mathbf{e}|V^{(i)}, \mathcal{C}_k) = \arg\max_{\mathbf{e}}\ p(\mathbf{e}, V^{(i)}, \mathcal{C}_k) \\
&= \arg\max_{\mathbf{e}}\ p(V^{(i)}|\mathbf{e}, \mathcal{C}_k)p(\mathbf{e}, \mathcal{C}_k) = \arg\max_{\mathbf{e}} \prod_{x \in \mathcal{C}_k^{(i)}} p(\mathbf{x}|\mathbf{e})p(\mathbf{e}, \mathcal{C}_k) \\
&= \arg\max_{\mathbf{e}}\ \exp\left(-\sum_{x \in \mathcal{C}_k^{(i)}} \frac{\|\mathbf{x} - \mathbf{e}\|^2}{2\delta_k^2} - \frac{\|\Lambda_k^{-1/2}\mathrm{U}_k^T(\mathbf{e} - \bar{\mathbf{x}}_k)\|^2}{2}\right).
\end{aligned} \tag{4}$$

Substituting all possible frames $\mathbf{e} \in \mathcal{C}_k^{(i)}$ into eq. (4) and maximizing eq. (4), we accomplish learning the optimal K keyframes. The top-m keyframes can also be learned by adopting eq. (4) as the keyframe score. The keyframe selection strategy supported by eq. (4) is termed Bayesian keyframe learning, which effectively coordinates the co-occurrence statistics and individual representative capability of selected keyframes. So far, we have learned the temporal embedding of video sequence $V^{(i)}$, which we denote as $\mathcal{T}^{(i)}$. Its k-th component in cluster $\mathcal{C}_k$ is denoted as $\mathcal{T}_k^{(i)}$, and its constituent top-m keyframes are represented by $\mathbf{e}_{kj}^{(i)}(k = 1, \cdots, K, j = 1, \cdots, m)$. Within the same cluster $\mathcal{C}_k$, keyframes are well synchronized and highly homogeneous.

3 Spatial Embedding

We will further learn the spatial embedding over the learned temporal embedding to achieve the final STE according to each video. This is achieved by performing a novel supervised dimensionality reduction algorithm called Nonparametric

Discriminant Embedding (NDE). We show that NDE is superior to PCA and LDA, two well-known linear dimensionality reduction algorithms. Hence, NDE endows STEs with much greater discriminating power.

3.1 Nonparametric Discriminant Embedding (NDE)

LDA is a popular feature extraction technique which aims to maximize ratio of the determinant of the between-class scatter matrix to that of the within-class scatter matrix. Assume there are c different classes, let μ_i, μ be the class mean and overall mean, and n_i the number of samples in class C_i, the within-class scatter matrix and the between-class scatter matrix are defined as

$$\begin{aligned} \mathrm{S}_w &= \frac{1}{c}\sum_{i=1}^{c}\frac{1}{n_i}\sum_{j\in C_i}(\mathbf{x}_j-\mu_i)(\mathbf{x}_j-\mu_i)^T \\ \mathrm{S}_b &= \frac{1}{c}\sum_{i=1}^{c}(\mu_i-\mu)(\mu_i-\mu)^T. \end{aligned} \tag{5}$$

The LDA algorithm seeks to determine the optimal projection W which maximizes the ratio between the between-class matrix and the within-class matrix $|\mathrm{W}^T\mathrm{S}_b\mathrm{W}|/|\mathrm{W}^T\mathrm{S}_w\mathrm{W}|$.

Until now, numerous LDA-based methods have been proposed for face recognition [3][14]. However, an inherent problem with LDA arises from the parametric form of the between-class scatter matrix, which leads to several disadvantages. Firstly, LDA is based on the assumption that the discrimination information is equal for all classes. Therefore, it performs well under Gaussian class distributions, but not under non-Gaussian distributions. Secondly, the number of the final LDA features, f, has an upper limit of c-1 because the rank of the between-class matrix S_b is c-1 at most. It is not sufficient for complex data distribution if only c-1 features are used. Thirdly, due to the presence of outliers, the between class matrix S_b in LDA cannot capture the information of the boundary structure effectively, which is essential for different classes.

To overcome the above drawbacks, we propose a Nonparametric Discriminant Embedding (NDE) algorithm motivated by nonparametric discriminant analysis (NDA) [5]. The original NDA algorithm only deals with two-class pattern recognition tasks, whereas the proposed NDE algorithm is generalized to tackle multi-class pattern classification problem. The difference between NDE and LDA is in the definition of the scatter matrices. In NDE, we define the within-class and between-class scatter matrix as

$$\begin{aligned} \mathrm{S}_w^{\mathcal{N}} &= \frac{1}{c}\sum_{i=1}^{c}\frac{1}{n_i}\sum_{t=1}^{n_i}(\mathbf{x}_t^i-\mu_i(\mathbf{x}_t^i))(\mathbf{x}_t^i-\mu_i(\mathbf{x}_t^i))^T \\ \mathrm{S}_b^{\mathcal{N}} &= \frac{1}{c(c-1)}\sum_{i=1}^{c}\sum_{j=1,\,j\neq i}^{c}\sum_{t=1}^{n_i}\lambda(i,j,t)(\mathbf{x}_t^i-\mu_j(\mathbf{x}_t^i))(\mathbf{x}_t^i-\mu_j(\mathbf{x}_t^i))^T, \end{aligned} \tag{6}$$

where $\mathbf{x}_t^i$ denotes the t-th sample of class i, and $\mu_j(\mathbf{x}_t^i)$ is the mean of local z-NNs, defined as $\mu_j(\mathbf{x}_t^i) = \sum_{p=1}^{z}\mathbf{n}_p^j(\mathbf{x}_t^i)/z$ where $\mathbf{n}_p^j(\mathbf{x}_t^i)$ is the pth nearest

neighbor from class j to sample $\mathbf{x}_t^i$, and $\lambda(i,j,t)$ is a weighting function which is defined as

$$\lambda(i,j,t) = \frac{\min\left\{d^\beta(\mathbf{x}_t^i, \mathbf{n}_z^i(\mathbf{x}_t^i)), d^\beta(\mathbf{x}_t^i, \mathbf{n}_z^j(\mathbf{x}_t^i))\right\}}{d^\beta(\mathbf{x}_t^i, \mathbf{n}_z^i(\mathbf{x}_t^i)) + d^\beta(\mathbf{x}_t^i, \mathbf{n}_z^j(\mathbf{x}_t^i))}, \tag{7}$$

where β is a control parameter that can be empirically chosen between zero and infinity, and $d(\mathbf{v}_1, \mathbf{v}_2)$ is the Euclidean distance between two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$. The weighting function is used to place more emphasis on the boundary information.

The NDE algorithm seeks to determine the optimal projection $\mathrm{W}_{opt} \in \Re^{d\times f}$, which maximizes the ratio between the generalized between-class matrix and within-class matrix

$$\mathrm{W}_{opt} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_f] = arg\max_{\mathrm{W}} \frac{|\mathrm{W}^T \mathrm{S}_b^{\mathcal{N}} \mathrm{W}|}{|\mathrm{W}^T \mathrm{S}_w^{\mathcal{N}} \mathrm{W}|}. \tag{8}$$

The NDE projection contains eigenvectors of the matrix $\left(\mathrm{S}_w^{\mathcal{N}}\right)^{-1} \mathrm{S}_b^{\mathcal{N}}$. From eq. (6) we have a few observations: (1) If we select $z = n_i$ and set all the weighting functions to unit value, $\mu_j(\mathbf{x}_t^i)$ will become μ_j. It means the NDE is indeed a generalized version of LDA. (2) As opposed to the conventional LDA algorithms which usually can only extract c-1 discriminative features at most, the NDE algorithm does not suffer from such limitation. The number $f(< d)$ of extracted discriminative features can be specified as desired. (3) The NDE algorithm is more effective in capturing the information of the boundary structure for different classes in contrast to the conventional LDA algorithms.

We illustrate the power of NDE with a toy problem where 3D data points are sampled from two half-spheres. The data points with 2 labels are shown in Fig. 2. Since the problem is binary classification, PCA, LDA and NDE all reduce the dimensions of raw data to 2 dimensions. Note that the LDA embedding is intrinsically in 1 dimension, for the comparative purpose we add the second dimension with the same coordinates to the intrinsic one. From the embedding results shown in Fig. 3 - Fig. 5, we can clearly observe that the PCA and LDA embeddings of two classes of points partially overlap with each other, while points from different classes are well separated with each other in the embedding results provided by NDE (Fig. 5). This demonstrated that NDE can find a better subspace than LDA or PCA in the case of abundant training data. Better results can be achieved by nonlinear methods, but most of these nonlinear methods only work on training data. NDE can be generalized outside the training points to the entire input space.

3.2 Multiple NDE

Due to the fact that NDE has an advantage over LDA when encountering with abundant training data, we will apply NDE to handle the video-based face recognition scenario which is just the classification problem with many samples. In details, we conduct multiple NDE to extract discriminative features for training

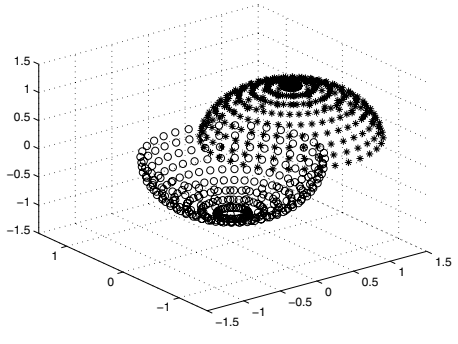

Fig. 2. The two half-sphere toy data points with 2 labels "∗" and "o"

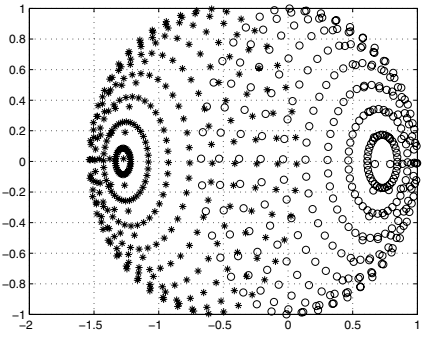

Fig. 3. PCA

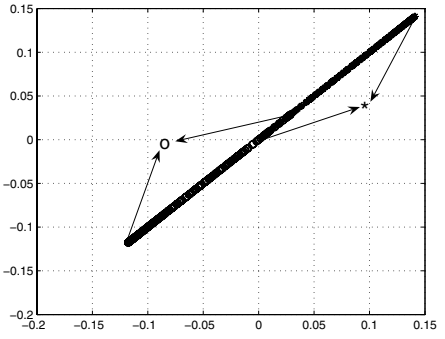

Fig. 4. LDA

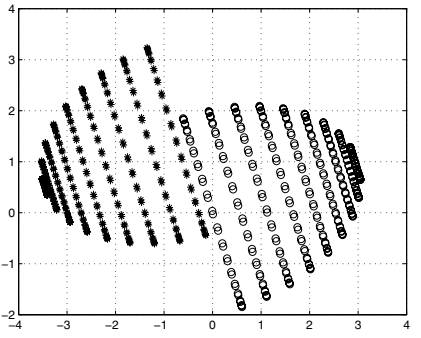

Fig. 5. NDE

videos. That is to run NDE on K slices $slice_k = \{\mathcal{T}_k^{(i)}\}_{i=1}^{M}$ (M is the total number of training video sequences), under which homogeneous keyframes belonging to synchronized and aligned clusters $\mathcal{C}_k = \{\mathcal{C}_k^{(i)}\}_i$ are input to NDE. We collect keyframes from different training videos presenting the same human identity to form one class in each slice. Ultimately, applying NDE on K slices leads to the target STEs, as well as K NDE projections $\mathrm{W}_k \in \Re^{d\times f}(k = 1, \cdots, K)$.

4 Statistical Recognition

In the testing stage, for any unidentified video frame $\mathbf{x} \in \Re^d$, we try to compute its statistical correlation to video sequences $V^{(i)}(i = 1, \cdots, c)$ in gallery which often has one video sequence for one human subject. Let the learned spatio-temporal embedding according to $V^{(i)}$ be $\mathcal{L}^{(i)} = \{\mathbf{y}_{kj}^{(i)} = \mathrm{W}_k^T \mathbf{e}_{kj}^{(i)} \in \Re^f | k = 1, \cdots, K, j = 1, \cdots, m\}$, and define $\mathcal{L}_k^{(i)} = \{\mathbf{y}_{kj}^{(i)} | j = 1, \cdots, m\}$. The statistical correlation is expressed as the posterior probability $p(\mathcal{L}_i|\mathbf{x})$.

Intuitively, we exploit the probabilistic fusion scheme to construct a MAP (maximum a posterior) classifier in terms of learned STEs, which settles on a

solution to the image-to-video face recognition problem. The MAP classifier is derived as follows

$$\begin{aligned}\max_{i\in\{1,\cdots,c\}} p(\mathcal{L}^{(i)}|\mathbf{x}) &= \max_i \sum_{k=1}^{K} p(\mathcal{L}^{(i)}, \mathcal{C}_k|\mathbf{x}) = \max_i \sum_{k=1}^{K} \frac{p(\mathcal{L}^{(i)}, \mathcal{C}_k, \mathbf{x})}{p(\mathbf{x})} \\ &= \max_i \sum_{k=1}^{K} \frac{p(\mathcal{L}^{(i)}|\mathbf{x}, \mathcal{C}_k)p(\mathbf{x}, \mathcal{C}_k)}{p(\mathbf{x})} \\ &= \max_i \sum_{k=1}^{K} p(\mathcal{L}_k^{(i)}|\mathbf{x}, \mathcal{C}_k)p(\mathcal{C}_k|\mathbf{x}). \end{aligned} \tag{9}$$

Since $p(\mathbf{x}, \mathcal{C}_k)$ has been modeled as eq. (3) through frame clustering on training videos, $p(\mathcal{C}_k|\mathbf{x})$ is calculated by $p(\mathcal{C}_k|\mathbf{x}) = p(\mathbf{x}, \mathcal{C}_k)/\sum_k p(\mathbf{x}, \mathcal{C}_k)$. Only the conditional probability $p(\mathcal{L}_k^{(i)}|\mathbf{x}, \mathcal{C}_k)$ is left to be inferred. To achieve that, we start by computing the asymmetric probabilistic similarity $S_k(\mathbf{y}, \mathbf{x})$

$$S_k(\mathbf{y}, \mathbf{x}) = \exp\left(-\frac{\|\mathbf{y} - \mathrm{W}_k^T\mathbf{x})\|^2}{2\sigma_k^2}\right), \tag{10}$$

where $\mathrm{W}_k^T\mathbf{x}$ is the k-th NDE, and parameter σ_k can be predefined or computed with respect to data distribution in the embedding space. Now we can formulate $p(\mathcal{L}_k^{(i)}|\mathbf{x}, \mathcal{C}_k)$ under the following stochastic selection rule

$$p(\mathcal{L}_k^{(i)}|\mathbf{x}, \mathcal{C}_k) = \frac{\sum_{j=1}^{m} S_k(\mathbf{y}_{kj}^{(i)}, \mathbf{x})}{\sum_{t=1}^{c}\sum_{j=1}^{m} S_k(\mathbf{y}_{kj}^{(t)}, \mathbf{x})}. \tag{11}$$

Once substituting eq. (10) and eq. (11) into eq. (9), we accomplish the image-to-video recognition task. It is noticeable that $p(\mathcal{L}_k^{(i)}|\mathbf{x}, \mathcal{C}_k)$ essentially behaves like a local analyzer $F_k(\mathbf{x}, i)$ as similarities given by eq. (10) are conducted. Specifically, $F_k(\mathbf{x}, i)$ correlates the testing frame $\mathbf{x}$ with the kth segment of spatio-temporal embedding of person i in subspace W_k. Our recognition solution eq. (9) merges these local analyzers into a global analyzer $G(\mathbf{x}, i)$ using the probabilistic fusion model $G(\mathbf{x}, i) = \sum_k p(\mathcal{C}_k|\mathbf{x})F_k(\mathbf{x}, i)$ which statistically fuses multiple NDEs of testing frame $\mathbf{x}$ with the probabilistic "weights" $p(\mathcal{C}_k|\mathbf{x})$.

Naturally, we can perform the probabilistic voting strategy to recognize a video sequence $V = \{\mathbf{x}_t\}_{t=1}^N$ in the probe videos. In details, combining the recognition confidences $\{G(\mathbf{x}_t, i)\}_i$ in every frame $\mathbf{x}_t$ to decide on the person identity i^* of probe video V, we thus realize video-to-video recognition as follows

$$\begin{aligned} i^* &= \arg\max_{i\in\{1,\cdots,c\}} p(\mathcal{L}^{(i)}|V) \\ &= \arg\max_i \sum_{t=1}^{N} p(\mathcal{L}^{(i)}|\mathbf{x}_t) = \arg\max_i \sum_{t=1}^{N} G(\mathbf{x}_t, i). \end{aligned} \tag{12}$$

5 Experiments

In this section, we conduct experiments on the XM2VTS face video database [11]. We select $294 * 4$ video sequences of 294 distinct persons across four different sessions. $294 * 3$ video sequences from the first three sessions are selected for training. The gallery set is composed of 294 video sequences from the first session. The probe set is composed of 294 video sequences from the fourth session. The persons in the video are asked to read two number sequences, "0 1 2 3 4 5 6 7 8 9" and "5 0 6 9 2 8 1 3 7 4".

5.1 Keyframes

In this paper we propose the Bayesian keyframe learning method for learning the temporal embeddings of videos, which complies with frame synchronization. In this section we will evaluate their performance on the XM2VTS face video database. Firstly we compare our keyframe learning method with the audio-guided keyframe extraction method proposed in [13]. The audio-guided method, called as "A-V Frame Synchronization", exploits the maximum points of audio information and strongly depends on the order of audio sentences which are spoken in video sequences. A-V synchronization extracts frames each of which corresponds to the waveform peak of audio signals.

Fig. 6(a) shows 10 cropped 72×64 keyframes learned by our method, which are the most likely frames, i.e. top-1, in $K = 10$ clusters provided by synchronized clustering on video frames without any additional information. From Fig. 6(a), we observe that keyframes extracted by our method bear rather distinct expression information, which will benefit face recognition as more expression variations are covered in the training data. An important advantage of our keyframe learning method is that it is fully automatic without relying on audio signals, which makes our recognition framework more general.

Fig. 6(b) illustrates our keyframe extraction from the audio-temporal perspective. It is surprising that these keyframes learned by our method correlate closely with the results obtained by A-V synchronization. 7 keyframes nearly lie on the peaks of audio signals, while the rest 3 frames are in the intermediate places between peaks. We thus conclude that our keyframe learning method is comparable to the audio-guided method confronted with orderly video data, but still robust to disordered video data.

5.2 Evaluate NDE

The first group of experiments is to compare the performance of the proposed NDE with the conventional LDA algorithm. For each video sequence, top-1 keyframes are selected for the experiments. So, the training set is composed of 294*3 facial images from the first three sessions. The gallery set contains 294 images from the first session and the probe set comprises 294 images from the fourth session. For the parameters z and β associated with NDE, we set $z = 1$ and $\beta = 2$ empirically.

The comparative results are shown in Fig. 7. The cumulative matching score is used for the performance measure. Instead of asking "is the top match correct",

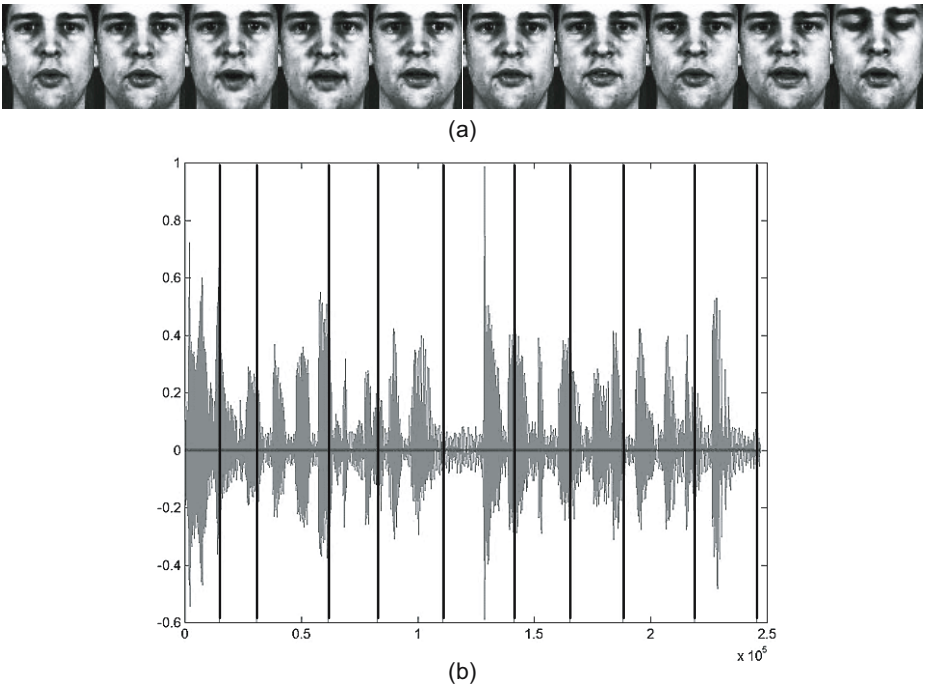

Fig. 6. Keyframes learned from one video sequence in XM2VTS. (a) Top-1 keyframes, each of which stands for a cluster in the sequence. (b) 10 keyframes shown in the temporal axis, compared with the speech signal.

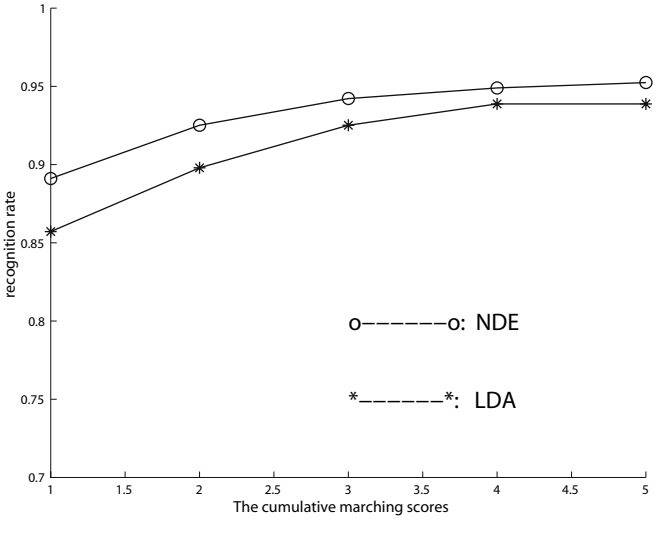

Fig. 7. NDE versus LDA

Table 2. Comparison of recognition results with existing video-based approaches

Video-Based Face Recognition Approaches	Recognition Rate
mutual subspace	79.3%
nearest frame	81.7%
nearest frame using LDA	90.9%
nearest frame using unified subspace analysis	93.2%
temporal embeddings + multi-level subspace analysis	98.0%
spatio-temporal embeddings + statistical classifier	99.3%

the cumulative matching score answers the questions of "is the correct answer in the top-n matches?", where the number n is called the rank. This lets one know how many images have to be examined to get a desired level of performance. The results clearly show the superiority of NDE over LDA.

5.3 Evaluate Statistical Recognition Performance

After 20 keyframes (top-2 keyframes from 10 clusters) are selected by means of synchronized frame clustering and Bayesian keyframe learning, we perform NDE on each slice containing 2*3*294 frames across 3 sessions from 294 persons and acquire 10 local analyzers with 6 training samples each person. Finally, based on the learned spatio-temporal embeddings (STEs), all these STE-based local analyzers are integrated using the probabilistic fusion model eq. (9) to accomplish image-to-video recognition followed by the probabilistic voting strategy eq. (12) which leads to the final video-to-video recognition results .

We compare our statistical recognition framework with existing video-based face recognition approaches. Here all approaches directly use image gray scale values as facial features. Using temporal embeddings learned by our keyframe learning method, we can perform multi-level subspace analysis proposed in our recent paper [13]. It is evident that both of our proposed approaches, temporal embeddings combined with multi-level subspace analysis and spatio-temporal embeddings incorporated into the statistical classifier, significantly outperform the existing video-based recognition approaches in Tab. 2. Further, our statistical recognition framework integrating STEs and statistical classification achieves the best recognition accuracy of 99.3%. Compared with the best performance of other recognition approaches, the error rate is still reduced by 65%, which is quite impressive and well validates the robustness and effectiveness of our framework.

6 Conclusion

This paper explores to seek a "good" spatio-temporal representation for each video sequence so that it could support the face recognition process. Considering both

co-occurrence statistics and representative capability of video frames, the temporal embedding spanned by synchronized keyframes is first learned for each sequence. Furthermore, the powerful NDE enforces discriminative cues over the learned temporal embeddings. So, learning in space-time gives rise to the intrinsic spatio-temporal embeddings (STEs). In this paper, we develop a statistical framework integrating several novel techniques including Bayesian keyframe learning, NDE, and the statistical classifier, all of which depend on each other and yield a synergistic effect. The success of the framework originates from not only synchronized but also discriminative spatio-temporal representations and statistical recognition.

Acknowledgement

The work described in this paper was fully supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region and a joint grant (N_CUHK409/03) from HKSAR RGC and China NSF. The work was done at the Chinese University of Hong Kong.

References

1. Baker, S., Kanade, T.: Limits on Super-Resolution and How to Break Them. *IEEE Trans. PAMI.* **24**:9 (2002) 1167-1183
2. Bar-hillel, A., Hertz, T., Shental, N., Weinshall, D.: Learning a mahalanobis metric from equivalence constraints. *J. of Machine Learning Research.* **6** (2005) 937-965
3. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Trans. PAMI.* **19**:7 (1997) 711-720
4. Duda, R., Hart, P., Stork, D.: Pattern Classification. Wiley, New York. (2000)
5. Fukunaga, K.: Statistical Pattern Recognition. Academic Press (1990)
6. Krueger, V., Zhou, S.: Exemplar-Based Face Recognition from Video. In *Proc. ECCV.* (2002) 732-746
7. Lee, K., Ho, J., Yang, M., Kriegman, D.: Video-Based Face Recognition Using Probabilistic Appearance Manifolds. In *Proc. IEEE Conf. CVPR.* (2003) 313-320
8. Liu, C., Shum, H., Zhang, C.: A Two-Step Approach to Hallucinating Faces: Global Parametric Model and Local Nonparametric Model. In *Proc. IEEE Conf. CVPR.* (2001) 192-198
9. Liu, X., Chen, T.: Video-Based Face Recognition Using Adaptive Hidden Markov Models. In *Proc. IEEE Conf. CVPR.* (2003) 340-345
10. Liu, W., Lin, D., Tang, X.: TensorPatch Super-Resolution and Coupled Residue Compensation. In *Proc. IEEE Conf. CVPR.* (2005) 478-484
11. Messer, K., Matas, J., Kittler, J., Luettin, J., Matitre, G.: XM2VTSDB: The Extended M2VTS Database. In *Proc. 2nd Int. Conf. Audio- and Video-Based Biometric Person Authentication.* (1999) 72-77
12. Satoh, S.: Comparative Evaluation of Face Sequence Matching for Content-based Video Access. In *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition.* (2000) 163-168

13. Tang, X., Li, Z.: Frame Synchronization and Multi-Level Subspace Analysis for Video Based Face Recognition. In *Proc. IEEE Conf. CVPR.* (2004) 902-907
14. Wang, X., Tang, X.: A Unified Framework for Subspace Face Recognition. *IEEE Trans. PAMI.* **26**:9 (2004) 1222-1228
15. Yamaguchi, O., Fukui, K., Maeda, K.: Face Recognition Using Temporal Image Sequence. In *Proc. Int. Conf. Face and Gesture Recognition* (1998) 318-323
16. Zhou, S., Krueger, V., Chellappa, R.: Probabilistic Recognition of Human Faces from Video. *Computer Vision and Image Understanding.* **91**:1 (2003) 214-245