Video-Based Face Recognition Using Spatio-Temporal Representations

John See[1], Chikkannan Eswaran[1] and Mohammad Faizal Ahmad Fauzi[2]
[1]*Faculty of Information Technology, Multimedia University*
[2]*Faculty of Engineering, Multimedia University*
Malaysia

1. Introduction

Face recognition has seen tremendous interest and development in pattern recognition and biometrics research in the past few decades. A wide variety of established methods proposed through the years have become core algorithms in the area of face recognition today, and they have been proven successful in achieving good recognition rates primarily in still image-based scenarios (Zhao et al., 2003). However, these conventional methods tend to perform less effectively under uncontrolled environments where significant face variability in the form of complex face poses, 3-D head orientations and various facial expressions are inevitable circumstances. In recent years, the rapid advancement in media technology has presented image data in the form of *videos* or *video sequences*, which can be simply viewed as a temporally ordered collection of images. This abundance and ubiquitous nature of video data has presented a new fast-growing area of research in video-based face recognition (VFR).

Recent psychological and neural studies (O'Toole et al., 2002) have shown that facial movement supports the face recognition process. Facial dynamic information is found to contribute greatly to recognition under degraded viewing conditionsand also when a viewer's experience with the same face increases. Biologically, the media temporal cortex of a human brain performs motion processing, which aids the recognition of dynamic facial signatures. Inspired by these findings, researchers in computer vision and pattern recognition have attempted to improve machine recognition of faces by utilizing video sequences, where temporal dynamics is an inherent property.

In VFR, temporal dynamics can be exploited in various ways within the recognition process (Zhou, 2004). Some methods focused on directly modeling temporal dynamics by learning transitions between different face appearances in a video. In this case, the sequential ordering of face images is essential for characterizing temporal continuity. While its elegance in modeling dynamic facial motion "signatures" and its feasibility for simultaneous tracking and recognition are obvious, classification can be unstable under real-world conditions where demanding face variations can caused over-generalization of the transition models learned earlier.

In a more general scenario, VFR can also be performed by means of image sets, comprising of independent unordered frames of a video sequence. A majority of these methods characterize the face manifold of a video using two different representations – (1) face subspaces[1], and (2)

[1] Sometimes termed as *local sub-manifolds* or *local models* in different works.

face exemplars, or representative images that summarizes face appearances in a video. While these methods are attractive due to their straightforward representation and computational ease, they are typically dependent on the effectiveness of extracting meaningful subspaces or exemplars that can accurately represent the videos. Certain works have incorporated temporal dynamics into the final classification step to some degree of success.

This chapter presents a new framework for video-based face recognition using spatio-temporal representations at various levels of the task. Using the exemplar-based approach, our spatio-temporal framework utilizes additional temporal information between video frames to partition training video data into local clusters. Face exemplars are then extracted as representatives of each local cluster. In the feature extraction step, meaningful spatial features are extracted from both training and test data using the newly proposed Neighborhood Discriminative Manifold Projection (NDMP) method. Finally, in order to facilitate video-to-video recognition, a new exemplar-driven Bayesian network classifier which promotes temporal continuity between successive video frames is proposed to identify the subject in test video sequences. In the next section, some related works in literature will be discussed.

2. Related work

By categorizing based on feature representation, recent methods in video-based face recognition (VFR) can be loosely organized into three categories: (1) direct modeling of temporal dynamics, (2) subspace-based representation, and (3) exemplar-based representation.

In video sequences, continuity is observed in both face movement and change in appearances. Successful modeling of temporal continuity can provide an additional dimension into the representation of face appearances. As such, the smoothness of face movement can also be used for face tracking. Simultaneous tracking and recognition by Zhou and Chellappa is the first approach that systematically incorporate temporal dynamics in video-based face recognition (Zhou et al., 2003). A joint probability distribution of identity and head motion using sequential importance sampling (SIS) was modelled. In another tracking-and-recognition work (Lee et al., 2005), a nonlinear appearance manifold representing each training video was approximated as a set of linear sub-manifolds, and transition probabilities were learned to model the connectivity between sub-manifolds. Temporal dynamics within a video sequence can also be modeled over time using Hidden Markov Models (HMM) (Liu & Chen, 2003). Likelihood scores provided by the HMMs are then compared, and the identity of a test video is determined by its highest score. Due to the nature of these representations, many of these methods lack discriminating power due to disjointed person-specific learning. Moreover, the learning of temporal dynamics during both training and recognition tasks can be very time-consuming.

Subspace-based methods represent entire sets of images as subspaces or manifolds, and are largely parametric in nature. Typically, these methods represent image sets using parametric distribution functions (PDF) followed by measuring the similarity between distributions. Both the Mutual Subspace Method (MSM) (Yamaguchi et al., 1998) and probabilistic modeling approaches (Shakhnarovich et al., 2002) utilize a single Gaussian distribution in face space while Arandjelovic et al. (Arandjelovic et al., 2005) extended this further using Gaussian mixture models. While it is known that these methods suffer from the difficulty of parameter estimation, their simplistic modeling of densities is also highly sensitive to conditions where training and test sets have weak statistical relationships. In a specific work on image sets, Kim et al. (Kim et al., 2007) bypass the need of using PDFs by computing similarity between subspaces using canonical correlations.

Exemplar-based methods offer an alternative model-free method of representing image sets. This non-parametric approach has become increasingly popular in recent VFR literature. Krüeger and Zhou (Krüeger & Zhou, 2002) first proposed a method of selecting exemplars from face videos using radial basis function network. There are some comprehensive works (Fan & Yeung, 2006; Hadid & Peitikäinen, 2004) that proposed view-based schemes by applying clustering techniques to extract view-specific clusters in dimensionality-reduced space. Cluster centers are then selected as exemplars and a probabilistic voting strategy is used to classify new video sequences. Later exemplar-based works such as (Fan et al., 2005; Liu et al., 2006) performed classification using various Bayesian learning models to exploit the temporal continuity within video sequences. Liu et al. (Liu et al., 2006) also introduced a spatio-temporal embedding that learns temporally clustered *keyframes* (or exemplars) which are then spatially embedded using nonparametric discriminant embedding. While all these methods have good strengths, none of these classification methods consider the varying influence of different exemplars with respect to their parent clusters.

Our proposed exemplar-based spatio-temporal framework for video-based face recognition can be summarized by the following contributions at each step of the recognition process:

1. **Clustering and Exemplar Extraction**: Motivated by the advantages of using hierarchical agglomerative clustering (HAC) (Fan & Yeung, 2006), a new spatio-temporal hierarchical agglomerative clustering (STHAC) method is proposed to exploit temporal continuity in video sequences. For each training video, the nearest faces to the cluster means are selected as exemplars.
2. **Feature Representation**: The newly proposed Neighborhood Discriminative Manifold Projection (NDMP) (See & Ahmad Fauzi, 2011) is applied to extract meaningful features that are both discriminative and neighborhood preserving, from both training and test data.
3. **Classification**: A new exemplar-driven Bayesian network classifier which promotes temporal continuity between successive video frames is proposed to identify the subject in test video sequences. Our classifier accumulates evidences from previous frames to decide on the subject identity. In addition, causal relationships between each exemplar and its parent class are captured by the Bayesian network.

Figure 1 illustrates the steps taken in our proposed framework, and where spatial and temporal information has been utilized.

3. Exemplar-based spatio-temporal framework

In this section, we elaborate in detail our proposal of an exemplar-based spatio-temporal framework for video-based face recognition.

3.1 Problem setting

The abundance of images in video poses a methodological challenge. While conventional still image-based face recognition is a straightforward matching of a test image to a gallery of training images *i.e.* an *image-to-image*[2] recognition task, it is an ill-posed problem for video

[2] In the abbreviated recognition settings, the first and third words denote data representation for each subject in the training/gallery and test/probe sets respectively.

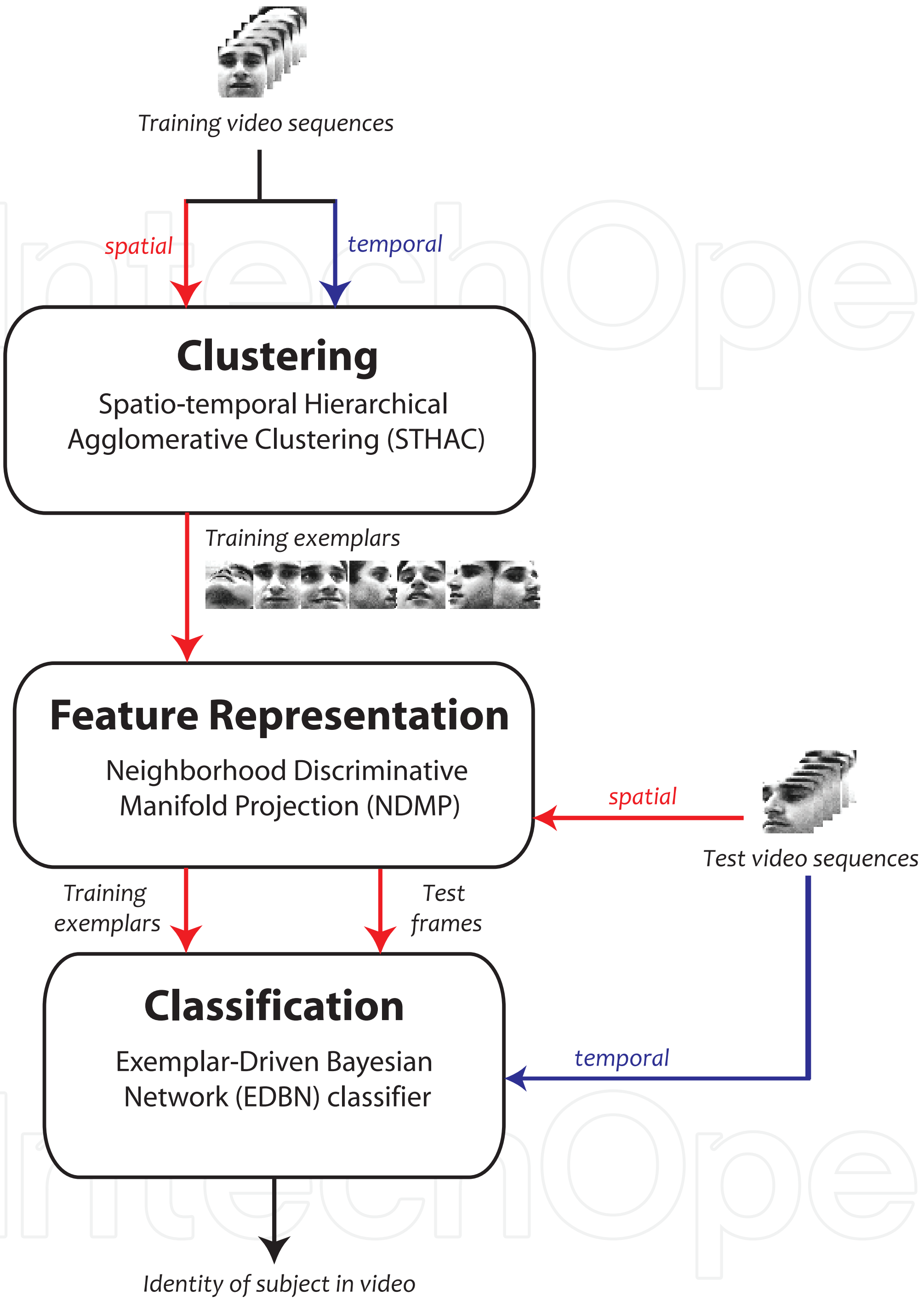

Training video sequences
spatial
temporal
Clustering
Spatio-temporal Hierarchical Agglomerative Clustering (STHAC)
Training exemplars
Feature Representation
Neighborhood Discriminative Manifold Projection (NDMP)
spatial
Test video sequences
Training exemplars
Test frames
Classification
Exemplar-Driven Bayesian Network (EDBN) classifier
temporal
Identity of subject in video

Fig. 1. The proposed exemplar-based spatio-temporal framework for video-to-video face recognition. The usage of spatial and temporal information are indicated in red and blue respectively in this diagram. The STHAC method in the clustering step and EDBN classifier in the classification step utilizes both spatio-temporal dynamics. Feature representation (NDMP method) takes only spatial information since the extracted training exemplars are used here, unlike subspace-based and temporal modeling-based methods (as discussed in Section 2).

sequences. Which image from the training video is to be matched with images from the test video?

To accomplish a complete *video-to-video* setting, one possible configuration used by exemplar-based approaches is to simplify it to a *image-to-video* recognition task, whereby each training video is represented by a set of exemplars (Fan & Yeung, 2006; Hadid & Peitikäinen, 2004). The availability of multiple image frames in the test video provides a good platform for utilizing temporal information between successive frames. For general notation, given a sequence of face images extracted from a video,

$$X_c = \{x_1^c, x_2^c, \ldots, x_{N_c}^c\} \ , \tag{1}$$

where N_c is the number of face images in the video. Assuming that each video contains the faces of the same person and c is the subject identity of a C-class problem, $c \in \{1, 2, \ldots, C\}$, its associated exemplar set

$$E_c = \{e_1^c, e_2^c, \ldots, e_M^c\} \ , \tag{2}$$

where $E_c \subseteq X_c$ and the number of exemplars extracted, $M \ll N_c$. Thus, the overall exemplar-class set can be succintly summarized as

$$E = \{e_{c,m} | c = 1, \ldots, C; m = 1, \ldots, M\} \ , \tag{3}$$

in which there are a total of $C \times M$ unique exemplar-classes. In cases where more than one training video of a particular class is used, image frames from all similar-class videos are aggregated to extract M exemplars.

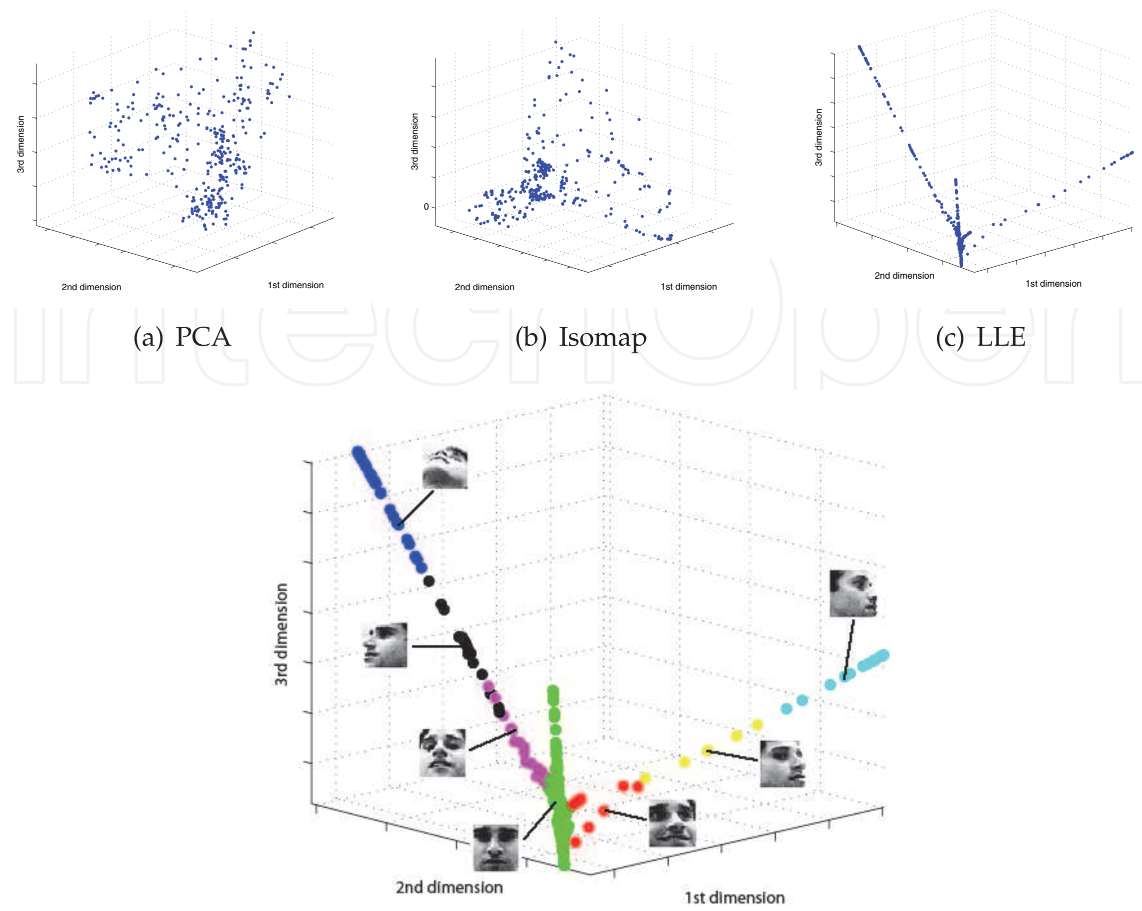
3.2 Embedding face variations

Considering the large amount of face variations in each training video sequence, a suitable dimensionality reduction method is necessary to uncover the intrinsic structure of the data manifold which may originally lie on a complex hyperplane. Recent nonlinear dimensionality reduction techniques such as Locally Linear Embedding (LLE) (Roweis & Saul, 2000) and Isomap (Tenenbaum et al., 2000) have been proven effective at seeking a low-dimensional embedding of a data manifold. LLE in particular, is known for its capability in modeling the global intrinsic structure of the manifold while preserving local neighborhood structures to better capture various face variations such as pose, expression and illumination. Conventional unsupervised methods such as Principal Component Analaysis (PCA) (Turk & Pentland, 1991) and Multidimensional Scaling (MDS) (Cox & Cox, 2001) have the tendency of overestimating the intrinsic dimensionality of high-variability data.

For each training video in our method, we apply the LLE algorithm to embed the face variations in a low-dimensional space, in preparation for the subsequent clustering step. Fig. 2(a), 2(b) and 2(c) shows the embedding of image data from a single video in PCA, Isomap and LLE spaces respectively. From Fig. 2(d), we can clearly see that the LLE method is able to uncover linear patches in its embedded space after performing spectral clustering, where each cluster represents a particular face appearance.

3.3 Spatio-temporal clustering for exemplar extraction

In the next step, clustering is performed on the LLE-space by extracting M number of clusters that group together faces of similar appearances. In many previous works (Fan et al., 2005; Hadid & Peitikäinen, 2004), k-means clustering is the primary choice for assigning data into different clusters due to its straightforward implementation. However, it has some obvious limitations – firstly, it is sensitive to the initial seeds used, which can differ in every run, and secondly, it produces suboptimal results due to its inability to find global minima.



(a) PCA (b) Isomap (c) LLE

(d) Linear cluster patches (in different colors) in LLE-space.

Fig. 2. The embedding plots of data taken from a sample video sequence, embedded using (a) PCA, (b) Isomap, and (c) LLE dimensionality reduction methods. Each data point represents an image in the embedded space. In the case of LLE, the formation of linear patches enables the spreading of different face appearances across its spectral projection (d).

3.3.1 Hierarchical Agglomerative Clustering (HAC)

Hierarchical Agglomerative Clustering (HAC) is a hierarchical method of partitioning data points by constructing a nested set of partitions represented by a cluster tree, or *dendrogram* (Duda et al., 2000). The agglomerative approach works from "bottom-up" by grouping smaller clusters into larger ones, as described by the following procedure:

1. Initialize each data point (of all N_c points) as a singleton cluster.
2. Find the nearest pairs of clusters, according to a certain distance measure $d(\Phi_i, \Phi_j)$ between clusters Φ_i and Φ_j. Commonly used measures are such as single-link, complete-link, group-average-link and Ward's distance criterion. Merge the two nearest clusters to form a new cluster.
3. Continue distance computation and merging (repeat Step 2), and terminate when all points belong to a single cluster. The required number of clusters, M is selected by partitioning at the appropriate level of the dendrogram.

3.3.2 Spatio-Temporal Hierarchical Agglomerative Clustering (STHAC)

Our proposed Spatio-Temporal Hierarchical Agglomerative Clustering (STHAC) differs from the standard HAC in terms of the computation of the nearest pair of clusters (Step 2). A *spatio-temporal distance measure* is proposed by fusing both spatial and temporal distances to influence the location of data points in time-space dimension. Since clustering procedures generally only utilize the distances between points, perturbations can be applied to the original spatial distances without cumbersome modeling of points in time-space dimension.

While spatial distance is measured by simple Euclidean distance between points, temporal distance is measured by the time spanned between two frame occurrences (x_i and v_j) in a video sequence,

$$d_T(x_i, x_j) = \left| t_{x_i} - t_{x_j} \right| \tag{4}$$

where t is a discretized unit time. This formulation is intuitive enough to quantify temporal relationships across sequentially ordered data points. The matrices containing pairwise spatial Euclidean distances $D_S(x_i, x_j)$ and temporal distances $D_T(x_i, x_j)$ between all data points, are computed and normalized. We present two varieties of fusion schemes, one which functions at the global structural level, and another at the local neighborhood level.

The Global Fusion (STHAC-GF) scheme blends the contribution of spatial and temporal distances using a temporal tuning parameter, α. The tuning parameter adjusts the perturbation factor defined by its upper and lower bounds, p_{max} and p_{min} respectively, which acts to increase or reduce the original distances. GSTF defines the spatio-temporal distance as

$$D_{ST,global} = (p_{max} - \alpha)D_S + (\alpha + p_{min})D_T, \qquad 0 \leq \alpha \leq 1\ . \tag{5}$$

The Local Perturbation (STHAC-LP) scheme perturbs spatial and temporal distances based on local spatio-temporal neighborhood relationships between a point and its neighbors. For each point x_i, a temporal window segment,

$$S_{x_i} = \{x_{i-w}, \ldots, x_i, \ldots, x_{i+w}\}\ , \tag{6}$$

of length $(2w + 1)$ is defined as its temporal neighborhood. The spatial neighborhood of point x_i,

$$Q_{x_i} = \{x_1, x_2, \ldots, x_k\} \tag{7}$$

is simply a set containing k-nearest neighbors of x_i computed by Euclidean distance. A point x_j is identified as a *common spatio-temporal neighbor* (CSTN) of point x_i if it belongs to both spatial and temporal neighborhood point sets, hence the criterion,

$$CSTN_{x_i} = S_{x_i} \cap Q_{x_i}\ . \tag{8}$$

With that, we introduce a perturbation affinity matrix containing multipliers that represent the degree of attraction and repulsion betwee

$$P_{ij} = \begin{cases} 1 - \lambda_{sim}, & \text{if } x_j \in CSTN_{x_i} \\ 1 + \lambda_{dis}, & \text{if } x_j \in (S_{x_i}\ CSTN_{x_i}) \\ 1, & \text{otherwise} \end{cases} \tag{9}$$

where λ_{sim} and λ_{dis} are the similarity and dissimilarity perturbation constants respectively, taking appropriate values of $0 < \{\lambda_{sim}, \lambda_{dis}\} < d(x_i, x_j)$. To simplify parameter tuning, we use a single perturbation constant, that is $\lambda = \lambda_{sim} = \lambda_{dis}$. In short P_{ij} seeks to accentuate the similarities and dissimilarities between data samples by artificially reducing and increasing spatial and temporal distances between samples. By matrix multiplication, STHAC-LP defines the spatio-temporal distance as

$$D_{ST,local} = P_{ij}(D_S + D_T) . \tag{10}$$

The linkage criterion chosen in our work for merging clusters is Ward's distance criterion,

$$d(\Phi_i, \Phi_j) = \frac{n_i n_j}{n_i + n_j} \left\| m_i - m_j \right\|^2 . \tag{11}$$

where m_i and m_j are means of cluster i and j respectively, while n_i and n_j are the cardinality of the clusters. Ward's criterion (Webb, 2002) is chosen due to its attractive sample-dependent mean squared error term, which is a good heuristic for approximating the suitable number of clusters via finding the "elbow" of the residual error curve (see Fig. 3).
After clustering the face data in each training video, face images that are nearest to each cluster mean are chosen as exemplars.

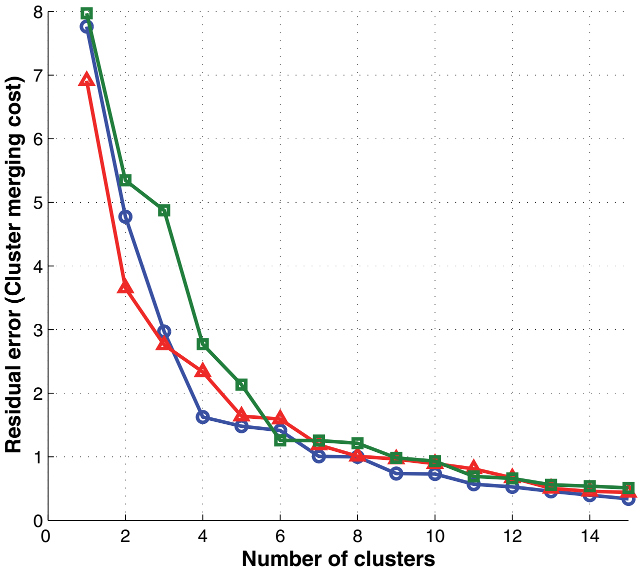


Fig. 3. Residual error curve of three different training videos (plotted with different colors) that were partitioned into different number of clusters. The "elbow" of the curve is approximately at 5 – 9 clusters.

3.4 Feature representation

Traditional linear subspace projection methods such as Principal Component Analysis (PCA) (Turk & Pentland, 1991) and Linear Discriminant Analysis (LDA) (Belhumeur et al., 1997) have been widely used to great effect in characterizing data in smooth and well-sampled manifolds. Recently, there has been a flurry of manifold learning methods such as Locality Preserving Projections (LPP) (He & Niyogi, 2003), Marginal Fisher Analysis (MFA) (Yan et al., 2007) and Neighborhood Preserving Embedding (NPE) (He et al., 2005), that are able to effectively derive optimal linear approximations to a low-dimensional embedding of complex manifolds. NPE in particular, has an attractive neighborhood preserving property due to its formulation based on LLE.

For improved extraction of features for face recognition, we apply the Neighborhood Discriminative Manifold Projection (NDMP) method (See & Ahmad Fauzi, 2011), which is our earlier work on feature representation. NDMP is a supervised discriminative variant of the NPE which seeks to learn an optimal low-dimensional projection by distinguishing between intra-class and inter-class neighborhood reconstruction. Global structural and local neighborhood constraints are imposed in a constrained optimization problem, which can be solved as a generalized eigenvalue problem:

$$(\mathbf{X}\mathbf{M}_{\mathbf{intra}}\mathbf{X}^{\mathbf{T}})\mathbf{A} = \Lambda(\mathbf{X}\mathbf{M}_{\mathbf{inter}}\mathbf{X}^{\mathbf{T}} + \mathbf{X}\mathbf{X}^{\mathbf{T}})\mathbf{A}, \tag{12}$$

where $\mathbf{X}$ denotes the face exemplar set in $\Re^D$, while $\mathbf{M}_{\mathbf{intra}}$ and $\mathbf{M}_{\mathbf{inter}}$ are the intra-class and inter-class orthogonal weight matrices respectively.

A new test sample $\mathbf{X}'$ can be projected to the embedded space in $\Re^{D'}$ by the linear transformation

$$\mathbf{Y}' = \mathbf{A}^{\mathbf{T}}\mathbf{X}' \tag{13}$$

where $D' \ll D$. More details on the theoretical formulation of the NDMP method can be found in (See & Ahmad Fauzi, 2011).

Fig. 4 shows the plots of face exemplar points in LDA, LPP and NDMP feature space. Note that among the three supervised methods, the NDMP dimensionality reduction method provides the best discrimination between classes. The insufficient amount of separation between different-class points (for LPP) and attraction within same-class points (for LDA and LPP) can potentially contribute towards erroneous classification in the next task.

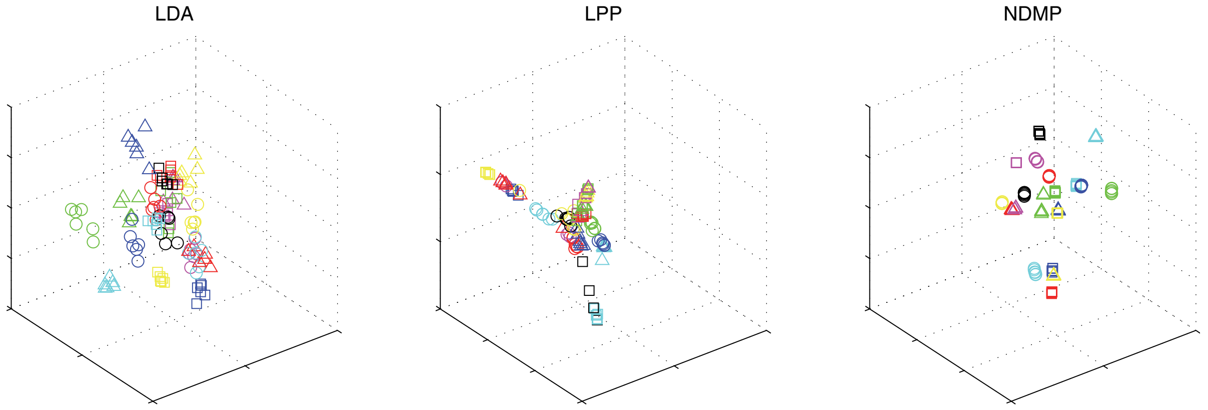


Fig. 4. Data points of a 20-class exemplar set plotted in LDA *(left)*, LPP *(center)*, and NDMP *(right)* feature space. Exemplars of each class are indicated by a different shape-color point.

3.5 Exemplar-driven Bayesian network classification

Many conventional still-image face recognition systems evaluate the performance of an algorithm by measuring recognition accuracies or error rates, which can simply be computed based on the number of correctly or incorrectly identified test images in a test set. In video-based evaluation, this is usually extended to a voting scheme, where the face in every video frame is identified independently, and then a voting method (typically majority vote or confidence-based methods such as sum rule and product rule) is performed to decide on the overall identity of the person in the sequence.

In this work, we propose a new classification method for video sequences using an exemplar-driven Bayesian network (EDBN) classifier, which introduces a joint probability function that is estimated by *maximum a posteriori* (MAP) decision rule within a Bayesian inference model. In comparison to other Bayesian methods (Fan et al., 2005; Liu et al., 2006), the EDBN classifier incorporates temporal continuity between successive video frames with consideration for the causal relationships between exemplars and their parent classes.

In a Bayesian inference model (Duda et al., 2000), the subject identity of a test video X can be found by estimating the MAP probability decision rule,

$$c^* = arg \max_C P(c|x_{1,\ldots,N_c}), \tag{14}$$

where the subscript notation of x succintly represents a sequence of N images.

In a typical Naive Bayes classifier, estimation based on MAP decision rule can be expressed as

$$P(c|x_{1,\ldots,N_c}) = \frac{P(c)P(x_{1,\ldots,N_c}|c)}{P(x_{1,\ldots,N_c})} = \frac{P(c)P(x_{1,\ldots,N_c}|c)}{\sum_c P(x_{1,\ldots,N_c}|c)P(c)}\ , \tag{15}$$

where $P(c)$ is the prior probability of each class, $P(x_{1,\ldots,N_c}|c)$ is the likelihood probability of x given class c and the denominator is a normalization factor to ensure that the sum of the likelihoods over all possible classes equals to 1.

Within an embedding space, the extracted exemplars can be irregularly located due to varying degree or magnitude of appearances. Thus, they should be weighted according to their influence or contribution in the subspace. Intuitively, contribution of exemplars that lie farther from the within-class exemplar subspace (more limited or uncommon) should be emphasized while exemplars that are near the within-class exemplar subspace (more likely found) should contribute at a lesser degree. To introduce causal relationship between exemplars and their parent classes, we formulate a joint probability function,

$$P(c, E, X) = P(X|c, E)P(E|c)P(c)\ , \tag{16}$$

which includes the exemplar-class set E as a new latent variable. The graphical model of the EDBN classifier is shown in Fig. 5. Thus, the MAP classifier is redefined by maximizing the *joint posterior probability* of the class c and exemplar-class E given observation X:

$$\begin{aligned} \max_C P(c, E|X) &= \max_C \frac{P(c, E, X)}{P(X)} \\ &= \max_C \sum_{j=1}^{M} \frac{P(X|c, e_{c,j})P(e_{c,j}|c)P(c)}{P(X)} \\ &= \max_C \sum_{j=1}^{M} \prod_{i=1}^{N_c} \frac{P(x_i|c, e_{c,j})P(e_{c,j}|c)P(c)}{P(x_i)}\ . \end{aligned} \tag{17}$$

The marginal probability $P(x_i)$ does not depend on both c and E, thus functioning only as a normalizing constant. Since the class prior probability $P(c)$ is assumed to be non-informative at the start of observation sequence X, using uniform priors is a good estimation. The conditional probability $P(e_{c,j}|c)$ represents the *exemplar prominence probabilitiy* while $P(x_i|c, e_{c,j})$ is the class likelihood probability.

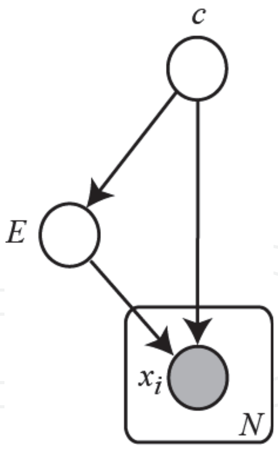


Fig. 5. Graphical model of the exemplar-driven Bayesian network (EDBN) classifier

3.5.1 Computation of class likelihood probability

Conventional Bayesian classifiers typically estimate the distribution of data using a multivariate Gaussian density function. However, accurate estimation of distribution can be challenging with the limited sample size in our problem setting. Alternatively, we can perform non-parametric density estimation by applying distance measures (or a kernel density estimator with uniform kernel), which are computationally inexpensive.

We define Frame Similarity Score (FSS) as the reciprocal of the ℓ^2-norm distance between the observed face image x_i and the j-th exemplar-class $e_{c,j}$,

$$S_i^{FSS}(x_i, e_{c,j}) = \frac{1}{d_{\ell^2}(x_i, e_{c,j})}. \tag{18}$$

The likelihood probability of the test face image x_i given the class c and exemplar-class e is determined by the ratio of FSS for exemplar-class $e_{c,j}$ to the total sum of FSS across all $C \times M$ exemplar-classes,

$$P(x_i|c, e_{c,j}) = \frac{S_i^{FSS}(x_i, e_{c,j})}{\sum_{k=1}^{C} \sum_{m=1}^{M} S_i^{FSS}(x_i, e_{k,m})}. \tag{19}$$

3.5.2 Computation of exemplar prominence

Causal relationship between exemplars and their parent classes can be represented by the exemplar prominence probability $P(e_{c,j}|c)$. Similar to the computation of likelihood probability, we avoid estimating density functions by representing the influence of an exemplar in its own class subspace using a normalized Hausdorff distance metric,

$$d_h(e_{c,j}, E_c) = \frac{1}{\kappa} \min_{e' \in E_c} \left\| e_{c,j} - e' \right\| , \tag{20}$$

where E_c is the exemplar set of class c and κ is a normalizing factor to ensure that distances are relatively scaled for each class.

By defining Exemplar Prominence Score (EPS) as the reciprocal of the distance metric,

$$S_{c,j}^{EPS}(E_c, e_{c,j}) = 1/d_h(e_{c,j}, E_c), \tag{21}$$

the exemplar prominence probability can be formulated as

$$P(e_{c,j}|c) = \frac{S^{EPS}_{c,j}(E_c, e_{c,j})}{\sum_{m=1}^{M} S^{EPS}_{c,j}(E_c, e_{c,m})}\,, \tag{22}$$

which can be pre-computed since it does not depend on input observation X.

4. Experimental setup

Unlike VFR, still image-based face recognition uses very standardized evaluation procedures and there exist many benchmark datasets up-to-date (Gross, 2004). Due to the different evaluation settings used for video-based face recognition, it is generally difficult to make direct comparisons between results reported in different works in the literature. Also, a large variety of datasets have been used or customized for the purpose of video-based face recognition.

4.1 Datasets

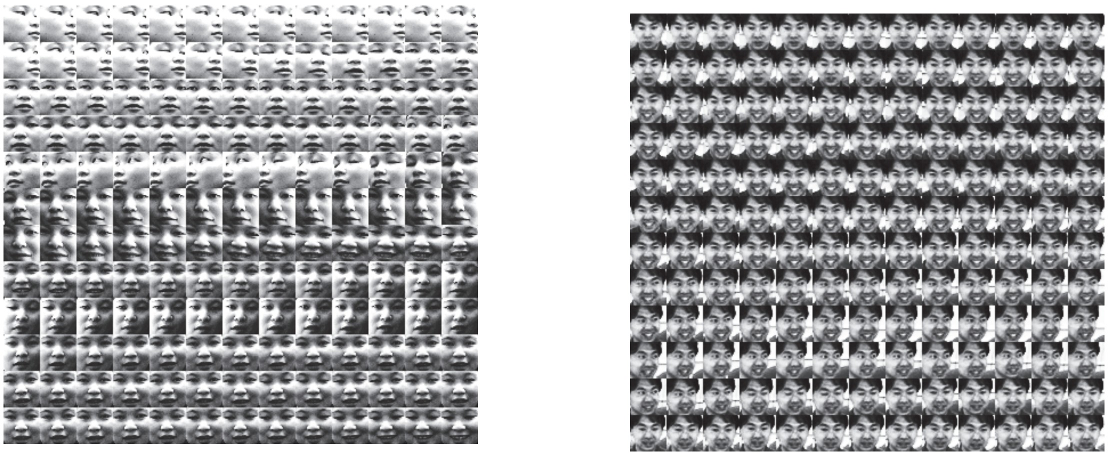


Fig. 6. Sample face images of a video sequence from the Honda/UCSD *(left)*, and CMU MoBo *(right)* datasets

In order to ensure a comprehensive evaluation was conducted, we use two standard video face datasets: Honda/UCSD Face Video Database (Lee et al., 2005) and CMU Motion of Body (MoBo) (Gross & Shi, 2001).

The first dataset, Honda/UCSD, which was collected for the purpose of VFR, consists of 59 video sequences of 20 different people (each person has at least 2 videos). Each video contains about 300-600 frames, comprising of large pose and expression variations and significant 3-D head rotations. The second dataset, CMU MoBo is another commonly used benchmark dataset customized for VFR, although it was originally intended for human motion analysis. It consists of 96 sequences of 24 different subjects (each person has 4 videos). Each video contains about 300 frames.

For both datasets, faces were extracted using the Viola-Jones face detector (Viola & Jones, 2001) and IVT object tracker (Ross et al., 2008) (which is very robust towards out-of-plane head movements), to ensure all frames with the presence of a face are successfully processed. The extracted faces are resized to grayscale images of 32 × 32 pixels, followed by histogram equalization to normalize lighting effects. Sample face images of a video sequence from both datasets are shown in Fig. 6.

4.2 VFR evaluation settings

For each subject, one video sequence is used for training, and the remaining video sequences for testing. To ensure extensive evaluation on the datasets, we construct our test set by randomly sampling 20 sub-sequences consisting of 100 frames from each test video sequence. We use 7 exemplars for Honda/UCSD and 6 exemplars for CMU MoBo[3]. Tuning parameters for STHAC-GF and STHAC-LP were set at $\alpha = 0.75$ and $\lambda = 0.2$ respectively.

5. Evaluation results and discussions

To evaluate our proposed exemplar-based spatio-temporal framework for video-based face recognition, we conduct a collection of experiments to test the effectiveness of novel spatio-temporal representations at various levels of the VFR task. Finally, some rank-based identification results are reported.

5.1 Exemplar extraction/clustering methods

Focusing our attention first to exemplar extraction/clustering methods, we perform experiments on the following exemplar-based methods:

- Random exemplar selection
- LLE+*k*-means clustering (used in (Hadid & Peitikäinen, 2004))
- Geodesic distances + HAC (used in (Fan & Yeung, 2006))
- LLE + HAC
- LLE + STHAC-GF
- LLE + STHAC-LP

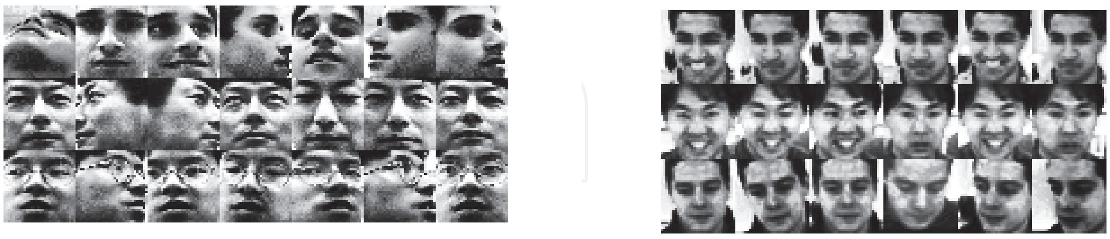


Fig. 7. Sample extracted exemplar sets of three different training videos (one in each row) from the Honda/UCSD *(left)*, and CMU MoBo *(right)* datasets.

To narrow the scope of our experiments, we only conduct this experiment on the Honda/UCSD dataset since it possesses much larger and more complex face variations compared to the CMU MoBo. Test sequences are evaluated using the proposed

[3] The number of exemplars selected from each video is heuristically determined using the "elbow" rule of thumb from the residual error curve of Ward’s distance criterion.

Exemplar-driven Bayesian Network (EDBN) classifier. Some sample extracted exemplars from both datasets are shown in Fig. 7.

Methods\Feature	PCA	LDA	NPE	NDMP
Random selection	63.68	64.81	65.68	66.09
LLE + *k*-means	68.54	70.43	65.36	73.66
Geodesic + HAC	73.69	71.30	66.07	76.75
LLE + HAC	66.18	71.20	71.70	86.90
LLE + STHAC-GF	74.89	76.94	80.68	95.33
LLE + STHAC-LP	81.91	87.21	90.84	94.52

Table 1. Average recognition rates (%) of different exemplar extraction methods on the Honda/UCSD

Table 1 demonstrates the superiority of our proposed spatio-temporal hierarchical agglomerative clustering (STHAC) technique in partitioning data into relevant clusters for exemplar extraction. The Global Fusion variant (STHAC-GF) performs slightly better than the Local Perturbation variant (STHAC-LP) when NDMP features are used. However, the STHAC-LP is notably more effective using the other tested features. This substantiates our hypothesis that the usage of the temporal dimension plays a vital role in processing continuous stream data such as video. Conventional spatial techniques such as HAC and *k*-means clustering do not utilize the temporal continuity to its good potential.

5.2 Feature representation methods

In our second evaluation, we examine the effectiveness of different feature representation methods in extracting meaningful features that can produce the best possible recognition results. Experiments were conducted on both Honda/UCSD and CMU MoBo datasets, with six different feature extraction techniques: PCA, LDA, LPP, MFA, NPE and our recently-proposed NDMP method. Training exemplars are selected using the STHAC-GF clustering method. Test sequences are evaluated using the proposed Exemplar-driven Bayesian Network (EDBN) classifier (EDBN).

Feature\Datasets	Honda/UCSD	CMU MoBo
PCA	74.89	85.32
LDA	78.04	89.79
LPP	75.00	91.41
MFA	58.88	80.68
NPE	81.62	91.80
NDMP	95.33	95.55

Table 2. Average recognition rates (%) of different feature representation methods on the evaluated datasets

The NDMP method is able to produce the best results among the tested state-of-art methods by extracting features that are discriminative, structure-constraining and neighborhood-preserving, as shown in Table 2. It is worth noticing that classic methods for still image-based recognition such as PCA and LDA struggle to find meaningful features within a complex face manifold. Our method also performed better than recent neighborhood-preserving methods such as MFA and NPE.

Overall, it can also be observed that evaluation on the CMU MoBo dataset yielded better recognition rates than the Honda/UCSD dataset. This is an expected outcome as the Honda/UCSD dataset contains much more challenging pose variations, head movements and illumination changes.

5.3 Classification methods

In our third evaluation, we fixed the methods used for the exemplar extraction/clustering and feature representation steps to our best options (STHAC-GF and NDMP respectively) while testing the recognition capability of an array of classification schemes. Our experiment implements the following schemes:

1. **Majority voting** (with nearest neighbor matching), where a vote is taken in each frame and the class with the majority vote is classified as the subject.
2. **Probabilistic voting**, where the likelihood probabilities of each frame (from Eq. 19) are combined cumulatively by simple sum rule. The class with the highest cumulative likelihoods is classified as the subject. The class with the largest sum of likelihoods is classified as the subject.
3. **Naive Bayes classifier** (from Eq. 15)
4. **Exemplar-driven Bayesian Network (EDBN) classifier** (from Eq. 17)

Classifier\Datasets	Honda/UCSD	CMU MoBo
Majority voting	78.68	92.67
Probabilistic voting	83.95	93.10
Naive Bayes	90.67	94.04
EDBN	95.33	95.55

Table 3. Average recognition rates (%) of different classification methods on the evaluated datasets

From the results in Table 3, it is clear that Bayesian classifiers performed better in the video-based setting where rapidly changing face pose and expression can easily cause recognition failure. It is no surprise that the superiority of the EDBN classifier is more pronounced on the Honda/UCSD dataset, where facial appearances change quickly. In Fig. 8, the posterior plot of a sample test sequence from the Honda/UCSD dataset demonstrates that the EDBN classifier is capable of arriving at the correct identity, even if initial frames were incorrectly classified.

Unlike conventional voting schemes, the major advantage of our framework is the ability to incorporate temporal dependencies between successive frames. Also, the establishment of causality between exemplars and their parent classes in our proposed Bayesian network slightly enhances recognition accuracy since exemplars that are more prominent are given more influence in the model and *vice versa*.

Our proposed framework can achieved the top recognition rate of more than 95% on both Honda/UCSD and CMU MoBo datasets using the best combination of spatio-temporal representations – exemplar extraction by clustering with STHAC-GF (Global Fusion scheme), representating the exemplar features with Neighborhood Discriminative Manifold Projection

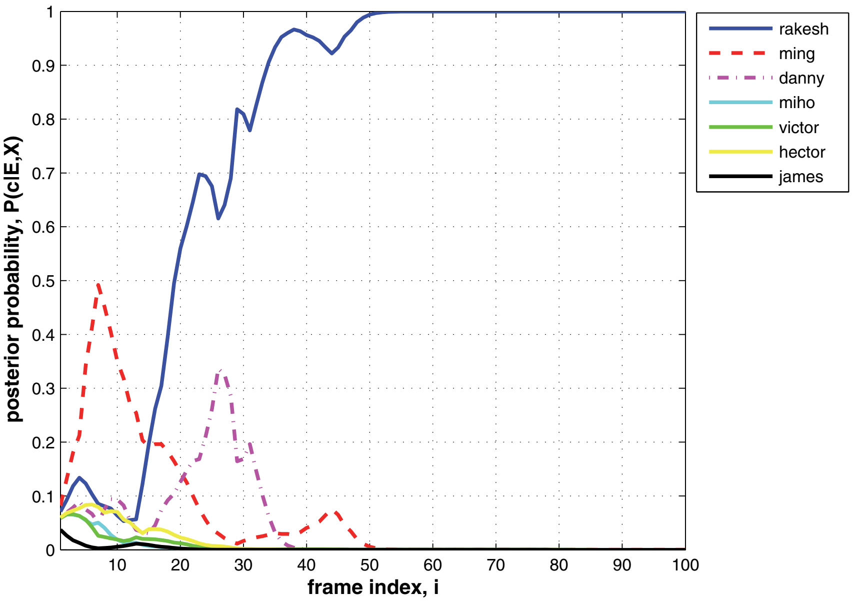


Fig. 8. Plot of posterior $P(c|E, X)$ versus frame index i of a sample 100-frame *rakesh* video subsequence from Honda/UCSD dataset. Posterior probabilities of the seven most probable subjects are shown in different colors. The subject *rakesh* (in blue line) is correctly identified for this test subsequence.

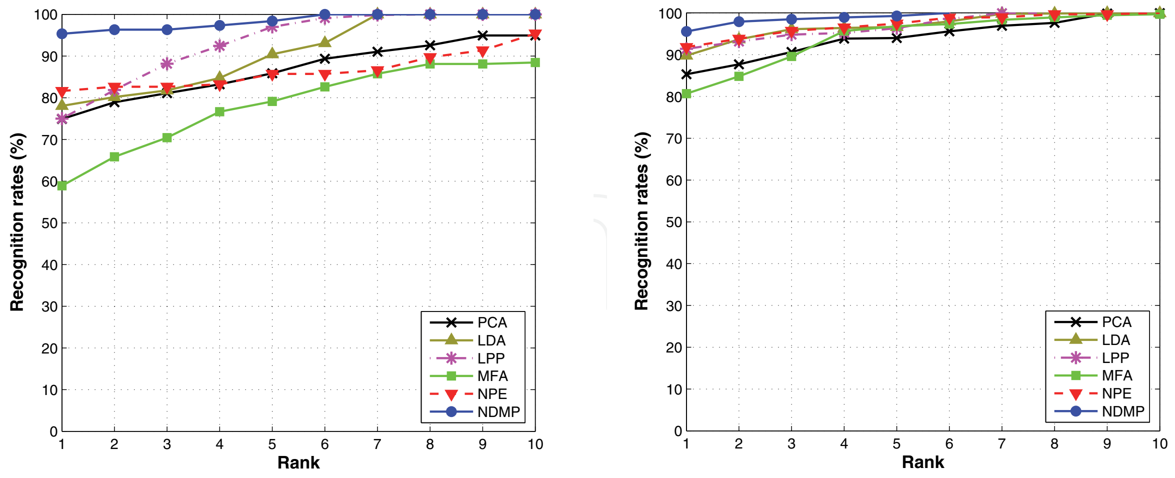
(NDMP) dimensionality reduction method, and classification with the Exemplar-Driven Bayesian Network (EDBN) classifier.

5.4 Rank-based identification

We further evaluate the reliability of the proposed spatio-temporal framework in a rank-based identification setting, by presenting its performance using a cumulative match curve (CMC). To accomodate this setting, we alter the MAP decision rule in Eq. 14 to take the top-n matches (insted of the maximum) based on the posterior probabilities. The subject in the test sequence is identified if it matches any class from the top-n matches. We gauge the rank-based performance of our proposed framework by comparing the effect of using different feature representation and clustering methods through their respective CMCs.

Figs. 9(a) and 9(b) shows the CMC of using different feature representation methods (PCA, LDA, LPP, MFA, NPE, NDMP) on both Honda/UCSD ad CMU MoBo datasets. At any given rank, the NDMP feature always yields better recognition rates compared to the other tested features. The NDMP feature is able to achieve a perfect recognition score of 100% as early as rank-6 on both datasets.

In the comparison of clustering methods in Fig. 10, the spatio-temporal HAC with Global Fusion (STHAC-GF) method is able to outperform other spatial methods such as k-means clustering and HAC in the rank-based identification setting. For both our methods, dimensionality reduction is performed using LLE. Interestingly, the Local Perturbation variant (STHAC-LP) is unable to improve its recognition rate even when the rank increases.



(a)

(b)

Fig. 9. Cumulative match curves of different feature representation methods on the (a) Honda/UCSD dataset, and (b) CMU MoBo dataset. On both datasets, it is clear that using NDMP features to represent exemplars consistently yield better recognition rates than using other features at any given rank. Assume STHAC-GF and EDBN classifier are used for clustering and classification respectively.

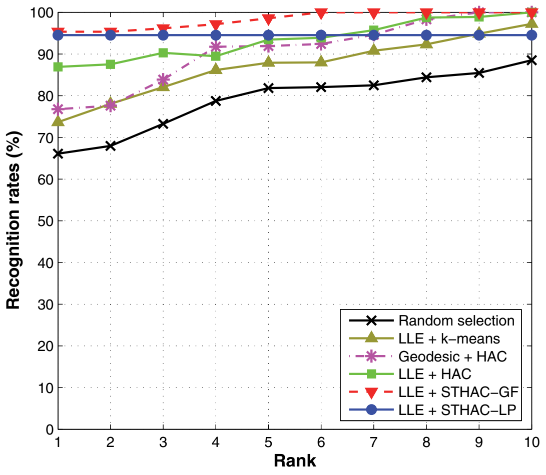


Fig. 10. Cumulative match curves of different clustering methods for exemplar extraction on the Honda/UCSD dataset. It can be observed that the proposed spatio-temporal HAC method with Global Fusion (STHAC-GF) is clearly superior to spatial clustering methods such as *k*-means and HAC. Assume NDMP and EDBN classifier are used for feature representation and classification respectively.

6. Future directions

While our proposed exemplar-based spatio-temporal framework has demonstrated promising results in our video-based face recognition evaluation, there are many avenues for future improvements in the following areas:

Clustering and Exemplar Extraction: While the straightforward STHAC-GF method is able produce good clusters for exemplar extraction, the STHAC-LP has shown unpredictable results especially in the rank-based identification setting in Section 5.4. It is possible that the optimum values for the λ_{sim} and λ_{dis} parameters in Eq. 9 have not been found, thus limiting its full potential for better performance. Future work can also extend beyond the exemplar-based approach by utilizing the already-extracted clusters as "local clusters" to create a dual exemplar-and-image set representation, with works by (Kim et al., 2007) and (Wang et al., 2008) being the primary motivations of this idea.

Feature Representation: The superior performance of the NDMP method compared to other existing state-of-art dimensionality reduction methods shows its potential usage for real-world applications such as biometric authentication, and video surveillance. Theoretically, the NDMP method can also be further generalized to a nonlinear form in kernel feature space, which may increase its robustness towards nonlinear manifolds.

Classification: The elegance of using a Bayesian network lies in its ability to encode causality between latent variables involved in the decision-making. While our EDBN model is the first step towards utilizing additional information to enhance classification, we can further formulate conditional probabilities between clusters, which can be easily simplified by computing canonical correlations.

Evaluation: With the luxury of temporal information in video, more comprehensive experiments can be further conducted to test the robustness of our proposed recognition framework in real-world scenarios, such as (1) having multiple identities in a same video sequence, (2) variable-length video sequences, and (3) degraded or low-quality video.

7. Conclusion

In this chapter, a novel exemplar-based spatio-temporal framework for video-based face recognition is presented. The paradigm of this framework involves identifying feasible spatio-temporal representations at various levels of a video-to-video recognition task. There are major contributions in all three stages of an exemplar-based approach – clustering for exemplar extraction, feature representation and classification. In the training stage, a new spatio-temporal hierarchical agglomerative clustering (STHAC) partitions data from each training video into clusters by exploiting additional temporal information between video frames. Face exemplars are then extracted as representatives of each cluster. In feature representation step, meaningful spatial features are extracted from both training and test data using the Neighborhood Discriminative Manifold Projection (NDMP) method. In the final classification task, a new exemplar-driven Bayesian network (EDBN) classifier which promotes temporal continuity between successive video frames is proposed to identify the subject in test video sequences. Extensive experiments conducted on two large video datasets demonstrated the effectiveness of the proposed framework and its associated novel methods compared to other existing state-of-art techniques. Furthermore, the reported results show promising potential for other real-world pattern recognition applications.

8. References

Arandjelovic, O., Shakhnarovich, G., Fisher, J., Cipolla, R. & Darrell, T. (2005). Face recognition with image sets using manifold density divergence, *Proceedings of IEEE Computer Vision and Pattern Recognition*, pp. 581–588.

Belhumeur, P., Hespanha, J. & Kriegman, D. (1997). Probabilistic recognition of human faces from video, *IEEE Trans on PAMI* 19: 711–720.

Cox, T. & Cox, M. (2001). *Multidimensional Scaling, 2nd ed.*, Chapman and Hall.

Duda, R., Hart, P. & Stork, D. (2000). *Pattern Classification*, John Wiley.

Fan, W., Wang, Y. & Tan, T. (2005). Video-based face recognition using bayesian inference model, *Proceedings of AVBPA*, Springer-Verlag, pp. 122–130.

Fan, W. & Yeung, D.-Y. (2006). Face recognition with image sets using hierarchically extracted exemplars from appearance manifolds, *Proceedings of IEEE Automatic Face and Gesture Recognition*, pp. 177–182.

Gross, R. (2004). Face databases, *in* S. Li & A. Jain (eds), *Handbook of Face Recognition*, Springer-Verlag, Berlin.

Gross, R. & Shi, J. (2001). The cmu motion of body (mobo) database, *Technical Report CMU CMU-RI-TR-01-18*, Robotics Institute, CMU.

Hadid, A. & Peitikäinen, M. (2004). From still image to video-based face recognition: An experimental analysis, *Proceedings of IEEE Automatic Face and Gesture Recognition*, pp. 813–818.

He, X., Cai, D., Yan, S. & H.J., Z. (2005). Neighborhood preserving embedding, *Proceedings of IEEE Int. Conf. on Computer Vision*, pp. 1208–1213.

He, X. & Niyogi, P. (2003). Locality preserving projections, *Proceedings of NIPS* 16: 153–160.

Kim, T., Kittler, J. & Cipolla, R. (2007). Discriminative learning and recognition of image set classes using canonical correlations, *IEEE Trans. PAMI* 29(6): 1005–1018.

Krüeger, V. & Zhou, S. (2002). Exemplar-based face recognition from video, *Proceedings of European Conference on Computer Vision*, pp. 732–746.

Lee, K., Ho, J., Yang, M. & Kriegman, D. (2005). Visual tracking and recognition using probabilistic appearance manifolds, *Computer Vision and Image Understanding* 99(3): 303–331.

Liu, W., Li, Z. & Tang, X. (2006). Spatio-temporal embedding for statistical face recognition from video, *Proceedings of Eur. Conf. on Computer Vision*, Springer-Verlag, pp. 374–388.

Liu, X. & Chen, T. (2003). Video-based face recognition using adaptive hidden markov models, *Proceedings of IEEE Computer Vision and Pattern Recognition*, pp. 340–345.

O'Toole, A., Roark, D. & Abdi, H. (2002). Recognizing moving faces: A psychological and neural synthesis, *Trends in Cognitive Sciences* 6(6): 261–266.

Ross, D., Lim, J., Lin, R.-S. & Yang, M.-H. (2008). Incremental learning for robust visual tracking, *Int. Journal of Computer Vision* 77(1): 125–141.

Roweis, S. & Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding, *Science* 290: 2323–2326.

See, J. & Ahmad Fauzi, M. (2011). Neighborhood discriminative manifold projection for face recognition in video, *Proceedings of Int. Conf. on Pattern Analysis and Intelligent Robotics (ICPAIR)*, to appear.

Shakhnarovich, G., Fisher, J. & Darrell, T. (2002). Face recognition from long-term observations, *Proceedings of European Conf. on Computer Vision*, pp. 851–868.

Tenenbaum, J., de Silva, V. & Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction, *Science* 290: 2319–2323.

Turk, M. & Pentland, A. (1991). Eigenfaces for recognition, *Journal of Cognitive Neuroscience* 3(1): 71–86.

Viola, P. & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features, *Proceedings of IEEE Computer Vision and Pattern Recognition*, pp. 511–518.

Wang, R., Shan, S., Chen, X. & Gao, W. (2008). Manifold-manifold distance with application to face recognition based on image set, *Proceedings of IEEE Computer Vision and Pattern Recognition*.

Webb, A. (2002). *Statistical Pattern Recognition, 2nd ed.*, John Wiley.

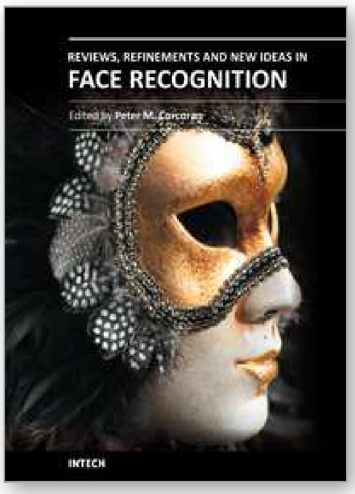
Yamaguchi, O., Fukui, K. & Maeda, K. (1998). Face recognition using temporal image sequence, *Proceedings of IEEE Automatic Face and Gesture Recognition*, pp. 318–323.

Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q. & Lin, S. (2007). Locality preserving projections, *IEEE Trans. PAMI* 29(1): 40–51.

Zhao, W., Chellappa, R., Phillips, P. & Rosenfeld, A. (2003). Face recognition: A literature survey, *ACM Computing Surveys* 35(4): 399–485.

Zhou, S. (2004). Face recognition using more than one still image: What is more, *Proceedings of 5th Chinese Conference on Biometric Recognition, (SINOBIOMETRICS)*, pp. 225–232.

Zhou, S., Krüeger, V. & Chellappa, R. (2003). Probabilistic recognition of human faces from video, *Computer Vision and Image Understanding* 91(1–2): 214–245.


REVIEWS, REFINEMENTS AND NEW IDEAS IN
FACE RECOGNITION
Edited by Peter M. Corcoran
INTECH