Exemplar Extraction using Spatio-Temporal Hierarchical Agglomerative Clustering for Face Recognition in Video

John See
Faculty of Information Technology
Multimedia University, Malaysia
johnsee@mmu.edu.my

Chikkannan Eswaran
Faculty of Information Technology
Multimedia University, Malaysia
eswaran@mmu.edu.my

Abstract

Many recent works have attempted to improve object recognition by exploiting temporal dynamics, an intrinsic property of video sequences. In this paper, a new spatio-temporal hierarchical agglomerative clustering (STHAC) method is proposed for automatic extraction of face exemplars for face recognition in video sequences. Two variants of STHAC are presented – a global variety that unifies spatial and temporal distances between points, and a local variety that introduces perturbation of distances based on a local spatio-temporal neighborhood criterion. Faces that are nearest to the cluster means are chosen as exemplars for the testing stage, where subjects in the test video sequences are recognized using a probabilistic-based classifier. Extensive evaluation on a face video database demonstrates the effectiveness of our proposed method, and the significance of incorporating temporal information for exemplar extraction.

1. Introduction

In the past few decades, work on face recognition algorithms has seen rapid developments mainly in the recognition of single still face images or "mugshot" images. Many algorithms [2, 8, 19] have since become core techniques in face recognition literature, and are able to achieve good success rates in most single still image scenarios. However, under unconstrained environments where significant face variations are unavoidable, they tend to perform poorly. In recent years, the abundance of video data has presented a rapidly growing area of research in video-based face recognition (VFR).

Notable psychological and neural studies [14] have shown that facial movement supports the face recognition process. Facial dynamic information is found to contribute greatly to recognition under degraded viewing conditions and 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.

Temporal dynamics can be exploited at various steps in a recognition process. Some common ways of utilizing temporal dynamics are simultaneous usage of spatial and temporal information through tracking and recognition, or modeling of transitions between face appearances. The sequential ordering of face images is essential for modeling temporal continuity. While its advantage lies in its ability to model and learn transitions between appearance variations and its usefulness in continuous stream processing, it may be unstable under real-world conditions where transitions between face variations such as pose, expression and illumination are likely to be demanding.

In a more general scenario, an image set containing multiple video frames can also be treated as an unordered set of independent observations, which are not necessarily acquired in order. A majority of methods represent the image set either as face subspaces or as *face exemplars*, a set of representative images that summarizes gallery video information. The main disadvantage is that classification performance is dependent on the effectiveness of clustering training data into meaningful subspaces or exemplars. Also, these methods appear to strip off temporal dynamics that is inherent in video sequences.

In this paper, our major contribution focuses on incorporating temporal dynamics through the formulation of a new spatio-temporal hierarchical agglomerative clustering method (STHAC). The locally linear embedding (LLE) algorithm [15], known for its capability to uncover nonlinear manifolds, is first applied to reduce the dimension of the training face images for each video. The proposed STHAC algorithm is then applied to the face images in LLE-space to automatically extract suitable exemplars for each training video. We present two variants of STHAC, designed to utilize temporal information at the global and local levels. A discriminative manifold learning algorithm is then applied to extract meaningful features in a lower-dimensional space for the classification task, where subjects in the test video sequences are identified using a probabilistic-based classifier. Extensive evaluations on a face video database

testify of the importance of temporal information in exemplar selection.

The rest of this paper is organized as follows. Section 2 discusses some previous related work. Section 3 describes the proposed method and its implementation in detail while Section 4 briefly elaborate on the feature representation and classification tasks in our setup. Experimental results and discussions are presented in Section 5, while Section 6 concludes this paper.

2. Related Work

By broadly categorizing based on input data, various recent methods in video-based face recognition (VFR) can be divided into methods that are based on modeling temporal continuity [12, 23], and methods that are based on multiple independent observations [1, 4, 5, 6, 11, 17, 21, 22]. Since this work focuses on unordered face image sets, the former category of works will not be elaborated.

Methods based on multiple independent observations can be loosely categorized into subspace-based and exemplar-based methods. Subspace-based methods represent entire sets of images as subspaces or manifolds, and are largely parametric in nature. Typically, these methods represent image sets by parametric distribution functions (PDF), and the similarity between two distributions is measured. Both the mutual subspace method (MSM) by Yamaguchi et al. [22] and a probabilistic modeling method by Shakhnarovich et al. [17] utilize a single Gaussian distribution on face space while Arandjelovic et al. [1] further extended this method using Gaussian mixture models. While it is known that these methods suffer from the difficulty of parameter estimation, they also easily fail when training and test sets have weak statistical relationships [3]. In a work on image sets, Kim et al. [10] bypass the need of using PDFs by computing similarity between subspaces using canonical correlations.

The use of exemplars offers an alternative model-free method of representing large image sets. This non-parametric setting has attracted much attention in many recent VFR works. Typically in many of these works, the number of exemplars extracted from the training stage is fixed to facilitate a systematic formulation in the subsequent feature extraction and classification stages.

Krüeger and Zhou [11] proposed a method of selecting exemplars from training face videos using radial basis function network. Hadid and Pietikäinen [6] proposed a view-based scheme which embeds the face manifold using LLE algorithm [15]. The data in embedded space is then partitioned using *k*-means clustering to extract cluster centers as exemplars. Fan et al. [4] also used the similar configuration except that classification is performed using a Bayesian inference model to exploit temporal dynamics. In another work, Fan and Yeung [5] reported better recognition performance by extracting exemplars using hierarchical agglomerative clustering (HAC) based on geodesic distances instead of embedded space. Wang and Chen [21] explored the use of a top-down hierarchical divisive clustering (HDC) for extracting exemplars, in which they claimed to be computationally more efficient than HAC. However, all these methods performed clustering in face space without considering temporal relationships between images in a sequence.

Nevertheless, there are a few interesting works relating to the usage of spatial-temporal data. Liu et al. [13] formulated a spatio-temporal embedding algorithm by modifying a classical *k*-means algorithm with a spatio-temporal objective function. A spatio-temporal extension to Isomap embedding [9] provided some insights into temporal relationships in local neighborhoods.

In this paper, we propose a novel spatio-temporal hierarchical agglomerative clustering (STHAC) method to automatically extract exemplars for face recognition in video. STHAC comes in two variants – a global variant called Global Spatio-Temporal Fusion (GSTF) that blends the contribution of spatial and temporal distances between points, and a local variant called Local Spatio-Temporal Perturbation (LSTP) that perturbs these distances based on a local spatio-temporal neighborhood criterion. Our recognition framework is similar to [5, 6] where exemplars are used to represent the face manifold of each training video for subsequent classification of test videos.

3. Proposed Method

3.1. Exemplar-based Representation

Training videos are summarized using an exemplar-based representation, similar to the problem setting described in [6]. For general notation, given a training video face sequence

$$V_c = \left\{ v_1^c, v_2^c, \ \ldots \ , v_{N_c}^c \right\} \tag{1}$$

of N_c image frames in $\mathbb{R}^D$ belonging to one of C subject classes $\left\{ c \mid c \in \{1, \ldots, C\} \right\}$, we want to select its associated exemplar set

$$E_c = \left\{ e_1^c, e_2^c, \ \ldots \ , e_{M_c}^c \right\}, \qquad E_c \subseteq V_c \tag{2}$$

where the number of exemplars extracted, $M_c << N_c$. In cases where more than one training video of a particular class is used, image frames from all training videos are aggregated to produce M_c number of exemplars.

To cope with the high dimensionality of nonlinear manifolds in each video, a suitable dimensionality reduction method is necessary to map the original face data to a low-dimensional embedded space. Nonlinear dimensionality reduction techniques such as Isomap [18]

and Locally Linear Embedding (LLE) [15] can effectively discover an underlying embedding of a nonlinear manifold unlike classical methods such as PCA [19] and MDS [3] that tend to overestimate the intrinsic dimensionality of face data sets. In our work, LLE is our choice of method as to its able to preserve local neighborhood relationships.

3.2. Spatio-Temporal Hierarchical Agglomerative Clustering (STHAC) and its Variants

From the embedded space, clustering is performed to extract K number of clusters that group together faces of similar appearances. In many previous works, k-means clustering [4] is the primary choice of assigning samples into different clusters due to its simplicity in implementation. However, it has some glaring 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.

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. The *agglomerative* approach works from "bottom up" by grouping smaller clusters into larger ones, as described in the following procedure:

1. Initialize each data point as a singleton cluster C_i. At the start, there are N_c clusters.
2. Find the nearest pair of clusters, C_i and C_j according to a certain distance measure between clusters. Commonly used measures are such as single-link, complete-link, group-average-link and Ward's criterion [3]. Merge the two nearest clusters to form a new cluster.
3. Continue 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 constructed dendrogram. This poses an advantage over the k-means method, which requires the number of clusters to be predetermined. Structurally, the HAC method solves the limitations of the k-means method, as initial selection of cluster seeds is not required while hierarchical merging ensures that it is not easily trapped in local minima [5].

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. We present two varieties of fusion schemes, one at the global structural level, and one at a local neighborhood level.

Prior to the schemes, we first introduce the distance measures. Spatial distance is measured by simple Euclidean distance between points. *Temporal distance* is measured by the time spanned between two frame occurrences (v_i and v_j) in a video sequence,

$$d_T(v_i, v_j) = |\, t_{v_i} - t_{v_j} \,| \tag{3}$$

where t is a discretized unit time. The temporal distance is sufficiently intuitive to quantify temporal relationships across sequentially ordered data samples. Firstly, the matrices containing pairwise spatial Euclidean distances $\mathbf{D}_S(v_i, v_j)$ and temporal distances $\mathbf{D}_T(v_i, v_j)$ between all samples, are computed and normalized.

Global Spatio-Temporal Fusion (GSTF). The first is a global variant that 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

$$\mathbf{D}_{ST,g} = (p_{max} - \alpha)\mathbf{D}_S + (\alpha + p_{min})\mathbf{D}_T, \quad 0 \le \alpha \le 1. \tag{4}$$

Local Spatio-Temporal Perturbation (LSTP). We also propose a local variant that perturbs spatial and temporal distances based on local spatio-temporal neighborhood relationships between a point and its neighbors. For each point v_i, a temporal window segment, $S_{v_i} = \{v_{i-w}, \ldots, v_i, \ldots, v_{i+w}\}$ of length $(2w+1)$ is defined as its temporal neighborhood. Meanwhile, the spatial neighborhood of point v_i, $Q_{v_i} = \{v_1, v_2, \ldots, v_k\}$ is simply a set containing k-nearest neighbors of v_i computed by Euclidean distance. A point v_j is identified as a *common spatio-temporal neighbor* (CSTN) of point v_i if it belongs to both spatial and temporal neighborhood point sets. Hence, the criterion defining the CSTN set of point v_i is

$$\mathrm{CSTN}_{v_i} = Q_{v_i} \cap S_{v_i}. \tag{5}$$

To perform perturbation between v_i and its CSTN and non-CSTN sets, we define a perturbation affinity matrix as

$$\mathbf{P}_{ij} = \begin{cases} 1 - \lambda_{sim}, & if\ v_j \in \mathrm{CSTN}_{v_i} \\ 1 + \lambda_{dis}, & if\ v_j \in (S_i \setminus \mathrm{CSTN}_{v_i}) \\ 1 & otherwise \end{cases} \tag{6}$$

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

$$\mathbf{D}_{ST,\ell} = \mathbf{P}_{ij}(\mathbf{D}_S + \mathbf{D}_T). \tag{7}$$

In our work, we set p_{min} = 0.5, p_{max} = 0.5 for GSTF, and k = 7, w = 7 for LSTP, all of which are suitable values determined through experiments.

The linkage criterion used in our experiments for merging clusters is Ward's distance criterion,

$$d_{ward}(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} \left\| m_i - m_j \right\|^2 \quad (8)$$

where m_i and m_j are means of cluster i and j respectively, while n_i and n_j are the number of points in their clusters. Justification of our choice is given in the Section 3.3. In the final step of exemplar extraction, face images that are nearest to each cluster mean are chosen as exemplars.

3.3. Heuristic Selection of Number of Clusters

In a true sense, the selection of "correct" or "optimum" number of clusters has little theoretical foundation and meaning. Very often, suitable heuristics are devised to provide good choices. In our case, the cluster merging cost of Ward's criterion from Equation (8) takes the form of a sum-of-squares term, which is known to provide a good statistical measure for residual error [3].

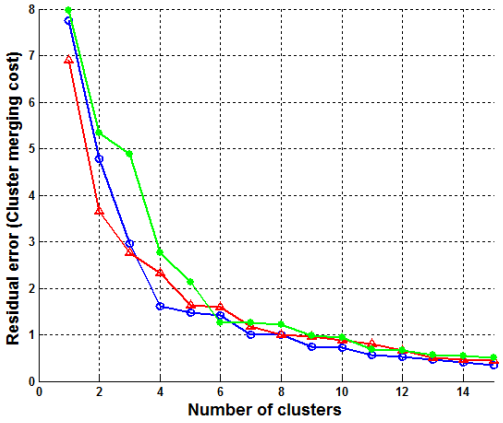

Figure 1: Residual error of three different training videos (drawn with different colors) partitioned with different number of clusters. The "elbow" of the curve is approximately at 5 – 9.

Figure 1 shows the residual error curve of three different training videos using different number of clusters. A simple but effective heuristic is to find the *elbow* of the curve, at which the curve stops decreasing significantly with the increase of clusters. To facilitate an unbiased representation of features for classification, the number of clusters for each class is fixed during training.

4. Face Recognition in Video

In this section, we briefly describe the subsequent tasks employed in our video-based face recognition setup. With the extraction of exemplars, a *video-to-video* recognition setting (where both training and test data consist of video sequences) is reduced to a *still-to-video* setting – where training exemplar faces are used as a gallery set for recognizing entire test video sequences. Methods appropriate for representing exemplar features and classification of test videos are described here.

4.1. Feature Representation

Traditional linear dimensionality reduction methods such as PCA [19] and LDA [2] have been widely used to great effect in characterizing data in smooth and well-sampled manifolds. More recently, manifold learning methods such as Locality Preserving Projections (LPP) [8] and Neighborhood Preserving Embedding (NPE) [7] are able to effectively derive optimal linear approximations to a low-dimensional embedding of nonlinear manifolds. NPE in particular, has an attractive neighborhood-preserving property due to its formulation based on LLE.

For better feature representation, we propose using a supervised discriminative variant of the NPE called Neighborhood Discriminative Manifold Projection (NDMP) [16], which seeks to learn an optimal low-dimensional projection by considering both intra-class and inter-class reconstruction weights. Global structural and local neighborhood constraints are imposed in a constrained optimization problem, which can be solved as a generalized eigenvalue problem:

$$(\mathbf{X}\mathbf{M}_{\text{intra}}\mathbf{X}^T)\mathbf{A} = \lambda(\mathbf{X}\mathbf{M}_{\text{inter}}\mathbf{X}^T + \mathbf{X}\mathbf{X}^T)\mathbf{A} \quad (9)$$

where $\mathbf{X}$ denotes exemplar face data in $\mathbb{R}^D$, while $\mathbf{M}_{\text{intra}}$ and $\mathbf{M}_{\text{inter}}$ are the intra-class and inter-class orthogonal weight matrices respectively. New test samples $\mathbf{X'}$ can be projected to embedded space in $\mathbb{R}^r$, by the linear transformation $\mathbf{Y'} = \mathbf{A}^T\mathbf{X'}$ where $r << D$. A more detailed theoretical formulation of NDMP can be found in [16].

4.2. Classification

In video-based classification, a majority voting scheme is typically used to decide on the identity of the subject in a video sequence by performing a majority vote on the subject identities in very frame. Similar to [16], we use a probabilistic-based strategy in the form of a Bayes *maximum a posteriori* (MAP) classifier. The subject identity in a test video V' is evaluated as

$$\hat{c} = \arg\max_C P(c \mid V') = \arg\max_C \frac{P(c)P(V' \mid c)}{\sum_{c'}^{C} P(V' \mid c')P(c')} \quad (10)$$

where $P(c \mid V')$ is the posterior probability of the Bayes classifier and $P(c)$ is the class prior. Assume that observations are *i.i.d.,* the class likelihood

$$P(V' \mid c) = \prod_{i=1}^{N} P(v'_i \mid c) \quad (11)$$

can be estimated using a normalized class probability score, computed for each *i*-th frame as

$$P(v'_i \mid c) = \frac{\sum_{j=1}^{M} 1/d_{L2}(v'_i, e_j^c)}{\sum_{c=1}^{C}\sum_{j=1}^{M} 1/d_{L2}(v'_i, e_j^c)} \quad (12)$$

where $d_{L2}(v'_i, e^c)$ is the *L2*-norm distance between test frame v'_i and training exemplar e^c. A probabilistic voting scheme can also be adopted by combining class probability scores across frames. Probabilistic strategies tend to produce more reliable and robust measures compared to simple majority voting method [4, 6].

5. Experiments and Discussion

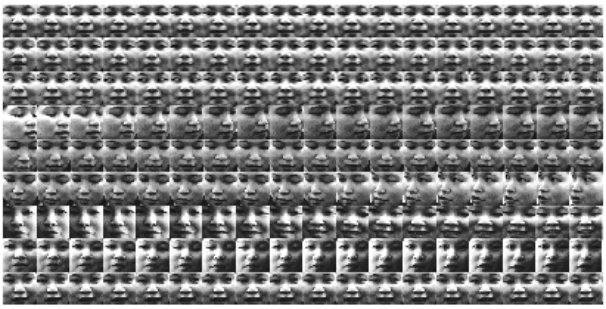

Figure 2: Sample faces of a same subject taken from the Honda/UCSD database.

We have performed extensive experiments on the Honda/UCSD data set (see Figure 2 for sample faces) [12]. We consider their first data set, which has 59 video sequences of 20 different people (each person has at least 2 videos). Each video contains about 300-600 frames, consisting of large pose and expression variations, and significant head rotations. Faces were extracted using the Viola-Jones cascaded face detector [20] and resized to 32×32 pixel grayscale images, followed by histogram equalization to remove illumination effects.

For each subject, one video sequence is used for training, and the remaining video sequences for testing. To ensure extensive evaluation on a single data set, we construct our test set by randomly sampling 50 video sub-sequences consisting of 200 frames from each testing video sequence. Based on overall observation of residual error curves, we select 7 exemplars per subject. Figure 3 shows some sample face exemplar sets. STHAC tuning parameters for GSTF and LSTP were set at $\alpha = 0.75$ and $\lambda = 0.2$, respectively.

For a comprehensive evaluation, we build the face recognition framework using four different dimensionality reduction algorithms – PCA, LDA, NPE and NDMP. The optimal number of feature dimensions for these methods were determined empirically. To focus our attention on the exemplar extraction methods, we perform experiments on the following exemplar-based methods for comparisons:

1. Random exemplar selection
2. LLE + *k*-means clustering [6]
3. Geodesic distances + HAC [5]
4. LLE + HAC
5. LLE + STHAC (both global and local variants)

Figure 3: Sample face exemplar sets extracted from training video sequences for three different subjects, showing the most representative faces selected using Global STHAC method.

Table 1: Average recognition rates (%) of various exemplar extraction methods using the MAP Bayes classifier

Method	**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	70.54	86.15
LLE + Global STHAC	74.89	76.88	80.10	95.04
LLE + Local STHAC	81.91	87.21	90.84	94.52

The summary of recognition rates in Table 1 clearly indicates better selection of exemplars using spatio-temporal approaches (Local STHAC, Global STHAC) compared to standard spatial approaches (HAC and *k*-means). The same performance can be observed across different dimensionality reduction methods used. This shows the importance of incorporating temporal information to enhance clustering of face video sequences. Among the two STHAC variants presented, the local variant (LSTP) slightly outperforms the global variant (GSTF). This can be attributed to the ability of CSTN in discovering the spatio-temporal relationships between data points within the local neighborhood. Although not the main focus here, it is worth mentioning that the NDMP dimensionality reduction method clearly outperforms the PCA, LDA and NPE methods, owing much to its elegant neighborhood discriminative formulation.

To further evaluate the performance of the methods in a rank-based identification setting, we directly adopt the class probability score from equation (12), cumulatively combined by simple sum rule. A vote is then taken in each frame and the class with the majority vote is classified as the subject of the test sequence. The cumulative match curve (CMC) in Figure 4 reinforces the effectiveness of the STHAC methods over spatial clustering methods in exemplar selection, especially across the top-most ranks.

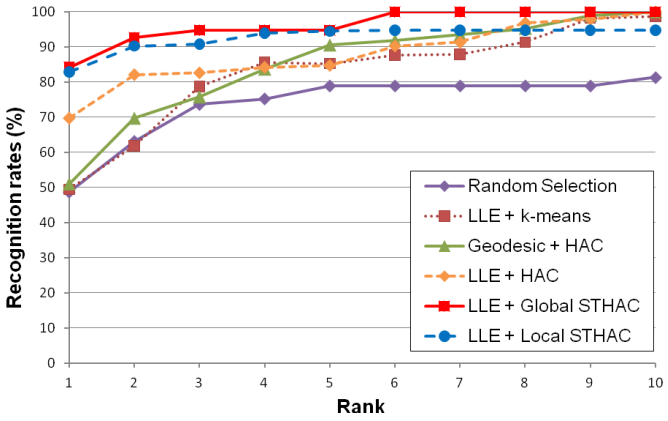

Figure 4: Cumulative match curve (CMC) of various exemplar extraction methods. Probabilistic voting is applied in the rank-based identification of face video sequences with NDMP-space features. Lines are shown here with different colors/patterns.

6. Conclusion

This paper presents a novel method of automatically selecting face exemplars from video sequences for face recognition. We proposed a spatio-temporal hierarchical agglomerative clustering (STHAC) algorithm that incorporates temporal information into the conventional HAC algorithm. Two variants were introduced, a global spatio-temporal blending method and a local spatio-temporal distance perturbation method. Faces nearest to the cluster means are selected as exemplars. The original data is first projected to a low-dimensional LLE embedded space before exemplar extraction. Finally, the subjects in the test video sequences are recognized using a probabilistic-based classifier. Promising experimental results on a face video database were obtained, showing the effectiveness of STHAC compared to other methods.

In future work, our focus is on finding good values for STHAC parameters to fully enhance the potential of the proposed method. Our work can also be extended beyond the use of exemplars only, by utilizing the extracted clusters for image set representations.

References

[1] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *Proc. of the CVPR,* volume 1, pages 581-588, 2005.

[2] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE PAMI*, 19:711-720, 1997.

[3] R. Duda, P. Hart and D. Stork. *Pattern Classification,* Wiley, 2000.

[4] W. Fan, Y. Wang and T. Tan. Video-based face recognition using Bayesian inference model. In *Proc. of AVBPA,* pages 122-130, 2005.

[5] W. Fan and D.-Y. Yeung. Face recognition with image sets using hierarchically extracted exemplars from appearance manifolds, In *Proc. of FGR,* pages 177-182, 2006.

[6] A. Hadid and M. Pietikäinen. From still image to video-based face recognition: An experimental analysis. In *Proc. of FGR,* pages 813-818, 2004.

[7] X. He, D. Cai, S. Yan and H. J. Zhang. Neighborhood preserving embedding. In *ICCV,* pages 1208-1213, 2005.

[8] X. F. He and P. Niyogi. Locality preserving projections. In *Proc. of NIPS*, 2003.

[9] O. C. Jenkins and M. J. Matarić. A spatio-temporal extension to Isomap nonlinear dimension reduction. In *Proc. of ICML,* pages 441-448, 2004.

[10] T. K. Kim, J. Kittler, R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE PAMI,* 29(6):1005-1018, 2007.

[11] V. Krüeger and S. Zhou. Exemplar-based face recognition from video. In *Proc. of ECCV,* pages 361-365, 2005.

[12] K. C. Lee, J. Ho, M. H. Yang and D. Kriegman. Visual tracking and recognition using probabilistic appearance manifolds. *CVIU,* 99(3):303-331, 2005.

[13] W. Liu, Z. Li and X. Tang. Spatio-temporal embedding for statistical face recognition from video. In *Proc. of ECCV,* pages 374-388, 2006.

[14] A. O'Toole, D. Roark, and H. Abdi. Recognizing moving faces: A psychological and neural synthesis. *Journal of Vision,* 2(7):604, 2002.

[15] S. T. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science,* 2900: 2323-2326, December 2000.

[16] J. See, M. F. Ahmad Fauzi. Neighborhood discriminative manifold projection for face recognition in video. In *Int. Conf. on Pattern Analysis and Intelligent Robotics*, 2011.

[17] G. Shakhnarovich, J. Fisher, and T. Darrell. Face recognition from long-term observations. In *Proc. of the CVPR,* volume 1, pages 511-518, 2001.

[18] J.B. Tenenbaum, V. de Silva and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319-2323, December 2000.

[19] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience,* 3(1):71-86, 1991.

[20] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of CVPR,* volume 1, pages 511-518, 2001.

[21] R. Wang and X. Chen. Manifold discriminant analysis. In *Proc. of CVPR,* pages 429-436, 2009.

[22] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *Proc. of FGR,* pages 318-323, 1998.

[23] S. Zhou, V. Krüeger and R. Chellappa. Recognition of human faces from video. In *CVIU,* 91(1):214-245, 2003.