Face Recognition with Image Sets Using Hierarchically Extracted Exemplars from Appearance Manifolds

Wei Fan, Dit-Yan Yeung
Department of Computer Science
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
{fwkevin,dyyeung}@cs.ust.hk

Abstract

An unsupervised nonparametric approach is proposed to automatically extract representative face samples (exemplars) from a video sequence or an image set for multiple-shot face recognition. Motivated by a nonlinear dimensionality reduction algorithm called Isomap, we use local neighborhood information to approximate the geodesic distances between face images. A hierarchical agglomerative clustering (HAC) algorithm is then applied to group similar faces together based on the estimated geodesic distances which approximate their locations on the appearance manifold. We define the exemplars as cluster centers for template matching at the subsequent testing stage. The final recognition is the outcome of a majority voting scheme which combines the decisions from all the individual frames in the test set. Experimental results on a 40-subject video database demonstrate the effectiveness and flexibility of our proposed method.

1. Introduction

The majority of the state-of-the-art face recognition algorithms [17] put emphasis on scenarios based on single-shot still images. Although these dominating approaches [14, 2, 6] have achieved a certain level of success under restrictive conditions (such as mugshot matching), they often fail to yield satisfactory performance when confronted with significant facial variations. A typical scenario is recognition in the context of visual surveillance and multimedia retrieval applications, where the appearance of a face may bear large pose, illumination and expression variations. In this paper, we propose a face recognition method that is especially tolerant of these factors, using hierarchically extracted exemplars or templates from multiple shots residing on a nonlinear face manifold.

Recently, there has been a significant trend in performing automatic face recognition based on multiple images [18, 9, 7, 16, 12, 8]. The underlying assumption is that a sequence or set of images can provide information about the variability in the appearance of the face that can be utilized to achieve more robust recognition. Two main strategies have been exploited by algorithms along this direction. Methods based on image sequences take consecutive video frames as input and then utilize visual dynamics or temporal consistency to enhance the recognition performance. On the other hand, methods based on image sets assume independence between face images in a set. This relaxed assumption allows them to be applicable even to sparse or unordered observations, rather than image sequences.

Some recent psychological and neural studies [10] indicate that the information for identifying a human face can be found both in the invariant structure of features and in idiosyncratic movements and gestures. However, most works in the computer vision literature simply combine the two cues in an *ad hoc* manner. Furthermore, they often assume continuous extraction of face regions in each video frame, which is a formidable challenge even to some state-of-the-art face detectors. This may explain why video sequences in FRVT 2002 did not improve the performance of the recognition task.

We believe the most essential features for face recognition still lie on the static facial configurations which are more stable and discriminating than the dynamic information. One possible approach for achieving video-based face recognition is to extract representative exemplars covering the dominant structural variability in the face appearance, and categorization for a single test sample can be readily performed via certain baseline method (e.g. PCA [14], LDA [2]) or simply through template matching. To determine the identity of a test set, we use the majority voting scheme over all frames in the set. The main contribution of this paper is to introduce a method for automatic acquisition of

representative exemplars from the training set. Motivated by the well-known nonlinear dimensionality reduction algorithm called Isomap [13], we use local neighborhood information to approximate the geodesic distances between face images, i.e., distances along the face manifold from which the face images are sampled. Using the dissimilarity matrix based on geodesic distances, we then apply a hierarchical agglomerative clustering (HAC) algorithm [4] to group similar face images according to their approximate locations on the appearance manifold and define the exemplars as cluster centers. Experimental results conducted on a medium-scale video database well support our assumptions and show high superiority of the newly developed method to its traditional counterparts based on image sets.

This paper is organized as follows. In section 2, we give a brief introduction to some recent video-based face recognition algorithms based on image sequences or image sets. Section 3 describes the problem setting of the proposed method and its implementation in detail. Experimental results and some further discussions are presented in section 4, followed by a conclusion in section 5.

2. Previous Work

An extensive survey of the face recognition literature can be found in [17]. In this section, we only briefly review some face recognition methods that are based on a set or a sequence of images.

Image sequence-based approaches use both spatial and temporal information simultaneously to enhance the recognition performance. In [18], Zhou *et al.* characterize the kinematics and identity using a motion vector and an identity variable, respectively, in a probabilistic framework. The sequential importance sampling (SIS) algorithm is developed to estimate the joint posterior distribution, and marginalization over the motion vector yields a robust estimate of the posterior distribution of the identity variable. Recently, hidden Markov models (HMM) [9] and probabilistic appearance manifolds [7] are both used to learn the transition probabilities among several viewing states embedded in the observation space.

Although facial dynamics, if properly modeled, are tolerate of appearance variations induced by changes in head pose orientation and expressions, they are not stable and discriminating enough for a real-world recognition system. In this paper, we are interested in a general scenario, in which the set of images may come from independent observations over an extended period of time and under different viewing conditions. It is often difficult to exploit temporal coherence in such isolated inputs. Two previous approaches to this problem are the mutual subspace method (MSM) [16] and the probabilistic modeling method in [12]. These methods propose rather simplistic modeling of face pattern variations, essentially representing the face space as a single linear subspace with a Gaussian density. This Gaussian assumption limits the kind of variability along the input sequence which can be effectively tolerated. The manifold density divergence method [1] takes a step further in modeling these densities as Gaussian mixture models (GMM) defined on low-dimensional nonlinear manifolds embedded in the image space, and evaluates the similarity between the estimated densities via the Kullback-Leibler divergence. Apart from these parametric approaches, Hadid *et al.* [5] first represent the face manifold using the locally linear embedding (LLE) [11] algorithm and then perform k-means clustering, setting the face models as the cluster centers.

Our work bears some resemblance to [5] in the sense that both methods utilize selected exemplars as local manifold models for video-based face recognition. However, in this paper, we do not explicitly calculate the low-dimensional embedding of all the training images to avoid the loss of information and the computational bottleneck with respect to the eigendecomposition problem in manifold learning. Essentially, only the geodesic distances between face images are estimated for the construction of the similarity matrix required by the subsequent hierarchical clustering algorithm. Another motivation is the benefit of the global embedding approach (Isomap [13]) in that it tends to give a more faithful representation of the global structure of the data, as opposed to local approaches (e.g., LLE [11], Laplacian Eigenmap [3]) which attempt to preserve only the local geometry of the data. In the following section, we describe the problem setting of the proposed method and its implementation in detail.

3. Proposed Method

3.1. Problem Setting

The problem that we focus on in this paper can be formulated as follows. Given a training face image sequence

$$G = \{g_1, g_2, \cdots, g_N\}, \tag{1}$$

we are interested in selecting the most representative samples

$$E = \{e_1, e_2, \cdots, e_K\}, \qquad K \ll N, \tag{2}$$

so that they can be considered as models for appearance-based face recognition. The desirable samples are those which summarize the content of the face sequence G. In other words, they should capture the within-class variability due to changes in illumination, pose, facial expression and other factors. With these selected samples as models, there is no need to compare all pairs of images exhaustively.

3.2. Geodesic Distance Estimation

In typical appearance-based methods, $m \times n$ face images are often represented by points in an mn-dimensional space. However, coherent structure in the appearance of human faces leads to strong correlation between them, generating observations that lie on or close to a low-dimensional manifold. When the face images are extracted from video sequences, it is reasonable to assume that the manifold is smooth and well-sampled. Figure 1 shows the first three principal components of a training set and a test set of images for one moving face, which are automatically detected from two short video clips. Notice the clear overlap between the two sets, and that the perceptually meaningful structure of the nonlinear face manifold has very limited independent degrees of freedom.

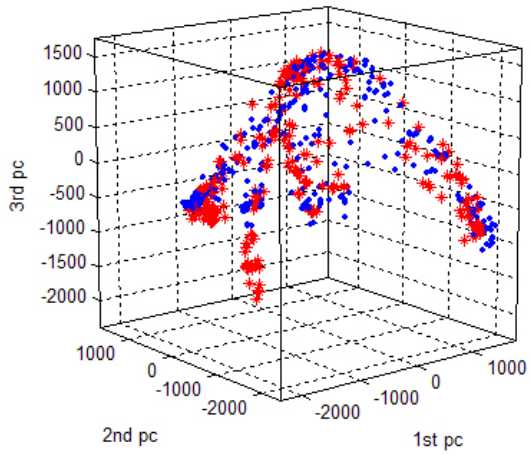

Figure 1. The first three principal components of a training set (blue dots) and a test set (red stars) of images for one moving face, which are automatically detected from two short video clips.

Unlike traditional linear dimensionality reduction techniques (e.g., PCA [14] and LDA [2]) which often overestimate the true degrees of freedom of the face data set, recently proposed nonlinear dimensionality reduction methods (e.g., Isomap [13] and LLE [11]) can effectively discover an underlying low-dimensional embedding of the manifold. As Euclidean distance between data points in the high-dimensional input space cannot reflect the true low-dimensional geometry of the manifold, we use the geodesic ("shortest path") distance instead, which is a key idea used in Isomap.

Given the Euclidean distances $d_X(i,j)$ between point pairs for n points (corresponding to n training face images of a particular person) in the input image space X.

1. Construct a neighborhood graph:
- Define a graph G over all n data points by connecting points i and j if their distance $d_X(i,j)$ is closer than ϵ (ϵ-Isomap) or if i is one of the k nearest neighbors of j or vice versa (k-Isomap).
- Set the edge lengths equal to $d_X(i,j)$.

2. Compute the shortest paths:
- Initialize $d_G(i,j) = d_X(i,j)$ if i and j are linked by an edge and $d_G(i,j) = \infty$ otherwise.
- For each $k = 1, \cdots, n$, replace all entries $d_G(i,j)$ by $\min(d_G(i,j), d_G(i,k) + d_G(k,j))$. ($D_G = [d_G(i,j)]$ contains the shortest-path distances between all point pairs in G.)[1]

The underlying assumption of step 2 is that, for neighboring points, Euclidean distance in the input space provides a good approximation of the geodesic distance, whilst for faraway points, the geodesic distance can be approximated by adding up a sequence of short hops between neighboring points based on Euclidean distance. Here we do not perform the final step of Isomap which constructs a low-dimensional embedding of the original data by performing multidimensional scaling (MDS) based on the matrix of geodesic distances, since it requires performing eigendecomposition and it can lead to some loss of information. One side product of this embedding step is a reasonable estimation of the intrinsic dimensionality of the face manifold using the residual variance. Figure 2 illustrates the difference between PCA and Isomap in estimating the intrinsic dimensionality of a training image set (Figure 6 or blue dots in Figure 1) corresponding to one person arbitrarily rotating his head in all directions. Notice that Isomap successfully discovers the three degrees of freedom in the rigid rotation by looking for the 'elbow' of the curve while PCA tends to overestimate it.

Having estimated the geodesic distances between face images in the training set, an affinity matrix can be easily computed with certain transformation, e.g., Gaussian radial basis function (RBF) kernel $W(i,j) = e^{-d_G(x_i,x_j)/2\sigma^2}$, with σ being a free parameter. One may then use this affinity matrix to perform spectral clustering by performing eigendecomposition on W to find meaningful clusters. In this paper, however, we apply HAC on the geodesic distance matrix directly without performing any transformation and eigendecomposition.

3.3. Hierarchical Agglomerative Clustering

Hierarchical clustering is a way to investigate grouping in the data set, simultaneously over a variety of scales, by

[1] This procedure, known as Floyd's algorithm, requires $O(n^3)$ operations. More efficient alternatives exist, such as Dijkstra's algorithm (with Fibonacci heaps) which requires $O(kn^2 \log n)$ operations.

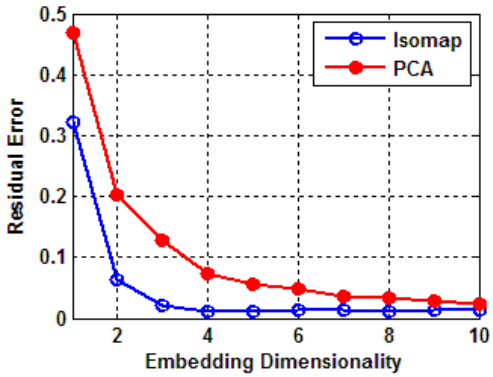

Figure 2. The residual variance of PCA (red line) and Isomap (blue line) on a training image set (Figure 6 or blue dots in Figure 1) corresponding to one person arbitrarily rotating his head in all directions.

creating a cluster tree called dendrogram. The tree does not represent a single set of clusters, but rather a multi-level hierarchy where clusters at one level are joined together as clusters at the next higher level. This allows one to decide what level or scale of clustering is most appropriate to the specific application at hand. Unlike k-means clustering which is sensitive to the initial seeds and may get trapped in local minima, HAC algorithm is more stable to the input data set.

To perform hierarchical cluster analysis on a data set using a certain distance measure, one can follow the following procedure:

- Initialize a set of clusters $C_i, i = 1, 2, \ldots, c$. One may either assign each data point as a distinct cluster or form some small initial clusters for seeding.

- Find the nearest clusters, say, C_i and C_j. Merge them into a new cluster and then repeat. The following measures are commonly used for the distance between C_i and C_j [4]:

$$
\begin{aligned}
d_{min}(C_i, C_j) &= \min_{x \in C_i, x' \in C_j} d_G(x, x') \\
d_{max}(C_i, C_j) &= \max_{x \in C_i, x' \in C_j} d_G(x, x') \\
d_{avg}(C_i, C_j) &= \frac{1}{n_i n_j} \sum_{x \in C_i} \sum_{x' \in C_j} d_G(x, x') \\
d_{mean}(C_i, C_j) &= \|m_i - m_j\|
\end{aligned}
$$

  where n_i, n_j are the numbers of points in C_i and C_j, respectively, and m_i, m_j are their cluster means.

- This procedure terminates when the specified number of clusters has been reached. The clusters are returned as sets of points. The mean or representative exemplar for each cluster can be computed as the average of the corresponding data points.

Through the above procedure, one actually gets a sequence of partitions of the n samples into c clusters, $c = 1, 2, \ldots, n$. Whenever two samples are in the same cluster at level k, they remain together at all higher levels. Figure 3 shows a dendrogram for a small data set containing eight samples.

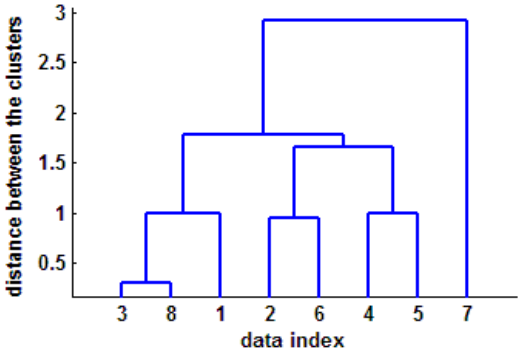

Figure 3. Dendrogram for a small data set with eight samples.

Given a training data set $G = \{g_1, g_2, \cdots, g_N\}$, we first initialize all the face images as singleton clusters. We then merge the two nearest clusters at each iteration using UPGMA (Unweighted Pair-Group Method with Arithmetic Mean) $d_{avg}(C_i, C_j)$, which computes the average distance between all pairs of objects in cluster i and cluster j. The procedure terminates when it reaches the number of clusters K specified beforehand by considering the length of the sequence. In our experiments, we set $K = 5 \sim 9$ based on the number of frames in the corresponding video clips (ranging from 250 to 800). The cluster centers $E = \{e_1, e_2, \cdots, e_K\}$ can be calculated as the representative exemplars summarizing the original data set. Since these exemplars are mean vectors, in general they may not correspond to real face images in the data set. An alternative is to find the image in the data set that is nearest to the cluster mean as exemplar. Figure 4 shows five exemplars extracted from the set of 250 training images based on the above two strategies (see Figure 6 for the original images and Figure 1 (blue dots) for the low-dimensional embedding). They seem to represent different head poses in the set, and we get almost the same recognition performance in the subsequent experiments using both strategies.

Figure 4. Five exemplars extracted from the set of 250 training images in Figure 1 (blue dots) corresponding to the cluster centers (first row) or nearest samples to their respective cluster centers (second row).

4. Experiments

To demonstrate the effectiveness of the proposed method, extensive experiments have been performed on a 40-subject video data set which bears large pose variation and moderate differences in expression and illumination. Each person is represented by one training clip and one testing clip, both captured using a CCD camera at 30fps for about 15-30 seconds. The faces are automatically detected from all frames using Viola and Jones' 'AdaBoost + Cascade' face detector [15]. All the detected faces are then resized to gray-level images of size 45×40, followed by a histogram equalization step to eliminate the lighting effects. The examples shown in Figure 5 are representative of the amount of variation in the data set.

Figure 5. Representative images of two subjects from the data set used in our experiments.

Since the focus of this paper is the automatic acquisition of exemplars, we simply build appearance-based face recognition systems based on a nearest neighbor criterion by template matching in the original image space or using traditional subspace methods, including PCA, LDA, and null space-based LDA (NLDA) [6]). For all the implemented methods, majority voting is adopted to combine the outputs from different frames in the test set. For a c-class problem $(\omega_1, \omega_2, \cdots, \omega_c)$, the test sequence contains K frames. If the kth frame is decided to belong to the ith class, denoted by the Kronecker delta δ_{ik}, the final recognition result is $h = \arg\max_{i=1}^{c} \sum_k \delta_{ik}$, i.e., the test sequence belongs to class ω_h.

Figure 6. Original images from the set of 250 training faces in Figure 1 (blue dots).

To demonstrate the effectiveness of the new method, we compare it with some traditional methods based on image sets, such as random selection of exemplars, applying k-means in the high-dimensional image space or in the PCA subspace. we also implement the 'LLE + k-means clustering' algorithm in [5] which has the same problem setting as ours. In the following experiments, the training video clip of each person is sent to different exemplar extraction procedures which result in 5-9 exemplars depending on the sequence length. The test set is constructed by randomly sampling from its corresponding testing video clip for 10 times with each set consisting of 30 independently and identically distributed (i.i.d.) samples. The recognition rates shown in Table 1 are the average results over all runs.

Table 1. Recognition rates (%) of different methods for selecting exemplars.

	Original	PCA	LDA	NLDA
Random selection	65.62	63.21	74.62	78.24
k-means	80.00	79.02	84.90	88.71
PCA + k-means	74.02	75.26	88.29	89.86
LLE + k-means	88.33	86.76	92.43	95.52
Isomap + k-means	87.14	84.71	93.91	96.38
Our method	89.74	88.10	94.14	96.52

The results clearly show that the approaches based on manifold learning (LLE + k-means, Isomap + k-means, and our method) can select better (more representative) exemplars than the traditional approaches (random selection, k-means, and PCA + k-means) since they yield better recognition rates. This observation is not unexpected as methods based on manifold learning can reveal the meaningful hidden structure in the nonlinear face manifold.

Another interesting finding is that our method slightly outperforms LLE and Isomap which are based on explicit embedding of the data. For the purpose of clustering (ex-

emplar selection), in fact there is no need to perform the last step (embedding) in LLE or Isomap. Doing so will not only require solving an eigendecomposition problem which is expensive for large data sets, but it can also lead to a certain degree of information loss. The reason we prefer a global embedding method (Isomap) to its local alternatives (LLE, Laplacian eigenmap) lies in its appealing property of explicitly preserving the global structure of a data set within a single coordinate system. As proved in the original paper, the estimated graph-based distance in Isomap asymptotically converges to the true geodesic structure of the manifold given sufficient data.

In summary, the success of our approach compared to other traditional methods (see Table 1) lies in the use of an elegant method (Isomap) for estimating geodesic distance and the subsequent direct transfer to the HAC clustering procedure without performing explicit embedding first.

5. Conclusion

This paper presents a novel method for selecting representative exemplars from video sequences or image sets and then uses it for building appearance-based face recognition systems. Our method consists of two main steps. First, based on local neighborhood information, geodesic distances between face images are estimated. Second, based on the geodesic distances estimated, distance-based clustering using the HAC algorithm is performed to group similar images to form clusters. The cluster centers are then identified as exemplars. When presented with a test face video sequence or image set, the final recognition result is obtained via a majority voting scheme by combining the decisions for the individual images in the test set. Experimental results on a medium-scale video database demonstrate the effectiveness of our proposed method.

Exemplar-based representation, as a reduced model, may not fully characterize the whole image set. In our future work, we will consider more flexible models (e.g., fitting a linear subspace to each cluster) and perform discriminant analysis on the local neighborhood of the face manifold.

Acknowledgments

The research reported in this paper is supported by Competitive Earmarked Research Grant (CERG) HKUST621305 from the Research Grants Council of the Hong Kong Special Administrative Region, China. We also thank the National Laboratory of Pattern Recognition (NLPR) at the Institute of Automation of the Chinese Academy of Sciences for providing us with the NLPR video dataset.

References

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

[2] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Trans. on PAMI*, 19(7):711–720, July 1997.

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, June 2003.

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

[5] A. Hadid and M. Pietikainen. From still image to video-based face recognition: an experimental analysis. In *Proc. of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 17–19, 2004.

[6] R. Huang, Q. Liu, H. Lu, and S. Ma. Solving the small size problem of LDA. In *Proceedings of the Sixteenth International Conference on Pattern Recognition*, volume 3, pages 29–32, August 2002.

[7] K. Lee, J. Ho, M. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *Proceedings of the CVPR*, pages 313–320, 2003.

[8] S. Z. Li, X. Peng, X. Hou, H. Zhang, and Q. Cheng. Multi-view face pose estimation based on supervised isa learning. In *Proc. of IEEE Internation Conf. on Face and Gesture Recognition*, pages 107–112, Washington, D.C. USA, May 2002.

[9] X. Liu and T. Chen. Video-based face recognition using adaptive hidden Markov models. In *Proceedings of the CVPR*, pages 340–345, 2003.

[10] A. O'Toole, D. Roark, and H. Abdi. Recognizing moving faces: A psychological and neural synthesis. *Journal of Vision*, 2:604a, July 2002.

[11] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, December 2000.

[12] G. Shakhnarovich, J. Fisher, and T. Darrell. Face recognition from long-term observations. In *Proceedings of the ECCV*, volume 3, pages 851 – 868, 2002.

[13] J. Tenenbaum, V. Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000.

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

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

[16] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *Proc. of IEEE Internation Conf. on Face and Gesture Recognition*, pages 318–323, 1998.

[17] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, 2003.

[18] S. Zhou and R. Chellappa. Probabilistic human recognition from video. In *Proceedings of the ECCV*, volume 3, pages 681–697, 2002.