Video-based Face Recognition Using Bayesian Inference Model

Wei Fan[1], Yunhong Wang[1,2], and Tieniu Tan[1]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100080, P.R.China
[2] School of Computer Science and Engineering, Bei Hang University, Beijing, 100083, P.R.China

Abstract. There has been a flurry of works on video sequence-based face recognition in recent years. One of the hard problems in this area is how to effectively combine the facial configuration and temporal dynamics for the recognition task. The proposed method treats this problem in two steps. We first construct several view specific appearance sub-manifolds learned from the training video frames using locally linear embedding (LLE). A general Bayesian inference model is then fit on the recognition task, transforming the complicated maximum likelihood estimation to some elegant distance measures in the learned sub-manifolds. Experimental results on a middle-scale video database demonstrate the effectiveness and flexibility of our proposed method.

1 Introduction

A majority of state-of-the-art face recognition algorithms [1] put emphasis on still image-based scenarios either by holistic template matching [2,3] or geometric feature-based methods [4]. Although these dominating approaches have achieved a certain level of success in restricted conditions such as mug-shot matching, they often fail to yield satisfactory performance when confronted with large pose, illumination and expression variations.

Recently, there is a significant trend in performing video-based face analysis [5,6,7,8], aiming to overcome the above limitations by utilizing visual dynamics or temporal consistence to facilitate performance of the recognition task. These approaches take root in relevant psychological and neural studies [9] which indicate that information for identifying a human face can be found both in the invariant structure of features and in idiosyncratic movements and gestures. As illustrated in Fig. 1, the *dynamic information* in terms of human face recognition can be typically divided into three categories: rigid head motions, no-rigid facial movements and the combination of both. Several researchers in this area have conjectured that if expressive dynamic information can be properly extracted, they will surely give a favorable improvement to video-based face recognition.

With this motivation, a new phase of recognition strategies that use both spacial and temporal information simultaneously has started. In [5], an identity surface for each subject is constructed in a discriminant feature space from

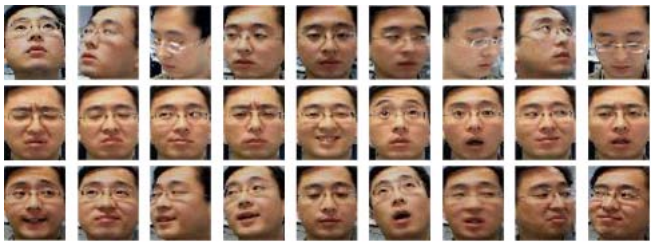


Fig. 1. Facial dynamics can be typically divided into three categories: rigid head motions (Top), no-rigid facial movements (Middle), and the combination of both (Bottom).

one or more learning sequences, and recognition is performed by computing distances between object trajectory and a set of model trajectories which encode the spatio-temporal information of a moving face. Zhou *et al* [6] simultaneously characterize the kinematics and identity using a motion vector and an identity variable respectively in a probabilistic framework. 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) [7] and probabilistic appearance manifolds [8] are both used to learn the transition probabilities among several viewing states embedded in the observation space. Hadid *et al* [10] compared the joint spatio-temporal representation (e.g. the HMM) with classical ones based on static images (e.g. PCA/LDA) for performing dynamic face recognition, and pointed out that the former model outperforms its counterparts in most experiments.

Although facial dynamics, if properly modelled, are tolerate to appearance variations induced by changes in head pose orientation and expressions (see Fig. 1 as an example), the most essential features for recognition still lie on those *static facial configurations*. Thus dynamic information, which provides us with some unstable behavioral characteristics, should be only treated as an assistant cue to the recognition task under non-optimal viewing conditions. The proposed approach in this paper is an attempt to somewhat balance the attention to static facial configurations for video-based recognition scenario.

To this end, we view sets of face images as high dimensional points whose underlying degrees of freedom is far less than the actual number of pixels per image. A well-known manifold learning algorithm, locally linear embedding (LLE) [11,12], is used to detect low dimensional structure in the image sequences for different individuals. As all human faces are similar patterns, we may anticipate under identical viewing conditions, e.g. rotation from left to right profiles [13], the manifolds of different individuals are often fairly close and parallel. Thus view specific sub-manifolds can be well constructed using classic clustering techniques on an individual's low dimensional embedding, assuming there is sufficient data (such that the manifold is well-sampled). Face images extracted from other training videos are sequentially assigned to its corresponding sub-

manifolds under the nearest "distance-from-feature-space" (DFFS) criteria [14]. To exploit the temporal coherence among successive video frames, we fit a general Bayesian inference model on the recognition task, transforming the complicated maximum likelihood estimation to some elegant distance measures in the learned view specific sub-manifolds. Experimental results conducted on a middle-scale video database strongly support our assumption and show high superiority of the newly developed method to its traditional still image-based counterparts.

2 View Specific Sub-manifolds Construction

2.1 Dimensionality Reduction Using LLE

In typical appearance-based methods, $m \times n$ face images are often represented by points in the mn-dimensional space. However, coherent structure in the facial appearance leads to strong correlations 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. Unlike traditional linear techniques, PCA and LDA, which often over-estimate the true degrees of freedom of the face data set, recent nonlinear dimensionality reduction methods, Isomap [15] and LLE [11,12], can effectively discover an underlying low dimensional embedding of the manifold. In this section, we use LLE to map the high-dimensional data to a single global coordinate system in a manner that preserves the neighboring relationships. An overview of the LLE algorithm is given in Table 1.

Table 1. An overview of the LLE algorithm [11,12]

INPUT:		$X = \{ x_1, x_2, \cdots, x_N \}$, where $x_i \in \mathbb{R}^D$.
OUTPUT:		$Y = \{ y_1, y_2, \cdots, y_N \}$, where $y_i \in \mathbb{R}^d, d \ll D$.
METHOD:		Repeat for each data point x_i:
	1	Find K nearest neighbors.
	2	Reconstruct x_i from its neighbors, minimizing the cost function $\varepsilon(W) = \left\| x_i - \sum_j W_{ij} x_j \right\|^2$ subjected to the additional constraints that $\sum_j W_{ij} = 1$ and $W_{ij} = 0$ if x_i and x_j are not neighbors.
	3	Define the embedding cost function $\varepsilon(y) = \left\| y_i - \sum_j \hat{W}_{ij} y_j \right\|^2$ where $\hat{W}_{ij}$ is the optimal result from step 2. Find the reconstructed vectors $\hat{y}_i = \arg\min_y \varepsilon(y), y_i \in \mathbb{R}^d$, with the additional constraints that $\sum_i y_i = 0$ and $\sum_i y_i y_i^T / N = I$.

2.2 View Specific Sub-manifolds

To illustrate the effectiveness of LLE, we applied it to a sequence of face images corresponding to a single person arbitrarily rotating his head. This data set contained $N = 788$ grayscale images at 23×28 resolution ($D = 644$). Fig. 2 shows the first three components of these images discovered by LLE (using $K = 12$ nearest neighbors) together with some representative frames. As we can see, the algorithm successfully revealed the meaningful hidden structure of the nonlinear face manifold.

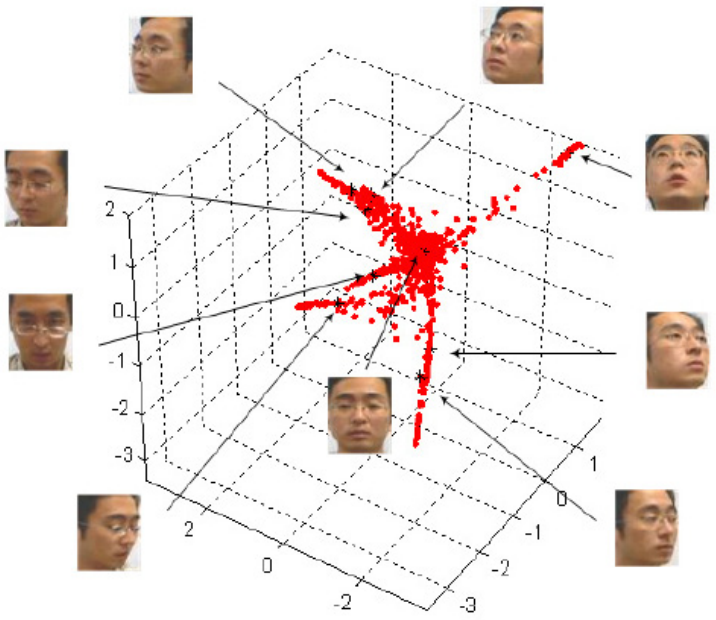


Fig. 2. LLE applied to a sequence of face images corresponding to a single person arbitrarily rotating his head.

To construct view specific sub-manifolds, we performed K-means clustering to points in the low-dimensional feature space given by LLE. The initial k cluster seeds were selected as some frames bearing distinct pose variations in the sequence (just like those shown in Fig. 2). Given large distances between the initial seeds and a moderate k, the output clusters could act as the expected view specific sub-manifolds in our method, which were further approximated by a set of linear subspaces ($P_i, i = 1, 2, \cdots, k$). Face images extracted from other training videos of different persons were sequentially assigned to its corresponding sub-manifolds under the nearest "distance-from-feature-space" (DFFS) criteria [14]. Thus in the training process, successive video frames will continuously update P_i, and provide an enhanced model of the view specific sub-manifolds.

3 Bayesian Inference Model for Recognition

In statistical pattern recognition [16], Bayesian inference model offers an efficient and principled approach for integrating prior knowledge and observed data to improve the classification performance. It is also an effective tool to characterize abundant temporal information in the video-based face recognition scenario [17]. Although these facial dynamics are by no means stable features as mentioned before, the temporal coherence among successive video frames still provides a significant aid in the recognition process.

Suppose w is the identity signature for a c class problem, i.e. $w \in \{1, 2, \cdots, c\}$. Given a sequence of face images $F = \{f_1, f_2, \cdots, f_N\}$ containing the appearances of the same but unknown person, Bayesian inference model aims to find the solution with the maximum a posterior probability (MAP)

$$\hat{w} = \arg \max_{\{1,2,\cdots,c\}} P(w|f_{1:N}) \tag{1}$$

According to the Bayesian theory

$$P(w|f_{1:N}) = \frac{P(w)P(f_{1:N}|w)}{P(f_{1:N})} \tag{2}$$

We further assume the prior probability $P(w)$ to be non-informative and neglect the normalization factor $P(f_{1:N})$ which is independent to the final decision. Thus the MAP solution is converted to an equivalent maximum likelihood (ML) estimation

$$\hat{w} = \arg \max_{\{1,2,\cdots,c\}} P(f_{1:N}|w) \tag{3}$$

As the face images tend to lie on or close to a non-convex low-dimensional manifold, it is hard to analytically capture its complexity in a universal or parametric solution. One possible way to tackle this problem is to build a view-based formulation with a set of subspaces $(P_i, i = 1, 2, \cdots, k)$ covering the whole manifold, as introduced in Section 2. Here we simply associate each image with a hidden view parameter θ, where $\theta \in \{P_1, \cdots, P_k\}$, and decompose (3) as follows:

$$\begin{aligned} P(f_{1:N}|w) &= \sum_{\theta_{1:N}} P(f_{1:N}|\theta_{1:N}, w)P(\theta_{1:N}) \\ &= \sum_{\theta_{1:N}} \prod_{t=1}^{N} P(f_t|\theta_t, w)P(\theta_t|\theta_{1:t-1}) \\ &= \sum_{\theta_{1:N}} \prod_{t=1}^{N} P(f_t|\theta_t, w)P(\theta_t|\theta_{t-1}) \end{aligned} \tag{4}$$

In the above derivation, we use two intuitive rules which are appropriate for video-based face recognition, namely (a) observational conditional independence: $P(f_{1:N}|\theta_{1:N}, w) = \prod_{t=1}^{N} P(f_t|\theta_t, w)$ and (b) the first-order Markov chain rule: $P(\theta_{1:N}) = \prod_{t=1}^{N} P(\theta_t|\theta_{1:t-1}) = \prod_{t=1}^{N} P(\theta_t|\theta_{t-1})$, $P(\theta_1|\theta_0) \doteq P(\theta_1)$. The following subsections show how to compute the two probabilities involved in (4).

3.1 Computation for $P(f_t|\theta_t, w)$

The term $P(f_t|\theta_t, w)$ denotes the probability of observing face image f_t at time t, given its corresponding identity w in the view sub-manifold θ_t. Typically a multivariate Gaussian density is fitted for this distribution when a large training set is available. Here we avoid directly estimating the particular density function for the limited training data, and convert it to some elegant distance measures related to the learned sub-manifolds.

From the definition of the *law of total probability* and the conditional probability we have

$$P(f_t|\theta_t, w) = P(\theta_t|f_t)P(w|f_t, \theta_t)\frac{P(f_t)}{P(\theta_t|w)P(w)} \tag{5}$$

As before, the prior $P(w)$ and evidence $P(f_t)$ are assumed non-informative. And $P(\theta_t|w)$ represents the likelihood of w being in sub-manifold θ_t at time t, which is related to the behavioral characteristic of subject w. To simplify the computational setting in our case, all three terms are treated as constants, thus

$$P(f_t|\theta_t, w) \propto P(\theta_t|f_t)P(w|f_t, \theta_t) \tag{6}$$

For a k sub-manifold problem, let $d_m(P_i, f_t)$ be the Euclidean distance between the ith sub-manifold and the test sample f_t (approximated by the DFFS measure [14]), an estimation of the probability $P(\theta_t|f_t)$ can be approximated as

$$P(\theta_t|f_t) = \frac{1/d_m(\theta_t, f_t)^2}{\sum_{i=1}^{k} 1/d_m(P_i, f_t)^2} \tag{7}$$

Similarly for a c class problem, let $d_c(j, f_t)$ be the distance between the jth class center and the test image f_t with all the related training data belonging to the sub-manifold θ_t. Thus the term $P(w|f_t, \theta_t)$ can be approximated as

$$P(w|f_t, \theta_t) = \frac{1/d_c(w, f_t)^2}{\sum_{j=1}^{c} 1/d_c(j, f_t)^2} \tag{8}$$

Here $d_c(j, f_t)$ is measured by the "distance-in-feature-space" (DIFS)criteria [14], and the feature space is constructed using null space-based linear discriminant analysis (NLDA) [18].

3.2 Computation for $P(\theta_t|\theta_{t-1})$

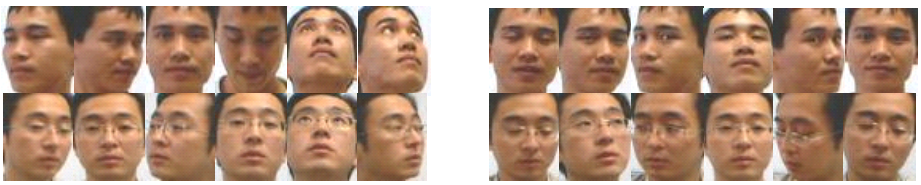
Motivated by the similar work of [8], the transition probability $P(\theta_t|\theta_{t-1})$ is defined by counting the actual transitions between different sub-manifolds P_i observed in all the training sequences:

$$P(\theta_t|\theta_{t-1}) = \frac{1}{\lambda}\sum_{t=2}^{k} \delta(f_{t-1} \in \theta_{t-1})\delta(f_t \in \theta_t) \tag{9}$$

where $\delta(f_t \in \theta_t) = 1$ if $f_t \in \theta_t$ and otherwise is 0. The normalization factor λ ensures $P(\theta_t|\theta_{t-1})$ to be a probability measure.

4 Experiments

To demonstrate the effectiveness of the proposed method, extensive experiments were performed on a 25-subject video dataset 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 in our lab with a CCD camera at 30 fps for about 15 seconds. The faces were manually cropped from all frames and resized to 23×28 pixel gray level images, followed by a histogram equalization step to eliminate lighting effects. The examples shown in Fig. 3 are representative of the amount of variation in the data.



(a) training video clips (b) testing video clips

Fig. 3. Representative examples for two subjects from the training and testing data used in the experiments. Note the significant pose variation in both sets.

Nine view specific sub-manifolds ($P_i, i = 1, 2, \cdots, 9$) are learned from all the training videos using strategies proposed in Section 2.2. The baseline image sequence for LLE modelling (Fig. 2) contains a person with abundant pose variations. Subsequent video frames of other subjects are sequentially absorbed by the relevant sub-manifolds. This step inevitably produces a few false assignments which are automatically detected by certain thresholds and manually corrected.

The MAP estimation for each testing sequence was evaluated by (2). Fig. 4 shows the computed posterior probabilities of the two persons in Fig. 3 as a function of time t. From the figure, it is obvious that the true signature (red line) always gives the largest posterior probability.

To illustrate the superiority of this newly developed method to its traditional still image-based counterparts, we implemented the LLE+clustering algorithm [10] which chose the cluster centers of each training sequence as extracted exemplars for template matching and took a vote to give the final decision. PCA and LDA were used as the classification methods in [10]. Here we also provide experimental result given by NLDA classifier [18]. Table 2 summarizes the recognition rates on our databet averaged among various sequence length (like the testing strategy in [7]) using different approaches mentioned above. The results clearly show that the proposed method outperforms all its still image-based counterparts, as it greatly profits from the Baysian inference model while other approaches use dynamic information only in its most crude form through voting.

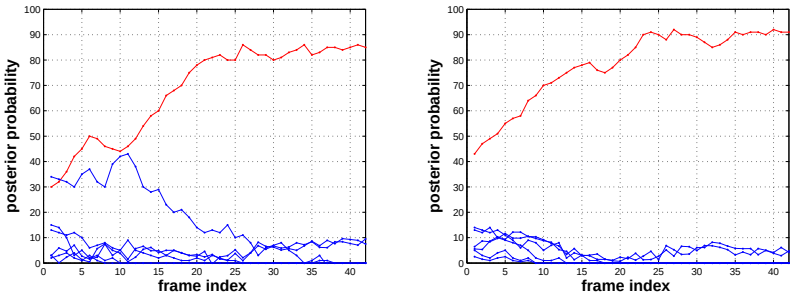


Fig. 4. Posterior probability $P(w|f_{1:N})$ of the two persons in Fig. 3 against time t. Only the seven most likely candidates are shown in this figure. From the figure, it is obvious that the true signature (red line) always gives the largest posterior probability.

Table 2. Recognition rate (%) for different methods using $k = 9$ clusters.

Method	LLE+PCA	LLE+LDA	LLE+NLDA	Our method
Recogition rate	82.62	87.21	91.62	95.24

5 Discussion and Conclusions

This paper presents a novel video-based face recognition method using both spacial and temporal information simultaneously. Unlike most other joint spatio-temporal representations which excessively rely on unstable facial dynamics for recognition, we exploit dynamic information in a moderate fashion, i.e. only those constraints of common transitions along the face manifold are modelled by the Baysian inference framework. More emphases are put on the construction of view specific sub-manifolds, which essentially convey relevant discriminating information, i.e. the static facial configurations, for the recognition task. As our work combines the major analytic features of the manifold learning algorithm LLE – precise preservation of the neighboring relationships in a single global coordinate system – with the flexibility to learn a moderate model of facial dynamics, it is especially suitable to the video-based face recognition scenario and exhibited satisfactory performance in a middle-scale video dataset.

Acknowledgements

This work is funded by research grants from the National Basic Research Program of China (No. 2004CB318110) and the National Natural Science Foundation of China (No. 60332010).

References

1. R. Chellappa, C.L. Wilson, and S. Sirohey, "Human and machine recognition of faces: a survey", Proceedings of the IEEE, Vol. 83, pp. 705-741, May 1995.
2. M. Turk and A. Pentland, "Eigenfaces for recognition", J. of Cognitive Neuroscience, Vol. 3, No. 1, pp. 71-86, 1991.
3. V. Belhumeur, J. Hespanda, and D. Kiregeman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. on PAMI, Vol. 19, No. 7, pp. 711-720, July 1997.
4. L. Wiskott, J.M. Fellous, N. Kruger and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 19, No.7, pp. 775-779, July, 1997.
5. Y. Li, S. Gong, and H. Liddell, "Video-Based Online Face Recognition Using Identity Surfaces", Proceedings of IEEE ICCV Workshop on Recogntion, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, pp. 40-46, 2001
6. S. Zhou and R. Chellappa. "Probabilistic human recognition from video". European Conference on Computer Vision (ECCV), May 2002.
7. X.Liu, and T.Chen, "Video-Based Face Recognition Using Adaptive Hidden Markov Models", Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition, pp. 340-345, 2003.
8. K.C.Lee, J.Ho, M.H.Yang, and D.Kriegman, "Video-Based Face Recognition Using Probabilistic Appearance Manifolds", In Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition, pp. 313-320, 2003.
9. Alice J. OToole*, Dana A. Roark, Herv Abdi, "Recognizing moving faces: A psychological and neural synthesis", Trends in Cognitive Sciences, 6, 261-266. Reed, CL, Stone, VE, Bozova, S., Tanaka, J. (2003).
10. A.Hadid, and M.Pietikainen, "From Still Image to Video-Based Face Recognition: An Experimental Analysis", Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 813-818, 2004.
11. S. T. Roweis and L. K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," Science 290, 2323-2326, 2000.
12. L. K. Saul and S. T. Roweis, "Think Globally, Fit Locally : Unsupervised Learning of Nonlinear Manifolds," Technical Report MS CIS-02-18, University of Pennsylvania, 2003.
13. Gong S, McKenna S J and Collins J J, "An Investigation into Face Pose Distributions", Second International Conference on Automated Face and Gesture Recognition, Vermont, USA, October 1996.
14. A. Pentland, B. Moghaddam, T. Starner, "View-based and modular eigenspaces for face recognition", Proceedings of IEEE, CVPR, 1994.
15. J. B. Tenenbaum, V. D. Silva, and J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," Science, 290(5500):2319–2323, 2000.
16. R. Duda, P. Hart, and D. Stork. Pattern Classification. WileyInterscience, 2001.
17. S. Zhou and R. Chellappa. "Probabilistic Identity Characterization for Face Recognition". In Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition, pp. 805-812, 2004.
18. Wei Fan, Yunhong Wang, Wei Liu, Tieniu Tan, "Combining Null Space-based Gabor Features for Face Recognition", In Proc. of the 17th International Conference on Pattern Recognition. pp. 330-333, Cambridge, UK, 2004.