Video-Based Face Recognition Using Probabilistic Appearance Manifolds

Kuang-Chih Lee[†] klee10@uiuc.edu
Jeffrey Ho[‡] jho@cs.ucsd.edu
Ming-Hsuan Yang[⋆] myang@honda-ri.com
David Kriegman[‡] kriegman@cs.ucsd.edu

[†]Computer Science
University of Illinois, Urbana-Champaign
Urbana, IL 61801

[‡]Computer Science & Engineering
University of California, San Diego
La Jolla, CA 92093

[⋆]Honda Research Institute
800 California Street
Mountain View, CA 94041

Abstract

This paper presents a novel method to model and recognize human faces in video sequences. Each registered person is represented by a low-dimensional appearance manifold in the ambient image space. The complex nonlinear appearance manifold expressed as a collection of subsets (named pose manifolds), and the connectivity among them. Each pose manifold is approximated by an affine plane. To construct this representation, exemplars are sampled from videos, and these exemplars are clustered with a K-means algorithm; each cluster is represented as a plane computed through principal component analysis (PCA). The connectivity between the pose manifolds encodes the transition probability between images in each of the pose manifold and is learned from a training video sequences. A maximum a posteriori formulation is presented for face recognition in test video sequences by integrating the likelihood that the input image comes from a particular pose manifold and the transition probability to this pose manifold from the previous frame. To recognize faces with partial occlusion, we introduce a weight mask into the process. Extensive experiments demonstrate that the proposed algorithm outperforms existing frame-based face recognition methods with temporal voting schemes.

1 Introduction

Face recognition has long been an active area of research, and numerous algorithms have been proposed over the years. However, most research has been focused on recognizing faces from a single image. Face recognition using video presents various challenges and opportunities. Typically, recognition using image sequences is done using a two-stage system: a tracking module and a recognition module. Given a video frame, a tracking module takes an estimate of the object's location in the previous frame and returns a subimage in the current frame that contains the object. A recognition module then operates on the subimage, perhaps integrating information/decisions from earlier frames. In a video, head pose may vary significantly. Therefore, successful video-based face recognition must be able to classify faces with a range of image plane and 3-D orientations. In addition, a good recognition method should be robust to misalignment errors introduced by inaccuracies from the tracking module. Meanwhile, partial occlusion poses another serious challenge, and this is likely to occur at some instants in unconstrained applications such as vision-based human computer interaction.

On the other hand, recognition in video offers the opportunity to integrate information temporally across the video sequence, which may help to increase the recognition rates. Our framework exploits temporal coherence in the following ways. First, our proposed appearance model is composed of a collection of pose manifolds, and a matrix of transition probabilities to connect them. The transition probabilities among the pose manifolds are learned from training videos each one characterizes the probability of moving from one pose to another pose between any two consecutive frames. We use the transition probability to implicitly infer the appropriate pose for each incoming video frame, and then integrate this information by Bayes' rule to perform face recognition. Therefore, our method effectively captures the dynamics of pose changes and thereby exploits the temporal information in a video sequence for recognition. Second, we use consecutive frames to define a mask whose elements represent the probability that a pixel corresponds to an occlusion. The mask is iteratively updated by analyzing the difference between the observed image at each time instance and the reconstructed image predicted from previous frame.

We have implemented the proposed method and evaluated it with numerous experiments. The experimental results show that our method is effective in recognizing faces in videos containing large variation of head motion as well as partial occlusions.

This paper is organized as follows. We briefly summarize the related literature which motivates this work in Section 2. In Section 3, we detail and contrast our algorithms with other existing work. Numerous experiments on a large and rather difficult data set are presented in Section 4. We conclude with remarks and future work in Section 5.

1063-6919/03 $17.00 © 2003 IEEE

2 Related Work

Most of the research work in the literature concentrates on representation and classification methods for recognizing faces in still and oftentimes single images [4, 24, 30]. Although there exist numerous face recognition algorithms operating on image sequences, they typically use temporal voting to improve identification rates [12, 26, 28]. We also note that there exist several algorithms that aim to extract 2-D or 3-D face structure from video sequences for recognition and animation [5, 14, 6, 7, 8, 9, 29, 11, 23]. However these methods require meticulous procedures to build 2-D or 3-D models, and do not fully exploit temporal information for recognition.

Among the few attempts aiming to truly utilize temporal information for face recognition in image sequences rather than simple voting, Li et al. presented a method to construct identity surfaces using shape and texture models as well as kernel feature extraction algorithms [16]. This approach estimates pose angle first in order to select an appropriate shape model for tracking and recognition. However, it does not fully take advantage of coherence information between consecutive frames except for a weighted temporal voting scheme to fit model parameters. Zhou and Chellappa [31] proposed a generic framework to track and recognize human faces simultaneously by adding an identity variable to the state vector in the sequential importance sampling method. They then marginalize over all state vectors to yield an estimate of the posterior probability of the identity variable. Though this probabilistic approach aims to integrate motion and identity information over time, it nevertheless considers only identity consistency in temporal domain and thus may not work well when the target is partially occluded. Furthermore, it is not clear how one can extend this work to deal with large 3-D pose variation. Krueger and Zhou [15] applied an on-line version of radial basis functions to select representative face images as exemplars from training videos, and in turn this facilitates tracking and recognition tasks. The state vector in this method consists of affine parameters as well as an identity variable, and the state transition probability is learned from affine transformations of exemplars from training videos in a way similar to [27]. Since only 2-D affine transformations are considered, this model is effective in capturing small 2-D motion but may not deal well with large 3-D pose variation or occlusion.

Recently, Li et al. [17] applied piecewise linear models to capture local motion and a transition matrix among these models to describe nonlinear global dynamics. They applied the learned local linear models and their dynamic transitions to synthesize new motion video such as choreography. Our work bears some resemblance to their method in the sense that both methods utilize local linear models, something advocated in several prior works [3, 1, 19], and both learn the relationships among these models [13, 20, 21, 25]. However in this paper, we consider propagating the probabilistic likelihood of the linear models through the transition matrix (i.e., utilizing temporal information) to recognize human identity. Furthermore, we exploit the information learned in the local models and transition matrix to infer missing data in recognizing partially occluded faces.

3 Probabilistic Appearance Manifold

Consider a recognition problem with N objects where the images of an object are acquired by varying the viewpoint. It is well understood that the set of images of an object under varying viewing conditions can be treated as a low-dimensional manifold in the image space as demonstrated in parametric appearance manifold work [19] or view-based Eigenspace approach [22]. The recognition task is straightforward if the appearance manifold M_k for each individual k is known: for a test image I, the identity k^* can be determined by finding the manifold M_k with minimal "distance" to I, i.e.,

$$k^* = \arg\min_k \; d_H(I, M_k). \tag{1}$$

Here, d_H denotes the L^2-Hausdorff distance between the image I and M_k. Let $x \in M_k$ denote a point on a manifold M_k where $dim(M_k) \leq dim(I)$. Given a point $x \in M_k$, let the corresponding reconstructed face image be denoted $\hat{I}_x$ where $dim(I) = dim(\hat{I}_x)$. If x^* is the point on M_k at minimal L^2 distance to I, then $d_H(I, M_k) = d(I, x^*)$ where $d(\cdot, \cdot)$ denotes the L^2 distance. Alternatively, x^* can be regarded as the result of some nonlinear projection of I onto M_k.

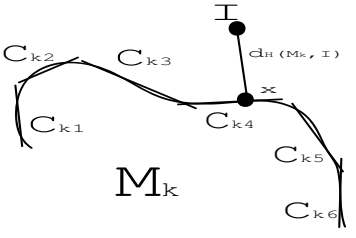

Figure 1: Appearance manifold. A complex and nonlinear manifold can be approximated as the union of several simpler pose manifolds; here, each pose manifold is represented by a PCA plane.

Probabilistically, Equation 1 is the result of defining the conditional probability $p(k|I)$ as

$$p(k|I) = \frac{1}{\Lambda} \; \exp(\frac{-1}{\sigma^2} d_H^2(I, M_k)). \tag{2}$$

where Λ is a normalization term, and for a given image I

$$k^* = \arg\max_k p(k|I). \quad (3)$$

In order to implement this recognition scheme, one must be able to estimate the projected point $x^* \in M_k$, and then the image to model distance, $d_H(I, M_k)$, can be computed for a given I and for each M_k. However, such distances can be computed accurately only if M_k is known exactly. In our case, M_k is usually not known and can only be approximated with samples. The main part of our algorithm is to provide a probabilistic framework for estimating x^* and $d_H(x^*, I)$. Note that if we define the conditional probability $p_{M_k}(x|I)$ to be the probability that among points on M_k, $\hat{I}_{x^*}$ has the smallest L^2-distance to I, then

$$d_H(I, M_k) = \int_{M_k} d(x, I) p_{M_k}(x|I) dx, \quad (4)$$

and Equation 1 is equivalent to

$$k^* = \arg\min_k \int_{M_k} d(x, I) p_{M_k}(x|I) dx. \quad (5)$$

The abovementioned formulation shows that $d_H(I, M_k)$ can be viewed as the expected distance between a single image frame I and a complex appearance manifold M_k. Clearly, if M_k were fully known or well-approximated (e.g., described by some algebraic equations), then $p_{M_k}(x|I)$ could be treated as a $\delta-$function at the set of points with minimal distance to I. When sufficiently many samples are drawn from M_k, the expected distance $d(I, M_k)$ will be a good approximation of the true distance. The reason is that $p_{M_k}(x|I)$ in the integrand of Equation 4 will approach a delta function with its "energy" concentrated on the set of points with minimal distance to I. In our case, M_k, at best, is approximated through a sparse set of samples, and so we will model $p_{M_k}(x|I)$ with a Gaussian distribution.

Since the appearance manifold M_k is complex and nonlinear, it is reasonable to decompose M_k into a collection of m simpler disjoint manifolds, $M_k = C^{k1} \cup \cdots \cup C^{km}$ where C^{ki} is called a pose manifold. Each pose manifold is further approximated by an affine plane computed through principal component analysis (called a PCA plane). We define the conditional probability $p(C^{ki}|I)$ for $1 \leq i \leq m$ as the probability that C^{ki} contains a point x with minimal distance to I. Since $p_{M_k}(x|I) = \sum_{i=1}^m p(C^{ki}|I) p_{C^{ki}}(x|I)$, we have,

$$\begin{aligned} d_H(I, M_k) &= \int_{M_k} d(x, I) p_{M_k}(x|I) dx \\ &= \sum_{i=1}^m p(C^{ki}|I) \int_{C^{ki}} d_H(x, I) p_{C^{ki}}(x|I) dx \\ &= \sum_{i=1}^m p(C^{ki}|I) d_H(I, C^{ki}). \end{aligned} \quad (6)$$

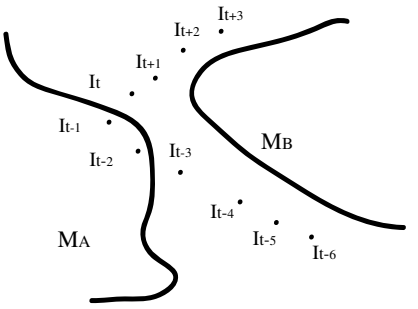

Figure 2: Difficulty of frame-based recognition: The two solid curves denote two different appearance manifolds, M_A and M_B It is difficult to reach a decision on the identity from frame I_{t-3} to frame I_t because these frames have smaller L^2 distance to appearance manifolds M_A than M_B. However, by looking at the sequence of images $I_{t-6} \ldots I_{t+3}$, it is apparent that the sequence has most likely originated from appearance manifold M_B.

The above equation shows that the expected distance $d(I, M_k)$ can be also treated as the average expected distance between I and each pose manifold C^{ki}. In addition, this equation transforms the integral to a finite summation which is feasible to compute numerically.

For face recognition from video sequences, we can exploit temporal coherence between consecutive image frames. As shown in Figure 2, the L^2 norm may occasionally be misleading during recognition. But if we consider previous frames in an image sequence rather than just one, then the set of closest points x^* will trace a curve on a pose manifold. In our framework, this is embodied by the term $p(C^{ki}|I)$ in Equation 6. In Section 3.1, we will apply Bayesian inference to incorporate temporal information to provide a better estimation of $p(C^{ki}|I)$, and thus $d_H(I, M_k)$ to achieve better recognition performance.

3.1 Computing $p(C_t^{ki}|I_t)$

For recognition from a video sequence, we need to estimate $p(C_t^{ki}|I_t)$ for each i at time t. To incorporate temporal information, $p(C_t^{ki}|I_t)$ should be taken as the joint conditional probability $p(C_t^{ki}|I_t, I_{0:t-1})$ where $I_{0:t-1}$ denotes the frames from the beginning up to time $t-1$. We further assume I_t and $I_{0:t-1}$ are independent given C_t^{ki}, as well as C_t^{ki} and $I_{0:t-1}$ are independent given C_{t-1}^{ki}. Using Bayes' rule we have the following recursive formulation:

$$\begin{aligned} p(C_t^{ki}|I_t, I_{0:t-1}) &= \alpha\, p(I_t|C_t^{ki}, I_{0:t-1}) p(C_t^{ki}|I_{0:t-1}) \\ &= \alpha\, p(I_t|C_t^{ki}) \sum_{j=1}^m p(C_t^{ki}|C_{t-1}^{kj}, I_{0:t-1}) p(C_{t-1}^{kj}|I_{0:t-1}) \\ &= \alpha\, p(I_t|C_t^{ki}) \sum_{j=1}^m p(C_t^{ki}|C_{t-1}^{kj}) p(C_{t-1}^{kj}|I_{t-1} I_{0:t-2}) \end{aligned} \quad (7)$$

where α is a normalization term to ensure a proper probability distribution.

The temporal dynamics of the video sequence is captured by the *transition probability* between the manifolds, $p(C_t^{ki}|C_{t-1}^{kj})$. Note that $p(C_t^{ki}|C_{t-1}^{kj})$ is the probability of $x_t \in C^{ki}$ given $x_{t-1} \in C^{kj}$. For two consecutive frames I_{t-1} and I_t, because of temporal coherency, we expect that their projected points x_{t-1}^* and x_t^* should have small geodesic distance on M (See Figure 2). That is the transition probability $p(C_t^{ki}|C_{t-1}^{kj})$ is related implicitly to the geodesic distance between C^{ki} and C^{kj}.

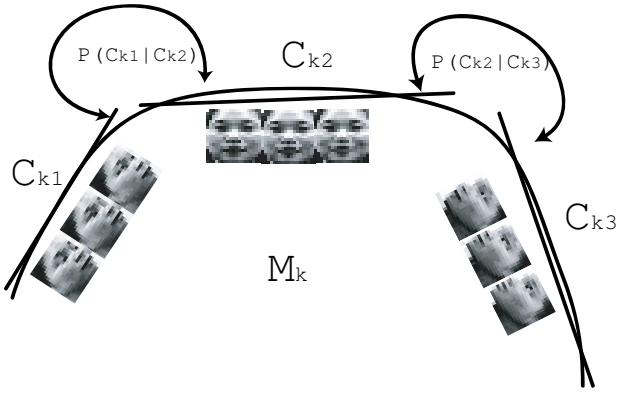

Figure 3: Dynamics among pose manifolds. The dynamics among the pose manifolds are learned from training videos which describes the probability of moving from one manifold to another at any time instance.

3.2 Learning Manifolds and Dynamics

For each person k, we collect at least one video sequence containing l consecutive images $S_k = \{I_1, \cdots, I_l\}$. We further assume that each training image is a fair sample drawn from the appearance manifold M_k. There are three steps in the algorithm. We first partition these samples into m disjoint subsets $\{S_1, \cdots, S_m\}$. For each collection S_{ki}, we can consider it as containing points drawn from some pose manifold C^{ki} of M_k, and from the images in S_{ki}, we construct a linear approximation to the C^{ki} of the true manifold M_k. After all the C^{ki} have been computed, we estimate the transition probabilities $p(C^{ki}|C^{kj})$ for $i \neq j$.

In the first step, we apply a K-means clustering algorithm to the set of images in the video sequences. We initialize m seeds by finding m frames from the training videos with the largest L^2 distance to each other. Then the general K-means algorithm is used to assign images to the m clusters. As our goal in performing clustering is to approximate the data set rather than to derive semantically meaningful cluster centers, it is worth noting that the resulting clusters are no worse than twice what the optimal center would be if they could be easily found [10].

Second, for each S_{ki} we obtain a linear approximation of the underlying subset $C^{ki} \subset M_k$ by computing a PCA plane L_{ki} of fixed dimension for the images in S_{ki}. Since the PCA planes approximate appearance manifold M_i, their dimension is the intrinsic dimension of M, and therefore all PCA planes L_i have the same dimension.

Finally, the transition probability $p(C^{ki}|C^{kj})$ is defined by counting the *actual* transitions between different S_i observed in the image sequence:

$$p(C^{ki}|C^{kj}) = \frac{1}{\Lambda_{ki}} \sum_{q=2}^{l} \delta(I_{q-1} \in S_{ki}) \delta(I_q \in S_{kj}) \quad (8)$$

where $\delta(I_q \in S_{kj}) = 1$ if $I_q \in S_{kj}$ and otherwise it is 0. The normalizing constant Λ_{ki} ensures that

$$\sum_{j=1}^{m} p(C^{ki}|C^{kj}) = 1. \quad (9)$$

where we set $p(C^{ki}|C^{ki})$ to a constant κ. A graphic representation of a transition matrix with $m = 5$ learned from a training video is depicted in Figure 4.

With C^{ki} and its linear approximation L_{ki} defined, we can define how $p(I|C^{ki})$ can be calculated. We can compute the L^2 distances $\hat{d}_{ki} = d_H(I, L_{ki})$ from I to each L_{ki}. We treat $\hat{d}_{ki}$ as an estimate of the true distance from I to C^{ki}, i.e., $d_H(I, C^{ki}) = d_H(I, L_{ki})$. $p(I|C^{ki})$ is defined as

$$p(I|C^{ki}) = \frac{1}{\Lambda_{ki}} \exp\left(\frac{-1}{2 * \sigma^2} \hat{d}_{ki}^2\right) \quad (10)$$

with $\Lambda_{ki} = \sum_{i=1}^{m} \exp(\frac{-1}{2*\sigma^2} \hat{d}_{ki}^2)$.

Notice that we use a non-compact subspace L_{ki} to approximate a compact pose manifold C^{ki}. The infinite extent of L_{ki} might be better captured by the underlying Gaussian, and similar work has been done by Moghaddam et al.[18]. However, our experiment shows that the recognition result using this more elaborate algorithm is no better than the one proposed in the paper. This can be explained by the fact that although the linear subspaces are non-compact, the test images will almost always be drawn from a compact subset of the image space. This effect makes the subspaces functionally compact in our algorithm. In other words, the subspaces behave as they only have finite extent.

3.3 Face Recognition from Video

Given an image I from a video sequence, we compute for each person k the distance $d_H(I, M_k)$ using the Equation 6. Note that $p(C^{ki}|I)$ has a temporal dependency, and it is computed recursively using Equation 7. Once all the $d_H(I, M_k)$ have been computed, the posterior $p(k|I)$ is computed by Equation 2 with appropriate σ, and the human identity is decided by Equation 5.

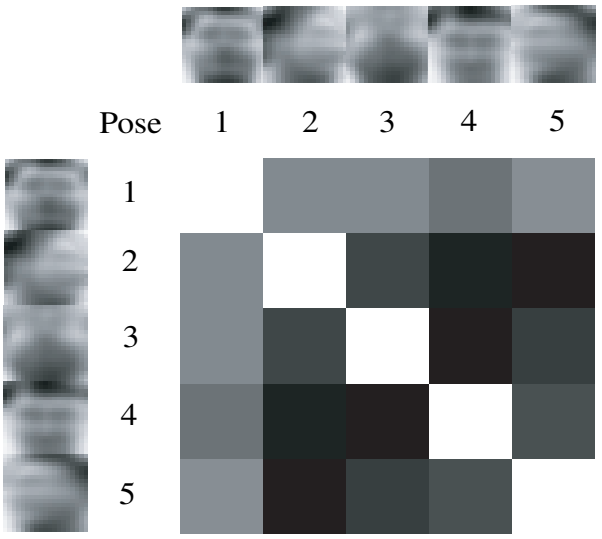

Figure 4: Graphic representation of a transition matrix learned from a training video. In this example, the appearance manifold is approximated by 5 pose subspaces. The reconstructed center image of each pose subspace is shown at the top raw and column. The transition probability matrix is drawn by the 5×5 block diagram. The brighter block means a higher transition probability. It is easy to see that the frontal pose (pose 1) has higher probability to change to other poses; the right pose (pose 2) has almost zero probability to directly change to the left pose (pose 3).

It is also worth mentioning that the proposed framework exploits the temporal coherence in the appearance of consecutive face images by integrating the manifold transition at the previous and current time instance. For face recognition with varying pose, our method ensures that the transitions between pose manifolds do not occur arbitrarily but rather in a constrained order. For example the appearance of one person's face cannot change immediately from left profile to right profile in two consecutive frames, but rather it must pass through some intermediate pose or orientation (See Figure 6). This process can also be considered as putting a first order Markov process or finite state machine over a piecewise linear structure. In contrast, simple temporal voting scheme has been commonly adopted in most video-based face recognition methods [16] [26].

3.4 Recognizing Partially Occluded Faces

Similar to our formulation exploiting temporal information for recognition, the same approach can be easily extended to deal with partial occlusion of a face by considering the previous frame as prior information. The original formulation for $d_H(C_t^{ki}, I_t)$ treats every pixel in image I_t with equal weight assuming that there is no occlusion anywhere in the image sequence. If we knew which pixels corresponded to occlusions, we would put lower weights on those pixels when computing $d_H(M_k, I_t)$. We introduce an image mask W, which defines the probability that a pixel is occluded, where W has the same dimension as image I, and its elements are initialized with a 1, i.e., assuming there is no occlusion at the first frame and no pixel is downweighted. The $d_H(M_{k*}, I_t)$ is then replaced by the weighted distance $d_H(M_{k*}, W_t. * I_t)$ where $.*$ denote element-by-element multiplication. Let the weighted projection of $W_t. * I_t$ on M_{k^*} be x^*, the mask W_t is updated in each frame I_t by the estimate at a previous frame W_{t-1} by

$$W_t^{(1)} = \exp(\frac{-1}{2 * \sigma^2}(\hat{I}_{x^*} - I_t). * (\hat{I}_{x^*} - I_t)) \qquad (11)$$

in the first iteration. Alternatively, W_t can be iteratively updated based on the $W_t^{(1)}$ and $\hat{I}_{x^*}^{(1)}$ (i.e., the reconstructed image based on $W_t^{(1)}$ and $d_H(M_{k*}, W_t^{(1)}. * I_t)$)

$$W_t^{(i+1)} = \exp(\frac{-1}{2 * \sigma^2}(\hat{I}_{x^*}^{(i)} - I_t). * (\hat{I}_{x^*}^{(i)} - I_t)) \qquad (12)$$

until the difference between $W_t^{(i)}$ and $W_t^{(i-1)}$ is below a threshold value at the i-th iteration.

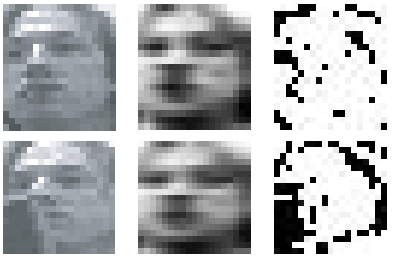

Figure 5: Top row: (left) an unoccluded face image, (center) a reconstructed image using corresponding pose manifold, and (right) a corresponding mask). Bottom row: (left) a face image partially occluded by one hand, (center) a reconstructed image using corresponding pose manifold, and (right) an updated mask.

Both the appearance manifold and mask information at previous frames are utilized to estimate the current occlusion mask in the equations above. We first perform the weighted projection to find a reconstructed image using the corresponding pose manifold and iteratively estimate the occlusion areas in the current frame. Once we get an updated mask W_t in frame I_t by Equation 11, we evaluate Equation 6 for face recognition by replacing $d_H(C_t^{ki}, I_t)$ with $d_H(C_t^{ki}, W_t. * I_t)$.

Figure 5 shows an example where a face is partially occluded by an object (lower left). The reconstructed image using the corresponding pose manifold is shown in the lower center. The updated mask is shown in the lower right where the values have been thresholded – a dark pixel denotes a probability of occlusion. Note that the updated mask matches the occluded region reasonably well. Note also that

the mask predicts that several pixels are occluded though in fact they are not. This is caused by the disagreement between the input image and the reconstructed image. Nevertheless, the regions that matter most for recognition (i.e., the central face region and the occluded region) are weighted appropriately. Our experimental results, presented in the next section, also demonstrate that the mask scheme is effective in recognizing partially occluded faces.

4 Experiments and Results

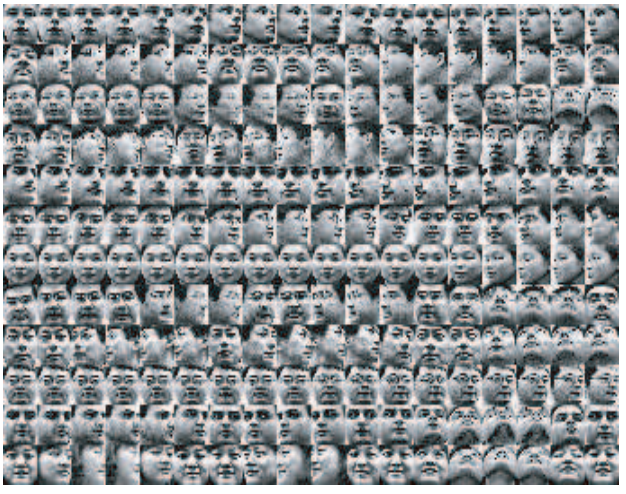

Figure 6: Sample gallery videos used in the experiments. Note the pose variation changed is rather large in this data set.

We performed numerous experiments and compared the proposed algorithm with other methods in the context of video-based recognition. Since there is no standard database that contains large 2-D and 3-D head rotation for video-based face recognition, we collected a set of 45 videos of 20 different people for experiments (This data set will be made available to the vision community in the near future.). Each individual in our database has at least two videos where each person moves in a different combination of 2-D and 3-D rotation, expression, and speed. Each video was recorded in an indoor environment and each one lasted for at least 20 seconds (with 30 color frames of 640×480 pixels per second). Some cropped frames from the videos are shown in Figure 6. A variant of the eigen-subspace tracker [2] was used to locate the face, and the results were inspected by humans. Each image was then downsampled to 19×19 pixels for computational efficiency.

To reduce the effect of misalignment caused by the tracker, we added small 2-D perturbations including translation (within 2 pixels in all directions), and scaling (within a scale from 0.9 to 1.1), to enlarge the training sets before applying the proposed probabilistic algorithm.

We evaluated the proposed algorithm on two sets of videos: one without any occlusion and one with partial occlusion. The overall recognition rate in the experiments is defined by the number frames where the identity is correctly recognized divided by the number of frames in all the test videos.

4.1 Number of Linear PCA Planes

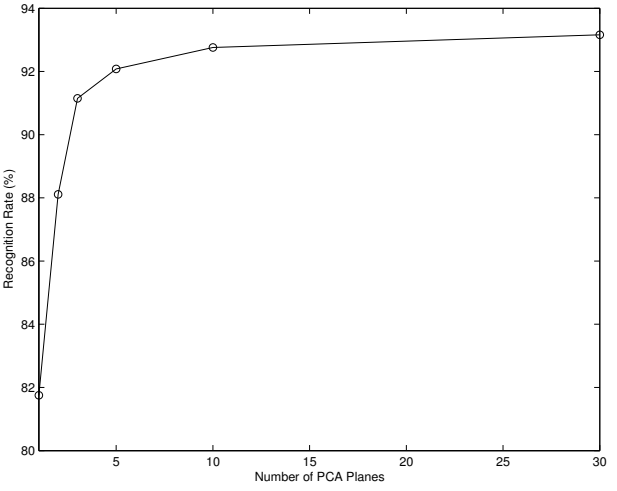

Figure 7: Recognition rate vs. number of piecewise linear PCA planes of our method. It shows that the proposed method is rather robust to parameter selection (i.e., the number of pose manifolds used in approximating appearance manifold.)

We first evaluate the proposed algorithm in the test set without occlusion, and analyze the number of PCA planes required to construct appearance manifolds yielding good recognition results. Figure 7 demonstrate that the average recognition rate does not change much when the number of PCA planes is varied from 5 to 30. The results suggest that the appearance manifold can be effectively approximated with a small number of PCA planes. The proposed algorithm performs well over a reasonably large range which shows that one can easily pick an appropriate number of PCA planes. Obviously, a smaller number of PCA planes is preferable for computational efficiency reasons. However, the recognition rate drops significantly and quickly when the number of manifolds is rather small (fewer than five for this data set). This is consistent with the claim that the appearance manifold is nonlinear and complex.

4.2 Transition Matrix $P(C^{ki}|C^{kj})$

In this set of experiments, we demonstrate that the transition matrix, $P(C^{ki}|C^{kj})$, in the proposed method capture the image dynamics sufficiently to improve recogni-

COMPARISON OF TEMPORAL STRATEGIES	
Temporal Strategy	Accuracy (%)
Proposed Method	92.1
Temporal Voting	84.2
Uniform Trans.	85.0

Table 1: Recognition results using various temporal strategies on a test set of videos without occlusion.

tion rates. Using the set of videos without occlusion, we compared our method with two different strategies, temporal voting and a uniform transition probability scheme. All three methods used the same number of manifolds for each person $m = 5$; they differ in their way of utilizing temporal information. The temporal voting scheme, commonly used in recognition methods is based on multiple frames, makes an identity decision by taking votes of the results of the previous f frames. In this case, 20 frames were used. The uniform transition scheme simply sets all the entries of transition matrix to 1, which means that no temporal dynamics are learned or utilized in the recognition process. The experimental results, shown in Table 1, demonstrate that our method outperforms others by a significant margin. In other words, learning transition probabilities among the pose manifolds does facilitate recognition which cannot be achieved by method using no dynamics information or a simple temporal voting scheme with a large window size.

4.3 Comparison with Single Frame Algorithms and the Effect of Occlusion

COMPARISON OF RECOGNITION METHODS		
Method	Accuracy (%)	
	Videos w/o occlusion	Videos with occlusion
Proposed Method	93.2	93.0
Ensemble of LPCA	82.2	20.9
Eigenface	75.5	28.4
Fisherface	75.4	20.5

Table 2: Recognition results using different methods. The results are based on the average recognition rates achieved by each method.

For completeness, we compared our method with several frame-based face recognition algorithms in the literature, and the results are shown in Table 2. All methods were trained with the exact same cropped images. We constructed 30 PCA planes and learn their dynamics from the training videos in the proposed algorithm. For the Ensemble of LPCA method, we used the same 30 PCA planes constructed in the proposed method but did not use the learned transition matrix. This method is, in spirit, similar to the view-based Eigenface method [22]. The dimensionality of Fisherface method is set to 19 (i.e., the number of classes minus 1) and the dimensionality for other methods is empirically set to 30. Though it may not seem to be fair to compare video-based and frame-based recognition algorithms, these baseline experiments suggest that frame-based methods may not work well in an unconstrained environment where there are large pose changes. For the test videos without occlusion, the Ensemble of LPCA method performs better than classic linear models (Eigenface and Fisherface methods) because an image sequence usually contain 2-D and 3-D rotations, which can not be effectively approximated by a global linear model. These results also show that the use of image dynamics by our method greatly helps face recognition in video. Except for the proposed method, all other methods performed poorly on the test videos where some faces were partially occluded. This result shows that appearance coherence between consecutive frames helps in predicting occlusions and in turn facilitates the recognition process.

5 Conclusion and Future Work

We have presented a novel framework for video-based face recognition. The proposed method builds an appearance manifold which is approximated by piecewise linear subspaces and the dynamics among them embodied in a transition matrix learned from an image sequence. It is worth noticing that the image sequences considered in this paper contains large 2-D and 3-D rotations as well as partial occlusions. These situations might occur in many vision-based human-computer interaction or surveillance applications. As experimentally demonstrated, our method approximates nonlinear appearance manifold well and achieves good recognition rates in video-based face recognition. Though the proposed model handles large motions well, it is nevertheless sensitive to large illumination changes, and our future work will address this.

Acknowledgments

Support of this work was provided by Honda Research Institute, and the National Science Foundation CCR 00-86094 and IIS 00-85980. This work was carried out at Honda Research Institute. We would like to thank the anonymous reviewers for their comments and suggestions, and all the people who help to record their faces in our video database.

References

[1] C. M. Bishop and J. M. Winn. Non-linear Bayesian image modelling. In *Proc. European Conf. on Computer Vision*, volume 1, pages 3–17, 2000.

[2] M. J. Black and A. D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *Int'l. J. Computer Vision*, 26(1):63–84, 1998.

[3] C. Bregler and S. Omohundro. Surface learning with applications to lipreading. In *Advances in Neural Information Processing Systems*, pages 43–50, 1994.

[4] R. Chellappa, C. L. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5):705–740, 1995.

[5] T. Cootes, C. J. Taylor, D. Cooper, and J. Graham. Active shape models - Their training and application. *Computer Vision and Image Understanding*, 61:38–59, 1995.

[6] D. DeCarlo, D. Metaxas, and M. Stone. An anthropometric face model using variational techniques. In *Proc. SIGGRAPH*, pages 67–74, 1998.

[7] G. J. Edwards, C. J. Taylor, and T. F. Cootes. Interpreting face images using active appearance models. In *Proc. IEEE Int'l. Conf. on Automatic Face and Gesture Recognition*, pages 300–305, 1998.

[8] G. J. Edwards, C. J. Taylor, and T. F. Cootes. Improving identification performance by integrating evidence from sequence. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 486–491, 1999.

[9] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.

[10] D. Hochbaum and D. Shmoys. A best possible heuristic for the k-center problem. *Mathematics of Operations Research*, 10:180–184, 1985.

[11] X. Hou, S. Li, H. Zhang, and Q. Cheng. Direct appearance models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 828–833, 2001.

[12] A. J. Howell and H. Buxton. Towards unconstrained face recognition from image sequences. In *Proc. IEEE Int'l. Conf. on Automatic Face and Gesture Recognition*, pages 224–229, 1996.

[13] M. Isard and A. Blake. A mixed-state Condensation tracker with automatic model-switching. pages 107–112, 1998.

[14] T. Jebara, K. Russell, and A. Pentland. Mixtures of eigen features for real-time structure from texture. In *Proc. Int'l. Conf. on Computer Vision*, pages 128–135, 1998.

[15] V. Krüeger and S. Zhou. Exemplar-based face recognition from video. In *Proc. European Conf. on Computer Vision*, volume 4, pages 732–746.

[16] Y. Li, S. Gong, and H. Liddell. Constructing facial identity surface in a nonlinear discriminating space. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 258–263, 2001.

[17] Y. Li, T. Wang, and H.-Y. Shum. Motion textures: A two-level statistical model for character motion synthesis. In *Proc. SIGGRAPH*, pages 465–472, 2002.

[18] B. Moghaddam and A. Pentland. Probabilistic visual learning for object recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):696–710, 1997.

[19] H. Murase and S. K. Nayar. Visual learning and recognition of 3-D objects from appearance. *Int'l. J. Computer Vision*, 14:5–24, 1995.

[20] B. North, A. Blake, M. Isard, and J. Rittscher. Learning and classification of complex dynamics. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(9):1016–1034, 2000.

[21] V. Pavlović, J. M. Rehg, T. J. Cham, and K. P. Murphy. A dynamic Bayesian network approach to figure tracking using learned dynamic models. In *Proc. Int'l. Conf. on Computer Vision*, pages 94–101, 1999.

[22] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 1994.

[23] S. Romdhani, V. Blanz, and T. Vetter. Face identification by fitting 3D morphable model using linear shape and texture error functions. pages 3–19, 2002.

[24] A. Samal and P. A. Iyengar. Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern Recognition*, 25(1):65–77, 1992.

[25] A. Schödl, R. Szeliski, D. H. Salesin, and I. Essa. Video textures. In *Proc. SIGGRAPH*, pages 489–498, 2000.

[26] G. Shakhnarovich, J. W. Fisher, and T. Darrell. Face recognition from long-term observations. In *Proc. European Conf. on Computer Vision*, volume 3, pages 851–865, 2002.

[27] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *Proc. Int'l. Conf. on Computer Vision*, volume 2, pages 50–59, 2001.

[28] H. Wechsler, V. Kakkad, J. Huang, S. Gutta, and V. Chen. Automatic video-based person authentication using the RBF network. In *Proc. Int'l. Conf. on Audio and Video-Based Biometric Person Authentication*, pages 177–183, 1997.

[29] W. Y. Zhao and R. Chellappa. Symmetric shape-from-shading using self-ratio image. *Int'l. J. Computer Vision*, 45(1):55–75, 2001.

[30] W. Y. Zhao, R. Chellappa, A. Rosenfeld, and J. P. Phillips. Face recognition: A literature survey. Technical Report CAR-TR-948, Center for Automation Research, University of Maryland, 2000.

[31] S. Zhou and R. Chellappa. Probabilistic human recognition from video. In *Proc. European Conf. on Computer Vision*, volume 3, pages 681–697, 2002.