IEICE TRANS. INF. & SYST., VOL.E92–D, NO.4 APRIL 2009

PAPER

Incremental Unsupervised-Learning of Appearance Manifold with View-Dependent Covariance Matrix for Face Recognition from Video Sequences

Lina[†a)], ***Student Member***, **Tomokazu TAKAHASHI**[††], **Ichiro IDE**[†], ***and*** **Hiroshi MURASE**[†], ***Members***

SUMMARY We propose an appearance manifold with view-dependent covariance matrix for face recognition from video sequences in two learning frameworks: the supervised-learning and the incremental unsupervised-learning. The advantages of this method are, first, the appearance manifold with view-dependent covariance matrix model is robust to pose changes and is also noise invariant, since the embedded covariance matrices are calculated based on their poses in order to learn the samples' distributions along the manifold. Moreover, the proposed incremental unsupervised-learning framework is more realistic for real-world face recognition applications. It is obvious that it is difficult to collect large amounts of face sequences under complete poses (from left sideview to right sideview) for training. Here, an incremental unsupervised-learning framework allows us to train the system with the available initial sequences, and later update the system's knowledge incrementally every time an unlabelled sequence is input. In addition, we also integrate the appearance manifold with view-dependent covariance matrix model with a pose estimation system for improving the classification accuracy and easily detecting sequences with overlapped poses for merging process in the incremental unsupervised-learning framework. The experimental results showed that, in both frameworks, the proposed appearance manifold with view-dependent covariance matrix method could recognize faces from video sequences accurately.

key words: *appearance manifold, view-dependent covariance matrix, incremental learning, video-based face recognition, eigenspace*

1. Introduction

Face recognition has long been an active area of research. Over the years, numerous works have been proposed that focus on recognizing 3D objects and human faces from still-images, such as [1]–[11]. Recently, video-based face recognition has attracted much attention since face recognition using video presents various advantages and also challenges over still-image based recognition.

For an efficient video-based face recognition process, first, every image (frame) in a video sequence is input in a feature extraction module, so that it becomes a low-dimensional vector in a feature space. One most widely used feature extractor in the pattern recognition field is the Principal Component Analysis (PCA) with its eigenspace representation [1], [2]. In a video, face images may vary significantly due to environmental changes, such as lighting condition, pose, facial expression, etc. In addition, various degradation effects might also influence the images in a video sequence, such as low-quality video and cropping errors due to inaccuracies of a tracking system. Therefore, a robust recognition system should be able to handle various image variations.

It is well known that an appearance manifold could capture image variations, especially pose changes, in eigenspace. Addressing various problems, many appearance manifold models have been proposed, such as the simple manifold [3], [4], the probabilistic appearance manifold [5], [6], the layer-transparent manifold [9], etc. Moreover, in our previous work, we have introduced various models of appearance manifold with view-dependent covariance matrix which could robustly recognize 3D objects from still-images under various degradation effects [11].

In the past years, several works have reported the use of appearance manifold for face recognition from video sequences. Among them, Raytchev and Murase [12] proposed a pairwise clustering method which calculates the interaction levels (attraction and repulsion) of every input sequence. A merging or splitting process is then applied according to the interaction's decision. Zhou et al. proposed a probabilistic approach which uses joint posterior distribution of motion vectors and estimates the temporal information of video sequences by propagating the identity variables over time [13]. Similar to [13], the work of Lee et al. in [14], [15], utilized local linear models and a transition matrix which propagates the probabilistic likelihood of the linear models to capture the temporal information. Meanwhile, Liu et al. processed the temporal information of video sequences using the adaptive Hidden Markov Models (HMM) method [16].

Proposing a different appearance manifold model from the previous works, the novelty of our approach lies in the scheme of embedding view-dependent covariance matrices in appearance manifolds for recognizing faces from video sequences in an incremental unsupervised-learning framework. The advantages of this model are, first, the appearance manifold with view-dependent covariance matrix model is robust to pose changes and also noise invariant since the embedded covariance matrices are calculated based on their poses in order to learn the samples' distributions along the manifold. Moreover, the proposed incremen-

Manuscript received July 10, 2008.
Manuscript revised December 3, 2008.

[†]The authors are with the Department of Media Science, Graduate School of Information Science, Nagoya University, Nagoya-shi, 464–8601 Japan.
[††]The author is with the Faculty of Economics and Information, Gifu Shotoku Gakuen University, Gifu-shi, 500–8288 Japan.
a) E-mail: lina@murase.m.is.nagoya-u.ac.jp

DOI: 10.1587/transinf.E92.D.642

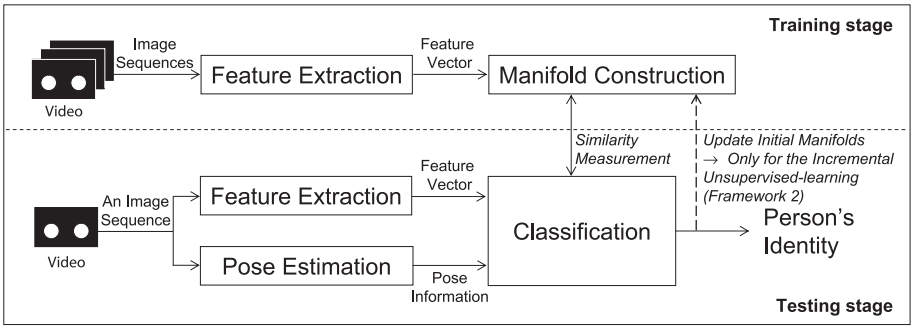


Fig. 1 Outline of the proposed video-based face recognition system.

tal unsupervised-learning framework is more realistic for a real-world face recognition application. It is obvious that it is difficult to collect large amounts of face sequences under complete poses (from left sideview to right sideview) for training purpose. Here, an incremental unsupervised-learning framework allows us to train the system with the available initial sequences, and later update the system's knowledge incrementally every time an unlabelled sequence is input.

In addition, the integration of a pose estimation system in the appearance manifold with view-dependent covariance matrix model also plays an important role in improving the classification accuracy, since, put in the terminology in [12], the images of the same object under a different viewing condition is different. Thus, it is clear that one is likely to obtain the classification results based on the availability (similarity) of a pose instead of its identity. Therefore, to increase the classification accuracy of the system, the most similar pose of a face in each manifold is searched (i.e. using a pose estimation system), before identifying the person. The pose estimation system also helps the merging process in the incremental unsupervised-learning framework to easily detect video sequences with overlapped poses.

In this paper, we address a video-based face recognition problem using the appearance manifold with view-dependent covariance matrix in two learning frameworks: the supervised-learning and the incremental unsupervised-learning. Figure 1 shows the outline of the proposed video-based face recognition system. In the supervised-learning framework, the training samples are labelled and used to represent the identity categories in the form of appearance manifolds with view-dependent covariance matrices. Meanwhile, in the incremental unsupervised-learning framework, the system first learns the initial categories through the initial manifolds and later updates its knowledge on identity categories by learning the unlabelled input sequences incrementally. Since the structure and the number of the category (class) changes every time a new pattern comes into the system, it is necessary to update the system's knowledge, either by constructing a new category or by modifying the structure of the existing categories through merging processes between sequences which have some overlapped poses with strong similarities.

We organized the rest of this paper as follows. Our appearance manifold model and the classification process for video-based face recognition in a supervised-learning framework is described in Sect. 2. In Sect. 3, the detailed classification process of video-based face recognition in an incremental unsupervised-learning framework is presented. Experimental results and discussions are described in Sects. 4 and 5, respectively. Finally, we concluded this paper and described the future works in Sect. 6.

2. Video-Based Face Recognition in a Supervised-Learning Framework (Framework 1)

Typically, a supervised video-based face recognition system consists of two stages: training and classification. In the training stage, a feature extraction module finds the appropriate features for representing the input patterns and the appearance manifolds are constructed to model the appearance variation of each object. Here, since the objects are human faces, the construction results of the appearance manifolds are called "face manifolds". Meanwhile, in the classification stage, the classifier assigns the unlabelled input sequence to one of the pattern categories which has the highest similarity based on a distance measurement. The detailed procedures of each module are described in Sects. 2.1 and 2.2.

2.1 Construction of Face Manifolds with View-Dependent Covariance Matrices in Eigenspace

In the pattern recognition field, it is well known that appearance-based approaches use sets of images in various poses to represent an object. Here, the role of a feature extraction module is to determine a low dimensional pattern representation compared with the image space. One widely used feature extractor is the Principal Component Analysis (PCA) [1], [2], that computes k-eigenvectors with the largest corresponding-eigenvalues to project the training samples onto an eigenspace. The linear transformation of the eigenspace representation is defined as:

$$\mathbf{q}_l^{(p)}(\theta) = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k]^T (\mathbf{x}_l^{(p)}(\theta) - \mathbf{c}) \tag{1}$$

where $\mathbf{x}_l^{(p)}(\theta)$ is the l-th sample image of person p with pose θ, $\mathbf{e}_i$ $(i = 1, 2, \cdots, k)$ is the eigenvector, $\mathbf{c}$ is the mean vector

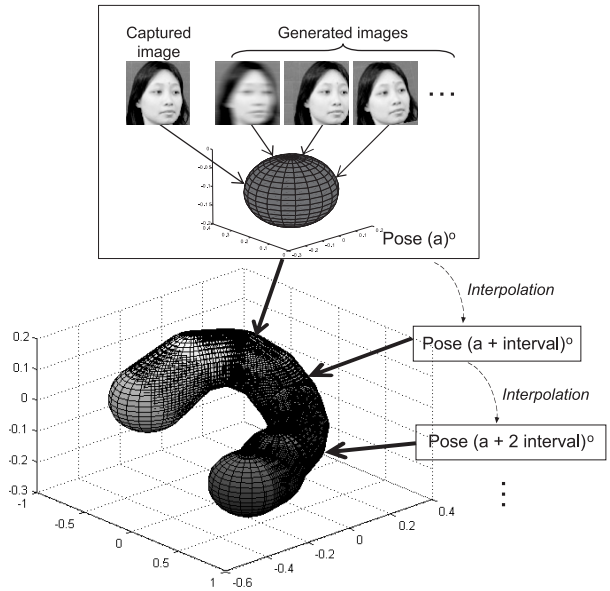


Fig. 2 Construction process of a face manifold with view-dependent covariance matrices.

of the training samples, and $\mathbf{q}_l^{(p)}(\theta)$ is the vector representation of image $\mathbf{x}_l^{(p)}(\theta)$ in the eigenspace. Note that the eigenvectors $\mathbf{e}_i$ in Eq. (1) are used only for image projections into the eigenspace, thus, are processed globally regardless to their poses. Meanwhile, later in the construction process of the face manifolds, the eigenvectors and the eigenvalues are view-dependent, since they are derived from the covariance matrix of each training-pose.

The construction process of a face manifold with view-dependent covariance matrices is shown in Fig. 2. The input for the construction process is $\mathbf{x}_l^{(p)}(\theta)$ which is the l-th sample image of person p with training pose θ. Next, the construction process of a face manifold with view-dependent covariance matrices consists of two steps:

Step 1. Calculation of covariance matrices: For each training pose, a mean vector and a covariance matrix is calculated. For this purpose, new images are generated by adding noise to each image in video-captured sequences. The type, level and number of the artificial noise are not limited and could be applied freely in various forms, such as geometric distortion (i.e. shift, rotation, etc.), quality degradation (i.e. blur, salt and pepper noise, etc.), illumination changes, etc.

Step 2. Interpolation of covariance matrices: In order to obtain the mean vectors and the covariance matrices of the untrained poses, interpolation processes are performed to each pair of mean vectors and covariance matrices of two consecutive training poses. For the mean vectors, the interpolation process is done by simply using one of the several existing interpolation algorithms. Meanwhile, the interpolation of the covariance matrices is done by interpolating the corresponding eigenvectors and eigenvalues of two consecutive training poses (see the VCEI method [11] for details).

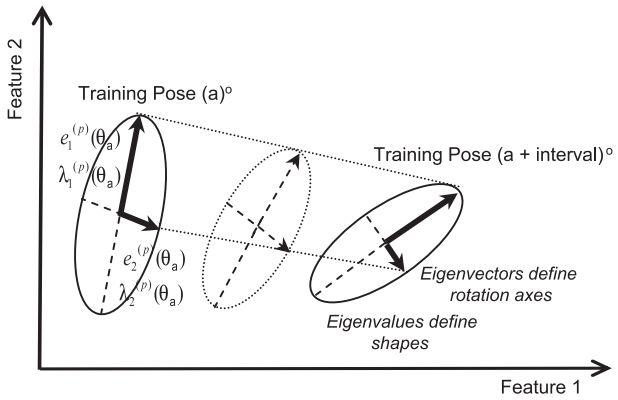


Fig. 3 Interpolation of the eigenvectors and the eigenvalues in a feature space.

A brief description of the eigenvectors and eigenvalues interpolation [11] is described as follows.

First, the matrices of eigenvectors $\mathbf{E}_0$ and $\mathbf{E}_1$ which consist of pairs of eigenvectors $\mathbf{e}_{0j}$ and $\mathbf{e}_{1j}$ $(j = 1, 2, \cdots, k)$ and the matrices of eigenvalues which consist of pairs of eigenvalues λ_{0j} and λ_{1j} $(j = 1, 2, \cdots, k)$ of covariance matrices $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$ are formed. Next, in order to correspond the axes, the eigenvectors of $\mathbf{E}_0$ and $\mathbf{E}_1$ are sorted based on their eigenvalues $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$ to form $\mathbf{E'}_0$ and $\mathbf{E'}_1$, respectively. The same task for the eigenvalues are then performed to form $\boldsymbol{\lambda}'_0$ and $\boldsymbol{\lambda}'_1$ from $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$. Then, check and invert if the eigenvectors satisfy $\mathbf{e'}_{0j}{}^T\mathbf{e'}_{1j} < 0$ so that the angle between corresponded axes is less than or equal to $\pi/2$.

For covariance matrix $\mathbf{\Sigma}_x$, the eigenvalues can be calculated by $\lambda_{xj} = \left((1-x)\sqrt{\lambda'_{0j}} + x\sqrt{\lambda'_{1j}}\right)^2$ $(j = 1, 2, \cdots, k)$ and $\mathbf{E}_x = \mathbf{R}(x\boldsymbol{\phi})\mathbf{E'}_0$ for the eigenvectors. Here, $\mathbf{R}$ represents an interpolated rotation when $0 \le x \le 1$ and $\boldsymbol{\phi} = [\phi_1, \cdots, \phi_m]$ represents the parameter vector of rotation angles to define the rotation matrix. Since the rotation angles always come in pairs in the complex conjugate roots process, then $m = \lfloor k/2 \rfloor$.

The rotation matrix is defined by $\mathbf{R}(\boldsymbol{\phi}) = \mathbf{E'}_1\mathbf{E'}_0{}^T$ and diagonalized with the Special Orthogonal (SO) rule by $\mathbf{R}(\boldsymbol{\phi}) = \mathbf{U}\mathbf{D}(\boldsymbol{\phi})\mathbf{U}^\dagger$ where $\mathbf{U}^\dagger$ represents a complex conjugate transpose matrix of $\mathbf{U}$. The complex conjugate roots are then processed by $\mathbf{D}(\boldsymbol{\phi}) = diag(\mathbf{e}^{i\phi_1}, \mathbf{e}^{-i\phi_1}, \cdots, \mathbf{e}^{i\phi_m}, \mathbf{e}^{-i\phi_m})$ if $n = 2m$. Meanwhile, if $n = 2m + 1$, then $\mathbf{D}(\boldsymbol{\phi}) = diag(1, \mathbf{e}^{i\phi_1}, \mathbf{e}^{-i\phi_1}, \cdots, \mathbf{e}^{i\phi_m}, \mathbf{e}^{-i\phi_m})$ where $\mathbf{e}^{i\phi} = \cos\phi + i\sin\phi$. Finally, interpolate the rotation matrix $\mathbf{R}(x\boldsymbol{\theta})$ and calculate the covariance matrix for the untrained poses using $\mathbf{\Sigma}_x = \mathbf{E}_x\mathbf{\Lambda}_x\mathbf{E}_x{}^T$ with $\mathbf{\Lambda}_x = diag(\lambda_x)$. Figure 3 shows the illustration of the interpolation of the eigenvectors and the eigenvalues in a 2D feature space.

The output of the construction process of the face manifold with view-dependent covariance matrix are the mean vectors $\boldsymbol{\mu}^{(p)}(\theta)$ and the covariance matrices $\mathbf{\Sigma}^{(\mathbf{p})}(\theta)$ of the training poses and poses obtained by the interpolation.

2.2 Classification of Face-Sequences

In the testing stage, unlike still-image recognition, video-based recognition needs to integrate the classification results of each frame to produce the decision of a sequence. Given a face sequence $\mathbf{S} = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_h]$ in an eigenspace, the classification process of a face image $\mathbf{f}_i$ $(i = 1, 2, \cdots, h)$ is based on its similarity to the trained face manifolds M_p. Meanwhile, the final sequence's classification decision is based on the minimal cumulative distance of each $\mathbf{f}_i \in \mathbf{S}$.

In this paper, we also propose the integration of a pose estimation system to provide pose information of each test image $\mathbf{f}_i \in \mathbf{S}$ for improving the classification accuracy and also easily detecting sequences with overlapped poses for the merging process later in the incremental unsupervised-learning framework. Here, various existing algorithms can be selected for developing the pose estimation system, since it is fully independent from the classification system. In this paper, the pose estimation system is based on the Nearest Neighbor algorithm which basically classifies an input feature vector to a class with the nearest distance in a feature space.

For classification purpose, first, a pose vector $\boldsymbol{\theta}^{(p)} = [\theta_1, \theta_2, \cdots, \theta_g]$ which consist of g training poses of a face manifold M_p is constructed. Then, the classification process is defined as follows:

$$\varphi_i = \text{pose\_estimation}(\mathbf{f}_i) \tag{2}$$

Once the pose φ_i of each unlabelled input image $\mathbf{f}_i$ $(i = 1, 2, \cdots, h)$ is determined by a pose estimation system, the normalization of the test image can be calculated by:

$$\mathbf{f'}_i = \mathbf{f}_i - \boldsymbol{\mu}^{(p)}(\varphi_i) \tag{3}$$

and the distance measurement of the input image is defined by:

$$df_i^{(p)} = \begin{cases} (\mathbf{f'}_i)^T \left(\boldsymbol{\Sigma}^{(p)}(\varphi_i)\right)^{-1} (\mathbf{f'}_i) & (\varphi_i \in \boldsymbol{\theta}^{(p)}) \\ 0 & (\text{otherwise}) \end{cases} \tag{4}$$

Equation (4) shows that the Mahalanobis distance is calculated only when the pose φ_i of an input image existed as a member of pose vector $\boldsymbol{\theta}^{(p)}$ of manifold M_p (defined as an overlapped pose). On the other hand, when the pose of the input image φ_i is not a member of $\boldsymbol{\theta}^{(p)}$, a zero value is given to $df_i^{(p)}$ as the distance of the input image $\mathbf{f_i}$ to the manifold M_p.

Finally, the classification decision of an input sequence $\mathbf{S}_i$ is made upon integrating the classification results of all input images $\mathbf{f}_i \in \mathbf{S}$. The identity p^* is determined by finding the manifold M_p with the minimal cumulative distance of all $\mathbf{f}_i \in \mathbf{S}$, as follows:

$$p^* = \arg\min_p \left(\sum_{i=1}^{h} df_i^{(p)} \Big/ Noo^{(p)} \right) \tag{5}$$

where $Noo^{(p)}$ is the number of overlapped poses between the input sequence $\mathbf{S}$ with manifold M_p.

3. Video-Based Face Recognition in an Incremental Unsupervised-Learning Framework (Framework 2)

Considering the practical interest, a face recognition system is expected to deal with unconstrain, dynamic and unpredictable environments. The system should be able to correctly identify every input and learn it to update the system's knowledge on identity categories (persons). The purpose of the incremental unsupervised-learning introduced in this paper is to assign a set of unlabelled face sequences into their corresponding identity categories (persons) in an unsupervised manner. The input sequence can be assigned to one of the existing categories or as a new identity category. The major difficulty of this approach is in finding the proper balance between not to overlook the existing category structure and at the same time not to superimpose a new structure. Our proposed approach is based on central clustering which classifies an unlabelled input sequence into an identity category according to its similarity to the central feature of each category with prior guidance of a pose estimation system. Figure 4 shows the summary of the identity classification (clustering) algorithm in an incremental unsupervised-learning framework, while the detailed processes are described in the following sections.

3.1 Identity Classification (Clustering)

The classification process of a face sequence $\mathbf{f}_i$ $(i = 1, 2, \cdots, h)$ in a sequence $\mathbf{S} = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_h]$ starts by measuring the similarity of each $\mathbf{f}_i \in \mathbf{S}$ using Eq. (4). Next, a threshold value β is defined in order to determine whether an input sequence is classified to one of the existing categories

Initialization:

$\mathbf{S} = [\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_h]$ = a face sequence consists of h images
$p = (1, 2, ..., P)$ = the existing identity categories
$df_i^{(p)}$ = the distance of image $\mathbf{f}_i$ to manifold M_p
$noo^{(p)}$ = the number of overlapped poses between $\mathbf{S}$ with manifold M_p
β = a threshold value

Begin

If satisfy the condition $df_i^{(p)} < \beta \quad (i = 1, ..., h)$

Then

Accepted:
1. Identify a sequence $\mathbf{S}$ according to the most similar identity category p
$$p^* = \arg\min_p \left(\sum_{i=1}^{h} df_i^{(p)} \Big/ noo^{(p)} \right)$$
2. Construct the manifold of $\mathbf{S}$ (M_s)
3. Merge parts of M_s and M_{p^*} which have overlapped poses

Else

Rejected:
1. Identify the sequence $\mathbf{S}$ as a new identity class
$$p^* = P + 1$$
2. Construct the manifold of $\mathbf{S}$ (M_s)

End

Fig. 4 Classification algorithm of an input sequence in an incremental unsupervised-learning framework.

(*accepted*) or assigned as a new identity category (*rejected*).

An input sequence is *accepted* being classified to one of the existing category if the distance of every image in the input sequence and its total distance are less than the threshold value β. Otherwise, the input sequence is *rejected* and assigned as a new identity category. The identity category is determined as follows:

$$p^* = \begin{cases} \arg\min_p \left(\sum_{i=1}^{h} df_i^{(p)} \Big/ Noo^{(p)} \right) & \text{(Accepted)} \\ P+1 & \text{(Rejected)} \end{cases} \quad (6)$$

The next process after the identity classification is to construct the face manifold of the input sequence in order to update the system's knowledge. Here, the face manifold of the input sequence is constructed using the same model with the initial face manifolds by embedding the view-dependent covariance matrix (see the construction process in Sect. 2.1). Moreover, for each *accepted* result, it is also necessary to perform a merging process between the manifold of the input sequence with the existing manifolds which have the same identity category. However, the manifold merging process is not performed for any *rejected* result. The details of the manifold merging process is described in the next section.

3.2 Manifold Merging

In the manifold merging process, only the overlapped parts of two face manifolds with view-dependent covariance matrix will be merged. Therefore, it is necessary to determine the overlapped parts of both manifolds by detecting the overlapped poses ω_i. Fortunately, the detecting process of the overlapped poses ω_i can be performed easily by the help of an integrated pose estimation system.

The merging process of two face manifolds with view-dependent covariance matrix is done as follows. First, the mean vectors $\boldsymbol{\mu}_i^{(p^*)}(\omega_i)$ of overlapped poses ω_i are merged through:

$$\boldsymbol{\mu}_{i,updated}^{(p^*)}(\omega_i) = \alpha \boldsymbol{\mu}_{i,old}^{(p^*)}(\omega_i) + (1-\alpha)\boldsymbol{\mu}_{i,new}^{(p^*)}(\omega_i) \quad (7)$$

Next, the covariance matrices $\boldsymbol{\Sigma}_i^{(p^*)}(\omega_i)$ are merged using:

$$\boldsymbol{\Sigma}_{i,updated}^{(p^*)}(\omega_i) = \alpha \boldsymbol{\Sigma}_{i,old}^{(p^*)}(\omega_i) + (1-\alpha)\boldsymbol{\Sigma}_{i,new}^{(p^*)}(\omega_i) \quad (8)$$

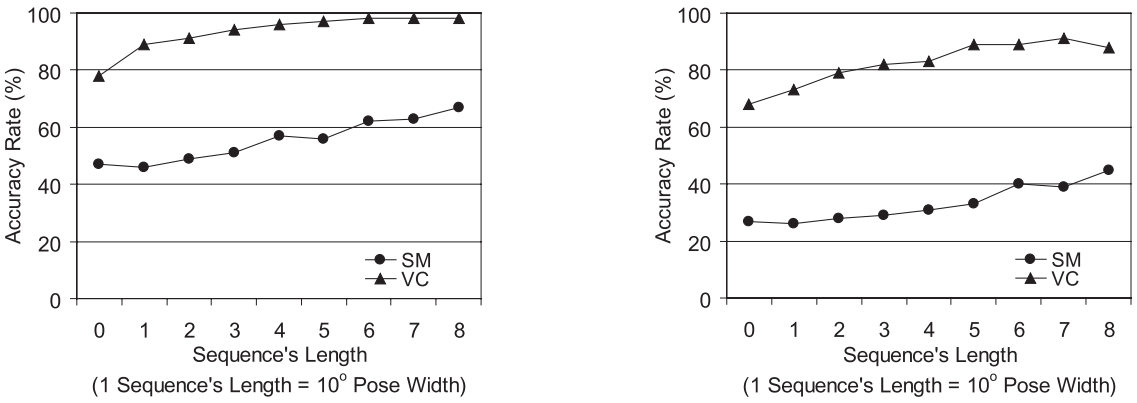
where α is an updating weight value, $\boldsymbol{\mu}_{i,old}^{(p^*)}(\omega_i)$ and $\boldsymbol{\Sigma}_{i,old}^{(p^*)}(\omega_i)$ are the mean vectors and the covariance matrices of the existing manifold p^* with overlapped poses ω_i. $\boldsymbol{\mu}_{i,new}^{(p^*)}(\omega_i)$ and $\boldsymbol{\Sigma}_{i,new}^{(p^*)}(\omega_i)$ are the mean vectors and the covariance matrices of the new constructed (input) manifold with overlapped poses ω_i.

4. Experiments and Analysis

We conducted several experiments to evaluate the performance of the proposed method in recognizing human faces from video sequences. For the experiments, we have collected 60 motion videos of 20 persons with pose changes from $-90°$ (left sideview) to $+90°$ (right sideview) from the frontal pose. For each person, three motion videos are taken in a different time under different conditions and are represented in three datasets. In the preprocessing step, first, the motion videos were trimmed with a frame rate of 30 frames/second, and images from a video sequence with $10°$ pose differences from each other were taken as a face sequence. Here, the sampling processes of the face sequences were performed in order to obtain a same frame-density condition within the face sequences, which is useful for fair evaluations of methods. However, in a real system, the sampling process is not necessary. Next, each image of the face sequences were manually cropped and downsampled to 32×32 pixels image size, however, in a real application system, a face tracking (cropping) system may be used to automatically detect and crop the face images.

Figure 5 shows the samples of face sequences of four persons from three datasets (Dataset 1: small face variations, Dataset 2: small face variations taken in a different time, and Dataset 3: severe face variations). It is clearly seen in Fig. 5 that the datasets contain many instances of noise data, in the form of pose variations, natural expression variations, and erroneous in face cropping (misalignments). In the experiments, face sequences of Dataset 1 are used for training, while face sequences of Dataset 2 and Dataset 3 are used as testing data.

Fig. 5 Samples of face sequences; for each person: first row: Dataset 1 (small face variations), second row: Dataset 2 (small face variations taken in a different time), third row: Dataset 3 (severe face variations).



(a) 10 identity categories (persons) task

(b) 20 identity categories (persons) task

Fig. 6 Accuracy rates in recognizing faces from video sequences of Dataset 2 with various sequence's lengths in a supervised-learning framework.

The experiments were conducted in two frameworks: supervised-learning and incremental unsupervised-learning, in which each of them has varying degrees of difficulty. In this paper, we compared the results of our proposed appearance manifold with view-dependent covariance matrix (VC) method with the simple manifold (SM) method (known as the Parametric Eigenspace method in [4]). Both of these methods used a manifold to capture pose variabilities of a face. However, in the Simple Manifold (SM) method, an appearance manifold was constructed based on the interpolations of the mean vectors of samples and used an identity matrix as the covariance matrix for each pose. Therefore, the SM method can only capture the pose changes. Meanwhile, in the proposed VC method, a view-dependent covariance matrix was also embedded to the appearance manifold. Therefore, the proposed VC method has the abilities to capture pose variability and also learn the sample's distribution of each pose. Here, both the VC and the SM methods were also combined with a pose estimation system using the same Nearest Neighbor algorithm. The experimental results are presented in the next sections.

4.1 Experiments in the Supervised-Learning Framework (Framework 1)

In the first experiment in the supervised-learning framework, we trained the system with 26 face sequences for each of the 10 different persons in Dataset 1. Among these 26 face sequences, only one sequence was captured by a motion video, while the other 25 sequences were generated by applying noise effects to the video-captured sequence. The noise effects we applied in this experiment are the motion blur effects which usually occur in the capturing process of moving objects, and the shift and the rotation effects which represent the erroneous face croppings and misalignments. For the testing set, we partially took the face sequences from Dataset 2 and Dataset 3 (which are different from the training set) and arranged them into 200 partial face sequences for each of the 9 different sequence's lengths. Here, one sequence's length represents 10° pose width. For face sequences with zero lengths, the identity recognition tasks were performed based on the classification result of an input image in the sequence (similar to the still-image based recognition).

Figure 6 (a) shows the accuracy rates of the VC and the SM methods when recognizing faces with 10 identity categories from video sequences of Dataset 2 with various sequence's lengths in a supervised-learning framework. The results in Fig. 6 (a) shows that the proposed VC method gave higher recognition accuracies in all categories compared with that of the SM method. Here, the longer the sequence's length, the higher accuracies could be achieved by the system, since a long sequence could usually give more appearance variation information compared with a short sequence. The highest accuracy rate for the VC method was 98%, while the highest accuracy rate for the SM method was only 67%. For a 20 identity category recognition task, as depicted in Fig. 6 (b), the accuracies of the recognition system decreased compared with that of the 10 identity category recognition task. However, the proposed VC method still outperformed the SM method, with 88% highest recognition accuracy for the VC method, while the SM method only achieved 45% as its highest recognition accuracy.

Furthermore, we also conducted an experiment to recognize faces from video sequences of Dataset 3 which contains severe image variations. Table 1 summarizes the accuracy rates of the 10 identity category face recognition task from video sequences with 8 sequence's length in two conditions: small face variation and severe face variation. It could be well understood that the recognition accuracies of all methods decreased when recognizing sequences with severe face variations. However, the VC method could still maintain its superiority on the SM method. Under severe face variation conditions, the highest recognition accuracy achieved by the VC method was 75%, while the SM method gave only 67% accuracy as its highest recognition result. For reference purpose, we also presented the accuracy rates of the recognition system for each method when the pose in-

Table 1 Accuracy rates of a 10 identity categories recognition task from video sequences with 8 sequence's length in two conditions: small face variation and severe face variation in a supervised-learning framework (The reference accuracy values, obtained when using pose information given by a human, are presented in the brackets).

Method	Accuracy of Dataset 2 with Small Face Variations (%)	Accuracy of Dataset 3 with Severe Face Variations (%)
Simple Manifold (SM)	67 (73)	67 (60)
View-dependent Covariance Matrix (VC)	98 (98)	75 (88)

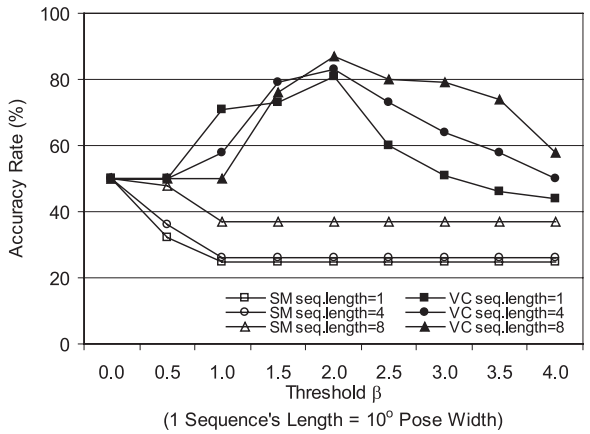


Fig. 7 Accuracy rates in recognizing faces from video sequences with various sequence's lengths using various threshold values.

formation was given by a human, as shown as values within brackets in Table 1.

All experimental results described above showed that the proposed VC method outperformed the SM method in the supervised-learning framework in various conditions, such as various numbers of identity category, various numbers of sequence's length, different levels of image variation (small or severe), and different pose estimation techniques (human estimator or system estimator).

4.2 Experiments in the Incremental Unsupervised-Learning Framework (Framework 2)

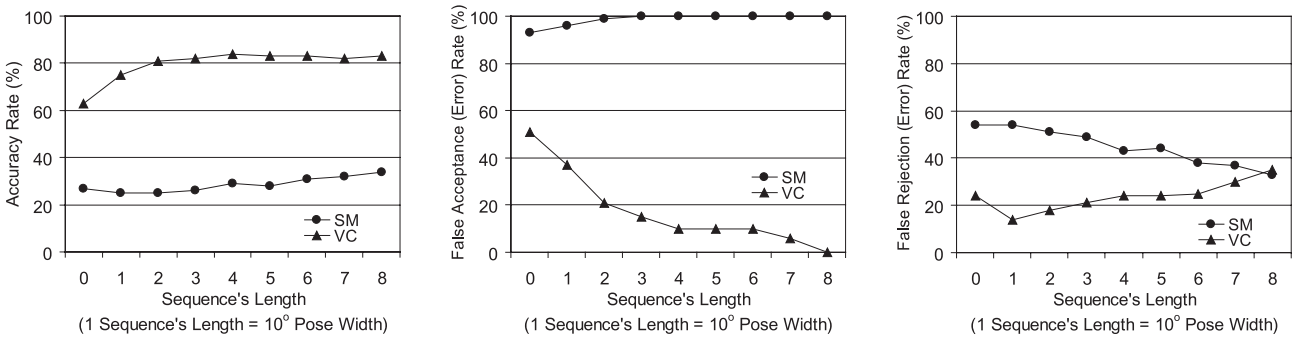
In the incremental unsupervised-learning framework, the system first learns the initial manifolds and later the existing manifolds are updated automatically by learning the sequences which are input incrementally into the system. Therefore, in this experiment, we first constructed 10 initial manifolds of 10 persons using 26 (1 original and 25 generated) face sequences of Dataset 1 per person. Next, several parameters such as a threshold value β (see Fig. 4) and a merging weight α (see Eq. (7) and Eq. (8)) are defined experimentally. Figure 7 shows the accuracy rates of the VC and the SM methods in recognizing short (1 sequence's length), medium (4 sequence's length), and long (8 sequence's length) sequences with a human pose estimator and using various threshold values. Based on the results in Fig. 7, we set $\beta = 2.0$ and $\alpha = 0.5$ as the optimal threshold and update weight values for our face database.

Figure 8 presents the classification rates for recognizing faces from video sequences in an incremental unsupervised-learning framework. The system was tested with 1,800 face sequences with various sequence's lengths which belong to 20 persons (10 initial persons and 10 new persons). For the experiments in the incremental unsupervised-learning framework, the evaluation parameters were the accuracy rate, the false acceptance rate, and the false rejection rate. The definitions of each evaluation parameter are defined as follows:

- Accuracy rate: the ratio of the number of classes that are correctly classified by the system to the total number of tests.
- False acceptance rate: the ratio of the number of pairs of different classes that are incorrectly matched by the system to the total number of match attempts.
- False rejection rate: the ratio of the number of pairs of the same class that are not matched by the system to the total number of match attempts.

In this paper, we presented each evaluation rate in a separate figure in Fig. 8 in order to give a clear presentation of the experimental results.

Figure 8 (a), Fig. 8 (b), and Fig. 8 (c) each show the accuracy rates, the false acceptance rates, and the false rejection rates of the VC and the SM methods when recognizing faces from video sequences in Dataset 2 with various sequence's lengths. The results in Fig. 8 (a) show that the proposed VC method gave higher accuracy rates compared with that of the SM method in all categories. Similar with the results in the supervised learning framework, the longer the face-sequences, the higher accuracy rates could be achieved by the system. For the error rates, it can be seen from Fig. 8 (b) that the false acceptance rates of the proposed VC method were lower than that of the SM method. Moreover, the false acceptance rates for the VC method decreased along with the increment of the sequence's length. On the contrary, the false rejection rates of the VC method increased along with the increment of the sequence's length. Tracing back the proposed identity classification (clustering) algorithm in Fig. 4 where the distance of every image in a face sequence should be less than the threshold value, it could be well understood that the difficulty level of fulfilling this criteria (being *accepted*) increases along with the increment of the sequence's length. On the other hand, the false acceptance rate of the SM method increased along the increment of the sequence's length, while the false rejection rate of the SM method decreased along the increment of the sequence's length. In more detail, the false acceptance rates for the SM were nearly 100% for all categories,



(a) Accuracy rates (b) False acceptance (error) rates (c) False rejection (error) rates

Fig. 8 Classification rates in recognizing faces from video sequences of Dataset 2 with various sequence's lengths in an incremental unsupervised-learning framework.

Table 2 Classification rates in recognizing faces from video sequences with 8 sequence's length in an incremental unsupervised-learning framework (The reference accuracy values, obtained when using pose information given by a human, are presented in the brackets).

Method	Accuracy (%)	False Acceptance (%)	False Rejection (%)
Simple Manifold (SM)	34 (37)	100 (100)	33 (27)
View-dependent Covariance Matrix (VC)	83 (87)	0 (0)	35 (26)

which means that the SM method was not able to differentiate new persons from the trained persons. This condition, on the contrary triggers the decrement of the false rejection rate.

Table 2 summarizes the classification rates for recognizing faces from video sequences of Dataset 2 with 8 sequence's length in an incremental unsupervised-learning framework, which also shows the highest recognition accuracies achieved by both methods. The accuracy rate achieved by the VC method was 83%, with 0% false acceptance rate and 35% false rejection rate. Meanwhile, the SM method only gave 34% accuracy rate with 100% false acceptance rate and 33% false rejection rate. For reference purpose, we also presented the accuracy rates of the recognition system for each method when the pose information was given by a human, as shown as values within brackets in Table 2.

It is clearly seen from all evaluation parameters in Fig. 8 and Table 2 that the proposed VC method outperformed the SM method in all categories in an incremental unsupervised-learning framework.

5. Discussion

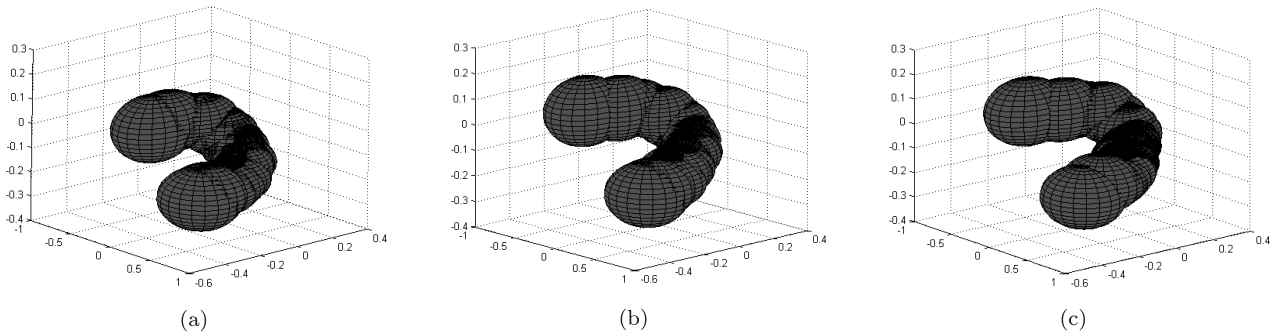
In Sect. 4, we have shown the performances of the proposed VC method and its comparisons with the SM method, where all results showed that the VC method outperformed the SM method in various conditions in both supervised and incremental unsupervised-learning frameworks. In this section, we focus and discuss in more detail the performance of the proposed VC method in incremental unsupervised-learning framework. As we have mentioned earlier, the advantage of an incremental unsupervised-learning framework is that the system could learn and update its knowledge automatically as the unlabelled sequences are input incrementally into the system. However, one critical point in the incremental unsupervised-learning is that the classification ability of the system is highly dependent on the initial settings (i.e. the threshold value, the updating weight, the number of the initial manifold, etc.) and the condition of the input sequences (i.e. the sequence's length, the number of overlapped poses, the input order, etc.). Thus, using different initial settings and/or processing different conditions of input sequences could give different classification results.

Table 3 presents the classification rates of two experiments which have the same initial settings but different conditions of input sequences. For the initial setting, we constructed 10 initial manifolds of 10 persons, defined $\alpha = 0.5$ and $\beta = 2.0$, and used the Nearest Neighbor algorithm as the pose estimation system. In the testing process, the sequence's length and the input order of the unlabelled sequences were set differently. In the first experiment, the system relatively processed longer input sequences within the range of 5–8 sequence's length than in the second experiment whose sequence's length was within the range of 1–8. The number of unlabelled input sequences were 100 sequences (randomly input from 5 sequences for 20 persons) which also shows how many times the system was updated. From Table 3, it can be seen that for both experiments, the VC method outperformed the SM method with 85% highest accuracy rate, 12% false acceptance rate and 18% false rejection rate. Meanwhile, the SM method only gave 18% accuracy rate, with 100% false acceptance and 64% false rejection rates for both experiments.

Next, Fig. 9 shows the visualization of face manifolds using the VC method with poses from −90° (left sideview) to +90° (right sideview) from the frontal pose. Figure 9 (a) shows the visualization of a face manifold of a person which was constructed from the sequences of Dataset 1 and Dataset 2 in a supervised-learning framework. Mean-

Table 3 Classification rates of two experiments which have the same initial settings but different testing conditions.

Exp	Method	Initial Class	Sequence's Length	Identity Class	Accuracy (%)	False Acceptance (%)	False Rejection (%)
1	Simple Manifold (SM)	10	5–8	20	18	100	64
	View-dependent Covariance Matrix (VC)	10	5–8	20	85	12	18
2	Simple Manifold (SM)	10	1–7	20	18	100	64
	View-dependent Covariance Matrix (VC)	10	1–7	20	78	22	22



(a) (b) (c)

Fig. 9 Visualization of face manifolds of a person using the proposed VC method, (a) obtained in a supervised-learning framework, (b–c) obtained in the incremental unsupervised-learning frameworks from different unlabelled input sequences.

Table 4 Accuracy rates of a 10 identity categories recognition task from video sequences with 30° training pose differences between frames (sparse training sequences).

Method	Training classes include samples of	Accuracy (%)
PCA + Nearest Neighbor (NN)	Training poses only	80
PCA + View-dependent Covariance (VC)	Manifold (Training poses + Interpolation)	92
RBF Kernel PCA + Nearest Neighbor (NN)	Training pose only	83
RBF Kernel PCA + View-dependent Covariance (VC)	Manifold (Training poses + Interpolation)	84

while, Fig. 9 (b) and Fig. 9 (c) show the visualization of face manifolds of a same identity (person) in the incremental unsupervised-learning frameworks (the initial manifolds were constructed from sequences of Dataset 1, and later the unlabelled sequences of Dataset 2 were input incrementally to update the initial manifold). The input sequences for Fig. 9 (b) had a 7 sequence's length, while, in Fig. 9 (c), shorter sequences with 3 sequence's length were input. From Fig. 9, it can be seen that the constructed manifolds in the incremental unsupervised-learning frameworks are similar to each other and also to the construction result of the manifold in the supervised-learning framework.

Finally, in order to emphasize the superiority of the proposed method, the accuracy rates of the PCA and nonlinear RBF Kernel PCA in combination with the simple Nearest Neighbor (NN) classifier and the appearance manifold with View-dependent Covariance (VC) are presented in Table 4. For the combinations with NN, the training classes included only the samples of the training poses. Meanwhile, for the combinations with VC, the manifolds with view-dependent covariance matrix were constructed. Here, constructing a manifold means attaining estimation of continuous poses by interpolating the classes (the mean vectors and the covariance matrices) of two consecutive training poses to obtain those of the untrained poses. Thus, the proposed VC method synthesized more pose variabilities than the NN method, since it has training classes with more complete poses (training poses + interpolation poses).

Due to the fact that the appearance of a person's face is highly dependent on its pose, it is obvious that an appearance-based method which can capture more pose variabilities can give more accurate recognition. Moreover, in the proposed VC method, the covariance matrices were embedded in the appearance manifold. Therefore, the advantage of the proposed VC method includes the abilities to capture pose variability and also learn the sample's distribution of each pose. As the consequence, the proposed VC method can give higher recognition accuracies than the NN method. Table 4 shows that for both the linear PCA and the RBF Kernel PCA feature extraction techniques, the proposed VC method gave higher recognition accuracies than that of the NN method. The results also show that the VC method worked better for both the linear PCA and the nonlinear RBF Kernel PCA feature extraction techniques.

Furthermore, the structure of a manifold in the proposed VC method is feature-space independent because a manifold is constructed only by the interpolation of classes of two consecutive training poses. As an interpolation technique can be applied to any feature-space, the structure of the constructed manifold is also not affected by the linearity

or non-linearity of the feature-space.

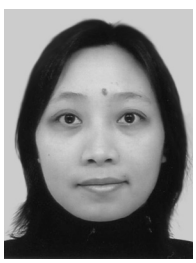
6. Conclusion

We have proposed the appearance manifold with view-dependent covariance matrix for face recognition from video sequences in two learning frameworks: the supervised-learning and the incremental unsupervised-learning. In the supervised-learning framework, the training samples are labelled and an appearance manifold with view-dependent covariance matrix is used to represent an identity category. Meanwhile, in the incremental unsupervised-learning framework, the system first learns the initial categories through the initial manifolds, then the unlabelled input sequences are incrementally learned in order to update the existing identity categories of the system. The advantages of this method are, first, the appearance manifold with view-dependent covariance matrix model is robust to pose changes and also noise invariant, since the embedded covariance matrices are calculated based on their poses in order to learn the samples' distributions along the manifold. Moreover, the proposed incremental unsupervised-learning framework is more realistic for a real-world face recognition application, since it allows us to train the system with the available initial sequences, and later update the system's knowledge incrementally every time an unlabelled sequence is input. We also integrated the appearance manifold with view-dependent covariance matrix model with a pose estimation system in order to improve the classification accuracy of the system and to easily detect the overlapped poses in video sequences which is useful for the merging process in the incremental unsupervised-learning framework. The merging process is performed in order to merge the manifolds which have some overlapped poses with strong similarities. The experimental results showed that in both frameworks, the proposed appearance manifold with view-dependent covariance matrix method outperforms the simple manifold model in recognizing faces from video sequences.

Our future work will concentrate on recognizing faces from continuous video sequences, with both vertical and horizontal pose directions, sudden pose changes, and other varying conditions.

References

[1] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve procedure for the characterization of human face," IEEE Trans. Pattern Anal. Mach. Intell., vol.12, no.1, pp.103–108, Dec. 1990.
[2] M. Turk and A. Pentland, "Face recognition using eigenfaces," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.586–591, Maui HI, USA, June 1991.
[3] H. Murase and S.K. Nayar, "Visual learning and recognition of 3-D objects from appearance," Int. J. Comput. Vis., vol.14, no.1, pp.5–24, Jan. 1995.
[4] H. Murase and S.K. Nayar, "Illumination planning for object recognition using parametric eigenspaces," IEEE Trans. Pattern Anal. Mach. Intell., vol.16, no.12, pp.1219–1227, Dec. 1994.
[5] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," IEEE Trans. Pattern Anal. Mach. Intell., vol.19, no.7, pp.696–710, July 1997.
[6] A.M. Martinez, "Recognition of partially occluded and/or imprecisely localized faces using a probabilistic approach," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol.1, pp.712–717, Hilton Head Island SC, USA, June 2000.
[7] Y. Chang, C. Hu, and M. Turk, "Probabilistic expression analysis on manifolds," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol.2, pp.520–527, Washington DC, USA, July 2004.
[8] C.M. Christoudias and T. Darrel, "On modeling nonlinear shape-and-texture appearance manifolds," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol.2, pp.1067–1074, San Diego CA, USA, June 2005.
[9] B.J. Frey, M. Jojic, and A. Kannan, "Learning appearance and transparency manifolds of occluded objects in layers," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol.1, pp.45–52, Madison WI, USA, June 2003.
[10] Lina, T. Takahashi, I. Ide, and H. Murase, "Appearance manifold with embedded covariance matrix for robust 3D object recognition," Proc. IAPR Conf. on Machine Vision Applications 2007, pp.504–507, Tokyo, Japan, May 2007.
[11] Lina, T. Takahashi, I. Ide, and H. Murase, "Construction of appearance manifold with embedded view-dependent covariance matrix for 3D object recognition," IEICE Trans. Inf. & Syst., vol.E91-D, no.4, pp.1091–1100, April 2008.
[12] B. Raytchev and H. Murase, "Unsupervised recognition of multi-view face sequences based on pairwise clustering with attraction and repulsion," Comput. Vis. Image Understand., vol.91, no.1-2, pp.22–52, 2003.
[13] S. Zhou and R. Chellappa, "Probabilistic human recognition from video," Proc. European Conf. on Computer Vision, vol.3, pp.681–697, Copenhagen, Denmark, May 2002.
[14] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Visual tracking and recognition using probabilistic appearance manifolds," Comput. Vis. Image Understand., vol.99, no.3, pp.303–331, 2005.
[15] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol.1, pp.11–18, Madison WI, USA, June 2003.
[16] X. Liu and T. Chen, "Video-based face recognition using adaptive hidden Markov models," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol.2, pp.26–33, Madison WI, USA, June 2003.



Lina received her B.S. degree (2001) from the Department of Computer Science, Tarumanagara University and a M.S. degree (2004) from the Department of Computer Science, University of Indonesia. Currently, she is pursuing a Ph.D. in the Department of Media Science at Nagoya University, Japan. Her research interests include still-image-based and video-based 3D object/face recognition.

Tomokazu Takahashi received his B.S. degree from the Department of Information Engineering at Ibaraki University, and a M.S. and Ph.D. from the Graduate School of Science and Engineering at Ibaraki University. His research interests include computer graphics and image recognition.

Ichiro Ide received his B.S. degree from the Deparment of Electronic Engineering, a M.Eng. degree from the Department of Information Engineering, and a Ph.D. from the Department of Electrical Engineering at the University of Tokyo. He is currently an Associate Professor in the Graduate School of Information Science at Nagoya University. His research interests include integrated media processing and video processing.

Hiroshi Murase received his B.S., M.S., and Ph.D. degrees from the Graduate School of Electrical Engineering at Nagoya University. He is currently a Professor in the Graduate School of Information Science at Nagoya University. He received the Ministry Award from the Ministry of Education, Culture, Sports, Science and Technology in Japan in 2003. He is a Fellow of the IEEE. His research interests include image recognition, intelligent vehicle, and computer vision.