VIDEO BASED FACE RECOGNITION USING FACE MANIFOLD WITH VIEW-DEPENDENT COVARIANCE MATRIX

Lina[*], Tomokazu Takahashi[+], Ichiro Ide[*], Hiroshi Murase[*]

*Graduate School of Information Science, Nagoya University, Japan
[+]Faculty of Economics and Information, Gifu Shotoku Gakuen University, Japan
E-mail: lina@murase.m.is.nagoya-u.ac.jp

ABSTRACT
Appearance variations in camera or video-captured images usually occur naturally. These variations might be caused by some changes in enviromental condition or by the erroneous of a preprocessing step. Therefore, a robust face recognition system should be able to deal with these variations in order to perform correct identity recognition of the input images. Unfortunately, relying on the simple manifold technique to deal with both pose and degradation problems is not sufficient. In this paper, we propose a face manifold with view-dependent covariance matrix method for video-based face recognition application. The view-dependent covariance matrices are obtained in an efficient way by interpolating only the eigenvectors and the eigenvalues of the covariance matrices of two consecutive training poses. Here the view-dependent covariance matrix plays an important role in providing the distribution information of samples in each class along the face manifold. Moreover, a pose estimation system is also integrated to the recognition system in order to handle pose variations. Experimental results showed that our proposed face manifold with view-dependent covariance matrix outperforms the well known simple manifold method.

KEY WORDS
Video-based face recognition, face manifold, covariance matrix, eigenspace

1. Introduction

In the early years of pattern recognition, many recognition methods have put their focuses on recognizing objects and human faces using still images, such as reported in [1]-[7]. Following the growth of multimedia technology, recently the research trend has moved to video-based face recognition. Similar to the still image-based recognition, in video-based face recognition, the face appearances in the captured 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.

It is well known that appearance-based methods have been proposed and successfully applied to many recognition systems, such as the simple appearance manifold (known as the Parametric Eigenspace method) [8]-[9], the appearance manifold with probabilistic techniques [5]-[6], and their modifications for video-based face recognition such as in [10]-[12]. Although the recognition processes of those manifolds in the previous works were different, however, their construction processes were all based on a simple manifold model proposed in [8]. The disadvantage of this model is that the simple manifold model only works well when the input images have not been affected by degradation. Unfortunately, this assumption is not realistic in real-world applications. Some degradation effects usually occur and contaminate the original images during the capturing and segmentation processes. Thus, relying on a simple manifold model to handle this problem is not sufficient.

To overcome this problem, we have shown that embedding covariance matrices to an appearance manifold is very useful, since the manifold could capture the pose changes and the embedded covariance matrix could define the sample distribution information of every pose along the manifold [13]-[14]. Moreover, since the appearance of an object in the captured image is different for each pose, the covariance matrix value is also different for every pose. Thus, it is necessary to construct a view-dependent covariance matrix.

In this paper, we address the problem of recognizing human faces from video sequences where the temporal correspondences between frames in each video sequence should be considered. In the training stage, a face manifold with view-dependent covariance matrix of each person is constructed by interpolating each pair of the eigenvectors and the eigenvalues of the training poses (known as the VCEI method in [14]). Meanwhile in the testing stage, the classification decision is based on the classification results of every frame in the video sequence.

Moreover, in a real-world application, two images of a same pose are likely to be identified as a same person, although they are actually images of two different persons. In [10], it is stated that the images tend to be identified according to the manifold which has the same pose rather than its identity. Therefore, we attempt to conduct a pose classification prior to the identity classification where the pose estimator will give the pose information of every frame in a video sequence to the recognition system, so that the similarity measurements are applied only to models with the appropriate poses.

The remainder of this paper is organized as follows: Section 2 and Section 3 describe the representation of the face manifold and the embedding process of the view-dependent covariance matrix. Meanwhile Section 4 and Section 5 present the experimental results and our conclusion.

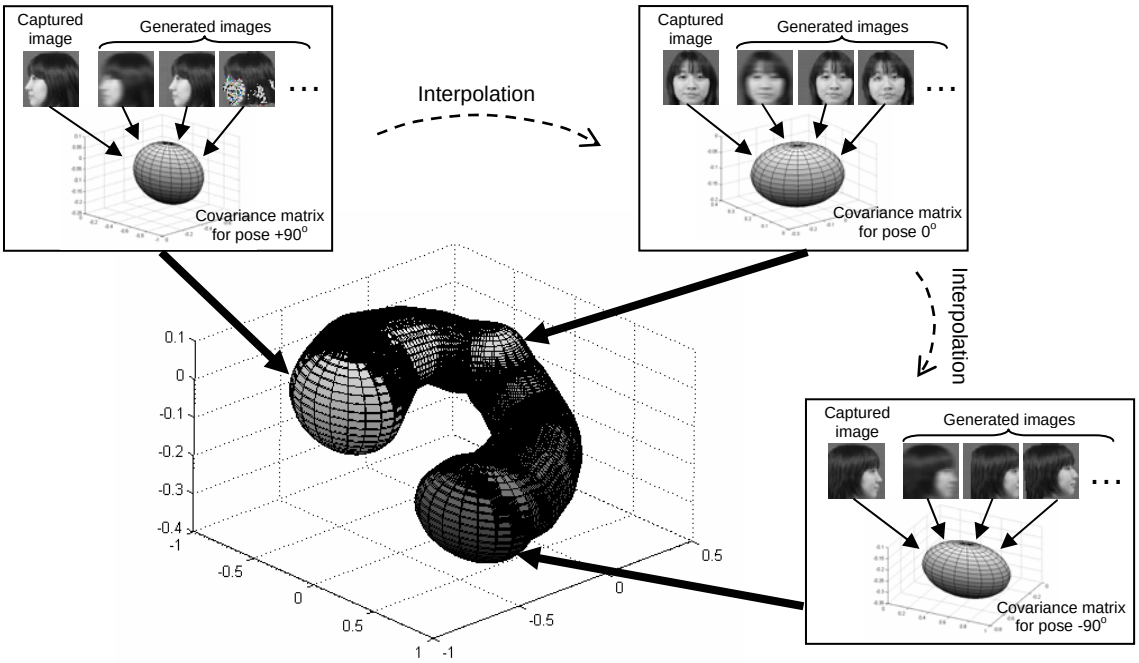


Figure 1. A face manifold with view-dependent covariance matrices in an eigenspace

2. Face Manifold

The appearance-based approaches usually deal with a set of learning images in various capturing conditions. These images are originally in high-dimension, thus, in a recognition application, a feature extraction module becomes necessary in order to transform the images into low-dimensional features. One well known feature extractor in the pattern recognition field is the Principal Component Analysis (PCA). Here, PCA is used to efficiently represent a collection of images by reducing their dimensionality.

PCA represents a linear transformation that maps the original *n*-dimensional image space onto a *k*-dimensional space (known as eigenspace) where normally $k << n$. In order to project *L* training samples of *P* persons in the eigenspace, first *k* eigenvectors which have the largest corresponding eigenvalues are selected. Then, the linear transformation of the eigenspace representation is defined by:

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

where $\mathbf{x}_l^{(p)}(\theta)$ is the *l*-th sample image of person *p* with pose θ. Here, $\mathbf{e}_i$ (i =1, 2, … , k) are the eigenvectors, $\boldsymbol{c}$ is the mean vector of the training samples, and $\mathbf{q}_l^{(p)}(\theta) \in \Re^k$ are the vector representations of images $\mathbf{x}_l^{(p)}(\theta)$ in the eigenspace. These eigenvectors $\mathbf{e}_i$ were obtained by solving the eigen decomposition $\lambda_i \mathbf{e}_i = \mathbf{Q}\mathbf{e}_i$, where $\mathbf{Q}$ is the auto-correlation matrix of the training set and λ_i is the eigenvalue associated with the eigenvector $\mathbf{e}_i$. Note that in this section, the eigenvectors and eigenvalues are used only to construct the eigenspace, where in later sections, the eigenvectors and eigenvalues are derived from a covariance matrix of each training-pose. In the Simple Manifold (SM) method (known as the Parametric Eigenspace method in [8-9]), the face manifold of a person can be obtained by interpolating the mean vector of *L* training images of the person from one pose to its consecutive poses using any interpolation algorithm. Meanwhile, the construction process of a face manifold with embedded view-dependent covariance matrix is described in Section 3.

3. Embedding View-dependent Covariance Matrices in a Face Manifold

This section describes the process of constructing the face manifold with embedded view-dependent covariance matrix in an eigenspace. Unlike the Simple Manifold (SM) method, in order to capture the image variation of untrained poses, the construction process of a face manifold with the View-dependent Covariance matrix (VC) method needs to interpolate both mean vectors and covariance matrices. In an eigenspace, a mean vector is used to represent the center point of samples in each learning pose, while a covariance matrix represents the distribution of samples in each pose. Here, the interpolation of the mean vector can be done simply by using one of the existing interpolation algorithms, while the interpolation of the covariance matrices is based on the interpolation of the eigenvectors and the eigenvalues of two consecutive training poses (known as the VCEI method in [14]).

Fig. 1 shows the construction process of a face manifold with view-dependent covariance matrices in an eigenspace. Here, we only use the horizontal pose parameter (θ) to construct the face manifold. The face manifold is constructed using several captured training images and generated images

by the addition of various types and levels of noise effect to the captured images. From each pose, a covariance matrix is calculated and visualized as a hyper-ellipsoid. In the eigenspace, a covariance matrix can be considered as a hyper-ellipsoid and its elements such as the eigenvectors and the eigenvalues can be considered as the axes directions and the lengths of the hyper-ellipsoid, respectively. Therefore, interpolating covariance matrices of two consecutive poses can be done by rotating the hyper-ellipsoids of the corresponding poses. The algorithm for interpolating the covariance matrices by the eigenvector and the eigenvalue interpolation which has been proposed in [14] is summarized as follows:

Input: $\mathbf{E}_0$ and $\mathbf{E}_1$ are matrices formed by aligning eigenvectors $\mathbf{e}_{0j}$ and $\mathbf{e}_{1j}$, while $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$ are matrices formed by aligning eigenvalues λ_{0j} and λ_{1j} (j =1, 2, … , k). The covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ represent the sample distribution of two consecutive poses.

1. Sort the eigenvectors $\mathbf{E}_0$ and $\mathbf{E}_1$ in the decreasing order according to their eigenvalues $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$ to obtain $\mathbf{E}'_0$ and $\mathbf{E}'_1$, and also $\boldsymbol{\lambda}'_0$ and $\boldsymbol{\lambda}'_1$
2. Check the angle between the corresponding axis so that it is less than or equal to 0.5π:
   if $\mathbf{e}'^{\,T}_{0j}\mathbf{e}'_{1j} < 0$ then invert $\mathbf{e}'_{1j}$ (j =1, 2, … , k)
3. For a covariance matrix $\boldsymbol{\Sigma}_x$, do the calculation for the eigenvalue interpolation with:
   $\lambda_{xj} = \left((1-x)\sqrt{\lambda'_{0j}} + x\sqrt{\lambda'_{1j}}\right)^2$
4. For a covariance matrix $\boldsymbol{\Sigma}_x$, do the interpolation for the eigenvectors with:
   $\mathbf{E}_x = \mathbf{R}(x\boldsymbol{\phi})\mathbf{E}'_0$
   where $\mathbf{R}$ represents an interpolated rotation when $0 \le x \le 1$ and $\boldsymbol{\varphi} = [\phi_1,...,\phi_m]$ represents the parameter vector of the rotation angles to define the rotation matrix. Here, $m = \lfloor n/2 \rfloor$ since the rotation angles always come in pairs in the complex conjugate roots process.
   (a) Define the rotation matrix by:
   $\mathbf{R}(\boldsymbol{\varphi}) = \mathbf{E}'_1\mathbf{E}'^{\,T}_0$
   (b) Diagonalize $\mathbf{R}(\boldsymbol{\varphi})$ with Special Orthogonal (SO) rule: $\mathbf{R}(\boldsymbol{\varphi}) = \mathbf{U}\mathbf{D}(\boldsymbol{\varphi})\mathbf{U}^+$
   where $\mathbf{U}^+$ is the conjugate transpose of $\mathbf{U}$
   (c) Process complex conjugate roots:
   if $n = 2m$ then
   $\mathbf{D}(\boldsymbol{\varphi}) = diag(\mathbf{e}^{i\phi_1}, \mathbf{e}^{-i\phi_1},..., \mathbf{e}^{i\phi_m}, \mathbf{e}^{-i\phi_m})$
   however, if $n = 2m+1$ then
   $\mathbf{D}(\boldsymbol{\varphi}) = diag(1, \mathbf{e}^{i\phi_1}, \mathbf{e}^{-i\phi_1},..., \mathbf{e}^{i\phi_m}, \mathbf{e}^{-i\phi_m})$
   where $\mathbf{e}^{i\phi} = \cos\phi + i\sin\phi$
   (d) Apply linear interpolation technique to obtain $\mathbf{R}(x\boldsymbol{\varphi})$

Output: The covariance matrix for untrained poses $\boldsymbol{\Sigma}_x = \mathbf{E}_x\boldsymbol{\Lambda}_x\mathbf{E}_x^{\,T}$ where $\boldsymbol{\Lambda}_x = diag(\boldsymbol{\lambda}_x)$

Finally, the output of the training stage is presented in the form of face manifolds with view-dependent covariance matrices which consist of mean vectors $\boldsymbol{\mu}^{(p)}(\theta)$ and covariance matrices $\boldsymbol{\Sigma}^{(p)}(\theta)$.

4. Face-sequence Classification

In the testing stage, we propose to attempt a pose classification prior to the identity classification. The pose classification is useful to determine an appropriate shape model for similarity measurement. Here, the pose estimator is fully independent from the identity classification system. To estimate the pose position of the testing image, one of the various existing algorithms can be selected, such as the *k*-Nearest Neigbor, Parametric Eigenspace, Nearest Feature Line, Back Propagation Neural Network, etc. In this paper, the pose estimation is based on the Nearest Neigbor algorithm which basically finds the nearest point of the test image to the center of every class. The output of the pose estimation system is a pose value φ_i of each test image $\mathbf{f}_i$ (i =1, 2, … , h) in a testing sequence.

It is well known that a video-based recognition system needs to integrate the classification results of every frame to produce the decision as a sequence. Given a face sequence $\mathbf{S} = [\mathbf{f}_1, \mathbf{f}_2,..., \mathbf{f}_h]$, the classification process of the testing sequence is based on a similarity measurement using the minimal cumulative distance of each frame $\mathbf{f}_i \in \mathbf{S}$ (i =1, 2, … , h) to the trained manifolds M_p. Once the pose φ_i of the test image $\mathbf{f}_i$ (i =1, 2, … , h) is determined by the pose estimation system, the distance measurement of a test image is defined by:

$$dfw_i^{(p)} = \left(\mathbf{f}'_i - \boldsymbol{\mu}^{(p)}(\varphi_i)\right)^T \left(\boldsymbol{\Sigma}^{(p)}(\varphi_i)\right)^{-1} \left(\mathbf{f}'_i - \boldsymbol{\mu}^{(p)}(\varphi_i)\right) \quad (4)$$

and the sequence's classification to determine the identity p^* can be processed as follows:

$$p^* = \arg\min_p \sum_{i=1}^{h} \left(dfw_i^{(p)}\right) \quad (5)$$

5. Experimental Results and Analysis

To evaluate the performance of our face manifold with view-dependent covariance matrix method, we developed a video-based face recognition system application. In the experiments, we recognized face sequences of 20 persons under various conditions. We have collected three video sequences for each person with pose changes from –90° from frontal pose (left sideview) until +90° from frontal pose (right sideview). In the preprocessing step, the motion videos were trimmed with a frame rate of 30 frames/second. Next, the images from a video sequence with 10° pose differences from each other were

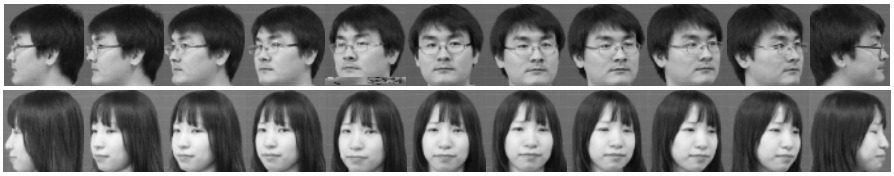


Figure 2. Samples of face sequences of Dataset 1 (steady head and normal expression)

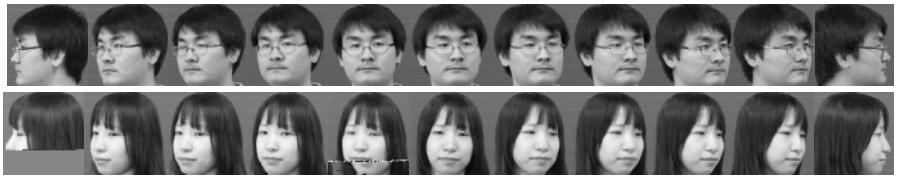


Figure 3. Samples of face sequences of Dataset 2 (steady head and normal expression, taken in a different time)

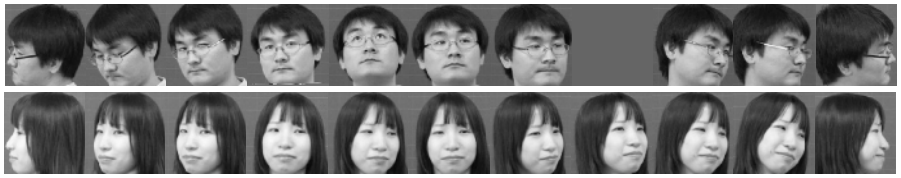


Figure 4. Samples of face sequences of Dataset 3 (free head movement and expression)

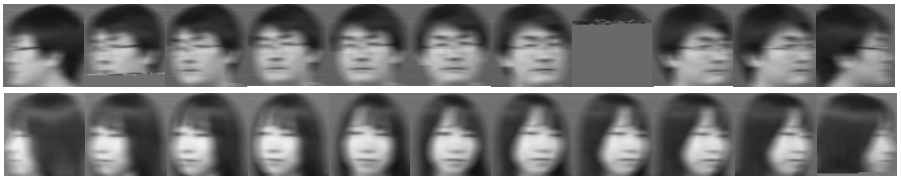


Figure 5. Samples of face sequences of Dataset 4 (low quality images)

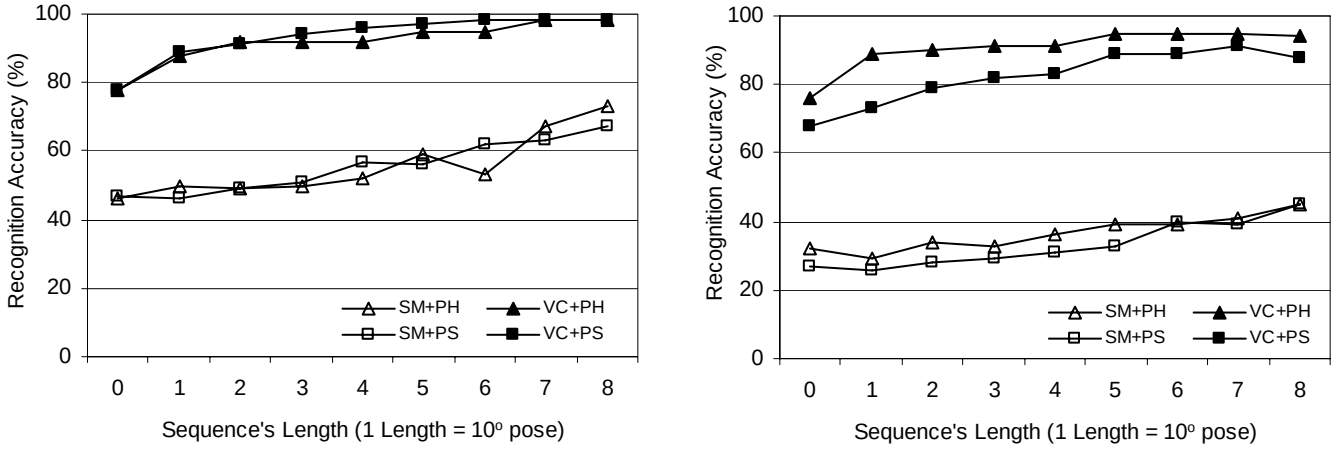
taken as a face sequence. This sampling process of each face sequence was performed in order to provide a fair evaluation condition where each face sequence contains a same frame-density condition. However, in a real system, this process is not necessary. Finally, the images of the face sequences were manually cropped in the face areas and down-sampled into 32 x 32 pixels of grayscale images.

In the experiments, four face datasets which represent different conditions were used as exemplified in Fig. 2, Fig. 3, Fig. 4, and Fig. 5. Dataset 1 which was used for training contains the face sequences of persons in steady head position and normal expression, while Dataset 2 contains the face sequences with the same condition with Dataset 1 but was taken in a different time. Dataset 3 presents the face sequences of persons with free head movement and expression, while Dataset 4 contains the face sequences of persons in steady head position and normal expression but with motion blur effects.

For constructing a face manifold of a person in the training stage, 26 face sequences which consist of a video-captured sequence of Dataset 1 and 25 generated sequences were used. The generated sequences were obtained by composing artificial noises, such as left and right translations (3, 6, 9, 12, 15 pixels), clockwise and counter-clockwise rotations (5^o, 10^o, 15^o, 20^o, 25^o), and motion blur (5%, 10%, 15%, 20%, 25%). Meanwhile, for the testing stage, the video sequences were split so that the sequences only contain partial poses. Here, every pair of the consecutive frames has 10^o pose differences and the pose width in a sequence is shown in the form of a sequence's length. A total number of 1,800 partial sequences with 9 different sequence's lengths were used for testing.

For performance evaluation purposes, we developed two different models of the face manifold with view-dependent covariance matrix: 1) the face manifold with view-dependent covariance matrix by eigenvector and eigenvalue interpolation with the pose information given by a human (VC+PH method), and 2) the VC method combined with a nearest neighbor pose estimation system (VC+PS method). To provide a fair comparison, we compare our proposed methods with the well known Simple Manifold method also with an integrated pose estimation system: 3) the SM method with a human pose estimator (SM+PH method) and 4) the SM method with a pose estimation system (SM+PS method).

Figure 6 shows the accuracy rates in recognizing faces from video sequences of Dataset 2 with various sequence's lengths. In every sequence, each frame has 10^o pose differences with the other consecutive frames. For a 10 identity categories task, the results in Fig. 6(a) shows that the proposed VC+PH and the VC+PS methods give higher recognition accuracies compared with that of the SM+PH and the SM+PS methods. The highest recognition was achieved by the VC+PH and the VC+PS methods with 98% accuracy for both methods.



(a) 10 identity categories (persons) task (b) 20 identity categories (persons) task

Figure 6. Recognition rates of face sequences of Dataset 2 with various sequence's lengths

Table 1. Recognition rates of face sequences of various Datasets with sequence's length = 8

Dataset	Pose Estimation	Recognition Rates (%)	
		Simple Manifold (SM)	View-dependent Covariance Matrix (VC)
Small Face Variations (Dataset 2)	PH	73	98
	PS	67	98
Severe Face Variations (Dataset 3)	PH	60	88
	PS	67	75
Low-quality Images (Dataset 4)	PH	71	98
	PS	65	98

*) PH = Pose estimation by Human, PS = Pose estimation by System

Meanwhile, the highest recognition result for the SM method was only 73% achieved by the SM+PH method. For all methods, the recognition accuracies increased along with the increment of the sequence's length. The reason is very obvious; the longer the sequence's length, the more images from various poses were available to represent the person's appearance, thus, the easier for the system to recognize the person's identity.

For a 20 identity categories recognition task, as depicted in Fig. 6(b), the accuracies of the recognition system decreased compared with that of the 10 identity categories recognition task. However, the proposed VC methods outperformed the SM methods, with 94% and 88% highest recognition accuracies for the VC+PH and the VC+PS methods, respectively. Meanwhile, the SM+PH and the SM+PS methods only achieved 45% as their highest recognition accuracies.

Furthermore, to evaluate the robustness of the proposed methods, we have conducted several experiments using various datasets and showed the results in Table 1. Here, the testing sequences' lengths were set equal to 8 and were chosen with consideration that the system gives accurate results from longer testing sequences. It can be seen from Table 1 that the recognition accuracies of the proposed VC+PH and the VC+PS methods were higher than that of the SM+PH and the SM+PS methods. For recognizing face sequences of Dataset 2, the highest recognition accuracy of 98% was achieved by both of the VC+PH and the VC+PS methods, while the SM+PH method gave the highest recognition accuracy with only 73% among the SM methods.

When recognizing testing sequences with severe face variations in Dataset 3, our proposed VC methods still could maintain their superiority over the SM methods. The highest recognition accuracy was achieved by the VC+PH method with 88%, while the highest recognition accuracy for the SM method was achieved by the SM+PS method with 67%.

Finally, we tested the system with Dataset 4 where the face images of the testing sequences were influenced with motion blur effects. Here, once again the VC methods proved their robustness with 98% highest recognition accuracy achieved by both of the VC+PH and the VC+PS methods, while 71% recognition accuracy was achieved by the SM+PH method as the highest recognition accuracy among the SM methods. It can be seen that although the face sequences were influenced with motion blur effects, the proposed VC methods could maintain their accuracies, while, the recognition accuracies of the SM methods decreased when the testing sequences were in low qualities.

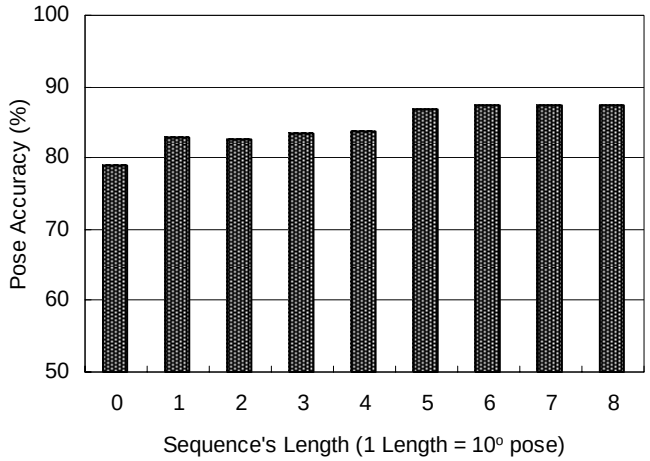


Figure 7. Accuracies of a pose estimation system using Nearest Neighbor algorithm for various sequence's lengths

We then observed the accuracies of a pose estimation system using Nearest Neighbor algorithm by comparing its results with the human estimator's as shown in Fig. 7. The pose accuracy achieved by the system with the nearest neighbor algorithm was 84% in average. Here, we tolerated a maximum of 10^{o} pose differences between a pose result from a pose estimation system with the result given by a human. We also assumed the pose information given by a human were 100% correct, thus, it was used as ground truth.

6. Conclusion

We proposed the construction of a face manifold with the embedded view-dependent covariance matrix for video-based face recognition application. In the construction process of the view-dependent covariance matrix, an interpolation process of each pair of the eigenvectors and the eigenvalues of two consecutive training poses is conducted in order to obtain the view-dependent covariance matrices of the untrained poses. The advantages of using the eigenvectors and the eigenvalues interpolation in the construction process of a face manifold with the view-dependent covariance matrix are its robustness and efficiency, since it only interpolates the eigenvectors and the eigenvalues without considering the number and the correspondence of each training image. Experimental results showed that our proposed face manifold with embedded view-dependent covariance matrix could recognize faces from video sequences accurately and outperforms the well known simple manifold method.

Our future work includes recognizing faces from continuous video sequences with both vertical and horizontal pose directions and developing an incremental learning framework for face manifold with view-dependent covariance matrix method so that the system could update its knowledge automatically in an unsupervised manner.

References

[1] M. Turk and A. Pentland, "Face Recognition Using Eigenfaces," Proc. 1991 IEEE Conf. Computer Vision and Pattern Recognition, pp. 586–591, 1991.

[2] L. Wiskott, J.-M. Fellous, and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 775–779, 1997.

[3] A. Lanitis, C.J. Taylor, and T.F. Cootes, "Automatic Face Identification System Using Flexible Appearance Models," Image and Vision Computing, vol. 13, no. 5, pp. 393–401, 1995.

[4] W. Zhao and R. Chellappa, "SFS Based View Synthesis for Robust Face Recognition," Proc.4th Conf. Automatic Face and Gesture Recognition, pp. 285–290, 2000.

[5] B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Representation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 696–710, 1997.

[6] A.M. Martinez, "Recognition of Partially Occluded and/or Imprecisely Localized Faces Using a Probabilistic Approach," Proc. 2000 IEEE Conf. Computer Vision and Pattern Recognition, vol. 1, pp. 712–717, 2000.

[7] W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld, "Face Recognition: a Literature Survey," ACM Computing Surveys, vol. 35, no. 4, pp. 399–458, 2003.

[8] H. Murase and S.K. Nayar, "Illumination Planning for Object Recognition Using Parametric Eigenspaces," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 16, no. 12, pp. 1219–1227, 1994.

[9] S.K. Nayar, S.A. Nene, and H. Murase, "Real-time 100 Object Recognition System," Proc. 1996 IEEE Conf. Robotics and Automation, vol. 3, pp. 2321–2325, 1996.

[10] B. Raytchev and H. Murase, "Unsupervised Recognition of Multi-View Face Sequences Based on Pairwise Clustering with Attraction and Repulsion," Computer Vision and Image Understanding, vol. 91, pp. 22–52, 2003.

[11] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Visual Tracking and Recognition Using Probabilistic Appearance Manifolds," Computer Vision and Image Understanding, vol. 99, pp. 303–331, 2005.

[12] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman, "Clustering Appearances of Objects under Varying Illumination Conditions," Proc. 2003 IEEE Conf. Computer Vision and Pattern Recognition, vol. 1, pp. 11–18, 2003.

[13] Lina, T. Takahashi, I. Ide, and H. Murase, "Appearance Manifold with Embedded Covariance Matrix for Robust 3D Object Recognition," Proc. 10th IAPR Conf. Machine Vision Applications, pp. 504–507, 2007.

[14] Lina, T. Takahashi, I. Ide, and H. Murase, "Construction of Appearance Manifold with Embedded View-dependent Covariance Matrix for 3D Object Recognition," IEICE Trans. Information and Systems, vol. E91-D, no. 4, pp. 1091–1100, 2008.