Combining Classifier for Face Identification at Unknown Views with a Single Model Image

Tae-Kyun Kim[1] and Josef Kittler[2]

[1] HCI Lab., Samsung Advanced Institute of Technology, Yongin, Korea
taekyun@sait.samsung.co.kr
[2] Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK
J.Kittler@surrey.ac.uk

Abstract. We investigate a number of approaches to pose invariant face recognition. Basically, the methods involve three sequential functions for capturing nonlinear manifolds of face view changes: representation, view-transformation and discrimination. We compare a design in which the three stages are optimized separately, with two techniques which establish the overall transformation by a single stage optimization process. In addition we also develop an approach exploiting a generic 3D face model. A look-up table of facial feature correspondence between different views is applied to an input image, yielding a virtual view face. We show experimentally that the four methods developed individually outperform the classical method of Principal Component Analysis(PCA)-Linear Discriminant Analysis(LDA). Further performance gains are achieved by combining the outputs of these face recognition methods using different fusion strategies.

1 Introduction

Face recognition has a benefit over other biometric techniques such as fingerprint and iris recognition in that humans can be identified without notice and at distance. However, to realize this potential, it is essential to counteract the degradation in performance exhibited by face recognition systems for views different from the frontal pose. View-changes can be learned from prototype faces and the learned models can be applied to other individuals.

Classically, a generic 3D model of a human face has been used to synthesize face images from different view points[8] and approximate models, such as a cylinder, have also been applied to face recognition. More recently, Vetter and Poggio[1] showed that the 2D image based technique is a viable method for view synthesis and recognition of face classes. In their work, face images are first represented in the view-subspace and the transformation matrix between the different view representations is computed in the sense of Least Square Error(LSE). Blanz[4] utilized a 3D morphable model and Yongmin Li[2] applied Kernel Discriminant Analysis and 3D Point Distribution Model for view-invariant face recognition. In spite of the recent successes, all the above methods have a strong drawback in requiring dense corre-

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 565–573, 2004.

spondence of facial features for image normalization. The step of feature detection or correspondence solving, which is needed for separating the shape and texture components of face images in these methods, is usually difficult itself. Errors in correspondences seriously degrade the overall face recognition performance of these methods as shown in [4]. Among other relevant works, Graham and Allinson[3] applied a Neural Network for learning the view transfer function of the normalized face images with a fixed eye position. Talukder[6] also proposed the method for the simply normalized images by using fixed eye points, which involves a linear view-transfer matrix obtained by the LSE method[1].

In this paper, we propose robust base classifiers for face identification at unknown views and a combining classifier for accuracy improvement. It is assumed that a single model image is given and face images are registered with reference to the eye positions. The classifiers differ in the way they model face view-changes. They can be categorized into methods based on statistical learning of face images at different poses and methods based on 3D face models. The two piecewise linear methods and the nonlinear kernel method are adopted as the base classifiers, which are based on statistical learning of face images. In addition, a computationally efficient approach, which stores the correspondence information of 3D face models at different views in a look-up table, is also developed for complementing the statistical learning methods. These base classifiers are quite different in their nature owing to different sources of information and architectures used. This motivates us to combine them for further accuracy improvement.

2 Base Classifier Design

There are a number of factors that cause the face data distribution of different poses to be nonlinear; this naturally motivates us to exploit the benefits of non-linear architectures. The existing view-invariant face recognition methods can generally be decomposed into three sequential steps: representation, view-transformation and discrimination function. First an input face image is projected into a view subspace via a function, **S,** which is obtained by linear or nonlinear subspace analysis of face images within a certain range of view-angles, as

$$\mathbf{b}_{v,i} = \mathbf{S}_v(\mathbf{x}_{v,i}, \mathbf{avg}_v) \qquad (1)$$

where $\mathbf{x}_{v,i}$ is the *i*-th face image in the set drawn from a certain small range of views, *v*. A linear matrix (LM)[1] or Neural Network[3] can be utilized to learn the transfer function **V** between the different view representations in the sense of LSE,

$$\min_{\mathbf{V}} \sum_{i=1}^{N} \left|\mathbf{b}_{f,i} - \mathbf{V}(\mathbf{b}_{r,i})\right|^2 \qquad (2)$$

where *f* and *r* denote a frontal-view and a rotated-view respectively and *N* is the number of images. The face images transformed to the frontal-view and the original frontal view faces are the input for learning a discriminant function **D**. LDA or Generalized Discriminant Analysis(GDA)[7] can be applied to learn the function mapping the pose corrected face images into discriminative feature vectors, i.e.

$$\mathbf{d}_{f,i} = \mathbf{D}(\mathbf{b}_{f,i}, \mathbf{avg}), \quad \mathbf{d}_{r,i} = \mathbf{D}(\mathbf{V}(\mathbf{b}_{r,i}), \mathbf{avg}) \tag{3}$$

where **avg** is a global mean. The final classification is based on the nearest neighbor matching of the feature vectors $\mathbf{d}$. As the performance of this system depends on the choice of each transformation function, various combinations of linear and nonlinear functions obtained by statistical learning have been compared in [10]. The study showed that the piecewise linear combinatorial method, "PCA(as a **S**)-LM(as a **V**)-LDA(as a **D**)" is one of the most accurate classifiers. As its computational cost is low, PCA-LM-LDA has been adopted as a base classifier in this study. However, it should be noted that this combinatorial method must yield a sub-optimal solution since each step is separately trained.

A novel nonlinear discriminant analysis, called "Locally Linear Discriminant Analysis(LLDA)", has been developed to provide a unified framework for the three stage structure. It concurrently finds the set of locally linear transformations to yield locally linearly transformed face classes that maximize the between-class covariance while minimizing the within-class covariance as shown in Figure 1.

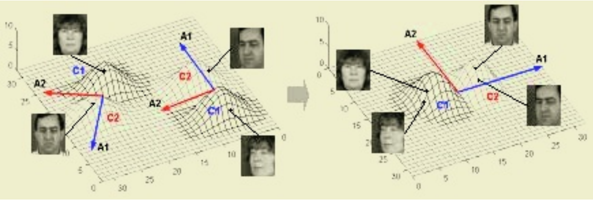

Fig. 1. LLDA for pose-invariant face classification; Left shows the original data distribution and the found components and right shows the transformed data distribution.

The solutions for the frontal and rotated face images found by this method, $\mathbf{U}_f$ and $\mathbf{U}_r$ respectively, correspond to the combined three stage transformation function as discussed above, i.e.

$$\begin{aligned} &\mathbf{d}_{f,i} = \mathbf{U}_f(\mathbf{x}_{f,i}, \mathbf{avg}_f), \quad \mathbf{d}_{r,i} = \mathbf{U}_r(\mathbf{x}_{r,i}, \mathbf{avg}_r), \\ &\to \ \mathbf{U}_f \equiv \mathbf{D}(\mathbf{S}_f(\cdot)), \quad \mathbf{U}_r \equiv \mathbf{D}(\mathbf{V}(\mathbf{S}_r(\cdot)) \cdot \end{aligned} \tag{4}$$

For details of the novel algorithm, LLDA, please refer to the study[5]. LLDA and PCA-LM-LDA will be discussed in more detail in the experimental sections.

Generalised Discriminant Analysis (GDA), which transforms the input space into a high-dimensional feature space by using a kernel function Φ and then linearly separates the data, is also developed. The major difference from the above piece-wise linear methods is that a single nonlinar transformation function $\mathbf{U}^{\Phi}$ such that

$$\mathbf{d}_{f,i} = \mathbf{U}^{\Phi}(\Phi(\mathbf{x}_{f,i}, \mathbf{avg})), \quad \mathbf{d}_{r,i} = \mathbf{U}^{\Phi}(\Phi(\mathbf{x}_{r,i}, \mathbf{avg})) \tag{5}$$

is applied to different view face images while different sets of linear functions are exploited for different view faces in the two piecewise linear classifiers in (3) and (4).

Although the methods based on statistical learning of 2D images are effective for capturing face view changes, a complementary benefit of the more classical method based on 3D face models has also been investigated. We propose to replace the processes of texture mapping, 3D rotation and rendering in graphics with a direct image transformation based on a look-up table(LUT) as shown in Figure 2. By using the average LUT, rotated face images are virtually generated from the frontal face images; Intensity of a pixel of a rotated face image $I_r(x,y)$ is obtained from that of the corresponding pixel of the frontal image I_f as $I_r(x,y) = I_f\left(\overline{\mathbf{LUT}}(x,y)\right)$, where $\overline{\mathbf{LUT}}$ is a funcntion which yields a stored coodinates. The view-transformation through the LUT is very fast. The rotation direction from the frontal to an arbitrary angle is more beneficial as most of the pixel information is kept. Each pose group has an average correspondence LUT. After transforming frontal faces to a certain view, LDA is applied to the pairs of the transformed and the original images at the same view yielding the output feature vectors **d**. Consequently, the proposed method deploys the view-specific discriminant functions of LDA.

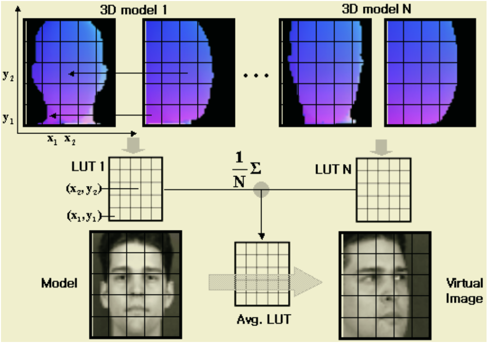

Fig. 2. Virtual View Generation by using the 3D correspondence LUT.

3 Combining Strategies and Experiments

3.1 Combining Techniques

Fusion at the confidence level is considered, where the matching scores reported by the individual classifiers are combined. We have tested the simple fixed combining rules such as the sum, product, maximum, minimum and median rule to access the viability of combining the pose-invariant face classifiers. The use of any trained combiner instead of the fixed rules, provided a suitable evaluation set is available, would be an extension to our work. The confidence value $C_{ij}(\mathbf{x})$ of the base classifier *j* for class *i* is the normalized Euclidean distance of the output vectors **d** produced by the base classifier. The confidence value is scaled by using the small independent evaluation set so that it is in the range of [0,1]. The combining classifier $\mathbf{Q}(\mathbf{x}) = \{\mathbf{Q}_i(\mathbf{x}), i = 1,...,c\}$ is defined as follows:

$$\mathbf{Q}_i = \prod_j \mathbf{C}_{ij}(\mathbf{x}) \quad \mathbf{Q}_i = \sum_j \mathbf{C}_{ij}(\mathbf{x}) \quad \mathbf{Q}_i = \max_j \mathbf{C}_{ij}(\mathbf{x})$$
$$\mathbf{Q}_i = \min_j \mathbf{C}_{ij}(\mathbf{x}) \quad \mathbf{Q}_i = \underset{j}{median}\,\mathbf{C}_{ij}(\mathbf{x}) \quad \mathbf{Q}_i = \sum_j w_j \mathbf{C}_{ij}(\mathbf{x}) \tag{6}$$

3.2 Experimental Setup

We used the two data sets, XM2VTS[19] as the main set and PIE[12] as the set for further comparison of the base classifiers. XM2VTS data set was annotated with pose labels of the face. The face database consists of 2950 facial images of 295 persons with 5 pose variations (F,R,L,U,D) and 2 different time sessions (S1,S2)(5 months time elapse). This may be the largest public data set which has a sizeable population of subjects taken in different poses. Each pose group has a small view range due to the unexpected error in personal pose. The images were normalized to 46*56 pixel resolution with a fixed eye position. The experimental sets consist of 1250 images of 125 persons, 450 images of 45 persons and 1250 face images of 125 persons for the training, evaluation and test respectively. The training, evaluation and test set have different face identities. The training set was utilized to learn the transformation functions of the base classifiers whereas the evaluation set served to adjust the parameters of the classifier: kernel parameters of GDA, the dimensionality of the output vectors and scaling parameters of the individual classifiers for combining. These were carefully chosen to achieve the best performance of each. The recognition performance is reported as the recognition rate on the test set. The frontal face F-S1 of the test set was selected as a gallery and the 9 rotated face images of the test set were exploited as queries. One more independent protocol was built based on the PIE data set: The selected PIE set consists of 15 images (3 poses x 5 illuminations) of 66 identities as shown in Figure 3. This was equally divided into the training and test set. The frontal face F1 of the test set was selected as a gallery and all the other images of the test set were exploited as queries.

3.3 Performance Comparison of the Base Classifiers

3.3.1 Performance of the Individual Classifiers

All the base classifiers have been tested on the XM2VTS DB. For the method of 3D correspondence LUT, 108 SNU 3D scanned facial models[11] were used. For GDA, an RBF kernel with an adjustable width was deployed. Of the proposed four base classifiers, the two methods, PCA-LM-LDA and 3D LUT, explicitly generate view-rotated images. The characteristics of the two transformation results are quite different as shown in Figure 4. While, the generalization performance of the transformation of the statistical learning based method, PCA-LM-LDA is much degraded for the non-trained individuals, the 3D model based method, 3D LUT, maintains its performance. On the contrary, LLDA and GDA implicitly represent the face images so that the rotated faces have a similar representation to that of the frontal view images. The recognition performance of the base classifiers is shown in Figure 5.

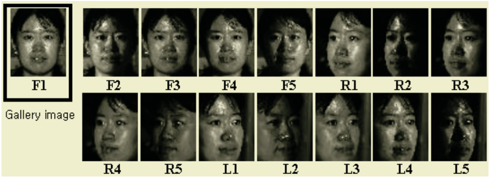

Fig. 3. Sample images of the PIE DB.

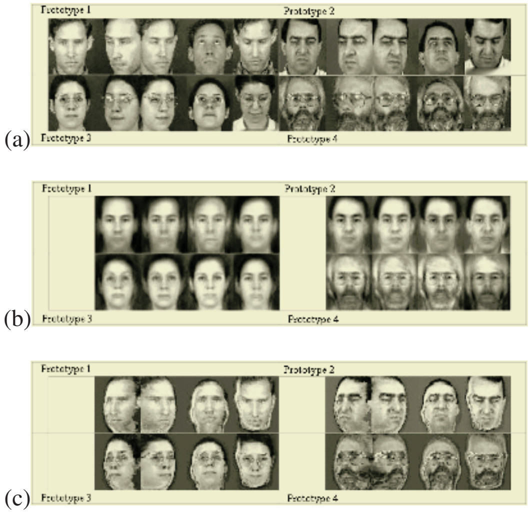

Fig. 4. Examples of the synthesized faces on XM2VTS DB. (a)5 views of the training faces (b)Tranformed faces to a frontal view by PCA-LM-LDA (c)Transformed frontal faces to a rotated view by 3D-LUT.

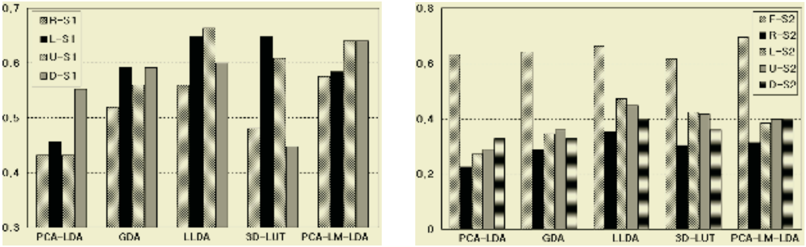

Fig. 5. Recognition rates (in %) of individual experts on XM2VTS DB.

All the four classifiers much outperformed the classical face recognition method, PCA-LDA, where the basis functions of LDA are learned from the eigenfeatures of the training set. The dimensionality of the feature vector at both PCA and LDA stages was carefully controlled to yield its best result. The LLDA method performed best, but the others were also comparable.

3.3.2 PCA-LM-LDA vs. LLDA

We look at the two methods, LLDA and PCA-LM-LDA more closely as the both are trained on the same sources of information and have similar architectures, which are piecewise linear. The PCA-LM-LDA method learns the representation, view-transformation and the discriminant function separately with an indirect objective function for classification. In contrast, in the LLDA, all the procedures are concurrently optimized directly for classification. Their difference is more apparent from the results on a dataset which varies in illumination as well as pose in Table 1. Please refer to the recognition results of the frontal faces by the conventional PCA-LDA for comparison. Illumination changes were relatively well compensated and generalized to novel test faces as compared with pose variations. Some results are ommited here and it is because the results were similar to those of the previous subsection.

Table 1. Recognition rates (in %) on PIE DB.

PCA-LDA		PCA-LM-LDA				LLDA			
		R1	59	L1	44	R1	59	L1	59
F2	85	R2	26	L2	22	R2	41	L2	30
F3	100	R3	56	L3	44	R3	56	L3	63
F4	100	R4	56	L4	37	R4	56	L4	56
F5	67	R5	30	L5	15	R5	30	L5	26
avg	88		45		33		48		47

3.4 Combining Results

Figure 6 shows the combining results of all the 4 base classifiers by the 6 different gating rules. All 6 different gating rules improved the average performance of the best base classifier.

The number of combined experts ranges from 1 up to 4. We first find the best expert, LLDA and then add the next best performing experts, the sequence of which is PCA-LM-LDA, 3D LUT, GDA, yielding the combined results by the sum and product rules. The results are shown in Figure 7. Interestingly, the recognition rate consistently improved as the number of different base classifiers increased. It is also noted that the improvement rate achieved with 2 experts was relatively low in the case of the different session experiment. This might be because the two combined base classifiers, LLDA and PCA-LM-LDA are the most correlated classifiers, due to the similar sources of information used and their piecewise linear structures. In conclusion, the performance improvement achieved by the proposed combining classifier is quite impressive compared with the conventional PCA-LDA method in face recognition: 46.8% → 73.2% for the same session and 34.8% → 54.4% for the different session respectively.

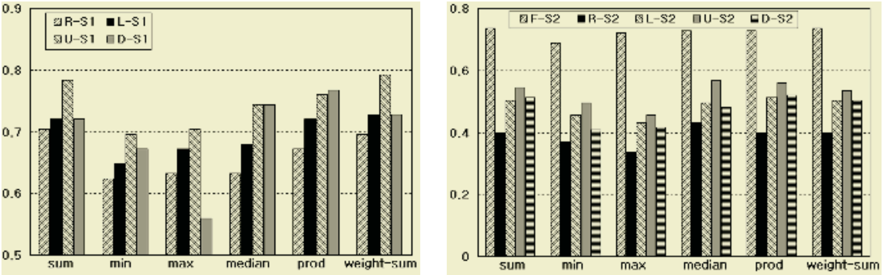

Fig. 6. Recognition rates (in %) of combining classifiers on XM2VTS DB.

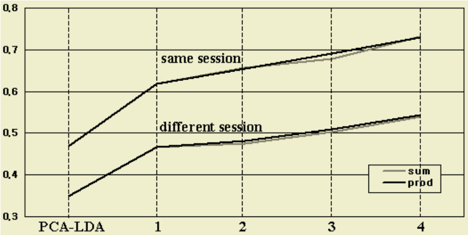

Fig. 7. Average recognition rate for the number of the combined base classifiers.

4 Conclusion

We have proposed a combining classifier based on the modelling of face view-changes. Robust base classifiers are obtained by learning the statistics of 2D images or fitting generic 3D models. The proposed base classifiers outperform the classical method of LDA. The fusion of the different classifiers yields an impressive performance improvement owing to their different characteristics in terms of sources of information exploited and architectures used. We intend to improve the performance of the proposed approach by exploiting dense facial feature correspondences for an image regularization step in the future. The current performance was obtained with the images registered with a fixed eye position and this can be seen as a poor basis of the image normalization for the method.

Acknowledgement

The authors would like to thank Hyun-Chul Kim, Hyung-Soo Lee and Daijin Kim in POSTECH for their helpful discussion. The assistance of Kyung-Ah Sohn, a colleague in SAIT, in implementing the algorithms based on the 3D face models is much appreciated

References

1. T.Vetter and T.Poggio, “Linear object classes and image synthesis from a single example image”, IEEE Trans. PAMI, vol. 19, no. 7, pp. 733-742, 1997.
2. Y.Li, S.Gong, and H.Liddell, “Constructing facial identity surfaces in a nonlinear discriminating space”, In Proc. of CVPR, 2001.
3. D.B.Graham, N.M.Allinson, “Automatic face representation and classification”, In Proc. of British Machine Vision Conference, 1998.
4. V.Blanz,S.Romdhani,T.Vetter, “Face identification across different poses and illuminations with a 3D morphable model”, Automatic Face and Gesture Recognition, 2002.
5. T.-K.Kim, J.Kittler, H.-C. Kim and S.-C.Kee, “Discriminant analysis by multiple locally linear transformations”, British Machine Vision Conference, pp 123-132, Norwich, UK, 2003.
6. A.Talukder and D. Casasent, “Pose-invariant recognition of face at unknown aspect views”, IEEE Joint Conf. on Neural Networks, vol. 5, pp. 3286-3290, 1999.
7. G.Baudat and F.Anouar, “Generalized discriminant analysis using a kernel approach”, Neural Computation, vol.12, pp.2385-2404, 2000.
8. A.C.Aitchison and I.Craw, “Synthetic images of faces – an approach to model-based face recognition”, British Machine Vision Conference, pp. 226-232, 1991.
9. D.Beymer and T.Poggio, “Face Recognition From One Example View”, In Proc. of ICCV, pp. 500-507, 1995.
10. T.-K.Kim, “View-transformation of face images in kernel space-comparative study”,Technical Report, Samsung AIT, 2003.
11. B.Choe, H.Lee, and H.-S.Ko, “Performance-driven muscle-based facial animation”, The Journal of Visualization and Computer Animation, vol. 12, pp. 67-79, 2001.
12. T.Sim, S.Baker and M.Bsat, “The CMU pose, illumination, and expression (PIE) database”, In Proc. of Automatic Face and Gestures Recognition, 2002.
13. K. Okada, C. v. d. Malsburg, “Analysis and synthesis of human faces with pose variations by a parametric piecewise linear subspace method”, In Proc. of CVPR, vol. 1 , pp. I -761-8, 2001.
14. A. Pentland, B. Moghaddan, T. Starner, “View-based and modular eigenspaces for face recognition”, In Proc. of CVPR, pp. 84-91, 1994.
15. R. Gross, I. Matthews, and S. Baker, “Eigen light-fields and face recognition across pose”, In Proc. of Automatic face and Geature Recognition, 2002.
16. A.S.Georghiades, P.N.Belhumeur, and D.J.Kriegman, “From few to many: illumination cone models for face recognition under varialbe lighting and pose”, IEEE Trans. on PAMI, vol. 23, no. 6, pp. 643-660, 2001.
17. J. Kittler, M. Hatef, R. P.W. Duin and J. Matas, “On combining classifiers”, IEEE Trans. PAMI, vol. 20, no. 3, pp. 226-239, 1998.
18. S. Li, J. Yan, X. Hou, Z. Li and H. Zhang, “Learning low dimensional invariant signature of 3-D object under varying view and illumination from 2-D appearances”, In Proc. of ICCV, vol. 1, pp. 635 –640, 2001.
19. K.Messer, J.Matas, J.Kittler, J.Lueettin and G.Maitre, “XM2VTSDB: The extended M2VTS database”, In Prof. of Audio and Video-based Biometric Person Authentication, 1999.