2008－CVIM－162 (75)
2008／3／11

Comparison of MDA and EMC in Robustness against Over-fitting for Facial Expression Recognition

Fan CHEN[1] Kazunori KOTANI[1]

School of Information Science,
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, 923-1211, Japan[1]
email: {chen-fan,ikko}@jaist.ac.jp

Abstract

Eigen-space Mehod based on Class-features (EMC), a variant of Multiple Discriminant Analysis (MDA), has been proposed and applied for automatic facial expression recognition. Although EMC was reported to outperform MDA in Ref. [1][2], no mathematical explanations for the difference of performance have been given. In the present paper, we will first refomualte MDA and EMC based on a new model of Maximum Log Likelihood (MLL) estimation. By using this model, we will explain from the perspective of statistical inference that the difference of the underlying mechanism locates in that EMC is a variant of MDA with lower degree of freedom by assuming the covariance to be sphered in all directions. A thorough comparison between EMC and MDA in robust recognition of facial expressions will also be made to verify our conclusion that EMC outperforms MDA because it is more robust against over-fitting due to its lower degree of freedom.

1 Introduction

MDA is an extension of Fisher discriminant analysis [3] to the multi-class case, which tries to maximize the discrimination among multiple clusters. Due to the simplicity of its implementation and clear physical meaning, MDA has been widely used in the literature of pattern recognition, which also includes facial expression recognition [4].

Various improvements have been proposed to further enhance the performance of MDA in pattern recognition. Beside improvements that focus on combining MDA and other methods, two important enhancements are EMC [1][2] and Heteroscedastic Discriminant Analysis(HDA) [5]. It is shown that HDA was more effective than MDA in speech recognition and text recognition, while EMC was proposed for facial expression recognition and was verified to outperform MDA in recognition rate without a mathematical analysis on the reason. [1][2] We will focus on the comparison of performance between EMC and MDA in facial expression recognition in the present paper.

If we put these methods in the framework statistical inference for deep consideration, we could make further mathematical explanations for the difference of MDA and EMC. By using MLL estimation, it is easy to show that MDA assumes that all clusters have the same covariance [6]. Therefore, most improvements are motivated by loosing or strengthening the constraint on the covariance. HDA was derived by removing the constraint of same covariance on all clusters, In contrast, We will show that EMC assumes sphered covariance for all clusters, which is thus a discriminant model with less degree of freedom than MDA. For facial expression recognition and similar tasks where the size of samples is much smaller than data dimensionality, over-fitting might occur due to the "curse of dimensionality"[7]. A model with less degree of freedom than MDA might be more robust against over-fitting while HDA with higher degree of freedom might be weak in over-fitting.

In the present thesis, we will deep into the comparison between MDA and EMC, give a mathematical analysis on their difference and make a thorough investigation on their relative performance in facial expression recognition. The paper is organized as follows: in Section 2, we will first derive MDA and EMC and compare their characteristics in a mathematical way. In Section 3, a thorough comparison of performance between MDA and EMC on facial expression recognition will be made. Finally, we conclude the present paper.

2 MDA and EMC in Facial Expression Recognition

2.1 Facial Expression Recognition by MDA or EMC

Given an matrix of N observed data $\mathbf{Y} = [\mathbf{y}_n | n = 1, \cdots, N]$ with $\mathbf{y}_n = [y_{nd} | d = 1, \cdots, D]^T$ be a D-dimensional vector from raster-scanned image, we are trying to recover the hidden factors $\mathbf{S} = [\mathbf{s}_n | n = 1, \cdots, N]$ by projecting observed data into a matrix of Q bases $\mathbf{W} = [\mathbf{w}_q | q = 1, \cdots, Q]$ which is obtained under a specified criterion, i.e., $\mathbf{s}_n = \mathbf{W}^T \mathbf{y}_n$. Without loss of generality, we assume that all observed data are mean-centered, i.e., $\sum_n \mathbf{y}_n = 0$. As a supervised method, the true class labels of all training data are known. We assume that all observed data $\{\mathbf{y}_n\}$ are from K classes, and $\mathbf{Z} = [z_n | z_n \in \{1, \cdots, K\}, n \in \{1, \cdots, N\}]$ represents the labelling of samples. We let $N_k = \sum_n \delta_{z_n,k}$ be the number of samples in class k, which also

satisfies $\sum_k N_k = N$.

$$\delta_{a,b} = \begin{cases} 1, a = b \\ 0, a \neq b \end{cases} \tag{1}$$

is the Kronecker delta function. Let $\widetilde{\mathbf{Y}} = [\widetilde{\mathbf{y}}_n | n \in \{1, \cdots, \widetilde{N}\}]$ be the matrix by putting all testing images into different columns, and let $\widetilde{N}$ be the number of samples in the testing set. We define their true classified labels as $\widetilde{\mathbf{Z}} = \{\widetilde{z}_n \in \{1, \cdots, K\} | n \in \{1, \cdots, \widetilde{N}\}\}$, and define a recognition rate as

$$r_c = \frac{1}{\widetilde{N}} \sum_n \delta_{\widetilde{z}_n, z_n^*}. \tag{2}$$

z_n^* is the estimated label value which belongs to the class whose center is the closest to the current feature vector, i.e.,

$$z_n^* = \arg\min_k \mathcal{D}(\widetilde{\mathbf{s}}_n, \mathbf{W}^T \overline{\mathbf{y}}_k) \tag{3}$$

and $\widetilde{\mathbf{s}}_n = \mathbf{W}^T \widetilde{\mathbf{y}}_n$. $\overline{\mathbf{y}}_k = \sum_{i=1}^N \delta_{z_n,k} \mathbf{y}_n / N_k$ is the mean of training samples belonging to class k. $\mathcal{D}(\mathbf{s}_a, \mathbf{s}_b)$ is a distance function between two feature vectors, and can be defined in various ways.

A kernel problem is to estimate the optimal matrix $\mathbf{W}$ that maximizes the ability of features in class discrimination. In MDA, $\mathbf{W}$ is optimized by solving a generalized eigen problem, i.e., $\lambda S_W \mathbf{W} = S_B \mathbf{W}$, while in EMC, the bases are recovered by solving $\lambda \mathbf{W} = (S_B - S_W)\mathbf{W}$. where S_B and S_W are the between-class and within-class scatter matrices, defined by Eq.(5).

$$S_B = \frac{1}{N} \sum_k N_k (\overline{\mathbf{y}}_k - \overline{\mathbf{y}})(\overline{\mathbf{y}}_k - \overline{\mathbf{y}})^T, \tag{4}$$

$$S_W = \frac{1}{N} \sum_{nk} \delta_{z_n,k} (\mathbf{y}_n - \overline{\mathbf{y}}_k)(\mathbf{y}_n - \overline{\mathbf{y}}_k)^T. \tag{5}$$

A block diagram for appearance based recognition of facial expression by using MDA or EMC is shown in Fig.1. Our recognition system is divided into two phases, i.e., a training phase and a testing phase. In training phase, we calcuate scatter matrics and recover de-mixing bases from the training data. In the running phase, all testing images will be projected into the space spanned by the de-mixing bases. The projected data will be used as a feature to make classification by using the criterion defined in Eq. 3.

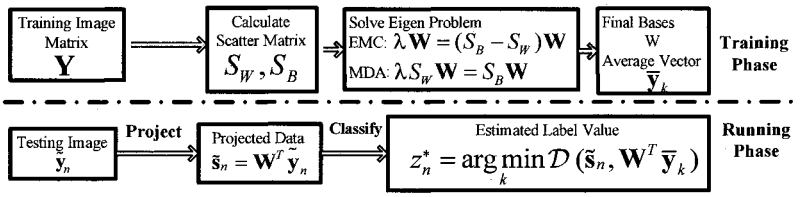

Fig. 1 Block Diagram for Appearance based Recognition with EMC/MDA.

2.2 MLL Derivation of MDA and EMC

We assume that all data from K clusters are distributed as K normal distribution, i.e.,

$$P(\mathbf{Y}|\mathbf{Z}) = \prod_n \prod_k \left\{ \frac{\exp[-(\mathbf{y}_n - \mu_k)^T \Sigma_k^{-1} (\mathbf{y}_n - \mu_k)/2]}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \right\}^{\delta_{z_n,k}}. \tag{6}$$

It is easy to derive optimal estimation for μ_k and Σ_k as

$$\mu_k = \sum_n \delta_{z_n,k} \mathbf{y}_n / N_k, \tag{7}$$

$$\Sigma_k = \sum_n \delta_{z_n,k} (\mathbf{y}_n - \mu_k)(\mathbf{y}_n - \mu_k)^T / N_k. \tag{8}$$

We form a full rank set $\mathbf{W}^F$ with D bases, with its topmost Q components taking from $\mathbf{W}$, i.e., $\mathbf{W}^F = [\mathbf{w}_q | q \in \{1, \cdots, D\}]$. Unit length and orthogality are assumed to those bases, i.e., $|\mathbf{w}_q| = 1, \mathbf{w}_p^T \mathbf{w}_q = \delta_{p,q}$. Let $\mathbf{a}_n = \sum_{q=1}^Q \mathbf{s}_{nq} \mathbf{w}_q$ be the reconstructed data from estimated factors, we have the relationship as $\mathbf{a}_n = \mathbf{y}^n - \sum_{q=Q+1}^D \mathbf{s}_{nq} \mathbf{w}_q$. We take a model for discriminant analysis which is different from that in Ref.[5]. In Ref.[5], in order to show that classification information is only included in $\mathbf{W}$, they assumed data extracted

by the topmost Q components of $\mathbf{W}^F$ to be cluster-wise distributed, and let data obtained from other $D-Q$ components be distributed on one normal distribution. We assume that the reconstructed data should satisfy cluster-wise Gaussian distribution as well as the original observed data. In other words, data extracted by $\mathbf{W}$ maintain the majority of classification information. Accordingly, we derive the log-likelihood for the recovered data $\mathbf{A} = \{\mathbf{a}_n | n \in \{1, \cdots, N\}$ as

$$\begin{aligned}
\mathcal{L} &= \log P(\mathbf{A}|\mathbf{Y}, \mathbf{Z}) \\
&= \log \prod_n \prod_k \left\{ \frac{\exp[-(\mathbf{a}_n - \mu_k)^T \Sigma_k^{-1} (\mathbf{a}_n - \mu_k)/2]}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \right\}^{\delta_{z_n,k}} \\
&= -\sum_{nk} \delta_{z_n,k} \frac{(\mathbf{a}_n - \mu_k)^T \Sigma_k^{-1} (\mathbf{a}_n - \mu_k)}{2} - \frac{ND}{2} \log(2\pi) - \sum_{nk} \delta_{z_n,k} \frac{1}{2} \log |\Sigma_k|. \qquad (9)
\end{aligned}$$

We maximize the above log-likelihood by minimizing the criterion for $\mathbf{W}$ which is defined as

$$\begin{aligned}
\mathcal{L}_W &= \sum_{nk} \delta_{z_n,k} (\mathbf{a}_n - \mu_k)^T \Sigma_k^{-1} (\mathbf{a}_n - \mu_k) \\
&= \sum_{nk} \delta_{z_n,k} (\mathbf{y}_n - \sum_{q=Q+1}^{D} s_{nq} \mathbf{w}_q - \mu_k)^T \Sigma_k^{-1} (\mathbf{y}_n - \sum_{q=Q+1}^{D} s_{nq} \mathbf{w}_q - \mu_k). \qquad (10)
\end{aligned}$$

2.2.1 EMC: $\Sigma_k = \text{diag}\{\sigma^2, \cdots, \sigma^2\}$

For a simple case, we assume an equal and diagonal covariance for all clusters. Furthermore, we assume equal variances in the direction of all axes, i.e., $\Sigma_k = \text{diag}\{\sigma^2, \cdots, \sigma^2\}$, we have

$$\begin{aligned}
\mathcal{L}_W^{EMC} &= \sum_{q=Q+1}^{D} \{ \sum_{nk} \delta_{z_n,k} [s_{nq}^2 - 2(\mathbf{y}_n - \mu_k)^T (s_{nq} \mathbf{w}_q)] \} \\
&= \sum_{q=Q+1}^{D} \mathbf{w}_q^T \{ \sum_{nk} \delta_{z_n,k} (-\mathbf{y}_n \mathbf{y}_n^T + 2\mathbf{y}_n \mu_k^T) \} \mathbf{w}_q \\
&= \sum_{q=Q+1}^{D} \mathbf{w}_q^T \{ \sum_{nk} \delta_{z_n,k} (\mu_k \mu_k^T - (\mathbf{y}_n - \mu_k)(\mathbf{y}_n - \mu_k)^T) \} \mathbf{w}_q \\
&= \sum_{q=Q+1}^{D} \mathbf{w}_q^T (S_B - S_W) \mathbf{w}_q. \qquad (11)
\end{aligned}$$

Optimizing $\mathcal{L}_W^{EMC}$ under constraints using a Lagrange multiplier λ, we have

$$\begin{aligned}
0 &= \frac{\partial}{\partial \mathbf{w}_q} \{ \mathcal{L}_W^{EMC} - \lambda (\mathbf{w}_q^T \mathbf{w}_q - 1) \} \\
&= (S_B - S_W) \mathbf{w}_q - \lambda \mathbf{w}_q, \qquad (12) \\
\Rightarrow \lambda \mathbf{w}_q &= (S_B - S_W) \mathbf{w}_q, \qquad (13)
\end{aligned}$$

which means that $\{\mathbf{w}_q\}$ are eigen-vectors of $S_B - S_W$. In orderto minimize $\mathcal{L}_W^{EMC}$, those bases dropped should correspond to the smallest $D-Q$ eigen-values.

2.2.2 MDA: $\Sigma_k = \Sigma$

In MDA, we assume equal covariance for all components, i.e., $\Sigma_k = \Sigma$. Accordingly, we have optimal value for Σ as

$$\Sigma = \sum_{nk} \delta_{z_n,k} (\mathbf{y}_n - \mu_k)(\mathbf{y}_n - \mu_k)^T / N = S_W. \qquad (14)$$

Since Σ is symmetrical and positive-definite, we have $\Sigma = \Sigma^{1/2} \Sigma^{1/2}$

$$\mathcal{L}_W = \sum_{nk} \delta_{z_n,k} [\Sigma^{-1/2} (\mathbf{a}_n - \mu_k)]^T [\Sigma^{-1/2} (\mathbf{a}_n - \mu_k)]. \qquad (15)$$

If we let $\hat{\mathbf{a}}_n = \Sigma^{-1/2} \mathbf{a}_n$, $\hat{\mu}_k = \Sigma^{-1/2} \mu_k$, $\hat{\mathbf{y}}_n = \Sigma^{-1/2} \mathbf{y}_n$, we use the same method in EMC to derive

$$\mathcal{L}_W^{MDA} = \sum_{q=Q+1}^{D} \mathbf{w}_q^T \Sigma^{-1/2} (S_B - S_W) \Sigma^{-1/2} \mathbf{w}_q \qquad (16)$$

Optimizing $\mathcal{L}_W^{MDA}$ under constraints using the Lagrange multiplier λ, we have

$$\begin{aligned} 0 &= \frac{\partial}{\partial \mathbf{w}_q}\{\mathcal{L}_W^{MDA} - \lambda(\mathbf{w}_q^T\mathbf{w}_q - 1)\} \\ &= \Sigma^{-1/2}(S_B - S_W)\Sigma^{-1/2}\mathbf{w}_q - \lambda\mathbf{w}_q, \qquad (17) \\ \Rightarrow (\lambda+1)\mathbf{w}_q &= S_W^{-1/2}S_B S_W^{-1/2}\mathbf{w}_q, \qquad (18) \end{aligned}$$

which means that $\{\mathbf{w}_q\}$ are eigen-vectors of $S_W^{-1/2}S_B S_W^{-1/2}$, which is equivalent to MDA. Furthermore, to minimize $\mathcal{L}_W^{MDA}$, those bases dropped should correspond to the smallest $D-Q$ eigen-values.

2.3 MDA vs. EMC

From the above models, we know MDA has a higher degree of freedom than EMC in modelling data. In Fig.2, we show the detailed modelling of MDA and EMC, i.e., MDA assumes normal distribution for all clusters with same covariance and different mean while EMC assumes same and sphered covariance for all clusters. Basically, MDA should provide a better fitting for given training data because of its higher degree of freedom. However, due to the "curse of dimensionality", over-fitting might occur during the training on facial images, as what we will show in next section. A model with lower degree of freedom should be more robust against overfitting in facial expression recognition, where EMC might outperform MDA as suggested by Ref.[1][2]. This will be a major topic we will discuss in the present paper.

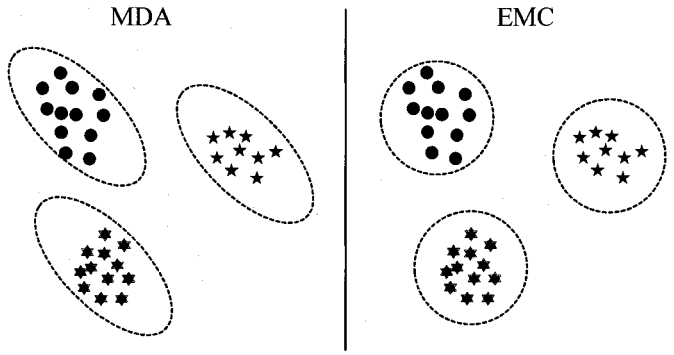

Fig. 2 Diagram of MLL models for MDA and EMC given in the present paper.

Fig. 3 Some normalized samples that are used in our numerical experiments from the Cohn-Kanade database. (a) Neutral (b) Happiness (c) Anger (d) Fear (e) Disgust (f) Sadness and (g) Surprise.

3 Numerical Experiments

We will focus our experiments on the comparison between the performance of EMC and MDA on facial expression recognition. Especially, their abilities against over-fitting will be discussed. We form a sub-database of sample images from Cohn-Kanade Database [9] for further recognition, which has 391 images from 71 subjects we have selected. For any subject, we have at most one image for one facial expression. All images are normalized in eye and nose-tip positions and resized to 70x80 pixels. Data vectors are further normalized to have unit lengths. Each time we randomly choose N images from the sub-database as the training set and use all images left for validation. Five different metrics will be investigated, which are L2(Euclidian), Cosine, Mahalanobis L2, Mahalanobis L1, and Mahalanobis Cosine. Details for the definition of those distance metrics could be found in Ref.[8]. In Fig.3, we shown some image samples from the Cohn-Kanade database for each facial expression.

Tabel 1 Comparison of recognition accuracy between MDA and EMC on the Cohn-Kanade Database. Different metrics have been used.

Trainning			EMC					MDA		
Size	L2	Cosine	MahL2	MahL1	MahCos	L2	Cosine	MahL2	MahL1	MahCos
100	**66.32**	59.45	62.20	61.86	62.54	62.54	57.04	58.42	59.11	55.33
120	**66.05**	59.41	63.84	63.84	61.25	59.78	55.72	53.14	53.14	49.45
140	**64.94**	61.35	60.56	60.56	**64.94**	63.75	60.16	56.18	58.17	50.20
160	66.23	58.87	64.94	68.83	65.37	**71.43**	60.11	68.40	66.67	59.74
180	65.40	68.24	67.30	68.25	**72.91**	64.45	59.72	58.29	59.71	56.40
200	63.87	66.49	66.49	65.97	68.06	**69.11**	65.45	56.02	54.97	60.21
220	68.42	68.42	70.76	**71.93**	66.67	66.08	57.31	54.97	59.65	60.23

Tabel 2 Comparison of recognition accuracy between MDA and EMC on the JAFFE and FEEDTUM-mixed Database. Different metrics have been used.

Trainning			EMC					MDA		
Size	L2	Cosine	MahL2	MahL1	MahCos	L2	Cosine	MahL2	MahL1	MahCos
60	39.60	35.64	43.56	40.59	**50.50**	46.53	49.50	43.56	44.55	47.52
70	53.85	53.85	52.75	52.75	**54.95**	**54.95**	**54.95**	45.05	47.25	46.15
80	**51.85**	44.44	48.15	45.68	48.15	49.38	48.15	45.68	48.15	50.62
90	**60.56**	56.34	53.52	54.93	56.34	54.93	53.52	53.52	49.30	53.52

We first list recognition accuracy r_c on the testing data under different training size N that we have tested in Table 1. We emphasize that sizes of all training datasets are small in the meaning of achieving robust facial expression recognition, therefore randomly selected training samples are biased in data distribution. Robustness against overfitting caused by this bias is a major topic in this paper. We find that EMC shows better results than MDA for almost all case due to its better robustness against over-fitting. Especially, for a smaller size of traing set, EMC outperforms MDA. We have also tested other databases, such as JAFFE [11] and FEEDTUM [12]. The results given in Table.2 show similiar trends in the relative performance of EMC and MDA. Since both of these two databases have less subjects than that in the Cohn-kanade database, we think that results on Cohn-Kanade database are more reliable.

Furthermore, if we show the topmost two components of projected training data from both EMC and MDA in Fig.4, it is easy to see that even for the largest training dataset, training data projected by MDA are heavily over-fitted. All data of the same class are clustered into one point in MDA, which means the covariance model in MDA has a much higher degree of freedom than required. Due to the possiblitiy of over-fitting, we could consider that in MDA, the strength in minimizing $\mathbf{w}^T S_W \mathbf{w}$ is much stronger than that in maximizing $\mathbf{w}^T S_B \mathbf{w}$, which is considered to cause the singularity in projected data. In EMC, the strength of $\mathbf{w}^T S_W \mathbf{w}$ is bounded. Therefore, it intends to focus on maximizing $\mathbf{w}^T S_B \mathbf{w}$, i.e., the distance between different clusters, which is thought to be more important in efficient classification.

In general, we consider improvements by increasing the complexity of a mathematical model so that it can deal with complex data structures. For tasks such as text recognition, a large database is possible and the groudtruth is trustable. In those cases, a model with higher degree of freedom such as HDA will fit the complex data better and enhance the discrimination ability of model, which will improve the overall performance.

However, for facial expression recognition, we still lack a large public database for unbiased training due to difficulties from both the various personal difference and the evaluation of groundtruth, which causes the stong database-dependency of training results in recognition performance. In this case, training is usually performed on a small database where the over-fitting might occur, and the factor that affects the performance most is the robustness against over-fitting. In the present paper, we argue that this kind of robustness could be achieved by reducing the degree of freedom. The reason for EMC ourperforming MDA in dealing with facial expression recognition is that EMC has a lower degree of freedom and thus is more robust against over-fitting.

4 Conclusions

In this paper, we focused on the comparison between MDA and EMC in both their mathematical meanings and their recognition performances of facial expressions. We proposed a new maximum log-likehood based model for both MDA and EMC. We further use this model to explain their difference in robustness against over-fitting. Although many improvements for pattern recognition focus on increasing the degree of freedom in modelling the data to achieve better fitting, we conclude that for image recognition where the dimensionality is much larger than the size of database, higher degree of freedom might cause over-fitting. Therefore, a

model with lower degree of freedom might perform better, just as Fisher recommended in Ref. [13] to use same covariance even for heteroscedastic data. Due to this reason, we explained that EMC was reported to outperform MDA in facial expression recognition because EMC has a lower degree of freedom than MDA which helps to prevent the occurrence of over-fitting. By making experiments on the Cohn-Kanade database, the mixtured Jaffe and FEEDTUM database under different metrics, we verify that EMC is more robust than MDA in dealing with biased training datasets for accurate facial expression recognition.

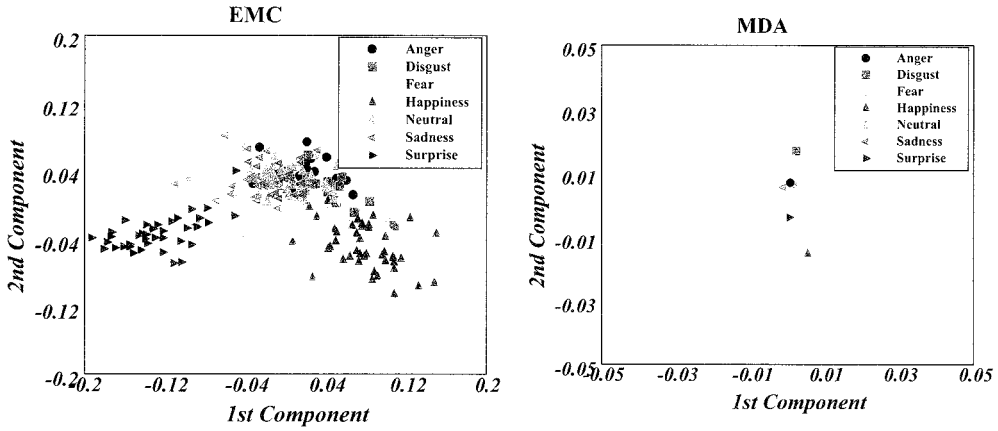

Fig. 4 Projected training data with sample size 220 from the Cohn-Kanade Database. Data are obtained from both EMC and MDA.

References

[1] T. Kurozumi, Y. Shinza, Y. Kenmochi, and K. Kotani, "Facial Individuality and Expression Analysis by Eigenspace Method or Multiple Discriminant Analysis," *IEICE Technical Report, Communication Systems*, Vol.98, No.482, pp.57-64, 1998.

[2] T. Kurozumi, Y. Shinza, Y. Kenmochi, and K. Kotani, "Facial Individuality and Expression Analysis by Eigenspace Method based on Class Features or Multiple discriminant Analysis," *IEEE Int'l Conf. on Image Processing*, 25PP6A, 1999.

[3] R.A. Fisher, "The statistical utilization of multiple measurements," *Annals of Eugenics*, Vol.8, pp.376-386, 1938.

[4] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, ""Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, No.7, pp.711-720, 1997.

[5] N. Kumar and A. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, Vol.26, pp.283-297, 1998.

[6] T. Hastie and R. Tibshrani, "Discriminant Analysis by Gaussian Mixtures," *Journal of the Royal Statistical Society*, Vol.58, Vol.158-176,1996.

[7] L. Kanal, and B. Chandrasekaran, "On dimensionality and sample size in statistical pattern classification," *Pattern Recognition*, pp.225-234, 1971.

[8] A.R. Webb, "Statistical Pattern Recognition," (2nd Ed.) *John Wiley and Sons*, 2002.

[9] T. Kanade, J. F. Cohn and Y. Tian, "Comprehensive database for facial expression analysis," *FG'00*, Vol.1, pp.46-53, 2000.

[10] V. Vapnik, "The Nature of Statistical Learning Theory," Springer, New York, 1995.

[11] M.J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," *FG'98*, Vol.1, pp.200-205, 1998.

[12] F. Wallhoff, "Database with Facial Expressions and Emotions from Technical University of Munich (FEEDTUM),"
http://www.mmk.ei.tum.de/ waf/fgnet/feedtum.html

[13] R.A. Fisher, "Contributions to Mathematical Statistics," John Wiley and Sons, New York, 1952.