International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:12, 2008

New Adaptive Linear Discriminante Analysis for Face Recognition with SVM

Mehdi Ghayoumi

***Abstract*—**We have applied new accelerated algorithm for linear discriminate analysis (LDA) in face recognition with support vector machine. The new algorithm has the advantage of optimal selection of the step size. The gradient descent method and new algorithm has been implemented in software and evaluated on the Yale face database B. The eigenfaces of these approaches have been used to training a KNN. Recognition rate with new algorithm is compared with gradient.

***Keywords*—** lda, adaptive, svm, face recognition.

I. Introduction

The need for adaptive image processing arises due to the need to incorporate adaptive aspects of biological vision into machine vision systems. In recent years, many algorithms have been presented to simulate the human process, in which various adaptive mechanisms are introduced such as neural networks, genetic algorithms, and support vector machines [1]. The most important part of a face recognition system is to handle all kinds of variations through modeling. There are many different kinds of variations of human faces due to the growth and the changes in lighting conditions, make-up, pose, illumination, expression, etc. The face recognition systems can achieve excellent performance when tested over a standard database, but they haven't this performance when they are operated in a practical environment. This is because the training set of face images will be either insufficient or inappropriate for future events. Even if a large amount of face images are available when constructing a face recognition system, all the variations that will happen in future cannot be considered in advance; thus high recognition performance in practical situations can hardly be expected with only a static database. A solution to this problem is to make face recognition systems learn continuously to adapt to incoming training samples. In many face recognition systems, the information processing consists of feature selection and classification.

On the other hand, one of the difficulties of classification methods is dimension of training samples. Choosing an appropriate set of features is a critical problem in classification systems. Recently, there has been an increased interest in eploying feature selection in applications such as face and gesture recognition [2]. In this paper, we use an adaptive LDA to subspace representation of database that gives high separability power and a SVM to classify features. Adaptive LDA algorithms have been used in on-line applications particularly for feature space dimensionality reduction [3]. Mao and Jain [4] proposed a two-layer network for LDA, each of which was a PCA network. Chatterjee and Roychowdhury [5] presented an adaptive algorithm and a self-organizing LDA network for feature extraction. Recently, Abrishami Moghaddam and Amiri Zadeh derived an accelerated convergence adaptive algorithm for LDA, based on the steepest descent optimization method [6].

The next section describes the fundamentals of LDA. In Section 3, the adaptive computation of the square root of the inverse covariance matrix $\Sigma^{-1/2}$ based on the gradient descent method is presented and its convergence is proved using the stochastic approximation theory. Section 4 is devoted to new accelerated adaptive $\Sigma^{-1/2}$ algorithms. A new recursive equation for on-line estimation of the covariance matrix is also presented in this section. Experimental results are introduced in Section 5 demonstrate the superior performance of the proposed methods. Finally, concluding remarks and plans for future works are given in section6.

II. Linear Discriminate Analysis

LDA criteria are mainly based on a family of functions of scatter matrices. For example, the maximization of $tr(\Sigma_w^{-1}\Sigma_b)$ or $|\Sigma_b|/|\Sigma_w|$ is used, where Σ_w, Σ_b are within and between-class scatter matrices, respectively. In LDA, the optimum linear transform is composed of $r(\leq n)$ eigenvectors of $\Sigma_w^{-1}\Sigma_b$ corresponding to its r largest eigenvalues. Alternatively, $\Sigma_w^{-1}\Sigma_m$ can be used for LDA, where Σ_m represents the mixture scatter matrix $(\Sigma_m = \Sigma_b + \Sigma_w)$. A simple analysis shows that both $\Sigma_w^{-1}\Sigma_b$ and $\Sigma_w^{-1}\Sigma_m$ has the same eigenvector matrix ϕ. In general, Σ_b is not full rank, hence Σ_m is used in place of Σ_b. The computation of the eigenvector matrix ϕ from $\Sigma_w^{-1}\Sigma_m$ is equivalent to the solution of the generalized eigenvalue problem $\Sigma_m\phi = \Sigma_w\phi\Lambda$, where Λ is the eigenvalue matrix [7]

III. Adaptive $\Sigma^{-1/2}$ Computation Algorithm

The following algorithm has been proposed for the adaptive computation of $\Sigma^{-1/2}$ [8]:

M.Ghayoumi, Islamic Azad University Shahr-e-ray Branch, P.O. Box 1653963113, Tehran, Iran (e-mail: ghayoumi_me@sr.iau.ac.ir).

$$\mathbf{W_{k+1}} = \mathbf{W_k} + \boldsymbol{\eta_k}\mathbf{G_k} \quad (1)$$

$$\mathbf{G_k} = \mathbf{I} - \mathbf{W_k}\mathbf{x_k}\mathbf{x_k^T}\mathbf{W_k} \quad (2)$$

Where $\mathbf{W_0} \in R^{n\times n}$ is symmetric and non-negative definite, and $\{\eta_k\}$ is a scalar gain sequence. According to the general form of adaptive algorithms [9] we have:

$$\mathbf{W_{k+1}} = \mathbf{W_k} + \boldsymbol{\eta_k}\mathbf{G}(\mathbf{W_k}, \mathbf{x_k}) \quad (3)$$

Where the update function $\mathbf{G}(\mathbf{W_k}, \mathbf{x_k})$ is the gradient of an objective function $J(\mathrm{W_k})$. Using the stochastic approximation theory and convergence analysis by Ljung [10], we may write:

$$\frac{\partial J(\mathbf{W_{k+1}})}{\partial \eta_k} = \frac{\partial J(\mathbf{W_{k+1}})}{\partial \mathbf{W_{k+1}}} \cdot \frac{\partial \mathbf{W_{k+1}}}{\partial \eta_k}, \quad (4)$$

$$\frac{\partial J(\mathbf{W_{k+1}})}{\partial \mathbf{W_{k+1}}} = \mathbf{I} - \mathbf{W_{k+1}}\boldsymbol{\Sigma_k}\mathbf{W_{k+1}},$$

$$\mathbf{G(W)} = \frac{\partial \mathbf{J(W)}}{\partial \mathbf{W}} = \mathbf{E}[\mathbf{G}(\mathbf{W_k}, \mathbf{x_k})] = \mathbf{I} - \mathbf{W}\Sigma\mathbf{W}$$

The gain sequence $\{\eta_k\}$ has an important role in the convergence of the algorithm and can be a constant or a decreasing sequence, satisfying the following conditions:

$(a) \sum_{k=0}^{\infty} \eta_k = \infty,$

$(b) \sum_{k=0}^{\infty} \eta_k^r \prec \infty (r \succ 1),$

$(c) \lim_{k\to\infty} \eta_k \to 0.$

For example, we can generate η_k as follows [6]:

$$\eta_k = \delta / k^{\alpha} \quad \delta \succ 0, \quad 1/2 \prec \alpha \le 1, \quad (5)$$

Where α, δ are selected, such that η_k satisfies the above stated conditions. The convergence of the algorithm has been proved [8] using the stochastic approximation theory [10]. That means:

$$\lim_{k\to\infty} \mathbf{W}_k = \boldsymbol{\Sigma}^{-1/2} \quad \text{with probability one,} \quad (6)$$

Where $\boldsymbol{\Sigma}$ is the correlation or covariance matrix of the random sequence $\{\mathrm{x_k}\}$.

IV. New Adaptive $\boldsymbol{\Sigma}^{-1/2}$ Algorithms

The adaptive computation of $\Sigma^{-1/2}$ using Eq. (1), suffers from a very slow convergence rate. Increasing η_k can accelerate the convergence of the algorithm, but large gain sequences may cause it to diverge or converge to a false solution. Choosing η_k as a monotonically decreasing function of the iteration number k may improve the convergence rate However; this cannot be considered as an optimal solution to the convergence problem. Noting that Eq. (6) is based on the gradient descent method, we developed new algorithms based on the steepest descent [6] in order to optimally determine the gain sequence in each iteration. The steepest descent method uses the following updating equations:

$$\mathbf{W_{k+1}} = \mathbf{W_k} + \boldsymbol{\eta_k}\mathbf{G_k} \quad (7)$$

$$\mathbf{G_k} = \mathbf{I} - \mathbf{W_k}\mathbf{x_k}\mathbf{x_k^T}\mathbf{W_k} \quad (8)$$

Where $\mathrm{x_k x_k^T}$ in Eq. (2) has been replaced by η_k which will be introduced later in this section. In the steepest descent method, instead of using a fixed gain sequence, η_k is calculated by the derivative of the cost function J with respect to η_k [6]:

$$\frac{\partial J(\mathbf{W_{k+1}})}{\partial \eta_k} = 0, \quad (9)$$

(10)

Using the chain rule:

Therefore,

$$(\mathbf{I} - \mathbf{W_{k+1}}\Sigma_k\mathbf{W_{k+1}}) \cdot (\mathbf{I} - \mathbf{W_k}\Sigma_k\mathbf{W_k}) = 0. \quad (12)$$

Replacing $\mathrm{W_{k+1}}$ with Eq. (12) and doing some mathematical operations we obtain the following quadratic equation:

$$a_k\eta_k^2 + b_k\eta_k + c_k = 0, \quad (13)$$

Where:

$$a_k = (\underline{\mathbf{G_k}\boldsymbol{\Sigma_k}\mathbf{G_k}}) \cdot \underline{\mathbf{G_k}},$$

$$b_k = (\underline{\mathbf{G_k}\boldsymbol{\Sigma_k}\mathbf{W_k}} + \underline{\mathbf{W_k}\boldsymbol{\Sigma_k}\mathbf{G_k}}) \cdot \mathbf{G_k},$$

$$c_k = -\underline{\mathbf{G_k}} \cdot \underline{\mathbf{G_k}}$$

and η_k is obtained as:

$$\eta_k = \frac{-b_k + \sqrt{b_k^2 - 4a_kc_k}}{2a_k}. \quad (14)$$

In Eq. (14), we use the positive sign in order to minimize the objective function $J(\mathrm{W_{k+1}})$. As will be shown in experimental results, the computation of η_k according to Eq. (14) accelerates the convergence of the adaptive algorithm. Furthermore, in the adaptive computation of $\Sigma^{-1/2}$ a fixed gain sequence may cause divergence problems in the case of non-stationary input data. Dynamic determination of η_k can overcome the problem while the convergence is guaranteed under different conditions. The on-line estimation of the covariance matrix η_k is obtained using the following recursive equation [6]:

$$\boldsymbol{\Sigma_{k+1}} = (1 - \theta(k/k+1))\mathbf{x_{k+1}}\mathbf{x_{k+1}^T} + \theta(k/k+1)\boldsymbol{\Sigma_k}, \quad (15)$$

Where $\theta \in]0,1]$ is a forgetting scalar factor. If $\{\mathrm{x_k}\}$ comes from a stationary process, $\theta = 1$ is used. On the other hand, if $\{\mathrm{x_k}\}$ comes from a non-stationary process, $0 < \theta < 1$ is

selected. This equation is applied to obtain an effective window of size $1/(1-\theta)$. This effective window ensures that the past data samples are down-weighted with an exponentially fading window. The exact value of θ depends on the specific application. In general for slow time varying $\{x_k\}$, θ is chosen close to one to obtain a large effective window, whereas for fast time varying $\{x_k\}$, θ is chosen near zero for small effective window [9].

V. Experimental Results

Performance of methods is evaluated on Yale face database B [11]. The background of images is cut out, and the images are resized to 13×13 pixels with wavelet. For this experiment, 2 class recognition experiments are performed over 36 pairs of subjects. For each pair of subjects, a training data set is constructed from the 200 image of each subject. The dimensionality of the training subspace is reduced to 50 prior to recognition. The training and test images were histogram equalized and mean centered before classification. The convergence of the new $\Sigma^{1/2}$ algorithm for facial database is illustrated in Fig. 1. Convergence behaviour of the new $\Sigma^{-1/2}$ algorithm compared to the gradient descent method using facial database. Fig. 2 shows the resulting linear basis images.

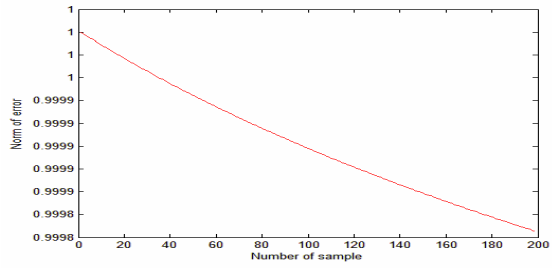


Fig.1 Convergence Behaviour of the Gradient Descent Method Using Facial Database

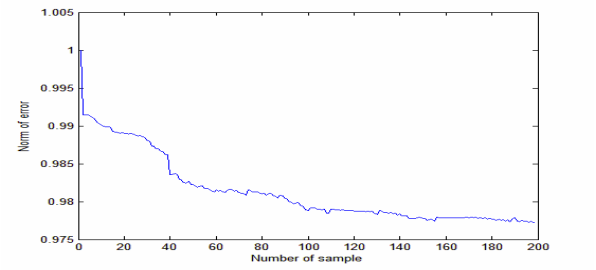


Fig.2 Convergence Behaviour of the New $\Sigma^{-1/2}$ Algorithm Method Using Facial Database

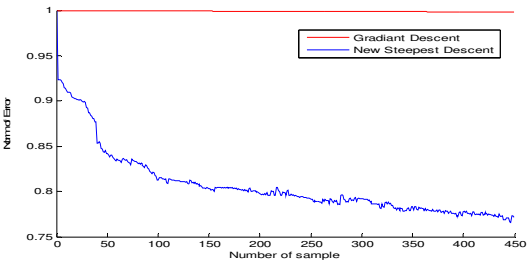
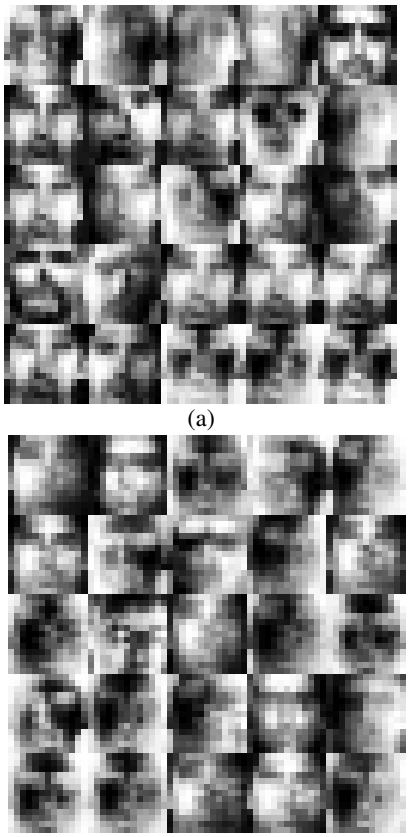


Fig. 3. Convergence Behavior of the New $\Sigma^{-1/2}$ Algorithm Compared to the Gradient Descent Method Using Facial Database



(a)

(b)

Fig.4 (a) images of gradient method (b) images of new adaptive method

Table .1 Classification results for Yale face

Method	Mean of Margin	Mean of SV
Adaptive lda	0.0390	48.4
New adaptive lda	0.0390	48.4

VI. Concluding Remarks

In this paper, we applied a new approach to constructing adaptive face recognition systems. The new algorithm has the

advantage of optimal selection of the step size. The eigenfaces of gradient descent method and new algorithm have been used to training a SVM. Experimental results demonstrate very fast convergence and robustness of the new algorithms with equal recognition rate and justify its advantages for on-line pattern recognition applications. Finding and performing good methods to improvement of SVM and its relative merits are subject for future research.

REFERENCES

[1] L.C. Jain., U. Halici, I. Hayashi and S.B. Lee,"Intelligent Biometric Techniques in Fingerprint and Face Recognition," *CRC Press*, 1999.
[2] Z. SunG. Bebis and R. Miller, "Object Detection Using Feature Subset Selection," *Pattern Recognition*, Elsevier, vol. 37, no. 11, pp.2165-2176, 2004.
[3] C .Lee and J. Hong,m "Optimizing Feature Extraction for Multiclass Cases," *IEEE International Conference on Computational Cybernetics and Simulations*, pp.2545–2548, 1997.
[4] J. Mao and A.K.Jain, "Discriminant Analysis Neural Networks," IEEE *International Conference on Neural Networks, San Francisco*, pp.300–305, 1993.
[5] C. Chatterjee, Z. Kang and,V.P. Roychowdhury, "Algorithms for Accelerated Convergence of Adaptive PCA," *IEEE Transaction of Neural Networks,* vol. 11, no. 2,pp.338–355,2000.
[6] H. Abrishami Moghaddam and Kh. Amiri-Zadeh, "Fast Adaptive Algorithms and Networks for Class-separability Features," *Pattern Recognition*, vol. 36,pp.1695–1702, 2003.
[7] K Fukunaga, "Introduction to Statistical Pattern Recognition," Academic Press, New York, 2nd edition, 1990.
[8] C. Chatterjee and V.P Roychowdhury, "On Self-Organizing Algorithm and Networks for Class-separability Features," *IEEE Transaction of Neural Networks*, vol. 8, no. 3,pp.663–678, 1997.
[9] Benveniste, A. M. Metivier and Priouret, P.:Adaptive Algorithms and Stochastic Approximations, Springer, Berlin, (1990)
[10] L. Ljung, "Analysis of recursive stochastic algorithms," *IEEE Transaction of Automat*. vol. 22, no. 4,pp.551–575, 1997.
[11] S. Georghiades, N. Belhumeur,and D. J. Kriegman,"From few to many: illumination cone models for face recognition under variable lighting and pose," *IEEE Transaction of Pattern Anal, Machine Intell*, pp.643-660, 2001.

Mehdi Ghayoumi working as a lecturer in Islamic Azad University,Shahr-e-Ray Branch. His latest education was in 2004: M.Sc. in, Computer Engineering (Artificial Intelligent and Robotic). His research interest are: Image Processing, Pattern Recognition, Machine Vision, Intelligent Agents,Fuzzy Logic and Genetic Algorithm..