Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Manifold Elastic Net for Sparse Learning

Tianyi Zhou
School of Computer Engineering
Nanyang Technological University
Singapore, 639798
ZHOU0144@ntu.edu.sg

Dacheng Tao
School of Computer Engineering
Nanyang Technological University
Singapore, 639798
dctao@ntu.edu.sg

***Abstract*—In this paper, we present the manifold elastic net (MEN) for sparse variable selection. MEN combines merits of the manifold regularization and the elastic net regularization, so it considers both the nonlinear manifold structure of a dataset and the sparse property of the redundant data representation. Face based gender recognition has received much attention in the psychophysical and video surveillance literatures. Most of existing works apply the appearance based information for data representation. A face image with size 40 by 40 could be seen as a point in a linear space with 1600 dimensions. For gender recognition, we have two classes (male and female) in total, so it is essential to find a small number of variables for representation to generalize duly. MEN can duly find the intrinsic structure of a dataset for separating males from the females. Sufficient experimental results on FERET and UMIST datasets suggest that MEN is more effective in selecting discriminative variables for face based gender recognition compared to principal component analysis, sparse principal component analysis, and discriminative locality alignment.**

***Keywords*—Least Angle Regression (LARS), Elastic Net, Manifold Learning, and Manifold Elastic Net (MEN)**

I. Introduction

Gender recognition, a popular problem, has been widely and deeply investigated in psychophysical science, biometrics and video surveillance. To solve this problem, a dozen of approaches have been proposed in the recent years. We can classify them into the following two groups: the shape-based [15] and the appearance-based [12].

Shape-based approaches usually utilize the active shape model (ASM) [15] or the active appearance model (AAM) [15] to extract geometrical features to represent a face. For example, ASM is a statistical model which iteratively deforms template face to match the new image; AAM is a generalization of ASM, but uses all the information in the image region covered by the target face, rather than just that near modeled edges. All these approaches however could perform poorly because 1) the calculation of model always is very complex and time consuming, 2) they are not robust enough to illumination, pose and expression variations and 3) they depend on shape information and lose the important texture information. Therefore, appearance-based approaches have become popular recently.

In Appearance-based approaches, the features are chosen to be the pixel intensity values directly from the face images rather than complex models. For example, Golomb's two layers' neural network "SexNet" [10] and Brunelli's two radial basis function based competing networks [11] were proposed for gender recognition. Moghaddam [12] investigated support vector machine (SVM) with the radial basis function kernels to classify genders from the face images. Jain and Huang [13] combined the independent component analysis (ICA) and SVM to gain a high accuracy in classification. Appearance-based approaches need a large number of training samples to achieve a sufficiently stable performance, because raw images are in a high dimensional feature space. Therefore, dimensionality reduction is usually used for gender recognition to simplify the data and to obtain an accurate classification performance.

Dimensionality reduction is designed to project the original data from a high dimensional space to a low dimensional subspace. For example, PCA [4] aims to maximize the variance of projected data and minimize the average projection cost. LDA [5] is a supervised method which maximizes the trace of the between-class scatter matrix and minimizes the trace of within-class scatter matrix. However, these conventional methods are limited by their global linearity. Manifold learning offers a more effective framework to reveal the nonlinear structure of data both locally and globally. For example, LLE [6][14] assumes a given measurement can be linearly reconstructed by its *k* nearest neighbors, while ISOMAP [7] represents the local geometry by using the pairwise geometric distance. Algorithms based on continuum spectral theory, e.g., Laplacian eigenmaps (LE) and Hessian eigenmaps (HLLE) build weighted graph model of given measurements. LTSA and some advanced algorithms use tangent coordinates to describe the local geometry. In [8], manifold learning algorithms are unified under the patch alignment framework. Also another manifold learning method called DLA [16] is proposed and its efficacy compared to other methods is established. Therefore, manifold learning especially DLA could be very helpful in face representation for gender recognition.

Sparse learning [1][2][3] focuses on how to generate sparse solution to linear regression problem. Owing to its simper and clearer representation ability, sparse learning is attracting more attention in many significant fields. L-1 regularization is an effective and popular solution in sparse learning. For example, according to Lasso (Least Absolute Shrinkage and Selection Operator) [1], sparse solution of least square regression can be achieved using L-1 norm constraints by iterations of quadratic programming. LARS (Least Angle Regression) [2] is a more efficient modification of Lasso. However, LARS and Lasso are limited by the number of samples. Moreover, they cannot

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

group similar features together. Elastic Net [3] eliminates these disadvantages by adding weighted L-2 norm and L-1 norm penalties. By only making the most important variables nonzero, sparsity has the following three advantages if used in dimensionality reduction:

1) Sparsity can make dimensionality reduction more succinct and simpler. Hence following process of recognition in subspace can become more efficient.

2) Sparsity can control the weights of original features and decrease the variance caused by possible over-fitting with least increment of bias.

3) Sparsity performs good interpretation of the model, thus revealing a more explicit relationship between recognition goals and features. Especially when there are many possible related features and a few measurements.

We believe that the complementary integration of priorities of manifold learning and sparse learning can generate a completely new data representation which performs better than either of these two approaches alone. On one hand, sparse learning only focuses on linear regression without considering the intrinsic geometry of data, while manifold learning considers the nonlinear structure of data. On the other hand, in manifold learning, each feature in subspace is still a linear combination of all the original ones, while sparse learning can select the most significant ones and ignoring the rest. To combine the priorities of these two algorithms together, we present a new algorithm called Manifold Elastic Net (MEN), which can use manifold learning to sparsely reduce the dimensions of data and solve the recognition problem. MEN and several popular dimensionality reduction algorithms were used to recognize gender on face databases FERET and UMIST. Compared to other algorithms, MEN has higher recognition rate even with a smaller training set and the data that has been reduced to lower dimensional subspace. Furthermore, MEN is computationally more efficient comparing to Sparse PCA [9]. Moreover, compared with the bases drawn from other non-sparse algorithms, sparse bases of MEN have less noise and more interpretive effect.

The rest of the paper is organized as follows: Section 2 describes the proposed Manifold Elastic Net. In Section 3, we use MEN to solve gender recognition problem on FERET and UMIST and proved its efficacy against other popular dimensionality reduction algorithms. Finally, in Section 4, we conclude the key achievements of MEN.

II. MANIFOLD ELASTIC NET

In gender recognition problem, assume *n* is the number of face images and *p* is the number of pixels, which are the original features here. $Y = (y_1, y_2, y_3, \dots, y_n)^T \in R^{n\times 1}$ is the class label vector, where $y_i \in \{0,1\}$ is the class label of the *i* th sample. In the original data matrix $X = (x_1, x_2, x_3, \dots, x_p) \in R^{n\times p}$, $x_j \in R^{n\times 1}$ is the *j* th feature or variable. Each row of X is a long vector obtained by reshaping the image matrix.

We now use the idea of manifold learning to reduce the dimensions of the original face images. According to patch alignment framework presented in [8], firstly several local patches are built; each one consists of a measurement and its neighbors. The similarity of local geometry between original patch and the one in subspace is maximized by part optimization. Then the part optimizations of all the patches are unified by whole alignment. Finally the manifold learning can be transformed into a minimization of $f^T Lf$. Here f is the low dimensional representation of original data, and *L* depends on the kind of manifold. We use *L* of DLA introduced by [8] in MEN.

However, we cannot directly use f in gender recognition of given faces outside the training samples X, since f is a nonlinear and implicit projection, i.e., the exact projective function is impossible to obtain. Therefore a linear projection vector W is built to approach the nonlinear projection. To reduce the error introduced by this linear approach, we should minimize $\|f - XW\|_2^2$. Therefore the objective of manifold learning for gender recognition is to minimize:

$$f^T Lf + \|f - XW\|_2^2.$$

Note that W here is assumed to be a vector rather than a matrix. If we want to reduce the data to more than 1 dimension, firstly, we need to calculate the projection matrix column by column; secondly, orthogonal projection is necessary between columns.

By manifold learning and its linear approach, the original face images can be reduced to low dimensional subspace. Then this low dimensional data is input into a classifier to recognize gender. Here we consider the simplest one, the nearest neighbor classifier. Classification error should be minimized:

$$\|Y - XW\|_2^2.$$

Sparse learning is able to choose the most important variables in linear regression and make most of the coefficients zero. In Elastic Net [3], L-1 regularization is used to make the coefficient vector of variables sparse, and L-2 norm penalty is added to obtain grouping effect. The contribution of L-1 norm to sparsity is attributed to the singularity on the origin and similarity to L-0 norm. The strict convexity of the weighted L-1 and L-2 penalty cause a grouping effect [3]. To make the projection vector W sparse and to have a grouping effect here, weighted L-1 and L-2 penalties have been considered. Based on the discussion above, we need to minimize:

$$\|Y - XW\|_2^2 + \beta f^T Lf + \gamma\|f - XW\|_2^2 + \delta_1\|W\|_1 + \delta_2\|W\|_2^2 \quad (1)$$

Therefore, the priorities of manifold learning and sparse learning have been integrated here to solve gender recognition problem. By using the following Manifold Elastic Net algorithm, this optimization problem can be completely solved.

A. Manifold Elastic Net

There are two variables f and W in F. To eliminate one of them, we differentiate F with respect to f, i.e., $\partial F / \partial f = 0$:

$$\beta(L + L^T)f + 2\gamma(f - XW) = 2(\beta L + \gamma I)f - 2\gamma XW = 0.$$

Thus we can obtain

$$\mathrm{f} = \gamma(\beta \mathrm{L} + \gamma \mathrm{I})^{-1} \mathrm{XW}. \tag{2}$$

Substitute (2) in F, and we get

$$\mathrm{W^T X^T A X W} - \mathrm{W^T X^T Y} - \mathrm{Y^T X W} + \delta_1 \|\mathrm{W}\|_1 + \delta_2 \|\mathrm{W}\|_2^2 \tag{3}$$

$$\begin{aligned} \mathrm{A} = \beta(\gamma(\beta \mathrm{L} + \gamma \mathrm{I})^{-1})^{\mathrm{T}} \mathrm{L}(\gamma(\beta \mathrm{L} + \gamma \mathrm{I})^{-1}) + \mathrm{I} \\ + \gamma(\gamma(\beta \mathrm{L} + \gamma \mathrm{I})^{-1} - \mathrm{I})^{\mathrm{T}} (\gamma(\beta \mathrm{L} + \gamma \mathrm{I})^{-1} - \mathrm{I}) \end{aligned} \tag{4}$$

Current correlation C is defined as the negative gradient of F with respect to W (ignoring the L-1 penalty here):

$$\mathrm{C} = -\frac{\partial \mathrm{F}}{\partial \mathrm{W}} = 2\mathrm{X^T Y} - (\mathrm{X^T}(\mathrm{A} + \mathrm{A^T})\mathrm{X} + 2\delta_2 \mathrm{I})\mathrm{W}$$

$$\mathrm{H} = \mathrm{A} + \mathrm{A^T}.$$

Suppose eigenvalue decomposition of H is:

$$\mathrm{H} = \mathrm{UDU^T}.$$

The larger the current correlation $\mathrm{C_i}$, the more important the respective variable $\mathrm{x_i}$ is, and the larger the corresponding coefficient $\mathrm{W^*}_\mathrm{i}$ should be. In sparse learning, the important variables are added sequentially according to their correlation in loops, and then the direction and distance of coefficient vector of all the important variables are decided.

Let *A* be the active set of important variables whose coefficients are nonzero, while the other variables form inactive set *I*. Thus the sparsity can be controlled by the size of *A*. The correlations of variables in *A* should be kept largest at any time. At the beginning of algorithm all the coefficients are zero and all the variables are in inactive set *I*.

Using simple algebra, C can be written in a clearer form:

$$\mathrm{C} = \mathrm{X^{*T}}(\mathrm{Y^*} - \mathrm{X^* W^*}), \tag{5}$$

where

$$\mathrm{X^*} = (1 + \delta_2)^{-1/2} \begin{pmatrix} (\mathrm{UD}^{\frac{1}{2}})^{\mathrm{T}} \mathrm{X} \\ \sqrt{2\delta_2} \mathrm{I} \end{pmatrix} \tag{6}$$

$$\mathrm{Y^*} = \begin{pmatrix} 2\left(\mathrm{UD}^{\frac{1}{2}}\right)^{-1} \mathrm{Y} \\ 0 \end{pmatrix} \tag{7}$$

$$\delta = \delta_1 / \sqrt{1+\delta_2}$$

$$\mathrm{W^*} = \sqrt{1+\delta_2}\mathrm{W}.$$

To make W *K*-sparse, we need *K* loops. At the first step of each loop, the variable in the inactive set *I* with the largest correlation is added into the active set *A*:

$$\hat{\mathrm{C}} = \max_{\mathrm{j}}\{|\hat{\mathrm{c}}_\mathrm{j}|\}, \mathrm{A} = \{\mathrm{j}: |\hat{\mathrm{c}}_\mathrm{j}| = \hat{\mathrm{C}}\}. \tag{8}$$

Then the direction of coefficient vector is decided. To make the optimization more global and less greedy, we want the correlations of active variables to decrease equally in preferred direction. In the *k* th loop, if we assume the direction vector is ω, then:

$$\begin{aligned} \mathrm{C_k} &= \mathrm{X^*}_\mathrm{A}^{\mathrm{T}}(\mathrm{Y^*} - \mathrm{X^* W_k}) \\ &= \mathrm{X^*}_\mathrm{A}^{\mathrm{T}}(\mathrm{Y^*} - \mathrm{X^*}(\mathrm{W_{k-1}} + \rho\omega)) \\ &= \mathrm{C_{k-1}} + \rho \mathrm{X^*}_\mathrm{A}^{\mathrm{T}} \mathrm{X^*}_\mathrm{A} \omega_\mathrm{A}. \end{aligned}$$

Ignoring the constant ρ, the change of correlation at this step is $\mathrm{X^*}_\mathrm{A}^{\mathrm{T}} \mathrm{X^*}_\mathrm{A} \omega_\mathrm{A}$. Since the sign *s* of ω is supposed to be the same as $\mathrm{C_{k-1}}$, here we could assume that ω is nonnegative and return its sign after calculation. It is not hard to discover that $\mathrm{X^*}_\mathrm{A} \omega_\mathrm{A}$ is just an extended simplex with vertices defined by active variables. If $\mathrm{x^*}_\mathrm{i}^{\mathrm{T}} \mathrm{X^*}_\mathrm{A} \omega_\mathrm{A}$ is one element of $\mathrm{X^*}_\mathrm{A}^{\mathrm{T}} \mathrm{X^*}_\mathrm{A} \omega_\mathrm{A}$, $\mathrm{x^*}_\mathrm{i}^{\mathrm{T}} \mathrm{X^*}_\mathrm{A} \omega_\mathrm{A}$ could be seen as the projection on $\mathrm{x_i}$ coordinate of a vector in this simplex. Hence, only a unique vector could make each $\mathrm{x^*}_\mathrm{i}^{\mathrm{T}} \mathrm{X^*}_\mathrm{A} \omega_\mathrm{A}$ equal. Here it is the normal vector through the origin in simplex space, which is

$$\omega_\mathrm{A} = \mathrm{s}.* (\mathrm{X^*}_\mathrm{A}^{\mathrm{T}} \mathrm{X^*}_\mathrm{A})^{-1} 1_\mathrm{A} = \mathrm{s}.* (\mathrm{G_A})^{-1} 1_\mathrm{A}$$

$$\mathrm{G_A} = \mathrm{X^*}_\mathrm{A}^{\mathrm{T}} \mathrm{X^*}_\mathrm{A} \text{ is the Gram matrix of } \mathrm{X^*}_\mathrm{A}.$$

This result could also be obtained by minimizing squared distance between the point in the simplex and the origin, with a summation constraint of ω_A, as stated in LARS [2]. To make the change of correlation $\mathrm{X^*}_\mathrm{A}^{\mathrm{T}} \mathrm{X^*}_\mathrm{A} \omega_\mathrm{A}$ as a unit vector $\mathrm{u_A}$

$$\omega_\mathrm{A} = \mathrm{s}.* \mathrm{A_A}(\mathrm{G_A})^{-1} 1_\mathrm{A} \tag{9}$$

$$\mathrm{A_A} = (1_\mathrm{A}^{\mathrm{T}} (\mathrm{G_A})^{-1} 1_\mathrm{A})^{-1/2}$$

$$\mathrm{u_A} = \mathrm{X^*}_\mathrm{A} \omega_\mathrm{A}.$$

Let

$$\mathrm{a} = \mathrm{X^*}_\mathrm{A}^{\mathrm{T}} \mathrm{u_A}.$$

After deciding the direction of change of $\mathrm{W^*}$, the distance or magnitude of changes ρ_1 needs to be decided. We want the distance to be as large as possible since larger distance helps in faster optimization. However, we should guarantee that the correlation of active variables is always kept larger than correlations of any inactive variables. Therefore the distance can be increased until the correlation of one inactive variable becomes equal to the correlation of active variables:

$$\rho_1 = \min^{+}_{\mathrm{j} \in \mathrm{A^c}} \left\{ \frac{\hat{\mathrm{C}} - \hat{\mathrm{c}}_\mathrm{j}}{\mathrm{A_A} - \mathrm{a_j}}, \frac{\hat{\mathrm{C}} + \hat{\mathrm{c}}_\mathrm{j}}{\mathrm{A_A} + \mathrm{a_j}} \right\}. \tag{10}$$

To generate strict solution of the proposed problem, Lasso modification should be added according to the demonstration in LARS [2]. That is, argument of the distance ρ_1 should be stopped when coefficient of one of the active variables becomes zero first:

$$\mathrm{W}_{\mathrm{A}}^{*} = \mathrm{W}_{\mathrm{A}} + \rho \mathrm{s}_{\mathrm{A}} \mathrm{w}_{\mathrm{A}}$$

$$\rho_2 = \min{}^{+}\{- \mathrm{W}_{\mathrm{A}} / \mathrm{s}_{\mathrm{A}} \mathrm{w}_{\mathrm{A}}\}. \tag{11}$$

Therefore, the distance of change of W^{*} is ρ:

$$\rho = \min{}^{+}\{\rho_1, \rho_2\}. \tag{12}$$

In each loop, one new active variable is added using (8), the direction and distance of the coefficient vector are calculated according to (9) and (12). After *K* loops, there are *K* elements of W^{*} that are nonzero. According to the Elastic Net [3], the final W should be corrected by eliminating the double shrinkage:

$$\mathrm{W} = \sqrt{1 + \delta_2}\,\mathrm{W}^{*}. \tag{13}$$

To obtain several W to form a projection matrix, we need to do all the steps above several times with orthogonal projection between them:

$$\mathrm{X}_{\mathrm{k+1}} = \mathrm{X}_{\mathrm{k}} - \mathrm{X}_{\mathrm{k}} \mathrm{W}_{\mathrm{k}} 1^{\mathrm{T}}. \tag{14}$$

B. *Algorithm*

Based on the Manifold Elastic Net proposed above, we now formally present the Manifold Elastic Net algorithm:

TABLE I. MANIFOLD ELASTIC NET ALGORITHM

Input	Training data matrix $\mathrm{X} \in \mathrm{R}^{\mathrm{n} \times \mathrm{p}}, \mathrm{X}_1 = \mathrm{X}$; Gender label vector $\mathrm{Y} \in \mathrm{R}^{\mathrm{n} \times 1}$; $\mathrm{W} = \{\mathrm{W}_1, \mathrm{W}_2, \mathrm{W}_3, \ldots, \mathrm{W}_{\mathrm{r}}\} = 0^{\mathrm{p} \times \mathrm{r}}$, where r is the dimensions of subspace; k=0, m=0, K is the number of loops.
Output	Projection matrix $\mathrm{W} = \{\mathrm{W}_1, \mathrm{W}_2, \mathrm{W}_3, \ldots, \mathrm{W}_{\mathrm{r}}\} \in \mathrm{R}^{\mathrm{p} \times \mathrm{r}}$.

Step 1	Calculate $\mathrm{X}^{*}, \mathrm{Y}^{*}$ from $\mathrm{X}_{\mathrm{k+1}}$ and Y using (6) and (7);
Step 2	Calculate sparse basis W_{k}, m=m+1;
Step 3	Add the variable with largest correlation to active set using (8);
Step 4	Direction calculation using (9);
Step 5	Distance calculation using (12);
Step 6	Update W_{k} using (13), if m<K, go back to step 2, loop until m=K;
Step 7	Calculate $\mathrm{X}_{\mathrm{k+1}}$ by orthogonal projection using (14);
Step 8	k=k+1, go back to step 1, loop until k=r;

III. EXPERIMENTS

In this section, the performance of MEN is evaluated in comparison with three representative dimensionality reduction algorithms, i.e., PCA, DLA and SPCA, on two face image databases, i.e., FERET and UMIST. PCA is the widely used unsupervised linear dimensionality reduction algorithm. DLA [8] was recently proposed as a supervised manifold learning method. SPCA is sparse PCA which can generate sparse solution approaching PCA. All the images of faces in the two datasets are normalized to 40 × 40 pixels arrays with 256 grey levels per pixel. Each image is reshaped to one long vector. Thus, the original data has 1600 features.

There are 287 female samples and 413 male samples in FERET, while in UMIST the corresponding numbers are 95 and 480 respectively. Fig. 1 and Fig. 2 show example images from FERET and UMIST respectively. The datasets are randomly separated into two groups: training set and testing set. Training set is used to derive the projection matrix W. In training set the numbers of female and male images are identical. Testing set is used to test the recognition. During testing phase, the Nearest Neighbor classifier is used. In DLA, the first step is PCA projection. Since in our experiments the number of samples n is smaller than the number of features p, we reduce the data to n-1 dimensions using PCA at the beginning of DLA.

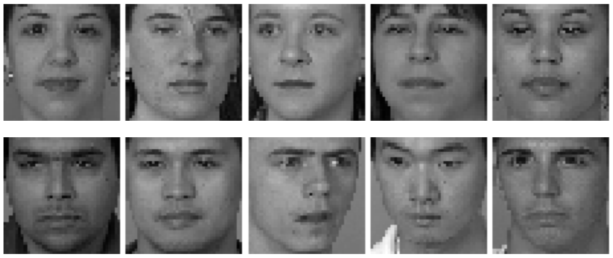

Figure 1. Female and Male sample images from FERET

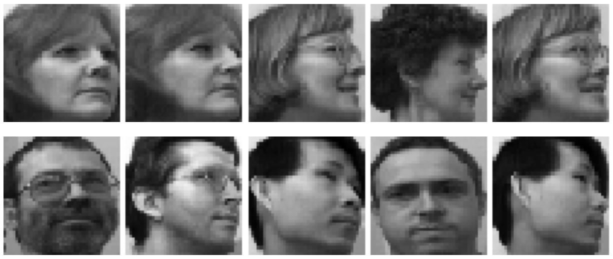

Figure 2. Female and Male sample images from UMIST

Two experiments test the algorithms on each dataset. The first one test all the algorithms in low dimensionality situation, the original data is reduced to 5 dimensional subspace, the number of given training samples were changed from 10 to 200 for FERET and from 5 to 80 for UMIST. The second experiment focuses on testing the algorithms in fixed training sets, 160 samples for FERET and 40 samples for UMIST were used, while the dimensions of subspace were changed from 1 to 100 for FERET and from 1 to 30 for UMIST. Both experiments were repeated 5 times, and then the average recognition rates were calculated.

Fig. 3 and Fig. 4 show the plots of recognition rates versus training set size on the two databases in the first experiment. It can be seen that MEN outperforms the other algorithms and it is more stable when the training set changes. Moreover, MEN has good recognition property even when training set is small. This is because the sparsity of MEN filters out unrelated features; which may bring unnecessary noise to classification; especially when number of classes is much smaller than the

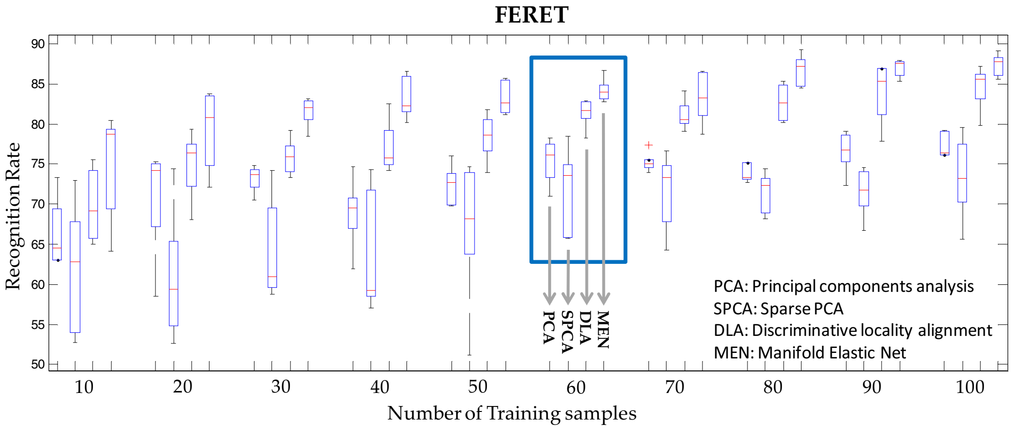

Figure 3. Recognition rate vs. number of training samples on FERET

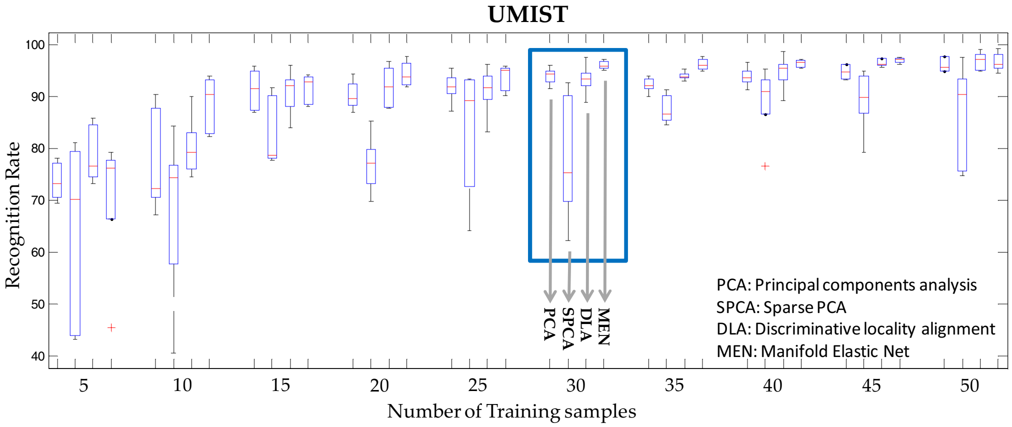

Figure 4. Recognition rate vs. number of training samples on UMIST

number of original features. SPCA is sparse but achieves the lowest recognition rate here because it is unsupervised training method which does not consider any label information.

Fig. 5 shows the plots of recognition rates versus dimension of subspace on the two databases obtained in the second experiment. It is shown that MEN gets the best classification effect. Furthermore, compared to other algorithms, the dimensionality of MEN has fewer limits. In DLA, as dimensionality increases after the optional rate, PCA plays a more important role while its own priorities have to be weakened. This is the reason why DLA's recognition rates decrease faster in high dimensionality situations. But MEN is stable even when dimensions change from small to large. Another advantage of MEN can be discovered from its first basis. In fact, using the first sparse basis only, we can achieve high recognition rate. In real applications preferring parsimony over accuracy, one basis is always sufficient, thus MEN is able to reduce the computational cost efficiently.

Fig. 6 and Fig. 7 show several bases of the 4 algorithms on FERET and UMIST respectively by reshaping them to 40 × 40 matrix. The bases of PCA are called Eigenfaces, while the bases of LDA are called Fisherfaces. Similar methods can be applied to SPCA, DLA and MEN here. Obviously the bases of MEN are sparser and have less noise than PCA and DLA, and more grouping than SPCA, because of MEN's supervised learning property, sparsity and grouping effect. Sparse bases lead to computational efficiency, simpler bases lead to clearer interpretation. In our gender recognition experiments, the bases generated by MEN on FERET indicate eyes and nose are significant for gender recognition of frontal faces, while the bases on UMIST suggest mouth and mustache are important features to distinct female and male profiles.

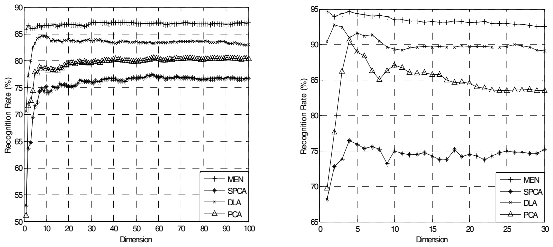

Figure 5. Recognition rate vs. Dimension on FERET and UMIST

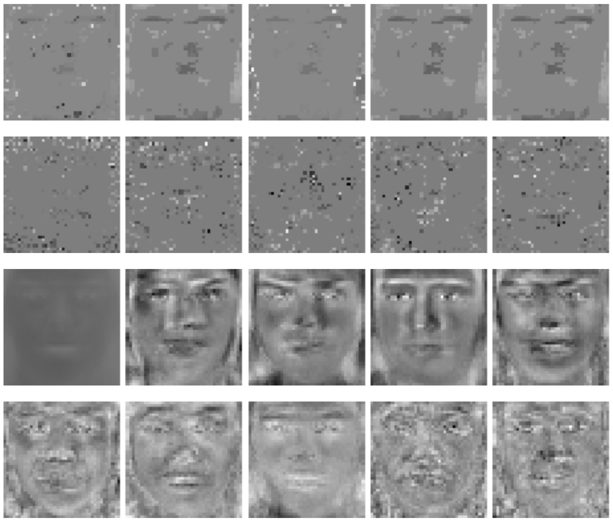

Figure 6. 5 bases of MEN, SPCA, PCA, DLA on FERET

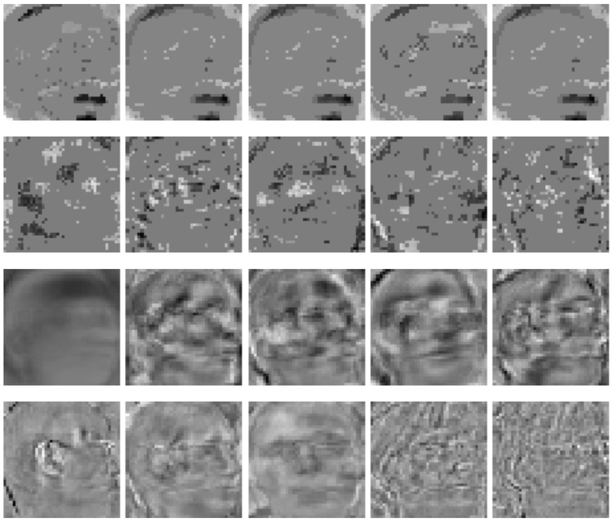

Figure 7. 5 bases of MEN, SPCA, PCA, DLA on UMIST

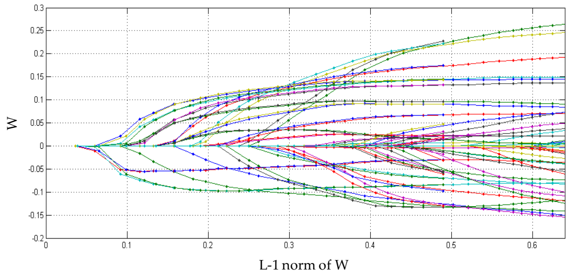

Figure 8. Coefficients' Paths of the first basis in MEN

Fig. 8 shows paths of 160 coefficients in MEN when calculating the first basis. This illustrates that the sequential augmentation of coefficients is one important reason for sparsity in MEN.

IV. Conclusions

In this paper, we have developed a new sparse nonlinear dimensionality reduction algorithm for gender recognition, called Manifold Elastic Net (MEN). It utilizes ideas of the manifold learning and supervised information for classification, and obtains sparse solution through L-1 penalty and grouping effect from the L-2 penalty. Comprehensive experiments on FERET and UMIST using MEN, PCA, SPCA and DLA, show that MEN has better classification performance than other algorithms in most situations, especially when the training set is small and dimension of subspace is low. Analysis of sparse bases of MEN is helpful in obtaining clearer interpretation of relationship between face features and gender recognition. Therefore MEN can achieve more accurate and more efficient results when it is used to solve problem of dimensionality reduction for classification and recognition.

Acknowledgment

Both authors would like to thank Mr. Bian for helpful discussions. This project was supported by the Nanyang Technological University Start-Up Grant (under project number M58020010).

References

[1] R. Tibshirani, “Regression shrinkage and selection via the lasso,” J. R. Statist. Soc. B, 58, pp. 267–288, 1996.

[2] B. Efron, T. Hastie and R. Tibshirani, “Least angle regression,” Ann. Statist. , 32, pp.68–73, 2004.

[3] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” J. R. Statist. Soc. B, 67, pp. 301–320, 2005.

[4] H. Hotelling, “Analysis of A Complex of Statistical Variables into Principal Components,” J. Edu. Phychology, vol. 24, pp. 417-441, 1933.

[5] R. A. Fisher, “The Use of Multiple Measurements in Taxonomic Problems,” Ann. Eugenics, vol. 7, pp. 179-188, 1936.

[6] S. T. Roweis and L. K. Saul, “Nonlinear Dimensionality Reduction by Locally Linear Embedding,” Science, vol. 290, pp. 2323-2326, 2000.

[7] J. Tenenbaum, V. Silva and J. Langford, “A Global Geometric Framework for Nonlinear Dimensionality Reduction,” Science, vol. 290, pp. 2319-2323, 2000.

[8] T. Zhang, D. Tao, X. Li, and J. Yang, “Patch Alignment for Dimensionality Reduction,” IEEE Trans. Know. & Data Eng., to appear.

[9] H. Zou, T. Hastie and R. Tibshirani, “Sparse Principal Component Analysis,” Journal of Comp. & Graphical Statistits, vol. 15, no. 2, pp. 265-286, 2006.

[10] B. Golomb, D. Lawrence, and T. Sejnowski “SEXNET: A neural network identifies sex from human faces,” NIPS, vol. 3, 1991.

[11] R. Brunelli and T. Poggio, “Hyberbf networks for gender recognition,” in Proc. DARPA Image Understanding Workshop, pp. 311-314, 1992.

[12] B. Moghaddam and M. H. Yang, “Learning gender with support faces,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 707-711, 2002.

[13] A. Jain and J. Huang, “Integrating independent components and linear discriminant analysis for gender recognition,” Proc. of the 6th IEEE Int’l Conf. Automatic Face and Gesture Recognition, pp. 159-163, 2004.

[14] X. Li, S. Lin, S. Yan, and D. Xu, “Discriminant Locally Linear Embedding With High-Order Tensor Data,” IEEE Trans. on Systems, Man, and Cybernetics, Part B, vol. 38, no. 2, pp. 342-352, 2008.

[15] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp. 681-685, 2001.

[16] T. Zhang, D. Tao, and J. Yang, “Discriminative Locality Alignment,” ECCV, vol. 1, pp. 725-738, 2008.