METAFACE LEARNING FOR SPARSE REPRESENTATION BASED FACE RECOGNITION

Meng Yang[a], *Lei Zhang*[a1], *Jian Yang*[b] *and David Zhang*[a]

[a]Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[b]Dept. of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, PR China

ABSTRACT

Face recognition (FR) is an active yet challenging topic in computer vision applications. As a powerful tool to represent high dimensional data, recently sparse representation based classification (SRC) has been successfully used for FR. This paper discusses the metaface learning (MFL) of face images under the framework of SRC. Although directly using the training samples as dictionary bases can achieve good FR performance, a well learned dictionary matrix can lead to higher FR rate with less dictionary atoms. An SRC oriented unsupervised MFL algorithm is proposed in this paper and the experimental results on benchmark face databases demonstrated the improvements brought by the proposed MFL algorithm over original SRC.

Index Terms— Face recognition, sparse representation, metaface learning

1. INTRODUCTION

Automatic face recognition (FR) has been, and remains being, one of the most visible and challenging research topics in computer vision, machine learning and biometrics. Although the facial images have a high dimensionality, they usually lie on a lower dimensional subspaces or sub-manifolds. Therefore, subspace learning and manifold learning methods have been dominantly and successfully used in appearance based FR [1-8], which includes Eigenface, Fisherface [1-3], locality preserving projection (LPP) [6], local discriminant embedding (LDE) [7], unsupervised discriminant projection (UDP) [8], etc.

The success of manifold learning implies that the high dimensional face images can be sparsely represented or coded by the representative samples on the manifold. Very recently, an interesting work was reported by Wright et al. [9] by using the sparse representation (SR) technique for robust FR. In Wright et al.'s pioneer work, the training face images are used as the dictionary of representative samples, and an input testing image is coded as a sparse linear combination of these sample images via l1-norm minimization. The results in [9] clearly validated the effectiveness of SR techniques in FR, which can not only lead to high classification accuracy, but also well handle the problem of face occlusion. In addition, Yang et al. [18] proposed Gabor SR technique with much better performance.

However, [9] use the original training samples as the dictionary. So an important issue that whether an optimal dictionary can be learned from training data needs to be further discussed. In [10], a set of "metagenes" are trained from the original gene expression data by using nonnegative matrix factorization and these "metagenes" provide a more robust clustering of the samples. Recently, in the field of image restoration a lot of efforts have been made on learning an over-complete dictionary of atoms from natural images and then using the learned dictionary for image analysis [11-13]. In face recognition, the original image samples have much redundancy as well as noise and trivial information that can be negative to the recognition. In addition, if the training samples are huge, the computation of SR will be time-consuming. So a more compact and/or robust set of bases, which are called metafaces in this paper, will be learned from the original images and then used as the dictionary to represent the input query image. The learned metafaces will be more representative for SR and more efficient in l_1-norm minimization.

The rest of the paper is organized as follows. Section 2 briefly reviews SRC. Then proposed metaface learning algorithm is presented in Section 3. Section 4 conducts experiments to validate the proposed method and Section 5 concludes the paper.

2. SPARSE REPRESENTATION BASED CLASSIFICATIION FOR FACE RECOGNITION

Denote by $A_i = [\boldsymbol{s}_{i,1}, \boldsymbol{s}_{i,2}, ..., \boldsymbol{s}_{i,n_i}] \in \mathbb{R}^{m \times n_i}$ the set of training samples of the i^{th} object class, where $\boldsymbol{s}_{i,j}$, j=1,2,…,n_i, is an m-dimensional vector stretched by the j^{th} sample of the i^{th} class. For a test sample $\boldsymbol{y} \in \mathbb{R}^m$ from this class, intuitively, $\boldsymbol{y}$ could be well approximated by the linear combination of the samples within A_i, i.e. $\boldsymbol{y} = \sum_{j=1}^{n_i} \alpha_{i,j} \boldsymbol{s}_{i,j} = A_i \boldsymbol{\alpha}_i$, where

[1] Corresponding author: cslzhang@comp.polyu.edu.hk.

$\boldsymbol{\alpha}_i = [\alpha_{i,1}, \alpha_{i,2}, ..., \alpha_{i,n_i}]^T \in \mathbb{R}^{n_i}$ are the coefficients. Suppose we have K object classes, and let A=[A_1,A_2,…, A_K] be the concatenation of the n training samples from all the K classes, where n=n_1+n_2+ n_K. If we use A to represent the input test image $\boldsymbol{y}$, there is $\boldsymbol{y}$=$A\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$=[$\boldsymbol{\alpha}_1$;…,$\boldsymbol{\alpha}_i$;…; $\boldsymbol{\alpha}_K$]. Since $\boldsymbol{y}$ is from the i^{th} class and $\boldsymbol{y}$=$A_i\boldsymbol{\alpha}_i$ holds well, a naturally good solution to $\boldsymbol{\alpha}$ will be that all the coefficients in $\boldsymbol{\alpha}_k$, k=1,2,…,K and k≠i, are nearly zero and only the coefficients in $\boldsymbol{\alpha}_i$ have significant values. In other words, the sparse non-zero entries in $\boldsymbol{\alpha}$ can well encode the identity of the test sample $\boldsymbol{y}$. The SRC algorithm [9] is summarized as follows.

1. Normalize the columns of A to have unit l_2-norm.
2. Solve the l_1-minimization problem:

$$\hat{\boldsymbol{\alpha}}_1 = \arg\min_{\boldsymbol{\alpha}} \left\{ \|A\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1 \right\} \quad (1)$$

   where λ is a positive scalar number that balances the reconstructed error and coefficients' sparsity.
3. Compute the residuals

$$r_i(\boldsymbol{y}) = \|\boldsymbol{y} - A\delta_i(\hat{\boldsymbol{\alpha}}_1)\|_2, \text{ for } i = 1, \cdots, k.$$

   where $\delta_i(\boldsymbol{\alpha}): \mathbb{R}^n \to \mathbb{R}^n$ is the characteristic function which selects the coefficients associated with the i-th class.
4. Output that identity($\boldsymbol{y}$)=argmin r_i($\boldsymbol{y}$).

3. METAFACE LEARNING

Inspired by the success of metagenes in gene expression data analysis [10] and dictionary learning (DL) in image processing [11-13], we propose to learn a set of metafaces, denoted by D_i, from the original training dataset A_i, and then use D_i to replace A_i in the SRC based FR.

For the convenience of expression, we denote by X=[$\boldsymbol{x}_1$, $\boldsymbol{x}_2$,…, $\boldsymbol{x}_n$]$\in \mathbb{R}^{m\times n}$ the training samples of the i^{th} object class, with each column of X being a sample vector. We want to learn a dictionary of metafaces $\varGamma = [\boldsymbol{d}_1, \boldsymbol{d}_2, ..., \boldsymbol{d}_p] \in \mathbb{R}^{m\times p}$ from X, where p≤n. It is required that each metaface $\boldsymbol{d}_j$, j=1,2,…,p, is a unit column vector, i.e. $\boldsymbol{d}_j^T\boldsymbol{d}_j = 1$. Our objective function in determining $\varGamma$ is as follows:

$$J_{\varGamma,\varLambda} = \arg\min_{\varGamma,\varLambda} \left\{ \|X - \varGamma\varLambda\|_F^2 + \lambda\|\varLambda\|_1 \right\} \text{ s.t. } \boldsymbol{d}_j^T\boldsymbol{d}_j = 1, \forall j \quad (2)$$

where $\varLambda$ is the representation matrix of X over the metafaces $\varGamma$, and parameter λ is a positive scalar number that balances the F-norm term and the l_1-norm term.

Eq. (2) is a joint optimization problem of the metafaces $\varGamma$ and the representation coefficient matrix $\varLambda$. Like in many multi-variable optimization problems, we solve Eq. (2) by optimizing $\varGamma$ and $\varLambda$ alternatively. The optimization procedures are described in the following Algorithm 1.

Algorithm 1. Meta-Face Learning

Step 1. Initialize $\varGamma$. We initialize each column of $\varGamma$ (i.e. each metaface) as a random vector with l_2-norm 1.

Step 2. Fix $\varGamma$ and solve $\varLambda$. By fixing $\varGamma$, the objective function in Eq. (2) will be reduced to

$$J_{\varLambda} = \arg\min_{\varLambda} \left\{ \|X - \varGamma\varLambda\|_F^2 + \lambda\|\varLambda\|_1 \right\} \quad (3)$$

The minimization of Eq. (3) can be achieved by some standard convex optimization techniques. In this paper, we use the algorithm in [14].

Step 3. Fix $\varLambda$ and update $\varGamma$. Now the objective function is reduced to

$$J_{\varGamma} = \arg\min_{\varGamma} \left\{ \|X - \varGamma\varLambda\|_F^2 \right\} \text{ s.t. } \boldsymbol{d}_j^T\boldsymbol{d}_j = 1, \forall j \quad (4)$$

We can write matrix $\varLambda$ as $\varLambda$=[$\boldsymbol{\beta}_1$;$\boldsymbol{\beta}_2$,…,$\boldsymbol{\beta}_p$], where $\boldsymbol{\beta}_j$, j=1,2,…,p, is the row vector of $\varLambda$. We update the metaface vectors one by one. When updating $\boldsymbol{d}_j$, all the other columns of $\varGamma$, i.e. $\boldsymbol{d}_l$, l≠j, are fixed. Then $J_\varGamma$ in Eq. (4) is converted into

$$J_{\boldsymbol{d}_j} = \arg\min_{\boldsymbol{d}_j} \left\| X - \sum_{l\neq j} \boldsymbol{d}_l\boldsymbol{\beta}_l - \boldsymbol{d}_j\boldsymbol{\beta}_j \right\|_F^2 \text{ s.t. } \boldsymbol{d}_j^T\boldsymbol{d}_j = 1 \quad (5)$$

Let $Y = X - \sum_{l\neq j} \boldsymbol{d}_l\boldsymbol{\beta}_l$, Eq. (5) can be written as

$$J_{\boldsymbol{d}_j} = \arg\min_{\boldsymbol{d}_j} \|Y - \boldsymbol{d}_j\boldsymbol{\beta}_j\|_F^2 \text{ s.t. } \boldsymbol{d}_j^T\boldsymbol{d}_j = 1 \quad (6)$$

Using Langrage multiplier, $J_{\boldsymbol{d}_j}$ is equivalent to

$$J_{\boldsymbol{d}_j,\gamma} = \arg\min_{\boldsymbol{d}_j} tr\left(-Y\boldsymbol{\beta}_j^T\boldsymbol{d}_j^T - \boldsymbol{d}_j \cdot \boldsymbol{\beta}_j Y^T + \boldsymbol{d}_j \cdot (\boldsymbol{\beta}_j\boldsymbol{\beta}_j^T - \gamma)\boldsymbol{d}_j^T + \gamma\right)$$

where γ is a scalar variable. Differentiating $J_{\boldsymbol{d}_j,\gamma}$ with respect to $\boldsymbol{d}_j$, and let it be 0, we have

$$\boldsymbol{d}_j = Y\boldsymbol{\beta}_j^T\left(\boldsymbol{\beta}_j\boldsymbol{\beta}_j^T - \gamma\right)^{-1} \quad (7)$$

So the solution of Eq. (7) under constrain $\boldsymbol{d}_j^T\boldsymbol{d}_j = 1$ is

$$\boldsymbol{d}_j = Y\boldsymbol{\beta}_j^T \Big/ \left\|Y\boldsymbol{\beta}_j^T\right\|_2 \quad (8)$$

where $\|\bullet\|_2$ is the l_2-norm.

Using the above procedures, we can update all the metafaces $\boldsymbol{d}_j$, and hence the whole set $\varGamma$ is updated.

Step 4. Go back to step 2 until the values of $J_{\varGamma,\varLambda}$ in adjacent iterations are close enough, or the maximum number of iterations is reached. Finally, output $\varGamma$.

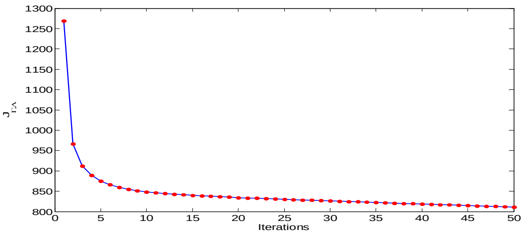

Figure 1. Example of the convergence of Algorithm 1.

It is straightforward that the above MFL algorithm converges because in each iteration $J_{\varGamma,\varLambda}$ will decrease, as shown in Fig. 1. By using Algorithm 1, we can learn a set of metafaces D_i for each class of face images A_i. Then all the K classes of metafaces can be concatenated into one dictionary $D=[D_1,\ldots,D_i,\ldots,D_K]$, which plays the role of A in Eq. (1).

4. EXPERIMENTAL RESULTS

We evaluated the performance of the proposed MFL on representative facial image databases: Extended Yale B [15-16], ORL, and AR [17]. In our experiment, Eigenfaces was first applied to reduce the dimensionality of face images, and then SRC with the proposed MFL is compared with the original SRC [9]. In addition, as in [9], the benchmark *nearest neighbor* (NN) classifier using cosine distance was also used in the experiments as a reference. The code of the proposed method can be downloaded at http://www4.comp.polyu.edu.hk/~cslzhang/code.htm.

1) Extended Yale B Database: The Extended Yale B database consists of 2,414 frontal-face images of 38 individuals, captured under various laboratory-controlled lighting conditions [15-16]. For each subject, we randomly selected half of the images for training (i.e. 32 images per subject), and used the other half for testing. In the SRC based FR, there is a parameter λ (refer to Eq. (1)) and we adjusted it to get the best performance. In original SRC without MFL, λ=5; and in SRC with MFL, λ=4 with 18 metafaces of each class. Fig 2 shows the recognition rates versus the dimension of features. We can see that SRC with the proposed MFL consistently outperforms the original SRC method, while the classical NN method performs the worst. This validates that the proposed MFL algorithm can not only reduce the number of representation samples, but also make these samples more representative so that the classification accuracy can be improved. Table 1 lists the maximal recognition rate of each method and the corresponding dimensionality. We see that SRC with the proposed MFL achieves the highest rate.

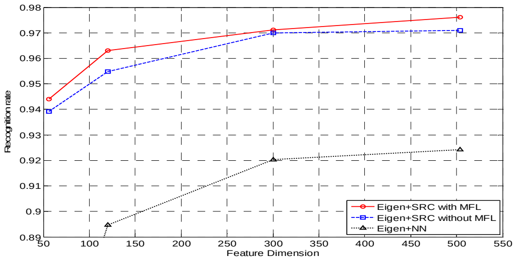

Figure 2. Recognition rates by different methods versus feature dimension on the Extended Yale B with features of Eigenfaces.

Table 1. The top recognition rates (%) of different methods on the Extended Yale B database and the associated dimension of features.

Method	SRC	MFL	NN
Rate	97.09	**97.62**	92.43
Dimension	504	**504**	504

In MFL, different numbers of metafaces (i.e. parameter p), will lead to different FR results. Thus, it is necessary to test the classification performance by varying p. Fig. 3 plots the curve of recognition rate versus p when the dimension of Eigenface features is 504. From Fig. 3 we can see that SRC with MFL has better performance when p is between 17 and 20. Particularly, on this database SRC with MFL can have satisfying recognition rate even when p is 8, a quarter of the number of samples (32) per class in SRC.

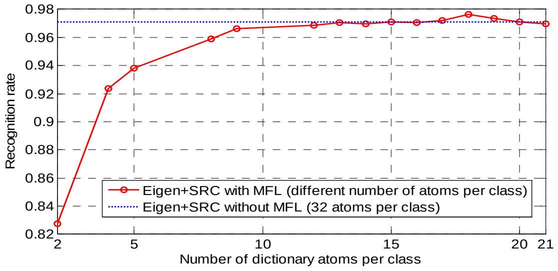

Figure 3. Recognition rates of SRC with MFL versus the number of metafaces per class on the Extended Yale B database.

2) ORL database: The ORL database (http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html) contains images from 40 individuals, each providing 10 different images. In the experiment, we used the first 6 images of each class for training, and the remaining 4 images for testing. The parameters λ and p were selected to gain the best performance for each method. In original SRC, λ=0.5; and in SRC with MFL, λ=4 with p=4. The curves of recognition rate versus the dimension of features are illustrated in Fig. 4, and the maximal recognition rate of each method with associated feature dimension is listed in Table 2.

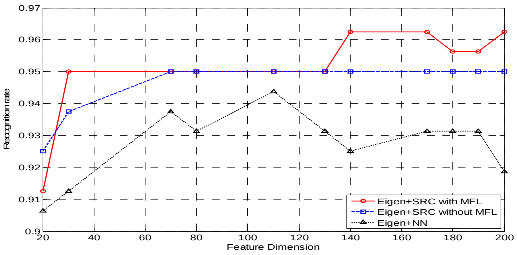

Figure 4. Recognition rates by different methods versus feature dimension on the ORL database with features of Eigenfaces.

Table 2. The top recognition rates (%) of different methods on the ORL database and the associated dimension of features.

Method	SRC	MFL	NN
Rate	95	**96.25**	94.37
Dimension	40	**140**	110

From Fig. 4 and Table 2 we can see that SRC with MFL outperforms the original SRC, and SRC with MFL by using the Eigenface features achieves the best recognition rate of 96.25%, while the highest rate of original SRC with

Eigenface features is 95%. Meanwhile, SRC with MFL use less dictionary atoms than SRC without MFL.

3) AR database: The AR database consists of over 4,000 frontal images from 126 individuals [17]. For each individual, 26 pictures were taken in two separate sessions. In the experiment, we chose a subset of the dataset consisting of 50 male subjects and 50 female subjects. For each subject, the seven images with illumination change and expressions from Session 1 were used for training, and the other seven images with illumination change and expression from Session 2 were used for testing. Parameters λ and p were selected to gain the best performance for each of the competing methods. In original SRC, λ=0.005; and in SRC with MFL, λ=0.005 with p=7. It should be noted that since the training samples per class is very limited (7 in this dataset), in this experiment the MFL actually does not reduce the number of representation samples. However, from Fig. 5 and Table 3 we can see that the learned metafaces are more representative so that the classification can be more robust and accurate.

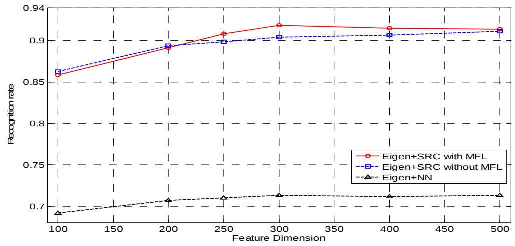

Figure 5. Recognition rates by different methods versus feature dimension on the AR database with Eigenfaces feature.

Table 3. The top recognition rates (%) of different methods on the AR database and the associated dimension of features.

Method	SRC	MFL	NN
Rate	91.14	**91.86**	71.29
Dimension	500	**300**	500

5. CONCLUSION

In this paper, we discussed the metaface learning (MFL) of face images when using sparse representation based classifier (SRC) for classification. Our experiments on Extended Yale B, ORL and AR face databases demonstrated that the proposed MFL algorithm can not only have higher accuracy than original SRC but also have less dictionary size. In learning the metafaces of each class, only the samples within that class were used. Therefore, the proposed MFL algorithm in this paper is an unsupervised learning method. In the future, we will investigate how to introduce the class label information into the MFL process so that a set of discriminative metafaces can be learned.

ACKNOWLEDGEMENT

This research is supported by the Hong Kong RGC General Research Fund (PolyU 5351/08E), the Program for New Century Excellent Talents in University of China and the NSFC under Grant No. 60973098.

6. REFERENCES

[1] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71-86, 1991.

[2] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriengman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *PAMI*, 19 (7):711-720, 1997.

[3] J. Yang, J.Y. Yang. Why can LDA be performed in PCA transformed space? *Pattern Recognition*, 36(2):563-566, 2003.

[4] J.B. Tenenbaum, V. deSilva, and J.C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500): 2319-2323, 2000.

[5] S.T. Roweis and L.K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500): 2323-2325, 2000.

[6] X. He, S. Yan, Y. Hu, P. Niyogi, and H.J. Zhang. Face recognition using laplacianfaces. *PAMI*, 27(3): 328-340, 2005.

[7] H.T. Chen, H.W. Chang, and T.L. Liu. Local discriminant embedding and its variants. In *CVPR*, 846-853, 2005.

[8] J. Yang, D. Zhang, J.Y. Yang, and B. Niu. Globally Maximizing, Locally Minimizing: Unsupervised Discriminant Projection with Applications to Face and Palm Biometrics. *PAMI*, 29(4):650-664, 2007.

[9] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust Face Recognition via Sparse Representation. *PAMI*, 31(2): 210 – 227, 2009.

[10] J.P. Brunet, P. Tamayo, T.R. Golun, and J.P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*, 101(12):4164-4169, 2004.

[11] R. Rubinstein, A.M. Bruckstein, and M. Elad. Dictionaries for Sparse Representation Modeling. To appear in Proceedings of *IEEE, Special Issue on Applications of Compressive Sensing & Sparse Representation*.

[12] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[13] M. Aharon, M. Elad, and A.M. Bruckstein. The K-SVD: An Algorithm for Designing of Overcomplete Dictionaries for Sparse Representation. *IEEE Trans. On Signal Processing*, 54(11):4311-4322, 2006.

[14] S.J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. A method for large-scale l1-regularized least squares. *IEEE Journal on Selected Topics in Signal Processing*, 1(4):606–617, 2007.

[15] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *PAMI*, 23(6):643-660, 2001.

[16] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *PAMI*, 27(5):684-698, 2005.

[17] A.M. Martinez and R. Benavente. The AR Face Database, *CVC Technical Report No. 24*, 1998.

[18] M. Yang and L. Zhang. Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary. In *ECCV 2010.*