Improving the Generalization of Fisherface by Training Class Selection Using SOM^2

Jiayan Jiang[1], Liming Zhang[1], and Tetsuo Furukawa[2]

[1]E.E. Dept. Fudan University, 220 Handan Road, Shanghai, China
jiangjiayan@citiz.net, lmzhang@fudan.edu.cn
[2]Kyushu Institute of Technology, Kitakyushu 808-0196, Japan
furukawa@brain.kyutech.ac.jp

Abstract. Fisherface is a popular subspace algorithm used in face recognition, and it is commonly believed superior to another technique, Eigenface, due to its attempt to maximize the separability of training classes. However, the derived discriminant subspace of the training set may not easily extend to unseen classes, as in the case of enrollment of new subjects. In this paper, we select some "representative" classes for Fisherface training using a recently proposed neural network architecture SOM^2. The experiment on ORL face database shows the proposed method can effectively reduce the performance variance and improve the generalization of Fisherface.

1 Introduction

Face recognition has become an active research topic for decades of years due to its value in both theory and application. To solve this problem, a great number of techniques have been developed, among which Eigenface, a PCA-based algorithm [1], and Fisherface, an LDA-based algorithm [2], are very popular ones.

Although it is argued that LDA may not always outperform PCA, especially when the training samples per class are insufficient or ill-sampled [6], it is a common belief that LDA is superior to PCA, since it tends to maximize the separability of the training classes [2]-[5]. However, in most previous work related to LDA, the classes are fixed during the training and testing phases, i.e. the subjects (not the images) being tested are always those involved in training phase [2]-[5].

On the other hand, training is a standalone process prior to the enrollment of subjects in some large-scale face recognition test-bed. The face image database is usually divided into a development set for training, a gallery which contains the images to be enrolled, and a probe set which comprises of unknown faces to be identified [7, 8]. It should be noted that the training set does not contain all the subjects in the gallery, just as in a real problem. Once training is accomplished, re-training is impractical because it requires updating millions of existed records [7].

It remains unclear whether Fisherface, an LDA-based algorithm especially tuned for training classes, can also perform well on unseen classes in the gallery. This is in fact a generalization problem. This paper aims at improving the generalization of

Fisherface by selecting some “representative” training classes using a recently proposed neural network architecture SOM^2 which has been applied in data class visualization and interpolation [9, 10].

The remaining of this paper is arranged as follows: a brief review of the Eigenface and FisherFace algorithms is given in Section 2; The algorithm of SOM^2 and its application in training class selection of Fisherface are described in Section 3; Section 4 gives experimental result on a publicly available face database, the ORL face database; Finally conclusion is drawn in Section 5.

2 Background

Eigenface is a classical subspace face recognition algorithm proposed in [1]. It is based on the observation that all face images (one image is denoted by a sample vector whose elements are pixels concatenated in a row-wise or column-wise manner) reside in a relatively small subspace, called “face space”, compared with the original image space. Thus a classical dimensionality reduction technique PCA is employed to derive such a subspace, which is spanned by the eigenvectors corresponding to the m largest eigenvalues of the samples’ covariance matrix. These eigenvectors are referred to as Eigenfaces because their appearances are like faces when displayed as images. Although the features derived from Eigenfaces capture most variances of the samples, they are not optimal for classification purposes, for the variances are caused by not only the intrinsic differences of faces (the identities) but also the unwanted extrinsic factors such as lighting conditions.

To overcome the drawback of Eigenface and make use of the label information of the training samples, several LDA-based algorithms are proposed [2]-[5], among which Fisherface is the most famous one. Assume that N training samples $\{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_N\}$ belong to I classes $\{X_1, X_2, \cdots, X_I\}$, the aim of LDA is to select the projection matrix W in order that the ratio of the between-class scatter and the within-class scatter is maximized, i.e.

$$\begin{aligned} W_{LDA} &= \arg\max_{W} \frac{\left|W^T S_B W\right|}{\left|W^T S_W W\right|} \\ &= \left[\vec{w}_1, \vec{w}_2, \cdots, \vec{w}_m\right] \end{aligned} \qquad (\mathbf{1})$$

The between-class scatter matrix is defined as $S_B = \sum_{i=1}^{I} N_i\left(\vec{u}_i - \vec{u}\right)\left(\vec{u}_i - \vec{u}\right)^T$ and the within-class scatter matrix is defined as $S_w = \sum_{i=1}^{I} \sum_{\vec{x}_j \in X_i} \left(\vec{x}_j - \vec{u}_i\right)\left(\vec{x}_j - \vec{u}_i\right)^T$, where N_i is the number of samples in X_i, and $\vec{u}_i$, $\vec{u}$ are the mean vector of the samples in X_i and the grand mean vector of all samples respectively. If S_w is nonsingular, the solution is given by the eigenvectors of $S_w^{-1} S_B$. However S_w is always singular in face recognition problems, therefore PCA is used to reduce the dimensionality of samples so that S_w is full ranked. In Fisherface, the dimension of this intermediate subspace is $N - c$, and then LDA is applied in this reduced space. Similarly, if these eigenvectors are treated as images, they are called Fisherfaces.

It can be seen that Eigenface aims at deriving a general face subspace. If the training samples are sufficient, a test face image can also be projected into this subspace effectively, and the classification is performed within it. On the other hand, the attention of Fisherface is mainly focused on deriving a subspace in which the separability is maximized between *training classes*, which generally results in better classification performance than Eiganface with regard to these classes. However, it is not evident that the separability can be easily extended to unseen classes, as in the case of enrollment of new subjects. Although a conjecture is proposed in [2] that "Fisherface methods, which tend to reduce within-class scatter for all classes, should produce projection directions that are also good for recognizing other faces besides the ones in the training set.", it is not theoretically sound and not validated by their experiments.

We notice that in practice the training classes of Fisherface are usually randomly selected from a large dataset [8]. In the worst case Fisherface may be trained on some "noise" (non-representative) classes, thus the discriminant performance will be poor when confronted with new classes. From a statistical viewpoint, since each time a random training set is used, the variance of performance can be large across different trials. Our idea is that if some "representative" training classes can be selected out of the whole dataset, the performance variance may be reduced and the generalization of Fisherface may be improved.

3 Training Class Selection Using SOM^2

Suppose that a face database includes N samples $\{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_N\}$ belonging to I classes $\{X_1, X_2, \cdots, X_I\}$, it is more often than not that only a subset of it can be used in Fisherface training. Our goal is to select some "representative" (prototype) classes out of the whole dataset to form the training set so that the generalization is improved compared with an arbitrary selection. Unfortunately, classical techniques of VQ family, such as K-Means, Neural Gas [11], or SOM [12], do not provide us any solutions to this problem, since they only induce some "reference (codebook) vectors" without any class formation of them. In our case we need some more high-level techniques which enable density approximation in terms of *classes* rather than *samples*. In this paper, we use SOM^2, a newly proposed neural network architecture which has been applied in data class visualization and interpolation, to achieve this end.

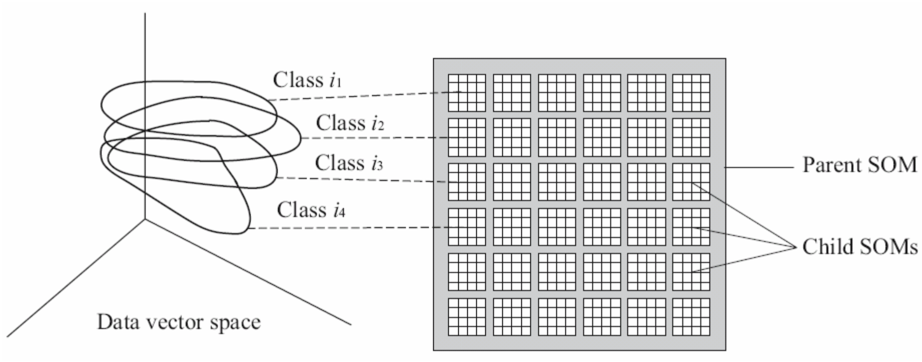

Fig. 1. The scheme and architecture of SOM^2 as "SOM of SOMs"

SOM^2 is short for "SOM of SOMs" [9, 10] which is a hierarchical structure of self-organizing maps, see Fig.1. The mapped objects in SOM^2 are called child SOMs, each trained to represent a data manifold. These child SOMs are interacting via a grand parent SOM, which finally generates a self-organizing map representing the distribution of data manifolds modeled by child SOMs. The algorithm of SOM^2 consists of three processes: the competitive process, the cooperative process, and the adaptive process. These processes are iterated until the result is converged or a maximum number of iterations is reached.

Suppose that SOM^2 comprises of K child SOMs, each of which has L codebook vectors $W^k = \left\{ \vec{w}^{k,1}, \cdots, \vec{w}^{k,L} \right\} \quad (k = 1, 2, \cdots, K)$. The competitive process includes the competition inside each child SOM and the competition between child SOMs (i.e. in the parent SOM). Let $l^{*k}_{i,j}$ denotes the "best matching unit (BMU)" in the k th child SOM for the j th sample of the i th data class. It is given by:

$$l^{*k}_{i,j} = \arg\min_{l} \left\| \vec{w}^{k,l} - \vec{x}_j \right\|^2 \quad \left(\vec{x}_j \in X_i \right) \tag{2}$$

Then the average quantization error of the k th child SOM for the i th class is determined as:

$$e_i^k = \frac{1}{N_i} \sum_{j=1}^{N_i} \left\| \vec{x}_j - \vec{w}^{k,l^{*k}_{i,j}} \right\|^2 \quad \left(\vec{x}_j \in X_i \right) \tag{3}$$

Finally, the "best matching map (BMM)", i.e. the child SOM which minimizes the average quantization error, is chosen for i th class:

$$k^*_{\ i} = \arg\min_{k} e_i^k \tag{4}$$

In the cooperative process, the learning rates for the parent SOM and the child SOMs are calculated. The normalized learning rate for the k th child SOM from the i th class is:

$$\phi_i^k = \frac{g\left[d\left(k, k^*_i\right), T\right]}{\sum_{i'=1}^{I} g\left[d\left(k, k^*_{i'}\right), T\right]} \tag{5}$$

And the normalized learning rate for the l th codebook vector from the j th sample of the i th class is:

$$\varphi_{i,j}^l = \frac{h\left[d\left(l, l^{**}_{i,j}\right), T\right]}{\sum_{j'=1}^{N_i} h\left[d\left(l, l^{**}_{i,j'}\right), T\right]} \tag{6}$$

Here $g[\cdot,\cdot]$ and $h[\cdot,\cdot]$ are the neighborhood functions of the parent and the child SOMs respectively, which shrink with the iteration T . $d(\cdot,\cdot)$ refers to the distance between two nodes in the map space, and $l^{**}_{i,j}$ denotes the BMU in the BMM for the j th sample of the i th class, i.e. $l^{**}_{i,j}$ □ $l^{*k^*_i}_{i,j}$.

In the adaptive process, all the codebook vectors of all the child SOMs are updated as follows:

$$\vec{w}^{k,l} = \sum_{i=1}^{I} \phi_i^k \sum_{\vec{x}_j \in X_i} \varphi_{i,j}^l \vec{x}_j \tag{7}$$

In our setting, SOM^2 is working in the “class density approximation” mode, i.e. the child SOMs are less than the data classes, or $K < I$. In this case, each child SOM learns the data vectors of the assigned classes in such a way that they can represent their average distribution. It can be regarded as an analogy of conventional SOM in the “point density approximation” mode: as the codebook vectors of a conventional SOM are “representative samples” (prototypes) of the training samples, the child SOMs of a SOM^2 form “representative classes” of the training classes. Inside each child SOM, it is still flexible whether to approximate or to interpolate the data distribution, depending on the number of codebook vectors and the number of samples.

The $K \times L$ codebook vectors (K classes with L samples per class) of SOM^2 are used as Fisherface training samples. Since these training classes are more “representative” than those randomly selected ones, they can be helpful to reduce the performance variance and improve the generalization of Fisherface, which will be validated in the next section.

4 Experiments

The ORL face database contains different images of 40 distinct subjects, with 10 images per subject. These images includes variations of lighting conditions, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All the subjects are in the upright, fontal position, with tolerance for some side movement. And all the images are grayscale with a resolution of 92×112. No preprocess is involved in this experiment. Ten images of one subject of the ORL database are shown in Fig. 2.

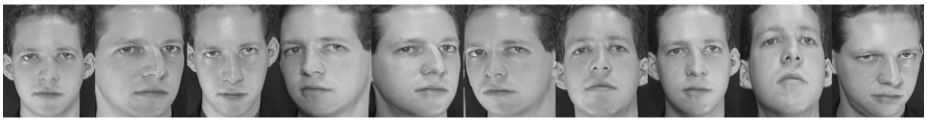

Fig. 2. Ten images of one sample subject from ORL face database

Please note that we are interested in the case of insufficient training classes and investigating the generalization of Fisherface. Thus we first divide the whole database into two partitions: a candidate training set which includes the first 5 images of all subjects, and a test set including the rest 5 images of all subjects. The candidate training set is used to train SOM^2, thus $I = 40$ and $N_i = 5 \quad (i = 1, \cdots, 40)$. There are $K(< I)$ child SOMs, and each one consists of 5 codebook vectors, i.e. $L = 5$. In this experiment, both the parent SOM and the child SOMs are one-dimensional maps. As mentioned before, these child SOMs are regarded as some "representative" classes for Fisherface training, and the codebook vectors within them are samples belonging to different training classes.

For comparison, another K classes are randomly selected from the candidate training set for Fisherface and Eigenface training. The ultimate feature dimensions are set to $(K-1)$ and $(K \times L - 1)$ for Fisherface and Eigenface respectively, since the best choices of feature dimensions for minimum recognition error are unknown beforehand in a real problem. Then the whole candidate training set serves as a gallery so that all the images in it are enrolled into the trained recognition system. At last, a Nearest Neighbor classifier is applied to determine the identity of each image in the test set based on the cos-similarity between a test image and the enrolled class centers. For each different K, 20 trials are conducted to determine the mean and standard deviations of recognition error rates. The results are listed in Table 1 and visualized in Fig. 3.

Table 1. Mean error rates and standard deviations of different K over 20 trials

K	**Proposed Method (Fisherface)**	Random Selection (Fisherface)	Random Selection (Eigenface)
10	**24.10%±2.52%**	27.30%±3.38%	25.00%±1.43%
15	**18.43%±1.09%**	19.03%±2.74%	20.40%±1.40%
20	**13.10%±0.87%**	15.43%±2.01%	18.25%±1.45%
25	**10.20%±1.34%**	12.53%±2.85%	17.13%±0.86%
30	**11.35%±0.69%**	11.73%±1.67%	16.82%±0.94%
35	**8.97%±1.04%**	11.43%±1.82%	16.00%±0.71%

Several interesting discoveries can be obtained from this experiment: 1) When the training classes are extremely insufficient ($K = 10$), Fisherface is inferior to Eigenface, then it outperforms Eigenface when more classes are involved in training. This phenomena is quite similar to that in [6], but the cause in [6] is insufficient data per class for training, rather than insufficient classes in our case; 2) The performance variances of Eigenface are always smaller than those of Fisherface, although the mean error rates are higher. This can be explained from a generalization perspective: since Eigenface is more successful in deriving a general face representation with the training samples, it is more statistically stable than Fisherface; 3) Our method effectively improves the generalization of Fisherface, which results in lower error rates, and reduces the performance variances, which are comparable to those of Eigenface.

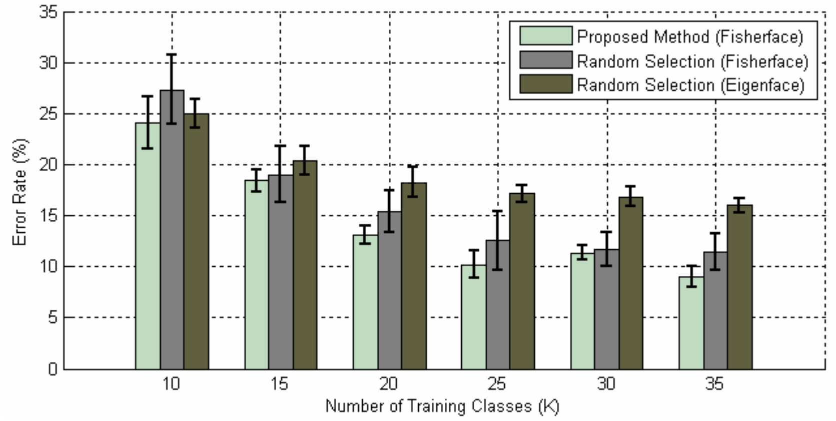

Fig. 3. Error rate versus number of training classes (K). The bars denote the mean error rates, and the lines denote the standard deviations over 20 trials

5 Conclusion

Although there have been a great number of papers published in the face recognition area, few of them investigate the impact of training set. A good start point is in [13], where some statistical properties of PCA (Eigenface) are studied. Along this line, we focus on the generalization problem of Fisherface in this paper. We first remind that the optimal discriminant subspace of the training set may not easily extend to unseen classes, as in the case of enrollment of new subjects; then we propose a method to reduce the performance variance and improve the generalization of Fisherface by selecting some "representative" training classes using a recently proposed neural network architecture SOM^2. The experiment on ORL face database validates this method.

In the future, some larger-scale face databases, such as FERET or CAS-PEAL-R1, will be used to investigate the statistical behavior of Fisherface. Another future work is to make use of some unique attributes of SOM^2, such as topological preservation in both parent SOM and child SOMs, and homology between different child SOMs.

Acknowledgement

The authors would like to thank AT&T Laboratories Cambridge for providing the ORL face database. This work was supported by NSFC 60571052.

References

1. Matthew A.T. and Alex P.P.: Face Recognition Using Eigenfaces. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (1991) 586-591
2. Peter N.B., Joao P.H., and David J.K.: Eigenface vs. Fisherface: Recognition Using Class Specific Linear Projection. IEEE Trans. on Pattern Anal. Machine Intell. Vol. 19 (1997) 711-720
3. Daniel L.S. and John W.: Using Discriminant Eigenfeatures for Image Retrieval. IEEE Trans. on Pattern Anal. Machine Intell., Vol.18 (1996) 831-836
4. Zhao W., Chellappa R., and Krishnaswamy A.: Discriminant Analysis of Principal Components for Face Recognition. Pro. IEEE Conf. on Automatic Face and Gesture Recognition (1998) 336-341
5. Chengjun L. and Harry W.: Robust Coding Schemes for Indexing and Retrieval from Large Face Databases. IEEE Trans. on Image Processing, Vol. 9 (2000) 132-137
6. Aleix M.M. and Avinash C.K.: PCA versus LDA. IEEE Trans. on Pattern Anal. Machine Intell. Vol. 23 (2001) 228-233
7. Phillips P.J., Hyeonjoon Moon, Rizvi S.A., and Rauss P.J.: The FERET Evaluation Methodology for Face-Recognition Algorithms. IEEE Trans. on Pattern Anal. Machine Intell. Vol. 22 (2000) 1090-1104
8. Bo C., Shiguang S., Xiaohua Z., and Wen G.: BaseLine Evaluation on the CAS-PEAL-R1 Face Database. Lecture Notes in Computer Science (LNCS3338), Advances in Biometric Person Authentication (2004) 370-378
9. Tetsuo F.: SOM of SOMs: Self-organizing Map Which Maps a Group of Self-organizing Maps. Proc. ICANN (2005) 391-396
10. Tetsuo F.: SOM^2 As “SOM of SOMs”. Proc. WSOM (2005)
11. Martinetz T.M., Berkovich S.G., Schulten K.J.: “Neural-Gas” Network for Vector Quantization and its Application to Time-Series Prediction. IEEE Trans. on Neural Networks, Vol. 4 (1993) 558-569
12. Kohonen T.: Self-Organizing Maps, 3rd.ed., Springer (2001)
13. Phillips P.J., Flynn P.J., Scruggs T., et al.: Overview of the Face Recognition Grand Challenge. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (2005) 947-954