Recognition from a Single Sample per Person with Multiple SOM Fusion

Xiaoyang Tan[1,2], Jun Liu[1], and Songcan Chen[1,2]

[1] Department of Computer Science and Engineering,
Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China
{x.tan,j.liu,s.chen}@nuaa.edu.cn
[2] Shanghai Key Laboratory of Intelligent Information Processing
Fudan University, Shanghai 200433, China

Abstract. One of the main challenges faced by the current face recognition techniques lies in the difficulties of collecting samples, and many existing face recognition techniques rely heavily on the size and representative of training set. Those algorithms may suffer serious performance drop or even fail to work if *only one training sample per person* is available to the systems. In this paper, we present a multiple-SOMs-based fusion method to address this problem. Based on the localization of the face, multiple Self-Organizing Maps are constructed in different manners, and then fused to obtain a more compact and robust representation of the face, through which the discrimination and class-specific information can be easily explored from the single training image among a large number of classes. Experiments on the FERET face database show that the proposed fusion method can significantly improve the performance of the recognition system, achieving a top 1 matching rate of 90.0%.

1 Introduction

The aim of face recognition is to identify or verify one or more persons from still images or video images of a scene using a stored database of faces. Many research efforts [1] have been focused on how to improve the accuracy of a recognition system. However, it seems that most of them ignore the potential problem that may stem from the face database at hand, where there may be only one sample image per person stored, possibly due to the difficulties of collecting samples or the limitations of storage capability of the systems, etc.

Under this condition, most of the traditional methods such as eigenface [3] and fisherface [4] will suffer serious performance drop or even fail to work, due to the absence of enough samples for a reliable covariation estimation. This problem, called the *one sample per person problem* (or, *one sample problem* for short), is defined as follows: Given a stored database of faces with only one image per person, the goal is to identify a person from the database later in time in any different and unpredictable poses, lighting, etc from the individual image.

Due to its challenge and significance for real-world applications, several researchers have recently made attempts to face the challenge. The methods in lit-

eratures include synthesizing virtual samples [5, 6], probabilistic matching [6],class-specific subspace[8] , neural network method [7], and so on.

In this paper, we attempt to address the above problems within a general framework based on the Self-Organization Map (SOM,[9]). The main idea behind is to extract latent local features that are invariant to appearance changes using the SOM network. In particular, a face image is partitioned into facial patches first, then they are mapped into a trained SOM topological space, where their distribution is analyzed. In this way, we can represent a face image more compactly, while providing enough local information for robust recognition. Our previous work [7] shows that this strategy is very successful in handling large variations contained in the dataset,such as large expression changes and partial occlusions.

In this paper, we further extend the framework by incorporating a multiple classifier fusion technique. We use multiple SOM maps constructed in totally different manners to explore local discrimination information as diverse as possible. The local information collected are then fused for the subsequent recognition. Experimental results on the FERET dataset show that the proposed hybrid method significantly improves the performance of the recognition system with one sample per person.

The paper proceeds as follows. After briefly reviewing the SOM network in section 2, we described the proposed method in section 3. The experiments are reported in section 4. Finally, conclusions are drawn in section 5.

2 The Self-Organization Maps

The Kohonen's Self-Organizing Map can be regarded as a Winner-Take-All unsupervised learning neural network. It stores prototypes w_{ij} of the input vectors at time t, x(t), in the neurons of a 2D map. Starting from complete disorder initial prototype vectors, topological order can be achieved by following two simple rules:

1)Locate the best-matching unit n_i , i.e. the closest prototype to the new input vector $x_i(t)$.

2)Updates the prototype vector of the winner neuron and its topological neighbors as follows:

$$\forall j : w_{ij}(t) = w_{ij}(t-1) + \eta(t)h(t)(x_i(t) - w_{ij}(t-1)) \tag{1}$$

where $\eta(t)$ denotes the learning rate at time t, which defines an attention field of a neuron that contributes a single stimulus. This allows different neurons to be trained for different types of data; $h(t)$ is a neighborhood function which governs both the self-organization process and the topographic properties of the map. Note that both the learning rate and neighborhood function decrease as the value of t, the time that was spent in the current context, increases.

3 The Proposed Method

3.1 Multiple SOM-based representation

Although the effectiveness and robustness of the basic SOM-face model has been revealed in our previous work[7], there is much room left to improve the performance of recognition system by extending the single-map to multiple-maps.

The single SOM scheme In the previous single map scheme, each face image is partitioned into M different local sub-blocks, each of which potentially represents specific local information of the image.For simplicity and efficiency, each sub-block is represented as a local feature vector (LFV) by concatenating the pixels of each sub-block. Such a gray-level-based local feature representation has been proven to be successful in both face detection and face recognition[1]. Actually, the obtained LFV's can not only encode the appearance of local regions but somehow preserve the spatial configuration of 2D face image as well. In addition, for high dimensional data, image partition is a powerful way to address the curse of dimensionality[2].

A self-organizing map (SOM) neural network is then trained and used to project all the LFVs onto a quantized lower dimensional space so as to obtain a compact but robust representation.Such a representation of face image is called "SOM-face". Its main advantage lies in that, in the "SOM-face", the information contained in the face is distributed in an orderly way and represented by several neurons instead of only one vector, so the common features of different classes can be easily identified.

However, in this scheme, it is the overall distribution of local features from every class that is emphasized. Therefore, the salient class-specific statistical features may be submerged in the other features. In addition, when the training sets grow large, the resulting map should also grow large enough to decrease the degree of overlapping between classes, thus increasing the computational cost.

The Multiple SOMs scheme To overcome those problems, we develop two multiple-map schemes, in which multiple rather than single maps are constructed based on the localization of face images, so as to deliberately specialize the representation of classes and local feature clusters, respectively.

In the first scheme, a separate SOM map for each class is trained using only the sub-blocks from the corresponding class. That is, each map is trained using class-designated samples so that the distribution of local features within each class is approximated. Thus we name this scheme cMSOM (class-dependent Multiple SOMs). An advantage of this scheme is that the robustness of recognition system can be improved due to the characterization of salient local features within each class. The computational cost of this scheme, however, is linear in relation to the number of classes to be recognized. When the number of classes is very large, the number of maps to be trained will be very large as well. Another multiple-map scheme is thus proposed to overcome this problem and is described below.

In the second scheme, a partition mechanism is carefully designed to divide the sample space into multiple training sets, one for each map. As illustrated in Fig.1, the training faces are divided into sub-blocks with equal size, and then the sub-blocks at the same position of each face are collected to form a separate training set, which is then used to train a SOM map for that set of facial features. Since each map's training set is constructed in a lateral way, thus the name lMSOM (lateral Multiple SOM). This scheme aims to characterize the distribution of similar local features between classes. In contrast to the cMSOM scheme, the totality of maps in the lMSOM scheme relies only on the number of sub-blocks of each face.

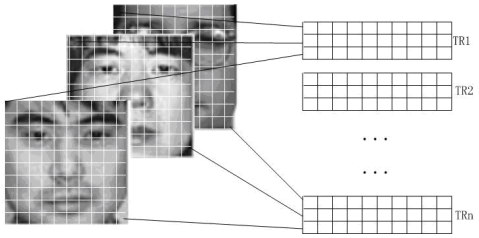

Fig. 1. Illustration of construction of training sets for each map in lateral MSOM scheme. (Left: the original images; Right: multiple training sets constructed).

In summary, the three schemes mentioned above (i.e. single SOM, cMSOM and lMSOM) are all trying to approximate the distribution of local features in sample space, based on the Kohonen maps, aiming to represent the salient statistical structures hidden in the high dimensional image data. We can visualize the original face image from the three different kinds of maps, and call the reconstructed face single-SOM-face, cMSOM-face and lMSOM-face respectively. See Fig.2 for an example.

Fig. 2. Example of an original image and its variant representations.

3.2 Multiple SOM-faces Fusion

The multiple SOM-faces models described above, as well as the basic single SOM-face model, are different from each other both in their construction manners and in the characterization of the salient local features. Intuitionally, such diversity makes it feasible for us to fuse the decisions of those models to improve the overall performance of the recognition system.

Actually, diversity among the classifiers in fusion has been widely recognized as an important factor in classifier combination[10]. In our case, the diversity is indeed a natural product of the three different SOM-face representation of face. Moreover, to further increase the extent of diversity, a disturbing faces are constructed for each testing face using E(PC)^{2}A technique [10]. An example of a disturbing face is shown in Fig.2 (e).

In contrast to most fusion methods where only the outputs for particular class are fused to make the decision, we take the outputs for all the classes into consideration. In this case, the outputs of each component classifier are interpreted as some support or confidence value for all the classes. This interpretation leads to a fuzzy aggregation method, whose high-level block diagram is shown in Fig.3.

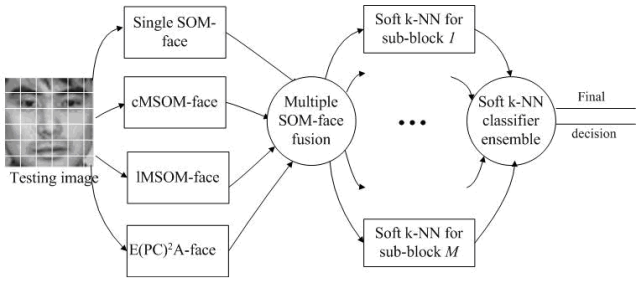

Fig. 3. Example of an original image and its variant representations.

As shown in Fig.3, the multiple SOM-face representations of the testing face are first fused, followed by a soft k-NN ensemble decision scheme to give the final class label.

To estimate the confidence c_{jk} that the j-th sub-block belongs to the k-th classes, we first calculate the L_2 pairwise distances between the j-th sub-block and all the "prototype sub-blocks" at the same position, according to the four models,respectively. Denote the obtained local pairwise distances as $\{d_{jk}^i\}_{k=1}^{k=N}, i = 1, 2, 3, 4$, where N is the total number of training samples. Note that in the situation of one sample per person, the number of training samples is equal to the number of training class.

To fuse them, a minimum aggregation rule is employed [10], that is, the fused distance value d_{jk} is given by,

$$d_{jk} = \min_{i=1,2,3,4}(d_{jk}^i) \tag{2}$$

Then we have:

$$c_{jk} = \frac{log(\min_{k=1...N}\{d_{jk}\} + 1)}{log(d_{jk} + 1)} \tag{3}$$

Clearly, the class with minimum distance to the testing sub-block will yield a confidence value closer to one, while a large distance produces a very small confidence value, meaning that it is less likely for the testing sub-block to belong to that class.

In our implementation, only the first K pairwise distances in $\{d_{jk}\}_{k=1}^{k=N}$ where used for confidence estimation. In other words, equation 2 is rewritten as

$$c_{jk} = \frac{log(\min_{k=1...K}\{d_{jk}\} + 1)}{log(d_{jk} + 1)} \tag{4}$$

and $c_{jk} = 0(\forall k > K)$. This helps to reduce the influence of outliers.

Finally, the label of the test image can be obtained through a linearly weighted voting scheme, as follows,

$$Label = arg \max_{k=1...N} \sum_{j=1}^{M} C_{jk} \tag{5}$$

4 Experiments

The experimental face database used in this work comprises 400 gray-level frontal view face images from 200 persons, with the size of 256×384. There are 71 females and 129 males. Each person has two images (fa and fb) with different facial expressions. The fa images are used as gallery for training while the fb images as probes for testing. All the images are randomly selected from the FERET face database [11]. No special criterion is set forth for the selection. So, the face images used in the experiments are very diversified, e.g. there are faces with different race, different gender, different age, different expression, different illumination, different occlusion, different scale, etc., which greatly increases the difficulty of the recognition task. See [5] for some concrete face samples.

Before the recognition process, the raw images are normalized according to some constraints so that the face area could be appropriately cropped. Those constraints include that the line between the two eyes is parallel to the horizontal axis, the inter-ocular distance (distance between the two eyes) is set to a fixed value, and the size of the image is fixed. Here in the experiments, the eyes are manually located, the cropped image size is 60×60 pixels and the inter-ocular distance is 28 pixels.

The aim of this paper is to illustrate the potentials of multiple SOM-faces as well as to study the behavior of classifiers constructed by using such SOM-faces. The first set of experiments focuses on the comparison between the proposed method and some other template-based approaches that deal with the *one image per person* problem, such as nearest neighbor (1-NN), eigenface, and E(PC)^{2}A, concerning the recognition performance. The details of the experiments are given below.

Although localizing the face image into sub-blocks itself is a way to improve diversity, here we do not try to discuss about optimal selection of the block size for the specific problem to achieve the perfect classification. Instead, we only choose block size of 3×3 for all experiments in the localization phase to verify effectiveness of the proposed methods. Then three kinds of SOM-face modes mentioned in section 3 are applied and their separate SOM maps are trained respectively. The training process is divided into two phases as recommended by [9], that is, an ordering phase and a fine-adjustment phase. 1000 updates are performed in the first phase, while 2000 times in the second one. The initial weights of all neurons are set to the greatest eigenvectors of the training data, and the learning parameter and the neighborhood widths of the neurons converge exponentially to 0 with the time of training. Finally, to label a testing image, its single-SOM-face, cMSOM-face, lMSOM-face and E(PC)^{2}A-face are constructed respectively and are fused, followed by a soft k-NN ensemble to make decisions about the class it belongs to.

Table 1 presents the performance of the proposed method with reference to the other three approaches mentioned above concerning the top 1 match rate. Table 1 reveals that the multiple SOM-face fusion method obtains the best performance among the compared methods. This promising result indicates that the fusion algorithm is able to combine information of different local features from different SOM-faces, resulting in a general performance improvement.

Methods	Accuracy(%)
1NN	84.0
Eigenface	83.0
EPC2A	85.5
Single SOM-face	88.5
cMSOM-face	87.5
lSOM-face	89.0
MSOM-faces fusion	**90.5**

Table 1. Comparison of recognition accuracies with different approaches

Next, we studied the behavior of the soft k-NN ensemble classifier constructed in the local feature spaces. Experimental results are presented in Figure 4. It is clear that the fusion algorithm outperforms any other individual schemes considered (i.e. single SOM, cMSOM and lMSOM schemes) consistently, concerning

the top 1 matching rate. Furthermore, it can be observed from Figure 4 that the performances of the three individual SOM-face schemes are both accurate and diverse at different k-value, making it possible to combine them to achieve a stronger classifier. This can be seen as a possible explanation to the superiority of the multiple SOM-faces fusion scheme in accuracy.

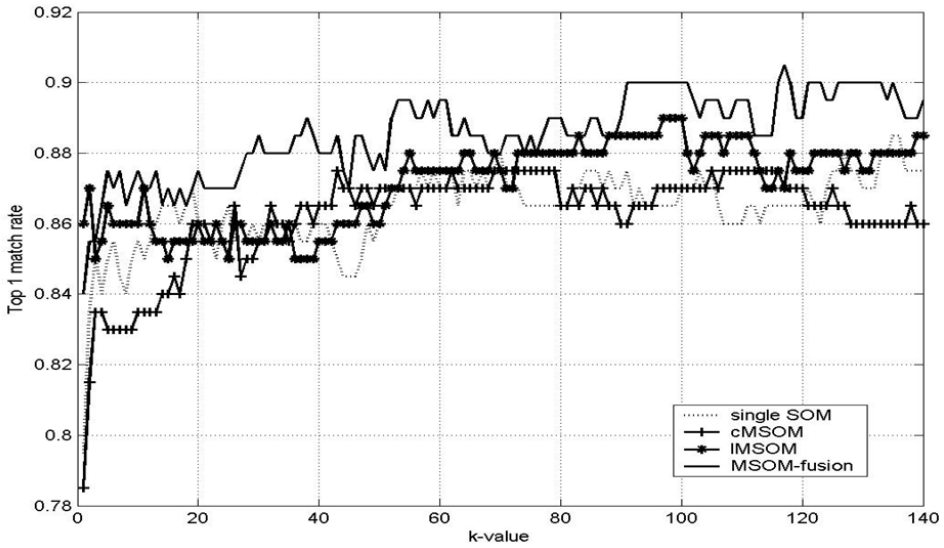

Fig. 4. Example of an original image and its variant representations.

5 Conclusions

In this paper, a novel face representation and recognition approach is discussed, where face images are first localized, then represented by their SOM-based proximities. This allow us to extract and exploit the salient statistical structures hidden in the high dimensional image data, under the situation of only one training image per person. Two different ways of building multiple SOM-faces are discussed. In the first approach, a separate SOM map for each class is trained using class-designated samples, thus the local features important for the specific class are encouraged. In the second approach, the sub-blocks at the same position of each face are grouped to form a separate training set to train one SOM map, such that local features important to distinguish one class from the others are emphasized. We conducted experiments on the FERET dataset for both approaches to illustrate their effectiveness in dealing with the one sample problem. The experimental results confirm the superiority in accuracy of the ensemble

approach. The proposed method is anticipated advantageous in other scenarios where samples are very small as well.

Acknowlegement

We thank Natural Science Foundation of China under Grant Nos. 60473035 for support.

References

1. Zhao, W. , Chellappa, R., Phillips, P. J. and Rosenfeld, A., "Face Recognition: A Literature Sur-vey, "ACM Computing Survey, December Issue, pp. 399-458, 2003
2. Jain, A.K. and Chandrasekaran, B. Dimensionality and sample size considerations in pattern recognition practice. In Handbook of Statistics, P.R. Krishnaiah and L.N. Kanal, Eds., vol 2, 835-855,1982.
3. Brunelli, R. and Poggio, T. (1993) Face recognition: features versus templates. IEEE Trans. Pattern Analysis and Machine Intelligence 15(10): 1042-1062.
4. Belhumeur, P., Hespanha, J. and Kriegman, D. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Analysis and Machine Intelligence 19(7): 711-720,1997
5. Chen S.C., Zhang, D.Q., and Zhou Z.-H. Enhanced (PC)2A for face recognition with one training image per person. Pattern Recognition Letter, 2004, 25:1173-1181
6. Martinez, A.M. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. IEEE TPAMI 748-763 2002.
7. Tan X.Y., Chen S.C., Zhou Z.-H., and Zhang F.Y., Recognizing partially occluded, expression variant faces from single training image per person with SOM and soft kNN ensemble. IEEE Transactions on Neural Networks, 16(4): 875-886,2005.
8. Shan S., Gao W., Zhao D., Face Identification Based On Face-Specific Subspace, International Journal of Image and System Technology, 13(1), pp23-32, 2003.
9. Kohonen T. Self-Organizing Map, 2nd edition, Berlin: Springer-Verlag, 1997
10. Kittler, J., Hatef, M., Duin, R.P.W. and Matas, J., On combining classifiers. IEEE Trans. Pattern Anal. Machine Intell. 20(3), pp. 226-239,1998
11. Phillips P. J., Wechsler H., Huang J., Rauss P. J. The FERET database and evaluation procedure for face recognition algorithms. Image and Vision Computing, 1998, 16(5): 295-306