Face Recognition Using Balanced Pairwise Classifier Training

Ziheng Zhou, Samuel Chindaro, and Farzin

School of Engineering and Digital Arts, University of Kent, Canterbury, UK
{z.zhou,s.chindaro,f.deravi}@kent.ac.uk

Abstract. This paper presents a novel pairwise classification framework for face recognition (FR). In the framework, a two-class (intra- and inter-personal) classification problem is considered and features are extracted using pairs of images. This approach makes it possible to incorporate prior knowledge through the selection of training image pairs and facilitates the application of the framework to tackle application areas such as facial aging. The non-linear empirical kernel map is used to reduce the dimensionality and the imbalance in the training sample set tackled by a novel training strategy. Experiments have been conducted using the FERET face database.format.

Keywords: face recognition, classification, aging.

1 Introduction

Face recognition (FR) has become one of the most important application areas for image analysis and understanding. Various feature-extraction techniques have been introduced and very often, large sets of features are extracted from a facial image (e.g., the Gabor features [1-3]) for classification. The question remains in how to best exploit the richness of such large feature sets and increasingly large training data sets to answer some of the challenging FR applications, such as compensating for the facial changes caused by aging [4].

The traditional approach to using large feature sets [5] is to find a subspace of the original feature space, which preserves as much subject-invariant information as possible. In other words, after a linear transformation from the original feature space to the subspace, the mapped feature points calculated from the images of the same person are made to be as close to each other as possible. By doing so, the intra-personal variations remaining in the feature sets are minimized and the distances between feature points are used to quantify the similarities between faces. Unlike other biometric modalities such as fingerprint and iris, human faces keep changing over our life time. In such a traditional FR approach, as the difference caused by aging becomes large, it becomes difficult to determine whether two images belong to the same person only by their measured distance.

Instead of ignoring the intra-personal variations within face images, we want to build an FR system that models the variations to handle facial aging. To do that, it needs to define a feature-extraction function $f\colon \mathfrak{I} \times \mathfrak{I} \rightarrow \mathbb{R}^m$ to explicitly characterize the difference between two faces. Here $\mathfrak{I}$ is the image domain and $\mathbb{R}^m$ denotes the

D. Weerasinghe (Ed.): ISDF 2009, LNICST 41, pp. 42–49, 2010.

feature space with dimension m. It can be seen that any set of features $\boldsymbol{x} = f(I_1, I_2)$, $I_1, I_2 \in \mathfrak{I}$ can be classified either as *intra-personal,* if images I_1 and I_2 are from the same person, or *inter-personal* otherwise. Therefore, the FR problem is turned into a supervised two-class classification problem. When someone uses the system, a feature vector is calculated from his (her) current captured facial image and a previously-enrolled stored image. The feature vector is then classified into one of the two classes, telling whether or not the two images are from the same person.

One attractive advantage of this approach is that the FR system can be built in such a way that our prior knowledge can be incorporated in the selection of the training pairs from which the features are extracted. The guiding knowledge for this selection can be any desired characteristic of the image pairs to be classified. For instance, when comparing two facial images, the information about when they are taken is a source of prior knowledge that can be used. Here the training set presented to the classifier is called the feature training set (FTS) and consists of the feature vectors calculated from the selected image pairs chosen from the training set (TS) of facial images. Note that although considering FR as a two-class classification problem [6] is not new, the idea of incorporating prior knowledge within such an approach as presented here is novel.

Although the classification methods can be used to tackle facial aging, there are difficulties in building such aging-adaptive FR systems. First, as mentioned above, the number of features extracted from images could be very large making the search for a classification solution in such a high-dimensional feature spaces very difficult. Secondly, since in the proposed approach image pairs are selected to calculate feature sets for training, there could be a large imbalance between the number of intra-personal pairs and the number of inter-personal pairs. Fig. 1 illustrates this imbalance.

In this paper, we propose a novel pairwise classification framework (PCF) that tackles the difficulties mentioned above. To demonstrate the framework's capability of handling high-dimensional feature vectors, the Gabor wavelets are used to calculate features from image pairs. To solve the two-class classification problem, we first reduce the dimensionality of the feature vectors using the non-linear empirical kernel map [7]. The Fisher's discriminant analysis is then used to find a classification

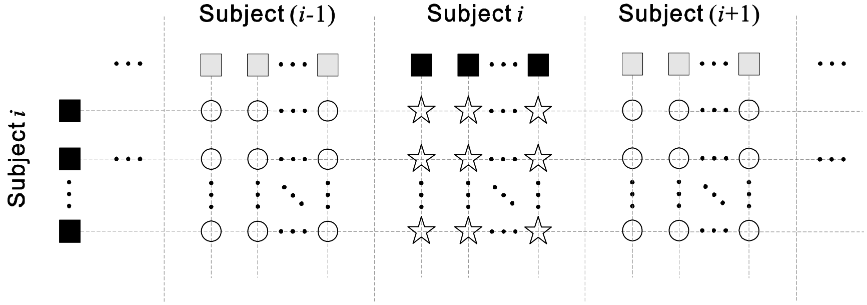

Fig. 1. An example showing the imbalance between the number of intra-personal pairs and the number of inter-personal pairs that can be sampled for Subject $\boldsymbol{i}$. The black squares represent images of Subject $\boldsymbol{i}$ and rest of the squares stand for other images. The stars mark all possible intra-personal image pairs for Subject $\boldsymbol{i}$, while the circles locate the inter-personal image pairs.

solution. A novel training strategy is proposed to tackle the imbalance within the training data. The FERET face database [8] is used for testing.

The rest of the paper is organized as following: Sections 2 and 3 describe the feature-extraction method and classifier training. Section 4 and 5 give details about the experiments and results with concluding remarks presented in Section 6.

2 Gabor Feature Extraction

Gabor wavelets have been successfully used in face recognition applications [1-3] and mathematically, can be defined as:

$$\Psi_{u,v}(\boldsymbol{z}) = \frac{\|k_{u,v}\|^2}{\sigma^2} e^{-\left(\|k_{u,v}\|^2\|z\|^2/2\sigma^2\right)}\left[e^{ik_{u,v}\,\boldsymbol{z}} - e^{-(\sigma^2/2)}\right] \tag{1}$$

where $u = 0, \dots ,7$ and $v = 0, \dots ,4$ define the scale and orientation of the wavelet, $\boldsymbol{z}$ is the xy coordinates and $k_{u,v}$ is defined as $k_{u,v} = 2^{-\frac{2+v}{2}}\pi\, e^{i\frac{u}{8}\pi}$. Features are obtained by convolving a given image with each of the Gabor wavelets.

Let $\boldsymbol{y}_{\mathrm{amp}}$ and $\boldsymbol{y}_{\mathrm{pha}}$ be the vectors that store the amplitude and phase values from an image using all the 40 wavelets defined by Equation 1. The total Gabor feature vector $\boldsymbol{y}$ is constructed as $\boldsymbol{y} = \left(\boldsymbol{y}_{\mathrm{amp}}^{\mathrm{T}}, |\boldsymbol{y}|_{\mathrm{pha}}^{\mathrm{T}}\right)^{\mathrm{T}}$. We then define the feature-extraction function f as:

$$\boldsymbol{x} = f(I_1, I_2) = |\boldsymbol{y}_1 - \boldsymbol{y}_2| \tag{2}$$

where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are the feature vectors calculated from image I_1 and I_2, and $\boldsymbol{x}$ is the feature vector characterizing the difference between image I_1 and I_2, which will be used to determine whether these two images are from the same person.

3 Classifier Training

In the proposed approach two main steps are included in the training process (Fig. 2): the empirical kernel map and the Fisher's discriminant analysis trained using an unbalanced FTS.

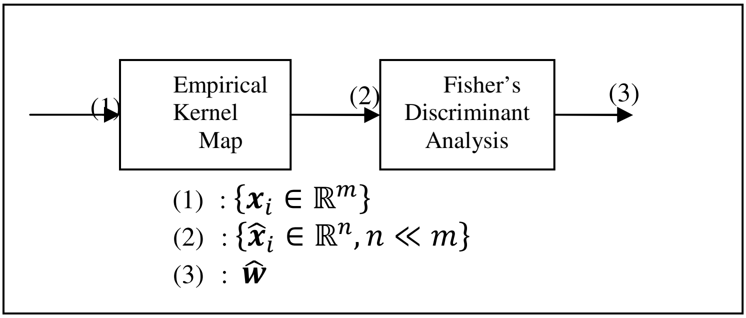
(1)
Empirical Kernel Map
(2)
Fisher's Discriminant Analysis
(3)
(1) : $\{\boldsymbol{x}_i \in \mathbb{R}^m\}$
(2) : $\{\hat{\boldsymbol{x}}_i \in \mathbb{R}^n, n \ll m\}$
(3) : $\widehat{\boldsymbol{w}}$

Fig. 2. Block diagram of classifier training

3.1 Empirical Kernel Map

The empirical kernel map (EKM) which is an approximation of the reproducing kernel map [7], is used to establish a non-linear map g: $\mathbb{R}^m \rightarrow \mathbb{R}^n$ which generates a new vector $\widehat{\boldsymbol{x}}$ which has a much lower dimension. Given a set of feature vectors $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}_i \in \mathbb{R}^m\}_{i=1}^n$, the empirical kernel map can be defined as:

$$\widehat{\boldsymbol{x}} = g(\boldsymbol{x}) = \begin{bmatrix} k(\boldsymbol{x}, \boldsymbol{x}_1) \\ \vdots \\ k(\boldsymbol{x}, \boldsymbol{x}_n) \end{bmatrix} \tag{3}$$

where k is a positive definite function. It can be seen that the size of $\boldsymbol{\mathcal{X}}$ decides the dimensionality of the new feature space. If $\boldsymbol{\mathcal{X}}$ includes all the training feature vectors, the training process shown in Fig. 2 is equivalent to the training process of the kernel discriminant analysis [7]. However, $\boldsymbol{\mathcal{X}}$ is chosen as a subset of all the training feature vectors in this work. It can be seen that the size of $\boldsymbol{\mathcal{X}}$, n decides the dimension of the new feature space which determines the computational complexity of the classification problem to be solved in the new feature space. Moreover, recent work [9] has shown that good performance can be achieved with lower values of n.

3.2 Training Strategy for Fisher's Discriminant Analysis on an Unbalanced Feature Training Set

Having reduced the feature dimension using EKM, the Fisher's discriminant analysis (FDA) is used in the new feature space to find the direction $\widehat{\boldsymbol{w}}$ along which the intra- and inter-personal feature points are maximally separated. As mentioned in Section 1, there could be an imbalance (see Fig. 1 for an example) in the FTS since the training samples (feature vectors) are calculated from image pairs rather than single images.

Since the system is to be tested on the standard FERET database, the image set defined in FERET for training is used as the TS. For this data set when selecting image pairs to compute the FTS, it is found out that the number of intra-personal samples, n^1 (approximately less than a thousand) is much less than the number of inter-personal samples, n^0 (approximately more than 500 thousands). If prior knowledge is incorporated (e.g., if it is known that the two images to be matched are taken in different sessions), a more restricted rule will be used to select image pairs (that is, every one of them will have two images taken in different sessions), possibly resulting in a much smaller n^1. Moreover, it is found that n^1 could also be much smaller than the dimension of feature vectors, n. Therefore, we need to perform the FDA under the condition that $n^1 \ll n \ll n^0$.

In the following, a novel training strategy is introduced to deal with the above-mentioned situation. It is well known that the solution of the FDA can be obtained by maximizing function $\mathcal{J}(\boldsymbol{w})$:

$$\mathcal{J}(\boldsymbol{w}) = \frac{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_{\mathrm{B}} \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} (\boldsymbol{S}_{\mathrm{w}}^1 + \boldsymbol{S}_{\mathrm{w}}^0) \boldsymbol{w}} \tag{4}$$

where $\boldsymbol{S}_{\mathrm{B}}$ is between-class scatter matrix and $\boldsymbol{S}_{\mathrm{w}}^1$ and $\boldsymbol{S}_{\mathrm{w}}^0$ are the scatter matrices for the intra- and inter-personal training samples, respectively (see [7] for the detailed definitions). Given that $n^1 \ll n^0$, it may be desirable not to misclassify any of the intra-personal samples, which means there will be a zero false reject rate on the training data. This can be achieved by restricting every possible $\boldsymbol{w}$ to be within the null space of $\boldsymbol{S}_{\mathrm{w}}^{1}$, $N(\boldsymbol{S}_{\mathrm{w}}^{1})$, or equivalently, $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{w}}^{1}\boldsymbol{w} = 0$. Let $\boldsymbol{V} = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_r]^{\mathrm{T}}$ be an orthonormal basis of the range of $\boldsymbol{S}_{\mathrm{w}}^{1}$ where r is the rank of $\boldsymbol{S}_{\mathrm{w}}^{1}$. The linear transformation from the feature space $\mathbb{R}^n$ onto $N(\boldsymbol{S}_{\mathrm{w}}^{1})$ can be established by using the matrix $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{V}\,\boldsymbol{V}^{\mathrm{T}}$ where $\boldsymbol{I}$ is the identity matrix. It is known that matrix $\boldsymbol{V}$ can be formed by the eigenvectors of $\boldsymbol{S}_{\mathrm{w}}^{1}$ with non-zero eigenvalues.

The problem of maximizing $\mathcal{J}(\boldsymbol{w})$ is then transformed into maximizing function $\Im(\boldsymbol{w}')$:

$$\Im(\boldsymbol{w}') = \frac{\boldsymbol{w}'^{\mathrm{T}}\boldsymbol{M}_{\mathrm{B}}\boldsymbol{w}'}{\boldsymbol{w}'^{\mathrm{T}}\boldsymbol{M}_{\mathrm{W}}^{0}\boldsymbol{w}'} \tag{5}$$

where

$$\boldsymbol{M}_{\mathrm{B}} = \boldsymbol{P}(\boldsymbol{\mu}^1 - \boldsymbol{\mu}^0)(\boldsymbol{\mu}^1 - \boldsymbol{\mu}^0)^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}} \tag{6}$$

$$\boldsymbol{M}_{\mathrm{W}}^{0} = \sum_i \boldsymbol{P}(\widehat{\boldsymbol{x}}_i^0 - \boldsymbol{\mu}^0)(\widehat{\boldsymbol{x}}_i^0 - \boldsymbol{\mu}^0)^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}} \tag{7}$$

Here $\boldsymbol{w}' \in N(\boldsymbol{S}_{\mathrm{w}}^{1})$, $\boldsymbol{M}_{\mathrm{B}}$ and $\boldsymbol{M}_{\mathrm{W}}^{0}$ are the corresponding within-class and inter-personal scatter matrices in $N(\boldsymbol{S}_{\mathrm{w}}^{1})$, $\{\widehat{\boldsymbol{x}}_i^1\}$ are the inter-personal feature vectors after the EKM and $\boldsymbol{\mu}^0$ and $\boldsymbol{\mu}^1$ are the means. Let $\widehat{\boldsymbol{w}'}$ be the vector that maximizes $\Im(\boldsymbol{w}')$. The solution of the FDA, $\widehat{\boldsymbol{w}}$ can be calculated as $\widehat{\boldsymbol{w}} = P\,\widehat{\boldsymbol{w}'}$.

4 Experimental Settings

The pairwise classification framework (PCF) was tested on two of standardized FERET experiments [8] which involved using some image subsets defined in the database, namely, the gallery set, FB probe set, Dup1 probe set and the training set (TS). The first experiment simulated an easy test scenario where every two image to be matched were taken around the same time. The second experiment represented a much harder scenario when every two images were taken on different days. The average time difference between two images is 251 days. The second experiment was designed to test the system's capability of handling facial aging and is used to demonstrate that by incorporating the prior knowledge a more effective FR system can be built to target such a difficult test scenario.

In the experiments, face images were normalized by the way described in [10], using the eye-coordinates provided in the FERET database. By using 40 Gabor wavelets defined by Equation (1) and (2), 1.64×10^5 features were calculated from each image pair. To reduce the dimension, a set of 2500 training samples were used to perform the EKM defined in Equation (3). Based on recent work [9] and the fact that $n^1 \ll n \ll n^0$, the set was constructed in such a way that all the intra-personal samples in the FTS were included and the rest were randomly sampled from the population of the inter-personal samples. Finally, the RBF kernel function, $\mathcal{k}(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|/2\sigma^2)$ was used in the EKM.

5 Results

In both experiments, the PCF was compared with other FR systems including the elastic bunch graph matching (EBGM) [1], the Gabor Fisher classifier (GFC) [2] and the Adaboost Gabor FR system (AGF) [3] using the features extracted by the same Gabor wavelets defined in (1). Table 1 shows the rank 1 recognition rate on the FB probe set (Results of EBGM and GFC were reported in [11]). It can be seen that all of the systems achieved recognition rates above 95% in this easy test scenario.

Table 1. Recogntion results on the FB probe set

FR System	Rank 1 Recognition Rate
EBGM	95.0%
GFC	96.3%
AGF	95.2%
PCF	97.8%

In the second experiment, systems were tested in a much harder test scenario. The prior knowledge used in selecting training image pairs is that the two images to be matched are taken on different days. To incorporate this knowledge, when selecting image pairs from the TS to calculate the FTS, we scanned all the possible image pairs and only used those whose images were taken on different days. Table 2 shows the rank 1 recognition results on the Dup1 probe set. It can be seen that the GFC outperformed the PCF. However, when closely looking at the TS, it was found out that only 244 intra-personal pairs satisfying the criterion could be formed from the images. Only 240 images from 60 subjects were involved in spite of the fact that there were 1002 images from 429 subjects in the TS. We believed that the small number of representative intra-personal image pairs caused the underperformance of the PCF system.

Table 2. Recogntion results on the Dup1 probe set

FR System	Rank 1 Recognition Rate
EBGM	59.1%
GFC	68.3%
AGFC	n/a
PCF	67.0%

To have more qualified intra-personal image pairs to train the PCF, we enlarged the TS by adding 72 images (10%) from the Dup1 probe set. The images were chosen in a way so that the number of intra-personal pairs that formed by two images taken on different days was maximized. In total, we managed to obtain 702 intra-personal image pairs using 320 images from 65 subjects. Instead of using all the images in the TS, the PCF was trained only on 320 images and tested on the reduced (90%) Dup1 probe set. To compare the results the system trained on the original TS was also tested on the reduced Dup1 probe set. Fig. 3 shows the cumulative match curves of the recognition results. The recognition rate for the PCF trained on 244 intra-personal samples was slightly higher (around 2%) from its figure (67%) in Table 2 due to the reduced size of the Dup1 probe set. For the PCF trained on the 703 intra-personal samples, the rank 1 recognition rate exceeded 75%. It can be seen that although the number of images used for training was much less than that in the original TS, the system performance was significantly improved in this difficult test.

6 Conclusion

We have presented a pairwise classification framework for face recognition. The novelty of this framework resides in providing a mechanism for incorporating prior knowledge through the selection of training data to tackle specific FR challenges such as facial aging and providing a novel training strategy to tackle the imbalance inherent in the training data available for pairwaise image classification. The experimental results have demonstrated the effectiveness of the proposed approach in incorporating prior knowledge, handling high-dimensional feature vectors and coping with the training data imbalance.

Acknowledgments. The authors would like to acknowledge the support of 3D FACE, a European Integrated Project funded under the European Commission IST FP6 program contract number 026845.

References

1. Wiskott, L., Fellous, J.-M., Küiger, N., von der Malsburg, C.: Face recognition by elastic bunch graph matching. IEEE Trans. PAMI 19(7), 775–779 (1997)
2. Liu, C., Wechsler, H.: Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition. IEEE Trans. IP 11(4), 467–476 (2002)
3. Yang, P., Shan, S., Gao, W., Li, Z., Zhang, D.: Face Recognition Using Ada-Boosted Gabor Features. In: Proc. IEEE Intl. Conf. Auto. Face and Gesture Recognition, pp. 356–361 (2004)
4. Patterson, E., Sethuram, A., Albert, M., Ricanek, K., King, M.: Aspects of Age Variation in Facial Morphology Affecting Biometrics. In: Proc. BTAS, pp. 1–6 (2007)
5. Cevikalp, H., Neamtu, M., Wilkes, M., Barkana, A.: Discriminative common vectors for face recognition. IEEE Trans. PAMI 27(1), 4–13 (2005)
6. Moghaddam, B., Wahid, W., Pentland, A.: Beyond eigenfaces: probabilistic matching for face recognition. In: Proc. IEEE Intl. Conf. Auto. Face and Gesture Recognition, pp. 30–35 (1998)
7. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. MIT Press, Cambridge (2001)
8. Phillips, P.J., Moon, H., Rauss, P.J., Rizvi, S.: The FERET evaluation methodology for face-recognition algorithms. IEEE Trans. PAMI 22(10), 1090–1104 (2000)

9. Zhou, Z., Chindaro, S., Deravi, F.: Non-Linear Fusion of Local Matching Scores for Face Verification. In: Proc. IEEE Intl. Conf. Auto. Face and Gesture Recognition (2008)
10. Beveridge, J.R., Bolme, D.S., Draper, B.A., Teixeira, M.: The CSU face identification evaluation system: its purpose, features and structure. Machine Vision and Applications 16(2), 128–138 (2005)
11. Zhang, B., Shan, S., Chen, X., Gao, W.: Histogram of Gabor phase patterns (HGPP): A novel object representation approach for face recognition. IEEE Trans. IP 16(1), 57–68 (2007)