Designing Multiple Classifier Systems for Face Recognition

Nitesh V. Chawla and Kevin W. Bowyer

Department of Computer Science and Engineering,
University of Notre Dame, IN 46556, USA
{nchawla, kwb}@cse.nd.edu

Abstract. Face recognition systems often use different images of a subject for training and enrollment. Typically, one may use LDA using all the image samples or train a nearest neighbor classifier for each (separate) set of images. The latter can require that information about lighting or expression about each testing point be available. In this paper, we propose usage of different images in a multiple classifier systems setting. Our main goals are to see (1) what is the preferred use of different images? And (2) can the multiple classifiers generalize well enough across different kinds of images in the testing set, thus mitigating the need of the meta-information? We show that an ensemble of classifiers outperforms the single classifier versions without any tuning, and is as good as a single classifier trained on all the images and tuned on the test set.

1 Introduction

Face recognition is becoming an increasing popular and relevant area of study. The Face Recognition Grand Challenge (FRGC) sponsored by various US Government agencies is a prime example of the growing importance of improving or benchmarking face recognition techniques [1, 2]. In this paper we focus on 2-D face recognition, which has been a subject of significant study [3, 4]. Two dimensional face images are usually represented as high-dimensional pixel matrices, where in each matrix cell is a gray-level intensity value. These raw feature vectors can be very large and highly correlated. Moreover, the size of the training data is usually small. To combat these issues of very high feature correlation, small sample size and computational complexity, the face images are often transformed into a lower dimensional manifold. One of the most popular techniques for linear transformation in feature space is Principal Component Analysis (PCA) [5, 6]. PCA reduces the dimensions by rotating feature vectors from a large highly correlated feature space *(image space)* to a smaller feature space *(face space)* that has no sample covariance between the features. After applying PCA to reduce the face space to a lower dimensional manifold, a single nearest neighbor classifier or a linear discriminant classifier is typically used.

We will now introduce some terms and notation from biometrics that will be used throughout the paper. **Subject**: A person or a subject in the training

N.C. Oza et al. (Eds.): MCS 2005, LNCS 3541, pp. 407–416, 2005.

set is similar to a class or concept in data. This person can be associated with multiple images in the training set; **Training set**: The training set is defined to be all the images of subjects that are available for constructing the face space; **Gallery set**: Gallery is the set of subjects enrolled in the database and can either be the same as the training set or different. Due to a lack of enough data, the gallery images are often used as the training set for constructing the face space. However, gallery images in this paper comprise of the same subjects (but images captured on a different date) and completely new subjects; **Probe set**: Probe set is the "testing" set. The imagesin the probe set are typically of the same subjects who are in the gallery set, but are taken at a later point in time. The goal is to project the probe set into the trained face space and correctly match it with the projected representative in the gallery.

Two dimensional face recognition presents a multitude of challenges when applied to conditions (including subjects) that weren't part of the training set. An example of this is a face space trained on a neutral expression if presented with a smiling expression face space. Ideally, the face recognition algorithm should be fairly insensitive to changes in the lighting direction and intensity or facial expression. In addition, even if we try to control the face space of the training session and the testing session to have the identical lighting and expression conditions, there still can be differences between the two caused by errors in normalization, slight pose changes, illumination variations etc. Even if the same controlled lighting environment is used, it can still cause illumination variations if the testing set image is captured on a different day, for example.

One may construct a single classifier by combining possible variations in the lighting direction and facial expression for constructing a face space. However, PCA can potentially retain the variation in lighting direction, illumination, and expression that is not relevant for recognition. The covariance matrix constructed will capture both inter-class and intra-class variance. To maximize the inter-class distance (across subjects) and minimize the intra-class distance (within subjects), Linear Discriminant Analysis [7, 8] (LDA) can be used. But LDA suffers from the small-sample size problem, and requires "enough" images of a subject [9, 7]. Typically, researchers have proposed using at least 10 images of each subject [10]. The goal is to correctly recognize a face, and not essentially distinguish between different variations of a face. Also another challenge in 2-D face recognition is that the subjects used in the testing or the probe set may not be present in the training set. So, essentially, we need a classifier that can generalize well enough, without overtraining on a specific face space.

We propose to utilize multiple classifier systems or ensembles in the biometric problem of 2-D face recognition. We randomly sample from the acquired images of a subject to construct face spaces. We construct 50 such face spaces for an ensemble. Given 4 images (different expression and lighting conditions) of each subject, we randomly sample 1, 2 and 3 images 50 times. We explain the data in the subsequent sections. In the sections that follow, we will compare different ways of defining the training set for using a classifier or a set of classifiers. We can formalize the objective of this paper as follows:

1. What is the best use of available multiple training images of a subject?
2. Can we construct a classifier or a set of classifiers that can be applied across probe images with different expressions and/or lighting conditions? The goal is to do as well if not better than the different single classifiers constructed specifically to represent particular lighting and expression conditions.

2 Classifiers

In this section, we discuss in brief the PCA methodology, the MahCosine distance metric as implemented in the CSU code [11], and the linear discriminant analysis classifier or LDA. For both nearest neighbor and LDA, PCA methodology is applied first. All the images are first normalized such that the pixel values have a zero mean and unit variance.

2.1 PCA

The raw feature vectors are a concatenation of the gray-level pixel values from the images. Let us assume there are m images and n pixel values per image. Let Z be a matrix of *(m,n)*, where m is the number of images and n is the number of pixels (raw feature vector). The mean image of Z is then subtracted from each of the images in the training set, $\Delta Z_i = Z_i - E[Z_i]$. Let the matrix M represent the resulting "centered" images; $M = (\Delta Z_1, \Delta Z_2, ..\Delta Z_m)^T$. The covariance matrix can then be represented as: $\Omega = M.M^T$. Ω is symmetric and can be expressed in terms of the singular value decomposition $\Omega = U.\Lambda.U^T$, where U is an m x m unitary matrix and $\Lambda = diag(\lambda_1, ..., \lambda_m)$. The vectors $U_1, ..., U_m$ are a basis for the *m-dimensional* subspace. The covariance matrix can now be re-written as

$$\Omega = \sum_{i=1}^{m} \zeta_i.U_i$$

The coordinate ζ_i, $i \in 1, 2, ...m$, is called the ζ_i^{th} principal component. It represents the projection of ΔZ onto the basis vector U. The basis vectors, U_i, are the principal components of the training set. Once the subspace is constructed, recognition is done by projecting a centered probe image into the subspace, and the closest gallery image to the probe image is selected as the match.

Before applying PCA, the images are normalized and cropped resulting in an image size of $130x150$. Unwrapping the image results in a vector of size 19,500. PCA reduces this to a basis vector count of $m - 1$, where m is the number of images. PCA approaches to face recognition typically drop some vectors to form the face space. A small number from the beginning and a larger number from the end.

2.2 Distance Measure

A popular and simple classification technique in 2-D face recognition is the nearest neighbor classifier. An image in the probe set is assigned the label that is closest in the gallery set. Various distance measures have been evaluated in the realm of face recognition [12, 13]. For our experiments, we utilized the MahCosine distance metric [11]. Our initial experiments showed that MahCosine significantly outperformed the other distance measures, such as Euclidean or Mahalanobis distance measures.

The MahCosine measure is the cosine of the angle between the images after they have been transformed to the Mahalanobis space [11]. Formally, the MahCosine measure between the images i and j with projections a and b in the Mahalanobis space is computed as:

$$MahCosine(i,j) = cos(\theta_{ij}) = \frac{|a||b|cos(\theta_{ij})}{|a||b|}$$

2.3 Linear Discriminant Analysis (LDA)

LDA tries to achieve a projection that best discriminates between the the different subjects. PCA can be used to reduce the dimensionality before applying LDA. The *Fisherface* is constructed by defining a d dimensional subspace in the first d principal components [14]. Fisher's method finds the projecting vectors W, such that the basis vectors in W maximize the ratio of the determinant of the inter-class scatter matrix S_b and the determinant of the intra-class scatter matrix S_w.

$$W = argmax \frac{|W^T S_b W|}{|W^T S_w W|}$$

Let us define the number of subjects to be m and the number of images (samples) per subject available for training to be s_i, where i is the subject index. Then S_b and S_w can be defined as:

$$S_w = \sum_{i=1}^{m} \sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T$$

$$S_b = \sum_{i=1}^{m} s_i(\mu_i - \mu)(\mu_i - \mu)^T$$

and where μ_i is the mean of vector of samples belonging to the class (or subject) i, μ is the mean vector of all the samples. S_w may not be well estimated if the number of samples is too small.

3 Data Collection

The data for this paper was acquired from that available from the University of Notre Dame[1] [2]. The subjects participate in the acquisition at week intervals

[1] http://www.cse.nd.edu/~cvrl

Fig. 1. Sample images of a subject in the training data

over a period of time. For the experiments in this paper, images were captured either with two side lights on (LF) or with two side lights and a center light on (LM). In addition, subjects were asked to have either a neutral expression (FA) or a smile expression (FB). The nomenclature is as used by FERET [15]. The data was acquired during Spring 2002, Fall 2002, and Spring 2003. Figure 1 shows sample images of a subject captured under the four conditions.

We divided the data into training, gallery, and probe sets. To run multiple trials, we randomly selected 10 times 121 subjects from an available pool of 484 subjects. For each of the 10 random runs, we utilized the same probe and gallery sets. We report the mean and standard deviation in the rank-one recognition rates on the probe and gallery sets. Each selected subject had four images for *FA-LF*, *FB-LF*, *FA-LM* and *FB-LM*. The training set images were captured at the first acquisition session. Then we took all the subjects that had at least three acqusition sessions. The 2nd session of acquisition became the gallery set and the last session became the probe set. This gave us 381 subjects for testing. This ensured that only a small subset of subjects in the probe set was used in the training set, and moreover there was a time-lapse element introduced in testing. The probe sets, however, comprised of completely different images (even if of same subjects) than the training set. There was no overlap whatsoever in the images between the training and probe sets. We tried to mimic a setting that may be used in a 2-D face recognition system — the subjects in the gallery may not always be in the training data. Our probe sets always had different images than the training set.

4 Multiple Classifier System

The applications of multiple classifier systems are becoming relevant in face recognition. Beveridge et al. [14] used bagging without replacement; they randomly sampled without replacement from their population of 160 subjects. They showed that replicates produced by sampling with replacement can cause problems with the scoring methodology. We also sample without replacement, albeit from the four different images available for each subject, thus always having at least one representative of each subject in the training set. Lu and Jain [10] randomly sampled within each class (or subject) to construct a set of the LDA classifiers. However, they had 10 images for each class. Wang and Tang [16] recently used random subspaces to improve the performance of LDA classifier. Lemieux and Parizeau [17] utilized a multi-classifier architecture also, but they used four different classifiers: HMM, DCT, EigenFaces and EigenObjects. We randomly sample from the set of images for each subject, and construct a set of one nearest neighbor classifiers using the MahCosine measure. To establish the generality of the classifiers, we evaluate on a varying set of expressions, lighting conditions, and subjects.

We included LDA as a comparison benchmark, even though we had a smaller set of images per subject than is typically used with LDA. We compared four techniques. Please note that the number of basis vectors after the PCA was $m - 1$, where m is the number of images considered as part of each of the techniques. For example, if there are 121 images in the training set, then the basis vector count is 120. (1) Single *specialized* face space: This is the face space trained on a particular expression and lighting combination. In this type, a single face space was constructed for each of the FA-LF, FA-LM, FB-LF, and FB-LM. Thus, it is called *specialized* as each one is representative of a particular lighting and expression combination. (2) *Complete* face space (*All-1NN*): This is constructed using a 1-nearest neighbor classifier on a training set of size 484 (121x4), where each subject has four representative images in the training set. We concatenated all the four images available of a subject and constructed a single training set. The face-space was then constructed on all the concatenated 484 images. (3) *All-LDA* classifier using the four images per subject. Again, for LDA we considered all the available images for each subject, giving us 121 classes (or subjects) with four images (or examples) each. (4) *Ensemble*: We randomly sampled ($num =$) one, two, and three images (from the four images) per subject and constructed multiple classifiers. These will be referred to as *Ensemble-1, Ensemble-2, Ensemble-3*. While we varied the number of images for each ensemble, we maintained the same size of 50 classifiers. Each of the aforementioned Ensembles had a different number of (randomly selected) training set images for each subject. Given 121 subjects, *Ensemble-1* had 121 images; *Ensemble-2* had 242 images; and *Ensemble-3* had 363 images. We can summarize our procedure as follows:

1. For each k=1,2,..,K (where K is the number of classifiers, set as 50 in our case.)
   (a) Randomly select without replacement *num* images for each subject.
   (b) Construct a face space, X^k. As we mentioned before, the number of basis vectors after PCA is $m - 1$. Thus, the number of basis vectors for *Ensemble-1* is 120; for *Ensemble-2* is 241; and for *Ensemble-3* is 362.
   (c) For each probe image, find the closest gallery image with X^k using the MahCosine measure. Each individual classifier (k) assigns a distance measure to the probe image.
2. Aggregate the distances assigned to each probe image by each X^k.
3. Rank order the images and compute the rank-one recognition rate. This is the final rank-one recognition rate of the ensemble.

5 Experiments

To test the suitability of multiple classifiers in this domain, we compare to classifiers specialized for the lighting and expression condition, and to classifiers that use all the available training images. In the specialized comparisons, our probe and gallery sets were used separately for each expression and lighting combination (FA-LF, FA-LM, FB-LF, and FB-LM). Thus, each classifier was tested four times and performances are shown in Table 1. The rows are the training face spaces and the columns are the probe face spaces. Besides each specialized classifier, we also indicate the performance obtained by All-1NN and All-LDA in Table 2.

We did not tune the individual classifiers by dropping eigen vectors either from "front" or "back" of the face space. Typically, the first couple of vectors are assumed to carry the illumination variation [13]. One can also drop some low variance eigen vectors from behind to further improve the individual classifiers. However, to maintain the same performance benchmark across all classifiers we retained all the eigen vectors. As part of our future work we propose to utlize a validation set to tune the face spaces before applying them to the testing (probe) sets. This is similar to the wrapper techniques deployed in feature selection wherein a validation scheme is introduced for selecting the appropriate subset of features. If the face space is tuned on the probe set, it can lead to overestimated accuracies; a bias is introduced in developing the nearest neighbor classifier.

As evident by Table 1, the specialized classifier usually performs better if the testing set comes from the corresponding set of conditions. However, we notice that the classifiers trained on the LF lighting condition tend to perform better on the LM lighting condition (than the corresponding LF lighting condition). It could be that the LF classifiers are potentially overfitting on their space, thus leading to a reduced accuracy. In addition, there can be implicit illumination variations in the probe set that were unaccounted for. Similar results were noted by Chang et al. [18]. Moreover, making a complete face space of all the available images performs better than the specialized classifier across the board. It is perhaps not surprising that *All-1NN* does better than any *Specialized* classifier, across all 4 conditions, since it has more representatives for each subject under varying conditions. It is very much possible that the images captured under exactly the same controlled environment, still have an implicit element of illumination and pose variation. Having a diverse set of images in the training set can help in such scenario. However, we expect that as the number of images in the training set increases, the face-space can be overfit. This can require tuning to get rid of the low variance vectors, as we are more interested in distinguishing between subjects than between different variations of a subject. Surprisingly, LDA does not perform as well as 1-NN with all the images. LDA's performance can be hurt by small-sample size in high dimensional spaces [7, 9]. We only have four samples per class. Not having enough images per subject, we also run into the curse of dimensionality problem. One may drop eigen vectors to improve the performance of the LDA classifiers.

Table 1. The rank-one recognition rates and the standard deviation for the *Specialized* classifiers. The columns are the probe and gallery sets, and the rows are the training sets

	FA-LF	FA-LM	FB-LF	FB-LM
FA-LF (Specialized)	0.660 ± 0.025	0.712 ± 0.01	0.649 ± 0.012	0.666 ± 0.017
FA-LM (Specialized)	0.649 ± 0.018	0.716 ± 0.009	0.637 ± 0.017	0.66 ± 0.014
FB-LF (Specialized)	0.603 ± 0.014	0.659 ± 0.014	0.711 ± 0.012	0.725 ± 0.007
FB-LM (Specialized)	0.583 ± 0.017	0.648 ± 0.015	0.699 ± 0.01	0.729 ± 0.015

Table 2. The rank-one recognition rates and the standard deviations of the *Ensemble* methods, *All-1NN*, and *All-LDA* across the probe sets with varying lighting and expression combinations (columns). The entries in bold indicate the best performances

	FA-LF	FA-LM	FB-LF	FB-LM
Ensemble-1	0.653 ± 0.017	0.717 ± 0.012	0.714 ± 0.013	0.738 ± 0.01
Ensemble-2	0.697 ± 0.014	0.739 ± 0.011	0.748 ± 0.011	0.76 ± 0.006
Ensemble-3	**0.707 ± 0.009**	**0.743 ± 0.012**	**0.756 ± 0.009**	**0.769 ± 0.01**
All-1NN	0.69 ± 0.01	0.73 ± 0.015	0.734 ± 0.0137	0.754 ± 0.01
All-LDA	0.569 ± 0.024	0.615 ± 0.021	0.604 ± 0.026	0.6601 ± 0.022

Table 2 shows the results of different sample sizes on the four different probe sets. Due to a lack of space, we only include the performance obtained at the iteration where the performance plateaued for the ensemble methods. Typically, that was by the 10th iteration. We notice a consistent trend in the Table: *Ensemble-2* and *Ensemble-3* are fairly comparable and outperforming the other classifiers. Moreover, both *Ensemble-2* and *Ensemble-3* generalize very well across different sets of images, and exceed the accuracy obtained by both the *Specialized* and *All-1NN* classifiers. *Ensemble-3* is statistically significantly better at 95% than *All-1NN* for FA-LM and FB-LF. And both the *Ensemble-2* and *Ensemble-3* methods are statistically significantly better (at 95%) than the *Specialized* classifiers tested on their corresponding face spaces. Ensemble power with fewer images exceeds the single classifier with all the images. This is in agreement with what is typically observed by the MCS community.

We note that *Ensemble-1* is consistently lower than the classifiers with more images, but (almost) always above the *Specialized* case. The FA-LF classifier is slightly better than *Ensemble-1*. Constructing multiple classifiers of one image for each subject may not be representative enough for each of the subsequent spaces, as the training set size will be small. Typically, a learning curve can be plotted to identify the "critical" amount of data for different domains as applicable for a classifier. Also, we believe that randomly sampled images for each subject are adding the "diversity" element in the ensemble. Various studies have shown that different classifiers follow a learning curve that typically grows with the amount of data and eventually plateaus [19, 20, 21]. Skurichina et al. show that bagging with linear classifiers does not work for very small datasets or large datasets [20].

6 Conclusions

We empirically evaluated various training set sizes by randomly sampling from the available images for each subject. We showed that the multiple classifier system of randomly sampled images achieves good performances across the different probe sets. We constructed our training and testing such that the testing set not only contained images that were captured at a different time than the training set images but also a set of unique subjects. This maintained the difficulty of testing sets. Moreover, we tested the set of classifiers across four different expressions and lighting conditions combinations. The changing environment of the new images is a very important problem. We quote from a recent article from the Government Security Newsletter: "It turns out that uncooperative subjects, poor lighting and the difficulty of capturing comparable images often make it difficult for face recognition systems to achieve the accuracy that government officials might seek in large-scale anti-terrorism applications. [22]" Hence, we tried to imitate that setting in our paper. Our results are indeed interesting in this scenario, as we show that multiple classifier systems generalize better across different kinds of images, without any explicit assumption, thus mitigating the need of specialized and tuned classifiers.

As part of future work, we plan to extend our study to include increasing number of subjects and study the effect of that on the face space as we resample. We believe that as the number of subjects increase the face space constructed from all the images might overfit, requiring a tuning by dropping eigen vectors from the front or back. We also aim to introduce diversity metrics in our system to understand the behavior of different classifiers in the ensemble. However, we would like to utilize a separate validation set for any tuning to make the results as generalizable as possible. We also propose to utilize more images of a subject and implement a resampling framework for LDA as by Lu and Jain [10]. We believe that multiple classifier systems will be generally applicable to the recognition task due to an improved generalization on out-of-time and out-of-sample data.

Acknowledgments

This work is supported by National Science Foundation grant EIA 01-20839 and Department of Justice grant 2004-DD-BX-1224.

References

1. "Face recognition grand challenge and the biometrics experimentation environment." available at http://bee-biometrics.org.
2. P. J. Flynn, K. W. Bowyer, and P. J. Phillips, "Assessment of time dependency in face recognition: An initial study," in *International Conference on Audio and Video Based Biometric Person Authentication*, pp. 44–51, 2003.
3. R. Chellappa, C. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proceedings of hte IEEE*, vol. 83, no. 5, pp. 705 – 740, 1995.

4. A. Samal and P. Iyengar, “Automatic recognition and analysis of human faces and facial expressions: A survey,” *Pattern Recognition*, vol. 25, no. 1, pp. 65 – 77, 1992.
5. M. Turk and A. Pentland, “Eigenfaces for recognition,” *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.
6. G. Shakhnarovich and G. Moghaddam, *Handbook of Face Recognition*, ch. Face recognition in subspaces. Springer-Verlag, 2004.
7. K. Fukunaga, *Introduction to Statistical Pattern Recognition.* New York: Academic Press, 1990.
8. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification.* New York: Wiley, 2nd ed., 2000.
9. A. M. Martinez and A. C. Kak, “Pca versus lda,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001.
10. X. Lu and A. K. Jain, “Resampling for face recognition,” in *International Conference on Audio and Video Based Biometric Person Authentication*, pp. 869 – 877, 2003.
11. D. Beveridge and B. Draper, “Evaluation of face recognition algorithms (release version 4.0).” available at http://www.cs.colostate.edu/evalfacerec/index.html.
12. V. Perlibakas, “Distance measures for pca-based face recognition,” *Pattern Recognition Letters*, vol. 25, no. 6, pp. 711–724, 2004.
13. W. Yambor, B. Draper, and R. Beveridge, “Analyzing PCA-based face recognition algorithms: Eigenvector selection and distance measures,” July 2000.
14. J. R. Beveridge, K. She, B. Draper, and G. Givens, “A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 535 – 542, 2001.
15. P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, “The FERET evaluation methodology for face-recognition algorithms,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.
16. X. Wang and X. Tang, “Random sampling LDA for face recognition,” in *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 259–265, 2004.
17. A. Lemieux and M. Parizeau, “Flexible multi-classifier architecture for face recognition systems,” *Vision Interface*, 2003.
18. K. Chang, K. W. Bowyer, and P. Flynn, “An evaluation of multi-modal 2d+3d face biometrics,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.
19. F. Provost, D. Jensen, and T. Oates, “Efficient progressive sampling,” in *Fifth International of Knowledge Discovery and Databases*, pp. 23–32, 1999.
20. M. Skurichina, L. Kuncheva, and R. P. W. Duin, “Bagging and boosting for the nearest mean classifier: Effects of sample size on diversity and accuracy,” in *Third International Workshop on Multiple Classifier Systems*, pp. 62 – 71, 2002.
21. N. V. Chawla, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, “Learning ensembles from bites: A scalable and accurate approach,” *Journal of Machine Learning Research*, vol. 5, pp. 421 – 451, 2004.
22. “GSN Perspectives –Grand challenge sets critical biometric face-off.” available at http://www.gsnmagazine.com/dec_04/grand_challenge.html.