Joint Boosting Feature Selection for Robust Face Recognition

Rong Xiao[1] Wujun Li[2] Yuandong Tian[3] Xiaoou Tang[1]
[1]Microsoft Research Asia, Beijing, 100080, P. R. China
[2]National Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, P. R. China
[3]Department of Computer Science and Engineering,
Shanghai Jiaotong University, Shanghai, 200030, P. R. China
rxiao@microsoft.com li_wujun@hotmail.com tydsh@hotmail.com xtang@microsoft.com

Abstract

A fundamental challenge in face recognition lies in determining what facial features are important for the identification of faces. In this paper, a novel face recognition framework is proposed to address this problem. In our framework, 3D face models are used to synthesize a huge database of realistic face images which covers wide appearance variations of faces due to various pose, illumination, and expression changes. A novel feature selection algorithm which we call Joint Boosting is developed to extract discriminative face features using this massive database. The major contributions of this paper are: (1) With the help of 3D face models, a massive database of realistic virtual face images is generated to achieve robust feature selection; (2)Because the huge database covers a wide range of face variations, our feature selection procedure only needs to be trained once, and the selected feature set can be generalized to other face database without re-training; (3) We propose a new learning algorithm, Joint Boosting Algorithm, which is effective and efficient in learning directly from a massive database without having to convert face images to intra-personal and extra-personal difference images. This property is important for applying our algorithm to other general pattern recognition problems. Experimental results show that our method significantly improves recognition performance.

1. Introduction

Robust face recognition under uncontrolled environment is a challenging pattern recognition problem. The main difficulty arises from the complicated appearance variations of faces due to varying lighting/illumination conditions, different head poses, and different facial expressions. In most publicly available face databases, only a small number of samples for each subject are available for training. Unfortunately, these small sets of samples cannot capture all the possible appearance variations of each individual face. This problem greatly limits the generalization ability of most face recognition methods.

To solve this problem, two types of solutions have been proposed. One is using virtual samples. Hu et al. [6] proposed an analysis-by-synthesis face recognition framework to address this problem. In the framework, they automatically reconstructed a 3D model for each input image, then based on this model virtual face images with different poses, illumination, and expression are synthesized. Finally, face recognition is performed using virtual face images. This framework significantly improved the recognition performance; however, there still exist three limitations: (1) fully automatic reconstruction of synthetic face images tends to bring in extra noise to the synthesized data; (2) the reconstruction and synthesis steps are required for each training subject, which significantly increases the computational cost; (3) the method cannot handle extremely large synthetic dataset thus still cannot cover the entire face variation space.

Another solution is using feature extraction and feature selection methods to improve the generalization ability by reducing the dimensionality of face image in the feature space. Laurenz et al. [8] proposed Elastic Bunch Graph Matching (EBGM) algorithm. With a bunch graph, a set of local Gabor features are selected to perform face recognition. In [7], a Jensen-Shannon boosting (JSBoost) algorithm is proposed to select the most discriminative local binary pattern (LBP) features. To address the imbalance problem between the amount of intra-personal samples and that of extra-personal samples, Yang et al. [21] proposed a resampling scheme for AdaBoost to select the most discriminative Gabor features. Compared with the conventional feature selection approaches, Boosting-based algorithms show

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

superior performance in learning from a massive data set with continuous feature values.

Most existing Boosting-based algorithms use difference images as in Bayesian face recognition [12]. By this way, a multi-class recognition problem is converted into a binary classification problem of the intra-personal and extra-personal spaces. By combining the feature selection with the classification procedure, Boosting-based algorithms have achieved the state-of-the-art performance in most databases including the FERET database [13]. However, since images of different face databases are usually captured under different environment, these algorithms tend to over-fit the environment of the training database. This results in the performance drop while the trained classifier is generalized to another face database without retraining. Another limitation is the computational cost. To re-train a boosting classifier for each face database requires a long training time. Furthermore, there exists a problem with the size of training dataset. For instance, for a face database with n samples, after converting the original face images into intra-personal and extra-personal difference images, the scale of the training set will be increased to $O(n^2)$. To cover all the variations in the training set, all the difference images of the training set need to be computed for training, This dramatically increases both computational time and memory space cost.

In this paper, we propose a novel face recognition framework called Joint boosting face recognition. In this framework, we adopt the analysis-by-synthesis strategy used in [6]. Using the 3D face models from USF Human ID 3-D database [1], we synthesize more than 600 realistic virtual face images with different poses and illumination conditions for each individual face. A novel Joint Boosting algorithm is proposed to efficiently select the most discriminative and robust features for face identification from the large scale database. Unlike conventional boosting algorithms, our Joint Boosting algorithm explicitly exploits multi-class (each subject corresponding to a class) information, and selects the most informative features that can separate each class from all the others.

The basic idea of Joint Boosting algorithm is built upon the sharing feature idea [16], which is used to learn sharing features across multiple object classes (classifiers) for the purpose of multiple object detection. The advantages of our approach can be summarized as follows: (1)The feature selection need to be performed only once using the synthetic face database, and the selected features can be generalized to recognize faces of other face databases; (2) Under the Joint Boosting framework, feature selection is applied directly on the feature space. There is no need to compute difference images, which significantly reduce the training cost. Our algorithm is very efficient thus can handle extremely large training dataset (over 60k images in our experiment); (3) This framework avoids the 3D reconstruction for gallery images. It improves recognition rate without extra computational cost; (4) Finally the proposed framework can be easily combined with most existing recognition algorithms and significantly improve the recognition rate.

2. Learning with Boosting

AdaBoost [14] is a well-known large margin classifier developed recently. The main methodology of this algorithm is ”boosting”, which combines the performance of many ”weak” classifiers to form a powerful ”committee” classifier. During the training stage, training samples are re-weighted according to the training error, and the weak classifiers trained later are forced to focus on the harder examples with higher weights.

In [5], the boosting procedure is formulated to fit an additive model $F(x) = \sum_{t=1}^{T} f_t(x)$, where $f_t(x)$ is a weak leaner.

By using different loss function and optimization strategies, several variants of Boosting algorithms were proposed, such as RealBoost, GentleBoost and LogitBoost[5].

The original Boosting algorithm is designed for binary class learning problems. Several algorithms, such as AdaBoost.MH, AdaBoost.MO and AdaBoost.MR [14], have been proposed to extend the original Booting algorithm to solve multi-class problems. To reduce the computational complexity of multi-class learning, Torralba et al. [16] proposed a novel boosting algorithm to exploit the common features that can be shared across different classes. However, the training cost of this algorithm is huge. The original algorithm needs to search all $2^C - 1$ possible sharing patterns. An improved algorithm is also proposed in [16] to reduce the computational cost to $C(C+1)/2$ with suboptimal performance.

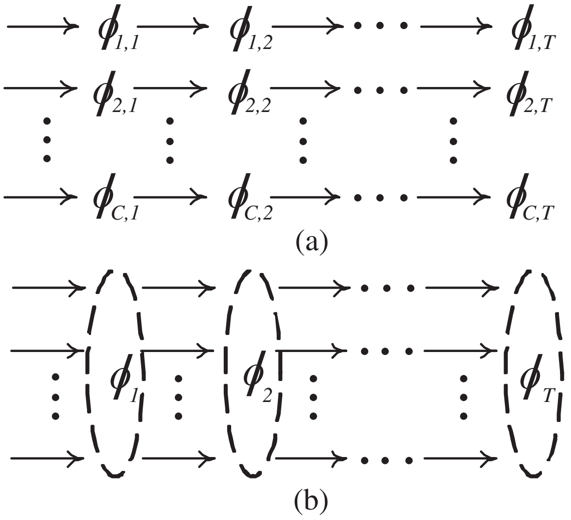
By building the analogy between weak classifiers and features, the ”boosting” procedure can also be interpreted as a greedy feature selection process [18].Different feature selection criteria lead to different boosting procedures. In [20],Bhattacharyya Distance is used in the RealBoost framework to do feature selection for face detection. In [9], KLBoost algorithm is proposed based on Kullback-Leibler divergence to select features for face detection. In [7], JSBoost algorithm is proposed based on symmetric Jensen-Shannon divergence (SJS), which is defined as follows:

$$SJS(r,s) = \int \Bigg(r(x) \log \frac{2r(x)}{r(x)+s(x)} + s(x) \log \frac{2s(x)}{r(x)+s(x)} \Bigg) dx \qquad (1)$$

where $r(x)$ and $s(x)$ are two distribution functions. The experimental results in [7] also show that JSBoost performs slightly better than KLBoost, RealBoost and GentleBoost for face recognition.

3. Joint Boosting for face recognition

Assume the sample points are given as $\{x_i, y_i\}_{i=1}^{N}$ where $x_i \in R^d$ is a training sample, $y_i \in L = \{1, 2, \ldots, C\}$ is a class label, and each individual $c \in L$ has n_c samples. Instead of directly processing the raw data, we map the input image into the feature space with projection functions $\phi_j \in \Phi : R^d \to R, j = 1, 2, , M$. Our goal is to select a subset of discriminative features $\{\phi_j(x)\}$ which effectively separate each class from all the others.



(a)

(b)

Figure 1. Feature selection procedure. (a) Multi Boosting. (b) Joint Boosting

Face recognition problem is essentially a multi-class classification problem, where each class contains the images of one individual. Face recognition problem can be straightforwardly formulated as multiple one-versus-the-rest binary classification problems, which can be formulated as a greedy feature selection process by fitting the following C additive models $F^c(x) = \sum_{t=1}^{T} f_{c,t}(x|w_t^c)$, where $c \in L$, $\phi_{c,t}$ is the feature for classifier $f_{c,t}$.

As shown in Fig. 1(a), each row represents one boosting feature selection procedure for each person based on the one-versus-the-rest strategy. The problem with this formulation is that it will generate an overwhelming number of features given the large number of training persons. In addition, the model tends to over fit to each individual, thus cannot be generalized to other datasets.

As faces have the same facial structure, the selected dominant features for different people may share the same properties. In order to capture both the common properties and the individual characteristics using only a manageable small set of features, we propose a new joint boosting method. The key assumption is that we can find a set of optimal features which are the same for all individuals, i.e. we assume:

$$\phi_{1,t} = \phi_{2,t} = \cdots = \phi_{c,t} = \phi_t \qquad (2)$$

Based on this assumption, we propose a Joint Boosting feature selection algorithm for face recognition, as shown in Fig. 1(b). Next we describe how to optimally find such a set of features.

Suppose $w(x)$ is the weight of a sample x. For any class c, the weighted distribution of positive samples on feature $\phi_j(x)$ is defined as:

$$h_j^{c,+}(x|w) = p(\phi_j(x)|y = c) * w(x|y = c)/W_j^{c,+}, \quad (3)$$

and that of the negative samples is:

$$h_j^{c,-}(x|w) = p(\phi_j(x)|y \neq c) * w(x|y \neq c)/W_j^{c,-}, \quad (4)$$

where $w = w(x)$, $W_j^{c,+}$ and $W_j^{c,-}$ are the normalization factors, $h_j^{c,+}(x|w)$ and $h_j^{c,-}(x|w)$ are distributions. Therefore the weak classifier of feature ϕ_j on class c is defined as:

$$f_j^c(x|w) = f^c(\phi_j(x)|w) = \frac{1}{2} \log \frac{h_j^{c,+}(x|w) + \epsilon}{h_j^{c,-}(x|w) + \epsilon} \qquad (5)$$

To evaluate the performance of feature ϕ_j in the c^{th} model, we define the cost function $G_j^c(x)$ on the weak classifier $f_j^c(x|w^c)$ as:

$$\begin{aligned} G_j^c(x) = & \quad \int_{x \in X} G(f_j^c(x|w^c)|y, w^c)dx, \\ = & \int_{x \in X} g(h_j^{c,+}(x|w^c), h_j^{c,-}(x|w^c))dx, \end{aligned} \qquad (6)$$

where function $g(r(x), s(x))$ is the measure of the classification error of logistic classifier defined on the weighted distributions $r(x)$ and $s(x)$.

Therefore, for each model c, the best feature $\phi_{c,t}$ for the t^{th}-step is selected by

$$j_{c,t} = argmin_j G_{t,j}^c(x)$$

All the feature selection procedures for the C different boosting models $F_c(x)$ can be combined into a joint procedure, which is called Joint Boosting. The best feature ϕ_t for the t^{th}-step of Joint Model is selected by:

$$t = argmin_j \sum_{c=1}^{C} G_{t,j}^c(x) \qquad (7)$$

Finally the proposed Joint Boosting algorithm is shown in following algorithm.

3.1. Cost functions for feature selection

In this paper, we propose to use the measure of Bayesian error for (6). The main reason to select this cost function

Algorithm 1 Joint boosting for feature selection

```
• Initialize the weights
  For $c = 1, 2, \ldots, C$ and $i = 1, 2, \ldots, N$
      $w^c_{1,i} = \begin{cases} 1/n_c & \text{if } y_i = c, \\ 1/(N - n_c) & \text{otherwise.} \end{cases}$
• For $t = 1, 2, \ldots, T$
    1. Normalize the weights $w^c_{t,i}$, and make $w^c_t$ is a probability distribution.
           $w^c_{t,i} \leftarrow w^c_{t,i} / \left(\sum_{i=1}^{N} w^c_{t,i}\right)$
    2. For $j = 1, 2, \ldots, M$ and $c = 1, 2, \ldots, C$
       Train the weak classifier $f^c_j(x|w^c_t)$ on feature $\phi_j(x)$, using (5) and evaluate the cost $G^c_{t,j}(x)$ using (6).
    3. Find the best feature $\phi_t$ using (7).
    4. For $c = 1, 2, \ldots, C$ and $i = 1, 2, \ldots, N$
       Update the weight:
           $w^c_{t+1,j} = w^c_{t,j} \exp(-\lambda^c_i f^c_t(x_i|w^c_t))$        (8)
       where $\lambda^c_i$=1, if $y_i = c$, otherwise $\lambda^c_i$=-1.
• The final strong classifiers are
      $F^c(x) = sign\left(\sum_{t=1}^{T} f^c_t(x|w^c_t)\right).$        (9)
  and the selected feature set is $\{\phi_t : t = 1, 2, \ldots, T\}$.
```

is because of its theoretical elegance and the low computational cost. For a binary classification problem for the classes ω_1 and ω_2 , the Bayesian error is defined as:

$$\begin{aligned} R = \quad & p(error) = \int_{x \in X} p(error|x)dx \\ = \quad & \int_{x \in X} min(p(x|\omega_1), p(x|\omega_2))dx, \end{aligned} \tag{10}$$

Substituting the probability distributions $p(x|\omega_1)$ and $p(x|\omega_2)$ with weighted distribution defined in (3) and (4), we have:

$$R^c_{t,j}(x) = \int_{x \in X} min(h^{c,+}_j(x|w^c_t), h^{c,-}_j(x|w^c_t))dx \tag{11}$$

Therefore, based on (11), Bayesian cost for (6) is defined as $BE(r, s) = \int_{x \in X} min(r(x), s(x))dx$.

Many different functions can be used for the cost measure $g(r, s)$ in (6), such as Kullback-Leibler divergence, Jensen-Shannon divergence and Bhattacharyya Distance. For Jensen-Shannon divergence and Kullback-Leibler divergence, we maximize (7) instead of minimizing it. The experimental results in [7] show that the Jensen-Shannon measure is better than other measures. However, the mathematical reason for choosing the Jensen-Shannon measure is not clear.

3.2. Look-up-table (LUT) weak classifier

Evaluating (5) and (11) directly is not straightforward. In this paper, we use $k-$bins histograms to discretize the distribution of the weighted distributions by partitioning the region $[min(\phi_i(x)), max(\phi_i(x))]$ into several disjoint bins $X^1_j, X^2_j, \ldots, X^k_j$. We define:

$$h^{c,+}_j(k) = \sum_{\phi_j(x_i) \in X^k_j \wedge y_i = c} (w_i / W^{c,+}_j) \tag{12}$$

$$h^{c,-}_j(k) = \sum_{\phi_j(x_i) \in X^k_j \wedge y_i \neq c} (w_i / W^{c,-}_j) \tag{13}$$

where $k \in \{1, 2, \ldots, K\}$, $W^{c,+}_j$ and $W^{c,-}_j$ are the normalization factors to make $h^{c,+}_j(k)$ and $h^{c,-}_j(k)$ distributions.

According to the definition of (12), we have:

$$h^{c,+}_j(x|w^c_t) \approx h^{c,+}_j(k), \text{when } \phi_j(x) \in X^k,$$

with equality when $K \to \infty$.

Therefore, $h^{c,+}_j(k)$ can be regarded as the discrete version of distribution $h^{c,+}_j(x|w^c)$. Similarly, $h^{c,-}_j(k)$ becomes the discrete version of distribution $h^{c,+}_j(x|w^c)$.

Substituting the LUT (look-up-table) functions defined in (12) and (13) to (5), the weak classifier can be defined as the LUT function:

$$f^c_j(k) = \frac{1}{2} log \left(\frac{\sum_{\phi_j(x_i) \in X^k_j \wedge y_i = c} (w_i / W^{c,+}_j) + \epsilon}{\sum_{\phi_j(x_i) \in X^k_j \wedge y_i \neq c} (w_i / W^{c,-}_j) + \epsilon} \right)$$

The discrete version of (11) is defined as:

$$\begin{aligned} R(f^c_j(k)) = \quad & BE(h^{c,+}_j(k), h^{c,-}_j(k)) \\ = \quad & \textstyle\sum_{k=1}^{K} \min(h^{c,+}_j(k), h^{c,-}_j(k)) \end{aligned} \tag{14}$$

Similarly, based on (1) , we have the discrete version of symmetric Jensen-Shannon divergence for weak classifier $f^c_j(k)$:

$$\begin{aligned} SJS(f^c_j(k)) = \sum_{k=1}^{K} \Bigg(& h^{c,+}_j(k) \frac{2h^{c,+}_j(k)}{h^{c,+}_j(k) + h^{c,-}_j(k)} \\ & + h^{c,-}_j(k) \frac{2h^{c,-}_j(k)}{h^{c,+}_j(k) + h^{c,-}_j(k)} \Bigg) \end{aligned} \tag{15}$$

4. Experimental results

We call the boosting method which uses the Bayesian error measure *BayesianBoost*, and the corresponding joint boosting method Joint *BayesianBoost*. Analogically, we call the Joint Boosting method which uses the Jensen-Shannon divergence *Joint JSBoost*. Considering the good performance of the Gabor features in face recognition [10][3][21][15][8], we use the boosting methods to select the most discriminative Gabor features. In our experiment, the Gabor parameters are set as described in [8]. After the convolution, 40 complex values are calculated for each pixel in the face image. Because the phase information of the transformation is time-varying, as in general Gabor feature-based applications, only its magnitudes are used to form the final face representation. Therefore, for a specific face image, the Gabor feature vector is 40 times of the original image size. We first use the boosting methods mentioned above to reduce the feature dimensionality. Later the selected Gabor features can be further processed by any kind of classification method, such as Principle Component Analysis (PCA) [17], Linear Discriminant Analysis (LDA) [4] and Bayesian Intra/Extrapersonal Classifier (BIC) [12]. To verify the performance of the joint boosting method, other feature selection methods, such as Chain AdaBoost learning method [21][15], are compared with our methods.

4.1. Data preparation

To model the appearance variations caused by varying lighting/illumination conditions and different head poses for each individual, a face database must contain enough training samples for each subject. However, most public databases including the FERET database [13] do not meet this requirement. In this paper, we use the 3D face models from USF Human ID 3-D database [1] which contains 100 laser-scanned heads of different individuals to synthesize a large number of images under various poses and lighting conditions for each individual. Two sample 3D models are shown in Fig. 2.

In the synthesized database, we simulate two kinds of pose rotations, yaw and pitch. For both rotations, we rotate the 3D face model from $-10°$ to $10°$ at the step of $5°$. Therefore, we have 25 different poses in total. For each pose, we use two light sources that are projected onto the 3D models. One is a fixed environment light; and the other is a dynamic spot light source. By changing the lighting directions of the spot light, 25 different lighting/illumination conditions are simulated. Therefore, for each 3D model, we obtain 625 face images under different poses and illumination conditions. Some examples of the synthetic images are shown in Fig. 3.

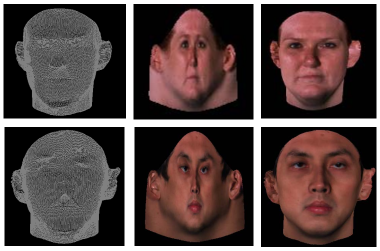


Figure 2. 3D face Models. Each row corresponds to a subject. The first column is the 3D meshes; the second column is the texture; and the third column is the 3D models with texture.

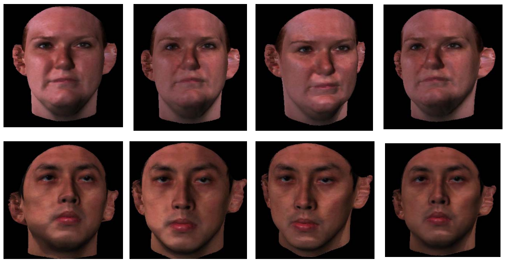


Figure 3. Synthetic images from 3D models

In our experiments, the face images are rescaled to $65*75$ pixels and normalized according to the preprocessing methods described in [2]. Some normalized examples from the database are shown in Fig. 4.

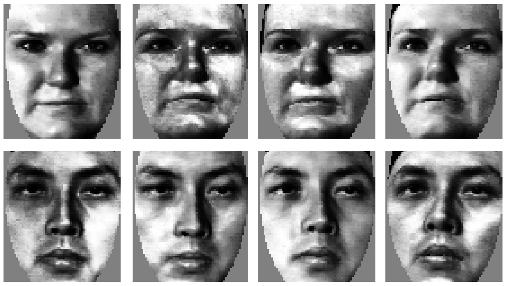


Figure 4. Normalized face images

4.2. Feature selection on the synthetic database

We have prepared $625 * 100$ face images to train the feature selection procedures mentioned above. In order to compare our method with JSBoost [7] and Chain AdaBoost [15], we randomly select 6 images from the 625 images for every one of the 100 subjects, thus we have 1500 intra-personal difference images and 178200 extra-personal difference images for training. For JSBoost, all the 1500 intra-personal difference images are used as positive samples and

100000 extra-personal difference images are randomly selected as negative samples to form the training set. For the Chain AdaBoost method, because it consists of several layers (in our experiment, 150 features are selected for every layer), all the 1500 positive samples and 3000 negative samples are selected by the method described in [15] for each layer.

For the feature extraction method, we select Gabor filter which is widely used in many face recognition algorithms with good recognition performance. The first four Gabor features selected by Joint JSBoost are shown in Fig. 5. We can observe that most discriminative features are located around the eyes and eyebrows. To evaluate the discriminative and generalization ability of our proposed algorithm, the recognition ratios are compared among BayesianBoost, Joint BayesianBoost, JSBoost, Joint JSBoost, and Chain AdaBoost. The experimentals are conducted on the FERET database [13] and the XM2VTS database [11].

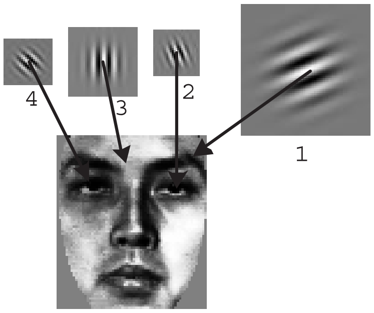


Figure 5. The first four features selected by the Joint JSBoost

4.3. Results on the FERET face database

We choose the same database configuration as that used in [15], which consists of the training set with 1002 images of 429 subjects, the gallery set FA with 1196 images of 1196 subjects, and the probe set FB with 1195 images of 1195 subjects. All images are rescaled to $65 * 75$ pixels and normalized by the preprocessing methods described in [2].

In these experiments, the feature selection algorithms are trained on either synthetic face database or the FERET training set, and the performance of face recognition is evaluated using the FERET Probe set FB. The City Block (L1) distance metric [2] and the Dual-Space LDA (DSLDA) algorithm [19] are used to achieve face recognition. The reasons why we choose these algorithms are: (1) L1 distance is simple and free from parameter tuning, and (2) DSLDA is known to be one of the best recognition algorithms.

In the first experiment, we compare the performance variation of the methods we mentioned above by changing the number of features used. The comparison of the recognition performance is shown in Fig.6. We can observe that the Joint boosting methods outperform the other methods in any dimensionality. Furthermore, the results show that when the number of selected features exceeds 1000, the recognition accuracy does not significantly improve for all these methods tested. Therefore, in the next experiment, we select 1000 features to make fair comparisons for all these methods using the DSLDA algorithm.

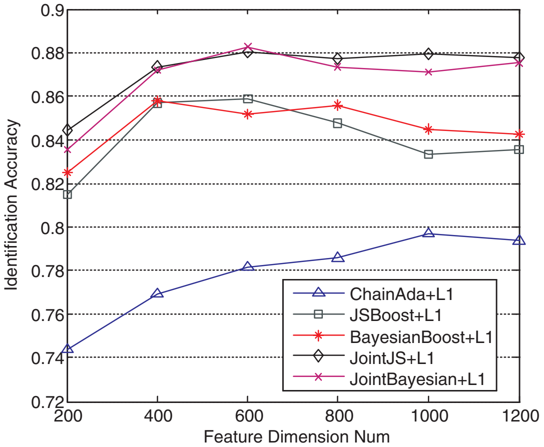


Figure 6. Face recognition performance on the FERET FB set with L1 distance. The two best results are achieved by our proposed methods: JointJS and JointBayesian.

In the second experiment, we use the DSLDA algorithm to compare the recognition performance of these feature selection algorithms. The accumulative matching scores of these feature selection algorithms are plotted in Fig.7. We can see that Joint JSBoost and Joint BayesianBoost are better than their corresponding binary versions. Compared with the Joint JSBoost algorithm, in both Fig.6 and Fig.7, the Joint BayesianBoost algorithm is slightly better. Also, we can see that the performance of the Chain Adaboost algorithm is not as good as the result reported in [15]. The main reason for this difference is that we use the synthetic face database for Boosting training instead of directly using the FERET training set.

In the third experiment, we use the FERET training set to train the Chain AdaBoost algorithm, and keep using the synthetic database for our Joint Boosting algorithms. As shown in Table 1 , the experimental results show that our Joint Boosting algorithms out-perform any other method. Among all Gabor-based methods, both Joint JSBoost and Joint BayesianBoost achieve the-state-of-the-art performance on the FETET FB dataset. This is to our knowledge, one of the best results on the FB data. It should

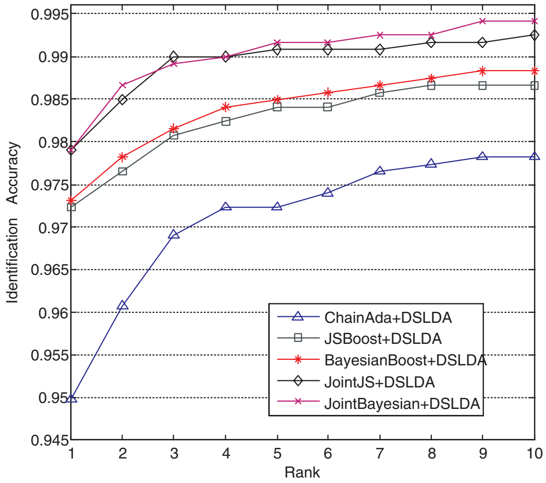


Figure 7. The accumulative matching scores of the proposed method when testing on the FERET FB set with DSLDA. The two best results are achieved by our proposed methods: JointJS and JointBayesian.

Table 1. Recognition performance on the FERET FB set with DSLDA (”ChainAda” is ”Chain AdaBoost”, ”JointJS” is ”Joint JSBoost” and ”JointBayesian” is ”Joint BayesianBoost”). Our new methods JointJS and JointBayesian achieve the best results. (FST = Feature selection training)

Method	FST-set	Error (%)
EBGM[8]	n/a	10.2
GFC[10]	n/a	3.7
AGFC[15]	FERET	2.8
ChainAda+DSLDA	Synthetic	5.0
ChainAda+DSLDA	FETET	2.4
JSBoost+DSLDA	Synthetic	2.8
BayesianBoost+DSLDA	Synthetic	2.7
JointJS+DSLDA	Synthetic	2.1
JointBayesian+DSLDA	Synthetic	2.1

also be the best result for using training data outside the FERET dataset. By comparing the results in Table 1 with the experimental results shown in Fig.6 and Fig.7, we can find that the performance of Chain AdaBoost is significantly improved in this experiment. This shows that the Chain Adaboost method is very effective when the training set and testing set satisfy the same distribution, while it shows poor performance when the testing set does not have the same distribution as the training set.

Table 2. Recognition performance on the XM2VTS database. Our new methods JointJS and JointBayesian achieve the best results. (FSTS = Feature selection training)

Method	FST-set	Error (%)
PCA	n/a	18.1
Bayesian ML	n/a	11.2
LDA	n/a	5.8
DSLDA	n/a	4.7
ChainAda + DSLDA	Synthetic	2.0
JSBoost + DSLDA	Synthetic	1.7
BayesianBoost+DSLDA	Synthetic	1.4
JointJS + DSLDA	Synthetic	1.0
JointBayesian+DSLDA	Synthetic	1.0

4.4. Results on the XM2VTS face database

In the fourth experiment we used the XM2VTS face database [11] for testing. In this database, there are 295 persons. For each person, four face images are captured from four different sessions. We select $295 * 3$ images of the 295 persons from the last three sessions for the training set. The gallery set is 295 images from the fourth session. And the probe set is 295 images from the first session. We use the Dual-Space LDA [19] to classify the 1000 features selected by the feature selection algorithms mentioned above, and compare the results with some conventional holistic feature based face recognition approaches, such as Principle Component Analysis (PCA) [17], Linear Discriminant Analysis (LDA) [4], and Bayesian Intra/Extrapersonal Classifier (BIC) [12], and DSLDA[19]. As shown in Table 2, the results clearly demonstrate the superiority of the joint boosting selection algorithms.

5. Conclusion

This paper has investigated the feature selection problem for face recognition and proposed a novel Joint Boosting algorithm for efficiently learning from a massive database. By adopting the analysis-by-synthesis framework, we construct a large database to model the face appearance variations under different lighting and pose conditions. Based on this dataset, a Joint Boosting algorithm is proposed to achieve the feature selection. The experimental results show that Joint Boosting algorithm is superior in extracting discriminative and robust features for face recognition than existing state of the art Binary Boosting algorithms. We observe that Bayesian Error is a good feature selection measure. Compared with the Jenson-Shannon divergence, Bayesian Error is simple and elegant with much lower computational cost.

Our future work will focus on two directions. One is how to synthesize more realistic virtual face images with more complicated lighting/illumination conditions and facial expression. Another one is to integrate the re-sample scheme used by Chain AdaBoost[15], which shows promising effect in learning for unbalanced dataset.

References

[1] V. BlanZ and T. Vetter. A morphable model for the synthesis of 3d-faces. In *Proc. of ACM SIGGRAPH*, pages 187–194, 1999.

[2] D. Bolme, R. Beveridge, M. Teixeira, and B. Draper. The csu face identification evaluation system: Its purpose, features and structure. In *Proc. of International Conf. on Vision Systems*, pages 304–311, 2003.

[3] J. Daugman. Complete discrete 2-d gabor transform by neural networks for image analysis and compression. In *IEEE Trans. On Acoustics, Speech and Signal Processing*, volume 36, pages 1169–1179, 1988.

[4] K. Etemad and R. Chellappa. Discriminant analysis for recognition of human face images. In *Journal of the Optical Society of America*, volume 14, pages 1724–1733, 1997.

[5] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. In *The Annals of Statistics*, volume 28, pages 337–374, 2000.

[6] Y. X. Hu, D. L. Jiang, S. C. Yan, L. Zhang, and H. Zhang. Automatic 3d reconstruction for face recognition. In *Proc. of the 6th IEEE International Conf. on Automatic Face and Gesture Recognition*, pages 843–848, 2004.

[7] X. Huang, S. Li, and Y. Wang. Jensen-shannon boosting learning for object recognition. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 144–149, 2005.

[8] W. Laurenz, F. Jean-Marc, K. Norbert, and M. Christoph. Face recognition by elastic bunch graph matching. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, volume 19, pages 775–779, 1997.

[9] C. Liu and H. Y. Shum. Kullback-leibler boosting. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 407–411, 2003.

[10] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. In *IEEE Trans. on Image Processing*, volume 11, pages 467–476, 2002.

[11] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Proc. of International Conf. on Audio- and Video-Based Person Authentication*, pages 72–77, 1999.

[12] B. Moghaddam, C. Nastar, and A.Pentland. A bayesian similarity measure for direct image matching. In *Proc. of the 13th International Conf. on Pattern Recognition*, volume 2, pages 350–358, 1996.

[13] P. J. Phillips, H. Moon, and et al. The feret evaluation methodology for face-recognition algorithms. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, volume 22, pages 1090–1104, 2000.

[14] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. In *Proc. of the 7th Annual Conf. on Computational Learning Theory*, pages 80–91, 1998.

[15] S. Shan, P. Yang, X. Chen, and W. Gao. Adaboost gabor fisher classifier for face recognition. In *Proc. of IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pages 278–291, 2005.

[16] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 762–769, 2004.

[17] M. Turk and A. Pentland. Eigenfaces for recognition. In *Journal of Cognitive Neuroscience*, volume 3, pages 71–86, 1991.

[18] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 511–518, 2001.

[19] X. Wang and X. Tang. Dual-space linear discriminant analysis for face recognition. In *Proc. of IEEE International Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 564–569, 2004.

[20] B. Wu, H. Z. Ai, and C. Huang. Lut-based adaboost for gender classification. In *Proc. of the 4th International Conf. on Audio- and Video-Based Biometric Person Authentication*, pages 104–110, 2003.

[21] P. Yang, S. Shan, W. Gao, S. Z. Li, and D. Zhang. Face recognition using ada-boosted gabor features. In *Proc. of the 6th IEEE International Conf. on Automatic Face and Gesture Recognition*, pages 356–361, 2004.