A Benchmark Study of Large-scale Unconstrained Face Recognition

Shengcai Liao, Zhen Lei, Dong Yi, and Stan Z. Li

Center for Biometrics and Security Research & National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun East Road, Beijing 100190, China
{scliao,zlei,dong.yi,szli}@nlpr.ia.ac.cn

Abstract

Many efforts have been made in recent years to tackle the unconstrained face recognition challenge. For the benchmark of this challenge, the Labeled Faces in the Wild (LFW) database has been widely used. However, the standard LFW protocol is very limited, with only 3,000 genuine and 3,000 impostor matches for classification. Today a 97% accuracy can be achieved with this benchmark, remaining a very limited room for algorithm development. However, we argue that this accuracy may be too optimistic because the underlying false accept rate may still be high (e.g. 3%). Furthermore, performance evaluation at low FARs is not statistically sound by the standard protocol due to the limited number of impostor matches. Thereby we develop a new benchmark protocol to fully exploit all the 13,233 LFW face images for large-scale unconstrained face recognition evaluation under both verification and open-set identification scenarios, with a focus at low FARs. Based on the new benchmark, we evaluate 21 face recognition approaches by combining 3 kinds of features and 7 learning algorithms. The benchmark results show that the best algorithm achieves 41.66% verification rates at FAR=0.1%, and 18.07% open-set identification rates at rank 1 and FAR=1%. Accordingly we conclude that the large-scale unconstrained face recognition problem is still largely unresolved, thus further attention and effort is needed in developing effective feature representations and learning algorithms. We thereby release a benchmark tool to advance research in this field.

1. Introduction

Due to a great value both in pattern recognition research and practical applications, face recognition has attracted a large attention over the last three decades, and so the performance of face recognition algorithms has advanced significantly. According to the Face Recognition Vendor Test (FRVT) 2006 [18] and Multiple Biometric Evaluation (MBE) 2010 [10], large-scale face recognition in controlled conditions has already achieved impressive performance, for example, with a verification rate over 99% at the false accept rate (FAR) of 0.1%. However, there still exist many challenges for face recognition in uncontrolled environments [11], such as partial occlusions, large pose variations, extreme ambient illumination, and low resolutions.

Research on unconstrained face recognition has attracted a recent focus [24, 13, 11, 16, 4]. For this research, the Labeled Faces in the Wild (LFW) database [13] has been widely used for benchmark evaluation, which includes 13,233 images of 5,749 subjects collected from the Internet. However, in the standard LFW benchmark, only 3,000 pairs of genuine matches and 3,000 pairs of impostor matches are considered for classification, which is very limited and does not fully exploit all the available data. Partially due to this limitation, the best performance by the standard LFW protocol has recently reached 97% by [21]. However, instead of an overall classification accuracy, biometric system evaluation generally measures both the verification rate and the FAR [20]. Regarding this, a 97% accuracy by the standard LFW protocol may be too optimistic because it may still imply a 3% FAR which is vulnerable for most practical systems. Furthermore, when focusing at low FARs such as FAR=0.1%, the standard LFW evaluation is not statistically sound because at such FAR only three impostor scores are available. However, the above limitations of the standard LFW benchmark have not been paid too much attention to in the literature. Until recently, there are some studies of new protocols for unconstrained face recognition, such as the open-set identification protocol proposed in [16] and [4].

In this paper, we consider to make a full use of the whole LFW database and design a new experiment protocol including both verification and open-set identification scenarios to benchmark a number of existing algorithms. The new protocol has a particular interest at low FARs. Based on this benchmark protocol, we evaluated three kinds of fea-

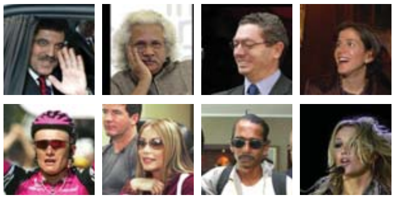


Figure 1. Sample images from the LFW database [13].

ture representations including the manually designed feature LBP [1], a learning based descriptor LE [6], and a well-aligned high-dimensional feature (HighDimLBP) [8], and seven kinds of learning algorithms including some recently developed metric learning algorithms [23, 9, 14, 15, 7]. The benchmark results show that the best approach achieves 41.66% verification rates at FAR=0.1%, and 18.07% open-set identification rates at rank 1 and FAR=1%. From the benchmark results, we conclude that the large-scale unconstrained face recognition problem is very challenging and still largely unresolved, thus further attention and effort is needed in developing effective feature representations and learning algorithms. From this study, we have developed a benchmark tool for large-scale unconstrained face recognition, which is made publicly available at http://www.cbsr.ia.ac.cn/users/scliao/projects/blufr/ to advance algorithm development along this direction.

2. Evaluation Database

The LFW database [13] is a large-scale unconstrained face image database, which is a very good source for the unconstrained face recognition evaluation. Images in LFW comes from the Faces in the Wild dataset [3], which is a large collection of Internet face images collected from the Yahoo News during 2002 to 2003. The LFW database includes 13,233 face images of 5,749 subjects. Face images in LFW were captured in uncontrolled environments. These images contain large variations in pose, illumination, expression, occlusion, and resolution. Fig. 1 shows some example images from this database.

There are two views with LFW for experiments, including View 1 for algorithm development and View 2 for algorithm training, evaluation, and performance report. The View 2 divides the dataset into 10 disjoint subsets of image pairs for cross validation, with each subset containing only 300 pairs of genuine matches and 300 pairs of impostor matches for classification. The standard LFW benchmark protocols include the image-restricted training and unrestricted training. The main difference is that the unrestricted protocol allows to form as many genuine and impostor pairs as possible beyond the restricted pairs for training. Recently, two new protocols are released in [12], namely the unsupervised and the use of outside data training protocols.

However, with all these training protocols, the test procedure is the same, that is, one must use the defined pairs of View 2 for algorithm evaluation and performance report. Therefore, in the whole benchmark only 3,000 pairs of genuine matching scores and 3,000 pairs of impostor matching scores can be computed for classification, which is very limited and does not fully exploit all the available data. As a result, performance evaluation at FAR=0.1% is not statistically sound because at such FAR only three impostor matching scores are available. Therefore, due to the limitation of the standard LFW benchmark protocol, we consider to make a full use of the whole LFW database and design a new experiment protocol to benchmark a number of existing algorithms, with a particular interest at low FARs.

3. Benchmark Protocol

3.1. Experimental Setting

With the LFW database, we designed our benchmark experiments as 10 random trials of training and test, and both face verification and open-set identification [20, 16, 4] were considered in the test phase. For algorithm development, we also designed a development set. The experimental setting is summarized in Table 1. Details are given below.

3.1.1 Development

A development set was randomly selected from the LFW database. The development set contains 521 images of 100 subjects for training, and 673 images of another 100 subjects for test. All classes in the training and test subsets have at least two face images. With the training subset, 3,925 sample pairs of true matches can be created, while with the test subset, 29,995 true matches can be formed.

This development set can be used if one algorithm needs to determine the optimal parameters. Alternatively, a cross validation procedure can also be done ***within*** a single trial of the training set and apply the selected model to the test set of the same trial. However, any parameter tuning depending on the test set is not allowed. This is to prevent parameter optimization with the test data. A good algorithm should have a good generalization ability of both models and parameters on unseen test data.

3.1.2 Training and Test

For each of the 10 trial, the whole LFW database was randomly divided into a training set and a test set. The training set of each trial includes 1,500 subjects, among which about 437 subjects have at least 2 face images. Each training set contains 3,524 images on average, and 85,341 genuine image pairs can be obtained on average. The test set of each trial contains the remaining 4,249 subjects, where 9,708 face images on average are available and about 1,243 subjects have at least 2 face images. In the test phase, the test

Table 1. Overview of the experimental setting for the new benchmark on the LFW database [13]. Numbers are averaged over the 10 trials.

Image set			No. Classes	No. Images	No. Genuine matches	No. Impostor matches
Development	Train		100	521	3,925	131,535
	Test		100	673	29,992	196,136
Evaluation	Train		1,500	3,524	85,341	6,122,185
	Test	All	4,294	9,708	156,915	46,960,863
		Gallery	1,000	1,000	-	-
		Genuine probe	1,000	4,350	-	-
		Impostor probe	3,249	4,357	-	-

set is used to compute the matching scores by face recognition algorithms. On average, 47,117,778 pairs of matching scores need to be computed in each trial. Then, these matching scores are used for the evaluation of both the face verification and open-set identification performance measures. Note that we did not adopt the 10 fold cross validation setting as usually applied. This is because we prefer to have a large test set for performance evaluation.

For the verification test, all the computed matching scores are used for performance evaluation, including about 156,915 genuine matching scores and 46,960,863 impostor matching scores in each trial on average. As for the open-set identification test, we randomly partitioned the test data into three subsets, the gallery set G, the genuine probe set P_G, and the impostor probe set P_N (explained in the next section). In each trial, 1,000 subjects in the test set were randomly selected to constitute the gallery set G, and only one image per subject was selected. To simulate real applications where the gallery image quality is usually good, we selected one face image per subject having the best image quality for that subject. Our selection of good quality face images is according to the following aspects: frontal or near frontal pose, normal lighting, neutral expression, no occlusion, and no blur. However, due to the unconstrained nature of LFW, it was very difficult to select one face image which satisfies all the above conditions, so we just did our best for the gallery face image selection. After the gallery image selection, the remaining face images of the 1,000 subjects were used to form the genuine probe set P_G, and all other images in the test set constituted the impostor probe set P_N. In each trial, the genuine probe set P_G contains 4,350 face images of 1,000 subjects, and the impostor probe set P_N contains 4,357 images of 3,249 subjects on average.

3.2. Performance Measures

Face verification and identification are two basic scenarios of face recognition, where verification is to decide whether two face images belong to the same identity, and (closed-set) identification is to determine the identity of the probe. In face verification, the Receiver Operating Characteristic (ROC) curve is used for performance measure, which is a plot of the verification rate versus the FAR by changing the decision thresholds. As for face identification, the Cumulative Matching Characteristic (CMC) curve is utilized for performance measure, which is a plot of cumulative matching score versus the rank of the probe [20].

As indicated in [20], the open-set identification task is more general, with the closed-set identification being its special case. Two sub-tasks, detection and identification, are involved in the open-set identification process. In the detection sub-task, the system decides whether the identity of the probe belongs to the gallery or not. In the identification sub-task, the system reports the identity of the accepted probe. Therefore, the task of open-set identification is to determine the identity of the probe or to reject the probe.

The performance evaluation of the open-set identification task involves three sets of face images. The first set is the gallery set G, which contains face images enrolled in the system. The other two are probe sets P_G and P_N. While P_G consists of subjects in the gallery set G but with different images, P_N includes subjects that are not present in G. Two performance measures, the detection and identification rate (DIR), and the FAR, are calculated for evaluation [20]. Let $id(g,p)$ be an indicator whether g and p belong to the same identity by the ground truth, that is,

$$id(g,p) = \begin{cases} 1, & g \text{ and } p \text{ belong to the same identity,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Let $s(\cdot,\cdot)$ be the similarity score function. Let

$$g^* = arg \max_{g \in G, id(g,p)=1} s(g,p), \tag{2}$$

that is, among all gallery images of the same identity as p, g^* reaches the maximum score. Further, let $rank(p|G)$ denote the rank order of $s(g^*,p)$ among matching scores between p and all gallery images. That is, $rank(p|G) = k$ means that $s(g^*,p)$ is the k^{th} largest similarity score. Then, the DIR and FAR measures are formulated as

$$DIR(\tau,k) = \frac{|\{p | p \in P_G,\ rank(p|G) \le k,\ s(g^*,p) \ge \tau\}|}{|P_G|}, \tag{3}$$

$$FAR(\tau) = \frac{|\{p | p \in P_N,\ \text{and} \max_{g \in G} s(g,p) \ge \tau\}|}{|P_N|}, \tag{4}$$

where τ is the decision threshold, and $|A|$ calculates the number of elements in the set A.

Given a rank level k, by changing the threshold τ, an ROC curve like in the verification scenario can be drawn by plotting DIR vs. FAR. Besides, given an FAR level, a CMC curve like in the closed-set identification scenario can also be drawn by first getting the threshold τ by Eq. (4), then plotting DIR vs. the rank k. Note that when FAR=1, the corresponding DIR is the traditional closed-set identification rate, with the gallery set G and the probe set P_G.

In the evaluation of both the verification and open-set identification, the performance measures of all the 10 random trials are averaged, and the standard deviation is also computed. The standard deviation of the ROC or CMC curves reflect the performance variation of one algorithm; a good algorithm should have a small standard deviation so that it performs stable under various conditions. In the face recognition literature, however, the standard deviation of the performance measure is usually omitted and most researchers focus on the average performance. One possible reason is that, in a performance reporting table, showing the standard deviation is not intuitive for ranking the compared algorithms, and with the ROC or CMC curves, the error bars are also not intuitive for comparison, not to say the figures can easily be made giddy with error bars. Therefore, we propose to use the $\mu - \sigma$ measure to fuse the two indicators. This fused measure can be directly used for algorithm ranking; a good algorithm should have a large average verification rate or identification rate, while having a small standard deviation. The $\mu - \sigma$ measure can also be understood as a kind of lower bound. If $N(\mu, \sigma)$ represents a normal distribution, this performance measure tells that on all runs of the benchmarked algorithm, about 84.15% of the time the algorithm performs above $\mu - \sigma$. As a result, using $\mu - \sigma$ to rank algorithms is similar as the max-min rule. We will use this measure throughout this paper for reporting figures and tables to enforce comparison of the standard deviation.

4. Face Recognition Approaches

4.1. Feature Representation

For feature representation, we utilized three kinds of features, namely the hand-crafted feature LBP [1], a learning based descriptor LE [6], and a high-dimensional LBP feature, denoted by HighDimLBP [8]. These three kinds of features have been extracted from the LFW database, and can be downloaded from the author's website of [8].

The LBP features were extracted by firstly dividing the face image into 10×10 non-overlapping sub-windows, then computing the 59-dimensional uniform LBP histogram on each of the sub-window, and finally concatenating all histograms [7]. This kind of feature has 5,900 dimensions.

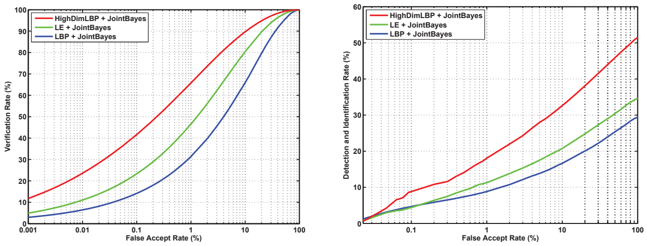
The LE descriptor is based on random projection tree learning to encode local image structures into discrete codes, which are considered to be uniformly distributed. Patch histogram is further applied with the learned encodings, resulting in 20,736 feature dimensions.

The HighDimLBP feature [8] is based on a recently developed accurate dense facial landmark detection [5]. In [8], 27 facial landmarks are detected, and the face image is aligned according to the detected facial landmarks. Then, 5 scales of local patches are sampled around each facial landmark, and each patch is further divided into 4×4 cells. A 59-dimensional uniform LBP histograms are exacted in each cell. Finally, all LBP histograms are concatenated, resulting in a 127,440-dimensional HighDimLBP feature.

4.2. Learning Algorithms

We evaluated seven learning algorithms, including the Principle Component Analysis (PCA) [22], Linear Discriminant Analysis (LDA) [2], Large Margin Nearest Neighbor (LMNN) [23], Information Theoretic Metric Learning (ITML) [9], Keep It Simple and Straightforward Metric Learning (KISSME) [14], Locally-Adaptive Decision Functions (LADF) [15], and Joint Bayesian (JointBayes) [7]. The LMNN algorithm [23] aims at learning a Mahalanobis distance metric for improving the k-nearest neighbor (kNN) classification, which can be solved by semidefinite programming. The ITML approach [9] considers minimizing the differential relative entropy between two multivariate Gaussians for learning the Mahalanobis distance function. The KISSME algorithm [14] applies the log likelihood ratio test of two Gaussian distributions for metric learning, and so a simplified closed-form solution can be derived, which is very similar to Moghaddam's Bayesian face approach [17]. The LADF algorithm [15] can be viewed as a joint model of a distance metric learning and a locally adapted thresholding rule, which is further formulated as an SVM-like problem. The JointBayes [7] algorithm revisits the famous Bayesian face approach [17], and considers concatenating every pair of samples for Bayesian modeling. A closed-form solution is obtained in [7] for JointBayes which can be efficiently solved. We implemented the JointBayes algorithm without EM [7] for simplicity, and we did not further implement the rotated sparse regression proposed in [7] because it leads to possible performance drop instead of improvement according to [7]. The LMNN [23], ITML [9], KISSME [14], and LADF [15] algorithms were downloaded from the authors' websites. Note that default parameters provided by the authors's codes were used, which may not reflect the best status of these algorithms. However, the JointBayes method we implemented has no parameters.

In each trial of the experiments, we reduced all the three kinds of features to 400 dimensions by PCA, and then applied the other learning algorithms except that LDA was directly applied to learn 400-dimensional subspaces.



(a) Verification (b) Open-set identification

Figure 2. Performances of the Joint Bayesian approach [7] with LBP [1, 7], LE [6], and HighDimLBP [8] features. (a) Verification ROC curves; (b) Open-set identification ROC curves at rank 1.

5. Benchmark Performance

Following the procedure described in the benchmark protocol, we evaluated 21 face recognition approaches (3 kinds of features × 7 learning algorithms) as introduced above. In this section, we report and discuss the benchmark performance of these methods for the large-scale unconstrained face recognition problem.

5.1. Verification Performance

5.1.1 Comparison of Features

We first compare the three kinds of features, LBP [1], LE [6], and HighDimLBP [8] under the verification setting. The Joint Bayesian approach [7] was applied for this comparison, which represents the best learning algorithm among the seven, as will be seen later. Fig. 2 (a) shows the resulting face verification ROC curves following the new benchmark procedure. From Fig. 2 (a) it is clear that the learning based descriptor LE is better than the hand-crafted feature LBP, and the HighDimLBP feature is the best one thanks to the well developed alignment and high dimensionality. These findings are consistent with [6] and [8]. At FAR=0.1%, the verification rates achieved by HighDimLBP, LE, and LBP are 41.66%, 23.31%, and 14.18%, respectively. Though great progress has been made by HighDimLBP+JointBayes (note that this approach has reached 93.18% classification rate in [8] with the standard LFW unrestricted protocol), our finding shows that, under the large-scale unconstrained face verification setting, the best performance today is still far from satisfactory.

5.1.2 Comparison of Learning Algorithms

We further compare the seven learning algorithms under the verification setting. Fig. 3 shows the ROC curves by using the LBP, LE, and HighDimLBP features, respectively. It can be observed that the JointBayes method is consistently the best performer, regardless of feature representation. By using the LBP feature, the difference between different algorithms is not too much; they all perform not very good. However, with better features LE and HighDimLBP, it is surprising that, the traditional LDA algorithm, being the second best performer, is still attractive compared to several recent-year developed metric learning algorithms.

The benchmark performances showing verification rates at FAR=0.1% and FAR=1% of the 21 evaluated approaches are summarized in Table 2. As aforementioned, these results are still not satisfactory for large-scale unconstrained face verification, leaving a large room for improvement. For example, new metric learning algorithms need to be developed for effective large-scale unconstrained face verification.

5.2. Open-set Identification Performance

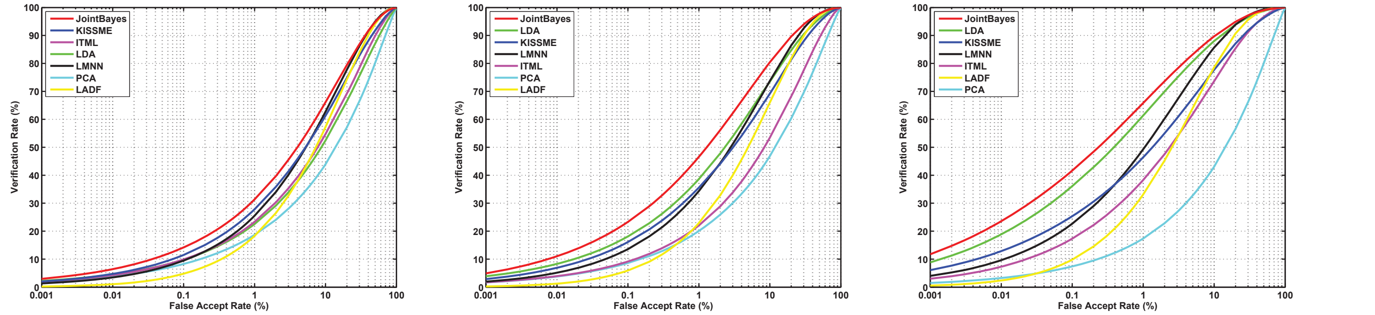
5.2.1 Comparison of Features

For the open-set identification scenario, we also compare the three features, LBP, LE, and HighDimLBP. Using the Joint Bayesian approach, the benchmark ROC curves at rank 1 are shown in Fig. 2 (b). The conclusion is similar with the verification case, that is, HighDimLBP is the best feature, followed by LE and LBP. However, the detection and identification rates are even lower than the verification scenario. For example, at FAR=1%, the detection and identification rates by using HighDimLBP, LE, and LBP are 18.07%, 11.26%, and 8.82%, respectively, and the corresponding rates at FAR=10% are 32.63%, 20.73%, and 16.61%. These results indicate that, by accepting 10% impostors, only 32.63% genuine probes can be correctly identified at rank 1 by the best algorithm.

5.2.2 Comparison of Learning Algorithms

We further compare the seven learning algorithms under the open-set identification setting. Firstly, we examine the traditional closed-set identification performance, in terms of CMC curves (corresponding to FAR=100% in the open-set identification setting). As shown in Fig. 4, closed-set identification performance measured by CMC curves is promising using the HighDimLBP feature. For example, all algorithms except PCA achieve over 80% identification rates at rank 100. LDA performs slightly better than JointBayes; they are the best two algorithms in this case, which achieve more than 50% identification rates at rank 1.

However, by considering the open-set identification scenario, the ROC curves at rank 1 using the LBP, LE, and HighDimLBP features respectively (shown in Fig. 5) show that the benchmark performance is still very poor. These benchmark results clearly show that the open-set identification performance significantly drops with the decreasing value of FAR. What is worse, according to the open-set identification CMC curves shown in Fig. 6, increasing the number of ranks helps very little in improving the performance. This is because genuine matching scores at lower ranks are hardly be larger than the decision threshold.



(a) LBP (b) LE (c) HighDimLBP

Figure 3. Verification ROC curves for the seven learning algorithms using the (a) LBP feature [1, 7], (b) LE feature [6], and (c) HighDimLBP feature [8].

Table 2. Benchmark performance of the 21 evaluated face recognition approaches for the verification scenario. The reported numbers are the mean verification rates (%) subtracted by the corresponding standard deviations over 10 trials.

Method	FAR=0.1%			FAR=1%		
	LBP	LE	HighDimLBP	LBP	LE	HighDimLBP
JointBayes [7]	**14.18**	**23.31**	**41.66**	**31.39**	**46.60**	**65.84**
LDA [2]	9.80	18.12	36.12	22.56	38.68	61.39
KISSME [14]	11.48	16.12	25.35	27.84	35.59	46.45
LMNN [23]	9.46	13.57	22.68	25.55	34.36	49.29
ITML [9]	9.87	9.16	17.32	23.37	22.06	38.32
LADF [15]	4.77	5.92	9.82	18.32	22.93	33.15
PCA [22]	8.28	8.61	7.41	18.69	20.03	17.38

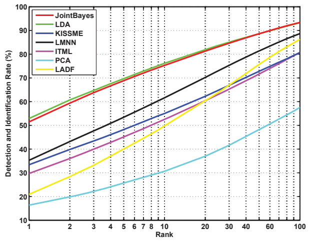
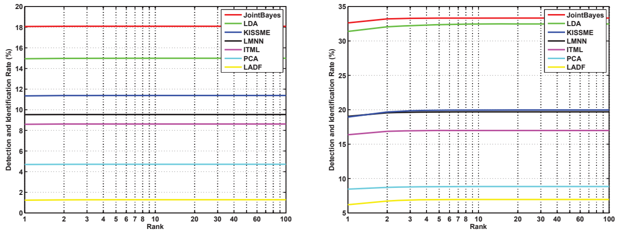


Figure 4. Closed-set face identification CMC curves for the seven learning algorithms using the HighDimLBP feature [8].

Nevertheless, the JointBayes approach is still consistently the best performer as observed from Fig. 5, regardless of feature representation. Different from the verification case, it can be observed that for open-set identification the superiority of the JointBayes approach against other learning algorithms is more obvious in using LBP than other features.

Table 3 summarizes the benchmark performances of the 21 evaluated methods for open-set identification at rank 1. These results show that the best unconstrained face recognition method today is still quite poor for open-set identification. On the other hand, it also indicates that the open-set unconstrained face identification problem is very challenging. It is more difficult than verification since both detection and identification should be satisfied. Therefore, special efforts may need to be spent in developing effective learning algorithms considering the open-set identification scenario.



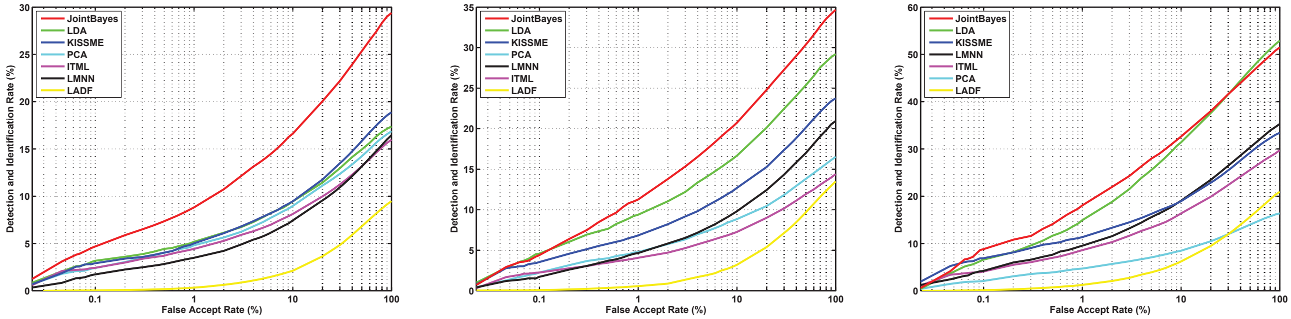
(a) FAR=1% (b) FAR=10%

Figure 6. Open-set identification CMC curves at (a) FAR=1% and (b) FAR=10% for the seven learning algorithms using the HighDimLBP feature [8].

An interesting finding from Tables 2 and 3 is that the top three learning algorithms, JointBayes, LDA, and KISSME, all have simple closed-form solutions which can be efficiently solved. This phenomenon may encourage the usage of simple models to efficiently handle large-scale learning.

5.3. Understanding of the Performance

To understand what can be achieved by the baseline algorithms evaluated in this paper, we conducted two additional experiments. The first one is with the FRGCv2 [19] database, where 16,028 controlled face images were used for experiments. We designed a similar protocol with these images, resulting in 10 trials of experiments, with each trial contains 3,448 images for training and 12,579 images for



(a) LBP

(b) LE

(c) HighDimLBP

Figure 5. Open-set identification ROC curves at rank 1 for the seven learning algorithms using the (a) LBP feature [1, 7], (b) LE feature [6], and (c) HighDimLBP feature [8].

Table 3. Benchmark performance of the 21 evaluated face recognition approaches for open-set identification at rank 1. The reported numbers are the mean detection and identification rates (%) at rank 1 subtracted by the corresponding standard deviations over 10 trials.

Method	FAR=1%			FAR=10%		
	LBP	LE	HighDimLBP	LBP	LE	HighDimLBP
JointBayes [7]	**8.82**	**11.26**	**18.07**	**16.61**	**20.73**	**32.63**
LDA [2]	5.18	9.38	14.94	9.45	16.66	31.39
KISSME [14]	5.02	6.83	11.34	9.45	12.69	18.94
LMNN [23]	3.49	4.66	9.53	7.44	9.81	19.04
ITML [9]	4.42	4.07	8.59	8.11	7.25	16.36
PCA [22]	4.81	4.77	4.70	8.97	8.84	8.46
LADF [15]	0.33	0.57	1.24	2.11	3.20	6.20

test on average. Face images were cropped and resized with three scales (120 × 120, 80 × 80, and 40 × 40 pixels), and the basic uniform LBP histogram was extracted and concatenated over 10 × 10 non-overlapping blocks in these images, resulting in 17,700 feature dimensions. These features were further reduced to 400 dimensions by PCA, and then the JointBayes approach was applied for face recognition. As a result, we got 94.36% verification rate @FAR=0.1%, and 82.71% open-set identification rates at rank 1 and FAR=1%.

Second, we also evaluated the JointBayes implementation with the HighDimLBP feature under the standard LFW unrestricted protocol. As a result, we got a 92.33% accuracy, compared to 93.18% reported in [8] with the same protocol. The difference is probably due to the use of EM in [7].

The above experimental results indicate that the implemented baseline method JointBayes performs well both under the new protocol with controlled face images and under the standard LFW unrestricted protocol with unconstrained face images. Besides, according to the LFW result page, the HighDimLBP+JointBayes approach is among the state of the art. Therefore, the low accuracy reported in this paper is mainly due to the new benchmark protocol designed with unconstrained face images, not the improper selection or failure of re-implementation of baseline algorithms.

In fact, the standard LFW protocol only considers the overall classification accuracy, but biometric system evaluation generally measures both the verification rate and the FAR [20]. regarding this, a 93% result by the standard protocol may still imply a 7% FAR. Such a high FAR is not useful for most practical systems, but this has not been paid too much attention to. While the low accuracy reported with the new benchmark protocol is partially due to tens of millions of matching scores evaluated, the main reason is the different FAR focusing. Our special interest is at low FARs, e.g. FAR=0.1%. Beyond this there is not much difference in performance between the new and the standard LFW protocols. For example, performance of the HighDimLBP+JointBayes under the new protocol shown in Fig. 2 (a) of this paper is comparable to other standard LFW ROC plots (e.g. Fig. 7 in [8]) at high FARs (e.g. 90% verification rate @FAR=10% vs. 95% in [8]), and in this informal comparison the main difference in the new protocol is the smaller training set and much larger test set. However, as discussed preciously, when focusing at FAR=0.1%, the standard LFW evaluation is not statistically sound because at such FAR only three impostor scores are available. Therefore, this motivate us for designing a new benchmark protocol to fully exploit the LFW database and evaluate performance at low FARs.

6. Conclusions

In this paper, we have developed a new benchmark protocol based on the largely studied LFW database for large-scale unconstrained face recognition evaluation under both verification and open-set identification scenarios, with a focus on low FARs. We also suggested the $\mu - \sigma$ measure to enforce comparing the standard deviation of performance measures over the 10 random trials. We have evaluated three kinds of feature representations and seven kinds of learning algorithms. The benchmark results show that HighDimLBP+JointBayes is the best approach, but achieving only 41.66% verification rates at FAR=0.1%, and 18.07% open-set identification rates at rank 1 and FAR=1%. Accordingly, we conclude that the large-scale unconstrained face recognition problem is still largely unresolved, thus further attention and effort is needed in developing effective feature representations and learning algorithms. From this study, we have developed a benchmark tool, which is made publicly available to advance algorithm development for large-scale unconstrained face recognition.

Acknowledgements

This work was supported by the NSFC Projects #61105023, #61103156, #61105037, #61203267, #61375037, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, and AuthenMetric R&D Funds. We thank the anonymous reviewers for providing some useful feedbacks and suggestions.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, July 1997.

[3] T. L. Berg, A. C. Berg, J. Edwards, and D. Forsyth. Whos in the picture. *Advances in neural information processing systems*, 17:137–144, 2004.

[4] L. Best-Rowden, H. Han, C. Otto, B. Klare, and A. K. Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. Technical Report MSU-CSE-14-1, Michigan State University, March 2014.

[5] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[6] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[7] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *European Conference on Computer Vision*. 2012.

[8] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[9] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007.

[10] P. Grother, G. Quinn, and P. Phillips. MBE2010: Report on the Evaluation of 2D Still-Image Face Recognition Algorithms. *NISTIR 7709, National Institute of Standards and Technology*, 2010.

[11] G. Hua, M.-H. Yang, E. Learned-Miller, Y. Ma, M. Turk, D. J. Kriegman, and T. S. Huang. Introduction to the special section on real-world face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1921–1924, 2011.

[12] G. B. Huang and E. Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, 2014.

[13] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. http://vis-www.cs.umass.edu/lfw/.

[14] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[15] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[16] S. Liao, A. K. Jain, and S. Z. Li. Partial face recognition: Alignment-free approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1193–1205, 2013.

[17] B. Moghaddam, T. Jebara, and A. Pentland. Bayesian face recognition. *Pattern Recognition*, 33(11):1771–1782, 2000.

[18] P. Phillips, W. Scruggs, A. O'Toole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale experimental results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):831–846, 2010.

[19] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[20] P. J. Phillips, P. Grother, and R. Micheals. Evaluation methods in face recognition. In S. Z. Li and A. K. Jain, editors, *Handbook of Face Recognition*, chapter 21, pages 551–574. Springer, 2nd edition, Sep. 2011.

[21] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deep-face: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[22] M. A. Turk and A. P. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, March 1991.

[23] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. *Advances in neural information processing systems*, 2006.

[24] S. K. Zhou, R. Chellappa, and W. Zhao. *Unconstrained face recognition*, volume 5. Springer, 2006.