Bayesian Face Revisited: A Joint Formulation

Dong Chen[1], Xudong Cao[3], Liwei Wang[2], Fang Wen[3], and Jian Sun[3]

[1] University of Science and Technology of China
chendong@mail.ustc.edu.cn
[2] The Chinese University of Hong Kong
lwwang@cse.cuhk.edu.hk
[3] Microsoft Research Asia, Beijing, China
{xudongca,fangwen,jiansun}@microsoft.com

Abstract. In this paper, we revisit the classical Bayesian face recognition method by Baback Moghaddam et al. and propose a new joint formulation. The classical Bayesian method models the appearance difference between two faces. We observe that this "difference" formulation may reduce the separability between classes. Instead, we model two faces jointly with an appropriate prior on the face representation. Our joint formulation leads to an EM-like model learning at the training time and an efficient, closed-formed computation at the test time. On extensive experimental evaluations, our method is superior to the classical Bayesian face and many other supervised approaches. Our method achieved 92.4% test accuracy on the challenging Labeled Face in Wild (LFW) dataset. Comparing with current best commercial system, we reduced the error rate by 10%.

1 Introduction

Face verification and face identification are two sub-problems in face recognition. The former is to verify whether two given faces belong to the same person, while the latter answers "who is who" question in a probe face set given a gallery face set. In this paper, we focus on the verification problem, which is more widely applicable and lay the foundation of the identification problem.

Bayesian face recognition [1] by Baback Moghaddam et al. is one of representative and successful face verification methods. It formulates the verification task as a binary Bayesian decision problem. Let H_I represents the intra-personal (same) hypothesis that two faces x_1 and x_2 belong to the same subject, and H_E is the extra-personal (not same) hypothesis that two faces are from different subjects. Then, the face verification problem amounts to classifying the difference $\Delta = x_1 - x_2$ as intra-personal variation or extra-personal variation. Based on the MAP (Maximum a Posterior) rule, the decision is made by testing a log likelihood ratio $r(x_1, x_2)$:

$$r(x_1, x_2) = \log \frac{P(\Delta|H_I)}{P(\Delta|H_E)}. \tag{1}$$

The above ratio can be also considered as a probabilistic measure of similarity between x_1 and x_2 for the face verification problem. In [1], two conditional probabilities in Eqn. (1) are modeled as Gaussians and eigen analysis is used for model learning and efficient computation.

Because of the simplicity and competitive performance [2] of Bayesian face, further progresses have been made along this research lines. For example, Wang and Tang [3] propose a unified framework for subspace face recognition which decomposes the face difference into three subspaces: intrinsic difference, transformation difference and noise. By excluding the transform difference and noise and retaining the intrinsic difference, better performance is obtained. In [4], a random subspace is introduced to handle the multi-model and high dimension problem. The appearance difference can be also computed in any feature space such as Gabor feature [5]. Instead of using a native Bayesian classifier, a SVM is trained in [6] to classify the the difference face which is projected and whitened in an intra-person subspace.

However, all above Bayesian face methods are generally based on the difference of a given face pair. As illustrated by a 2D example in Fig. 1, modeling the difference is equivalent to first projecting all 2D points on a 1D line (X-Y) and then performing classification in 1D. While such projection can capture the major discriminative information, it may reduce the separability. Therefore, the power of Bayesian face framework may be limited by discarding the discriminative information when we view two classes jointly in the original feature space.

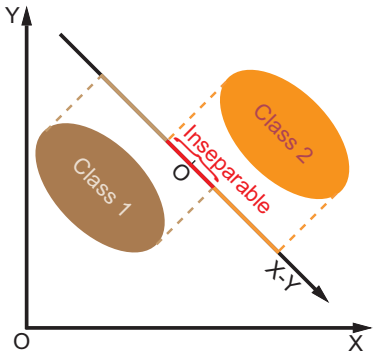

Fig. 1. The 2-D data is projected to 1-D by $x-y$. The two classes which are separable in joint representation are inseparable after the projecting. "Class1" and "Class2" could be considered as an intra-personal and an extra-personal hypothesis in face recognition.

In this paper, we propose to directly model the joint distribution of $\{x_1, x_2\}$ for the face verification problem in the same Bayesian framework. We introduce an appropriate prior on face representation: each face is the summation of two independent Gaussian latent variables, i.e., intrinsic variable for identity, and intra-personal variable for within-person variation. Based on this prior, we can effectively learn the parametric models of two latent variables by an EM-like algorithm. Given the learned models, we can obtain joint distributions of $\{x_1, x_2\}$

and derive a closed-form expression of the log likelihood ratio, which makes the computation efficient in the test phase.

We also find interesting connections between our joint Bayesian formulation and other two types of face verification methods: metric learning [7–10] and reference-based methods [11–14]. On one hand, the similarity metric derive from our joint formulation is beyond the standard form of the Mahalanobis distance. The new similarity metric preserves the separability in the original feature space and leads to better performance. On the other hand, the joint Bayesian method could be viewed as a kind of reference model with parametric form.

Many supervised approaches including ours need a good training data which contains sufficient intra-person and extra-person variations. A good training data should be both "wide" and "deep": having large number of different subjects and having enough images of each subject. However, the current large face datasets in the wild condition suffer from either small width (Pubfig [11]) or small depth (Labeled Faces in Wild (LFW) [15]). To address this issue, we introduce a new dataset, Wide and Deep Reference dataset (WDRef), which is both wide (around 3,000 subjects) and deep (2,000+ subjects with over 15 images, 1,000+ subjects with more than 40 images). To facilitate further research and evaluation on supervised methods on the same test bed, we also share two kinds of extracted low-level features of this dataset. The whole dataset can be downloaded from our project website `http://home.ustc.edu.cn/~chendong/JointBayesian/`.

Our main contributions can be summarize as followings:

- A joint formulation of Bayesian face with an appropriate prior on the face representation. The joint model can be effectively learned from large scale, high-dimension training data, and the verification can be efficiently performed by the derived closed-form solution.
- We demonstrate our joint Bayesian face outperforms the state of arts supervised methods, through comprehensive comparisons on LFW and WDRef. Our simple system achieved better average accuracy than the current best commercial system (face.com) [16][4].
- A large dataset (with annotations and extracted low-level features) which is both wide and deep is released.

2 Our Approach: A Joint Formulation

In this section, we first present a naive joint formulation and then introduce our core joint formulation and model learning algorithm.

2.1 A naive formulation

A straightforward joint formulation is to directly model the joint distribution of $\{x_1, x_2\}$ as a Gaussian. Thus, we have $P(x_1, x_2|H_I) = N(0, \Sigma_I)$ and $P(x_1, x_2|H_E) =$

[4] Leveraging an accurate 3D reconstruct and billions training images. But the details have not been published.

$N(0, \Sigma_E)$, where covariance matrixes Σ_I and Σ_E can be estimated from the intra-personal pairs and extra-personal pairs respectively. The mean of all faces is subtracted in the preprocessing step. At the test time, the log likelihood ratio between two probabilities is used as the similarity metric. As will be seen in later experiments, such naive formulation is moderately better than the conventional Bayesian face.

In above formulation, two covariance matrixes are directly estimated from the data statistics. There are two factors which may limit its performance. First, suppose the face is represented as a d-dimensional feature, in the naive formulation, we need to estimate the covariance matrix in higher dimension ($2d$) feature space of $[x_1\ x_2]$. We have higher chance to get a less reliable statistic since we do not have sufficient independent training samples. Second, since our collected training samples may not be complectly independent, the estimated Σ_E may not be a blockwise diagonal. But in theory, it should be because x_1 and x_2 are statistically independent.

To deal with these issues, we next introduce a simple prior on the face representation to form a new joint Bayesian formulation. The resulting model can be more reliably and accurately learned.

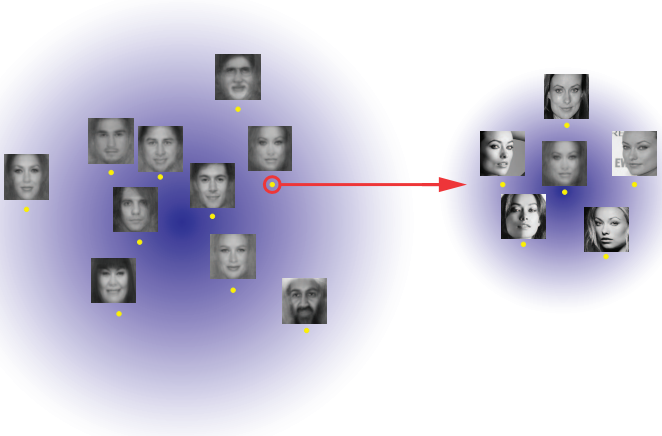

Fig. 2. Prior on face representation: both of the identities distribution (left) and the within-person variation (right) are modeled by Gaussians. Each face instance is represented by the sum of identity and the its variant.

2.2 A joint formulation

As already observed and used in previous works [17–20], the appearance of a face is influenced by two factors: identity, and intra-personal variation, as shown in Fig. 2. A face is represented by the sum of two independent Gaussian variables:

$$x = \mu + \varepsilon, \tag{2}$$

where x is the observed face with the mean of all faces subtracted, μ represents its identity, ε is the face variation (e.g., lightings, poses, and expressions) within the

same identity. Here, the latent variable μ and ε follow two Gaussian distributions $N(0, S_\mu)$ and $N(0, S_\epsilon)$, where S_μ and S_ϵ are two unknown covariance matrixes. For brevity, we call the above representation and associated assumptions as a face prior.

Joint formulation with prior. Given the above prior, no matter under which hypothesis, the joint distribution of $\{x_1, x_2\}$ is also Gaussian with zero mean. Based on the linear form of Eqn. (2) and the independent assumption between μ and ε, the covariance of two faces is:

$$\mathbf{cov}(x_i, x_j) = \mathbf{cov}(\mu_i, \mu_j) + \mathbf{cov}(\varepsilon_i, \varepsilon_j), \quad i, j \in \{1, 2\}. \tag{3}$$

Under H_I hypothesis, the identity μ_1, μ_2 of the pair are the same and their intra-person variations ε_1, ε_2 are independent. Considering Eqn.(3), The covariance matrix of the distribution $P(x_1, x_2|H_I)$ can be derived as:

$$\Sigma_I = \begin{bmatrix} S_\mu + S_\varepsilon & S_\mu \\ S_\mu & S_\mu + S_\varepsilon \end{bmatrix}.$$

Under H_E, both the identities and intra-person variations are independent. Hence, the covariance matrix of the distribution $P(x_1, x_2|H_E)$ is

$$\Sigma_E = \begin{bmatrix} S_\mu + S_\varepsilon & 0 \\ 0 & S_\mu + S_\varepsilon \end{bmatrix}.$$

With the above two conditional joint probabilities, the log likelihood ratio $r(x_1, x_2)$ can be obtained in a *closed form* after simple algebra operations:

$$r(x_1, x_2) = \log \frac{P(x_1, x_2|H_I)}{P(x_1, x_2|H_E)} = x_1^T A x_1 + x_2^T A x_2 - 2 x_1^T G x_2, \tag{4}$$

where

$$A = (S_\mu + S_\varepsilon)^{-1} - (F + G), \tag{5}$$

$$\begin{pmatrix} F + G & G \\ G & F + G \end{pmatrix} = \begin{pmatrix} S_\mu + S_\varepsilon & S_\mu \\ S_\mu & S_\mu + S_\varepsilon \end{pmatrix}^{-1}. \tag{6}$$

Note that in Eqn. (4) the constant term is omitted for simplicity.

There are three interesting properties of this log likelihood ratio metric. Readers can refer to the supplementary materials for the proof.

- Both matrix A and G are negative semi-definite matrixes.
- The negative log likelihood ratio will degrade to Mahalanobis distance if $A = G$.
- The log likelihood ratio metric is invariant to any full rank linear transform of the feature.

We will have more discussions on the new metric obtained by log likelihood ratio over the Mahalanobis distance in the Section 3.

2.3 Model learning

S_μ and S_ε are two unknown covariance matrixes which need to be learned from the data. We develop an EM-like algorithm to jointly estimate two matrixes in our model.

E-step: for each subject with m images, the relationship between the latent variables $\mathbf{h} = [\mu; \varepsilon_1; \cdots; \varepsilon_m]$ and the observations $\mathbf{x} = [x_1; \cdots; x_m]$ is:

$$\mathbf{x} = \mathbf{P}\mathbf{h}, \quad \text{where } \mathbf{P} = \begin{bmatrix} \mathbf{I} & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \end{bmatrix}. \tag{7}$$

And the distribution of the hidden variable $\mathbf{h}$ is

$$\mathbf{h} \sim N(0, \Sigma_h),$$

where $\Sigma_h = diag(S_\mu, S_\varepsilon, \cdots, S_\varepsilon)$. Therefore, based on Eqn. (7), we have the distribution of $\mathbf{x}$:

$$\mathbf{x} \sim N(0, \Sigma_x),$$

where

$$\Sigma_x = \begin{bmatrix} S_\mu + S_\varepsilon & S_\mu & \cdots & S_\mu \\ S_\mu & S_\mu + S_\varepsilon & \cdots & S_\mu \\ \vdots & \vdots & \ddots & \vdots \\ S_\mu & S_\mu & \cdots & S_\mu + S_\varepsilon \end{bmatrix}.$$

Given the observation $\mathbf{x}$, the expectation of the hidden variable $\mathbf{h}$ is:

$$E(\mathbf{h}|\mathbf{x}) = \Sigma_h \mathbf{P}^T \Sigma_x^{-1} \mathbf{x}. \tag{8}$$

Directly computing Eqn. (8) is expensive because the complexity in both memory and computation is $O(m^2d^2)$ and $O(d^3m^3)$ respectively, where d is the dimension of the feature. Fortunately, by taking the advantage of the structure of $\Sigma_\mathbf{h}$, $\mathbf{P}^T$ and Σ_x, the complexity in memory and computation can be reduced to $O(d^2)$ and $O(d^3 + md^2)$ respectively. Readers can refer to supplementary materials for the details.

M-step: in this step, we aim to update the values of parameters $\Theta = \{S_\mu, S_\varepsilon\}$:

$$S_\mu = \mathbf{cov}(\mu)$$
$$S_\varepsilon = \mathbf{cov}(\varepsilon)$$

where μ and ε are the expectations of the latent variables estimated in the E-step.

Initialization: In the implementation, S_μ and S_ε are initialized by random positive definite matrix, such as the covariance matrix of random data. We will have more discussions on the robustness of the EM algorithm in the next section.

2.4 Discussion

In this section, we discuss three aspects of our joint formulation.

Robustness of EM. Generally speaking, the EM algorithm only guarantees convergence. The quality of the solution depends on the objective function (the likelihood function of the observations) and the initialization. Here we empirically study the robustness of the EM algorithm by the following five experiments:

1. S_μ and S_ε are set to between-class and within-class matrixes (no EM).
2. S_μ and S_ε are initialized by between-class and within-class matrixes.
3. S_μ is initialized by between-class matrix and S_ε is initialized randomly.
4. S_μ is initialized randomly and S_ε is initialized by within-class matrix.
5. Both S_μ and S_ε are initialized randomly.

The above experiments are performed by using the training dataset (WDRef) and testing dataset (LFW, under unrestrict protocol). We use PCA to reduce the dimension of feature to 2,000. The rest of the experimental settings are the same with those in Section 4.2.

The experiment results in Table 1 show that: 1) EM optimization improves the performance (87.8% to 89.44%); 2) EM optimization is robust to various initializations. The estimated parameters converge to the similar solutions from quite different starting points. This indicates that our objective function may have very good "convex" properties.

Experiments	1(no EM)	2	3	4	5
Accuracy	87.80%	89.4%	89.34%±0.15%	89.38%±0.20%	89.39%±0.08%

Table 1. Evaluation of the robustness of EM algorithm. Results of 3-5 are averages over 10 trials.

Adaptive subspace representation. In our model, the space structures of μ and ε are fully encoded into the covariance matrixes S_μ and S_ε. Even if μ and ε are in lower dimensional intrinsic spaces, we do not need to pre-define intrinsic dimensions for them. The covariance matrixes will automatically adapt to the intrinsics structures of the data. Therefore our model is free from parameter tuning for the number of intrinsic dimension and able to fully exploit the discriminative information in the whole space.

Efficient computation. As the matrix G is negative definite, we can decompose it as $G = -U^T U$ after the training. In the test time, we can represent each face by a transformed vector $y = Ux$ and a scalar $c = x^T Ax$. Then, the log likelihood ratio can be very efficiently computed by: $c_1 + c_2 + 2y_1^T y_2$.

3 Relationships with Other Works

In this section, we discuss the connections between our joint Bayesian formulation and three other types of leading supervised methods which are widely used in face recognition.

3.1 Connection with metric learning

Metric learning [7–10, 21] has recently attracted a lot of attentions in face recognition. The goal of metric learning is to find a new metric to make two classes more separable. One of the main branches is to learn a Mahalanobis distance:

$$(x_1 - x_2)^T M (x_1 - x_2), \tag{9}$$

where M is a positive definite matrix which parameterizes the Mahalanobis distance.

As we can see from above Equation, this method shares the same drawback of the conventional Bayesian face. Both of them firstly project the joint representation to a lower dimension by a transform $[\mathbf{I}, -\mathbf{I}]$. As we have discussed, this transformation may reduce the separability and degrade the accuracy.

In our method, using the joint formulation, the metric in Eqn. (4) is free from the above disadvantage. To make the connection more clear, we reformulate Eqn. (4) as,

$$(x_1 - x_2)^T A (x_1 - x_2) + 2{x_1}^T (A - G) x_2. \tag{10}$$

Comparing Eqn. (9) and Eqn. (10), we see that the joint formulation provides an additional freedom for the discriminant surface. The new metric could be viewed as more general distance which better preserves the separability.

To further understand our new metric in Eqn. (4), we rewrite it here:

$$x_1^T A x_1 + x_2^T A x_2 - 2 x_1^T G x_2.$$

There are two components in the metric: the cross inner product term ${x_1}^T G x_2$ and two norm terms ${x_1}^T A x_1$ and ${x_2}^T A x_2$. To investigate their roles, we perform an experiment under five conditions: a) use both A and G; b) $A \to 0$; c) $G \to 0$; d) $A \to G$; e) $G \to A$. The experiment results are shown in Table 2.

Experiments	A and G	$A \to 0$	$G \to 0$	$A \to G$	$G \to A$
Accuracy	87.5%	84.93%	55.63%	83.73%	84.85%

Table 2. Roles of matrixes A and G in the log likelihood ratio metric. Both training and testing are conducted on LFW following the unrestricted protocol. SIFT feature is used and its dimension is reduce to 200 by PCA. Readers can refer to Section 4.4 for more information.

We can observe two things: most of discriminative information lies in the cross inner product term ${x_1}^T G x_2$; the norm terms ${x_1}^T A x_1$ and ${x_2}^T A x_2$ also play significant roles. They serve as the image specific adjustments to the decision boundary.

There are many works trying to develop different learning methods for Mahalanobis distance. However, relatively few works investigate the other forms of metrics like ours. A recent work [22] explores the metric based on cosine similarity by discriminative learning. Their promising results may inspire us to learn the log likelihood ratio metric in a discriminative way in the future work.

3.2 Connection with LDA and Probabilistic LDA

Linear Discriminant Analysis (LDA) [17] learns discriminative projecting directions by maximizing the between-class variation and minimizing within-class variation. The solution for the projections are the eigenvectors of an eigen problem. The discriminative power of the projections decreases along with the corresponding eigen value rapidly, which make the discriminative power distribution is very unbalance. This heterogeneous property make it hard to find an appropriate metric to fully exploit the discriminative information in all projections. Usually, LDA gets the best performance only with a few top eigenvectors, and its performance will decrease if more projections are added even though there are still useful information for discrimination.

Probabilistic LDA (PLDA) [18, 19] uses factor analysis to decompose the face into three factors: $x = B\alpha + W\beta + \xi$, i. e. identity $B\alpha$, intra-personal variation $W\beta$ and noise ξ. The latent variables α, β and ξ are assumed as Gaussian distributions. The covariance matrix of α and β are identity matrixes and the covariance matrix of ξ is diagonal matrix Σ_ξ. B, W and Σ_ξ are unknown parameters which needs to be learned. Similar to LDA, PLDA works well if the dimension of hidden variables is low. But its performance will rapidly degrade if the dimension of hidden variable is high, in which case, it is hard to get reliable estimation of the parameters.

In contrast to LDA and PLDA, we do not make the low dimension assumption and treat each dimension equally. Our joint model can reliably estimate the parameters in high dimension and does not require elaborate initialization. These two properties enable us to fully exploit the discriminative information of the high dimensional feature and achieve better performance.

3.3 Connection with Reference Based Methods

Reference-based methods [11–14] represent a face by its similarities to a set of reference faces. For example, in work [11], each reference is represented by a SVM "Simile" classifier which is trained from multiple images of the same person. The score of SVM is used as the similarity.

From the Bayesian view, if we model each reference as a gaussian $N(\mu_i, \Sigma)$, then the similarity from a face x to each reference is the conditional likelihood $P(x|\mu_i)$. Given n references, x can be represented as $[P(x|\mu_1), ..., P(x|\mu_n)]$. With this reference-based representation, we can define the similarity between two faces $\{x_1, x_2\}$ as the following log likelihood ratio:

$$\mathrm{Log}\left(\frac{\frac{1}{n}\sum_{i=1}^{n} P(x_1|\mu_i)\,P(x_2|\mu_i)}{\left(\frac{1}{n}\sum_{i=1}^{n} P(x_1|\mu_i)\right)\left(\frac{1}{n}\sum_{i=1}^{n} P(x_2|\mu_i)\right)}\right). \tag{11}$$

If we consider the references are infinite and independent sampled from a distribution $P(\mu)$, the above equation can be rewritten as:

$$\mathrm{Log}\left(\frac{\int P(x_1|\mu)\,P(x_2|\mu)\,P(\mu)d\mu}{\int P(x_1|\mu)\,P(\mu)d\mu \int P(x_2|\mu)\,P(\mu)d\mu}\right). \tag{12}$$

Interestingly, when $P(\mu)$ is a Gaussian, the above metric is equivalent to the metric derived from our joint Bayesian in Eqn. (4). (The proof can be found in supplementary). Hence our method can be considered as a kind of probabilistic reference-based method, with infinite references, but under Gaussian assumptions.

4 Experimental Results

In this section, we compare our joint Bayesian approach with conventional Bayesian face and other competitive supervised methods.

4.1 Dataset

Label Face in the Wild(LFW) [15] contains face images of celebrities collected from the Internet with large variations in pose, expression, and lighting. There are a total of 5749 subjects but only 95 persons have more than 15 images. For 4069 identities, just one image is available.

In this work, we introduce a new dataset, WDRef, to relieve this "depth" issue. We collect and annotate face images from image search engines by querying a set of people names. We need to emphasize that there is no overlap between our queries and the names in LFW. Then, the faces are detected by a face detector and rectified by affine transform estimated by five landmarks, i.e. eyes, nose and mouth corners. The landmarks are detected by [23]. We extract two kind of low-level features: LBP [24] and LE [25] in rectified holistic face.

Our final data set contains 99,773 images of 2995 people, and 2065 of them have more than 15 images. The large width and depth of this dataset will enable researchers to better develop and evaluate supervised algorithms which needs sufficient intra-personal and extra-personal variations. We will make the extracted features publicly available.

Fig. 3. Sample images in WDRef dataset

4.2 Comparison with other Bayesian face methods

In the first experiment, we compare conventional Bayesian face, unified subspace, Wang and Tang's unified subspace work [3], and our joint Bayesian. All methods are tested on two datasets: LFW and WDRef. When tested on LFW, all identities in WDRef are used for the training. We vary the identities number in the training data from 600 to 3000 to study the performance w.r.t training data size. When test on WDRef, we split it into two mutually exclusive parts: 300 different subjects are used for test, the others are for training. Similar to protocol in LFW, the test images are divided into 10 cross-validation sets and each set contains 300 intra-personal and extra-personal pairs. We use LBP feature and reduce the feature dimension by PCA to the best dimension for each method (2000 for joint Bayesian and unified subspace, 400 for Bayesian and naive formulation methods).

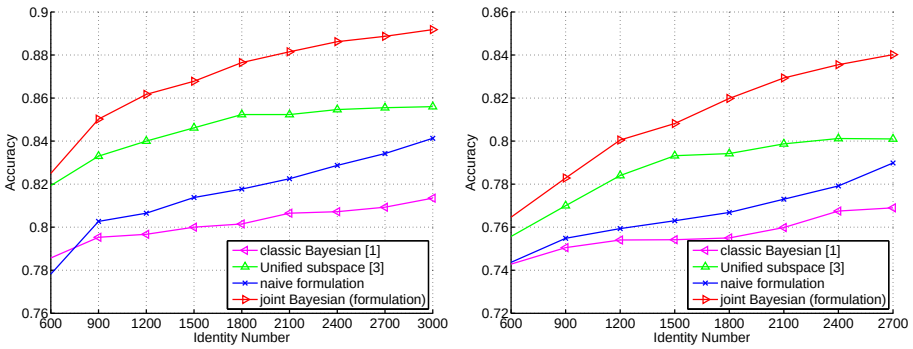

Fig. 4. Comparison with other Bayesian face related works. The joint Bayesian method is consistently better by using different training data sizes and on two databases: LFW(left) and WDRef(right).

As shown in Figure.(4), by enforcing an appropriate prior on the face representation, our proposed joint Bayesian method substantially better on various training data sizes. The unified subspace stands in the second place by taking the advantage of subspace selection, i.e. retaining the identity component and excluding the intra-person variation component and noise. We also note that when the training data size is small, the naive formulation is the worst. The reason is that it needs to estimate more parameters in higher dimension. However, as training data increasing, The performance of conventional Bayesian and unified subspace method(using the difference of face pair) gradually saturate. In contrast, The performance of the naive joint formulation keeps increasing as the training data increase. Its performance surpasses the performance of the conventional Bayesian method and is approaching the performance of unified subspace method. The trend of joint Bayesian method shares the same pattern as the naive joint formulation. The observation strongly demonstrates that joint formulation helps the discriminability.

4.3 Comparison with LDA and PLDA

We use the same experiment setting (trained on WDRef) as described in Section 4.2. LDA is based our own implementation and PLDA is from authors's implementation [19]. In the experiment, the original 5900 dimension LBP feature is reduced to the best dimension by PCA for each method (2000 for all methods). For LDA and PLDA, we further traverse to get the best sub-space dimension (100 in our experiment). As shown in Figure 5, our method significantly outperforms the other two methods for any size of training data. Both PLDA and LDA only use the discriminative information in such a low dimension space and ignore the other dimensions even though there are useful information in them. On the contrary, our method treats each dimension equally and leverages high dimension feature.

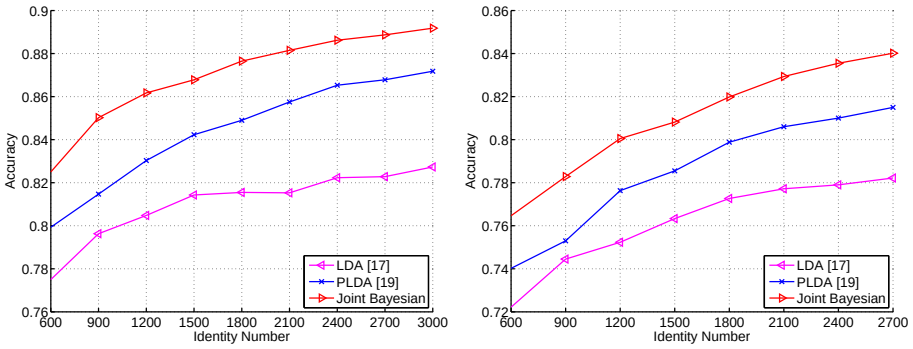

Fig. 5. Comparison with LDA and PLDA on LFW(left) and WDRef(right).

4.4 Comparison under LFW unrestricted protocol

In order to fairly compare with other approaches published on the LFW website, in this experiment, we follow the LFW unrestricted protocol, using only LFW for training. We combine the scores of 4 descriptors (SIFT, LBP, TPLBP and FPLBP) with a linear SVM classifier. The same feature combination could also be find in [12]. As shown in Table 3, our joint Bayesian method achieves the highest accuracy.

Method	LDML-MkNN [9]	Multishot [12]	PLDA [19]	Joint Bayesian
Accuracy	87.5%	89.50%	90.07%	**90.90%**

Table 3. Comparison with state of the arts method following the LFW unrestricted protocol. The results of other methods are from their original papers.

4.5 Training with outside data

Finally, we present our best performance on LFW along with existing state of the art methods or systems. Our purposes are threefold: 1) verify the generalization ability from one dataset to another dataset; 2) see what we can achieve using a limited outside training data; 3) compare with other methods which also rely on the outside training data.

We follow the standard unconstricted protocol in LFW and simply combine the four similarity scores computed under LBP feature and three types of LE [25] features. As shown by ROC curves in Figure 6, our approach with simple joint Bayesian formulation is ranked as **No.1** and achieved **92.4**% accuracy. The error rate is reduced by over 10%, compared with the current best (commercial) system which takes the additional advantages of an accurate 3D normalization and billions of training samples.

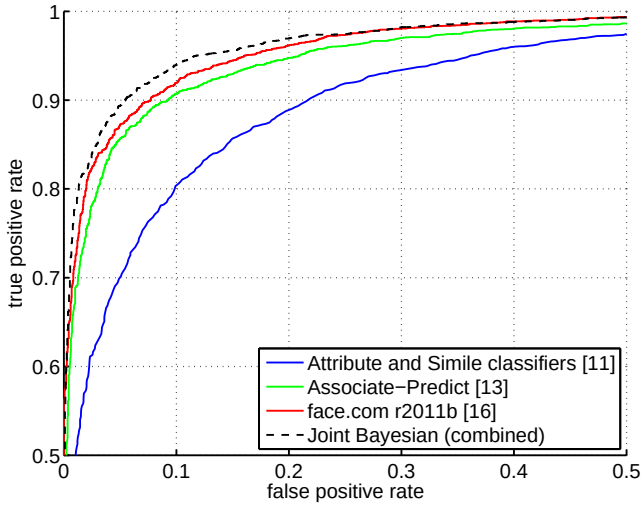

Fig. 6. The ROC curve of Joint Bayesian method comparing with the state of the art methods which also rely on the outside training data on LFW.

5 Conclusions

In this paper, we have revisited the classic Bayesian face recognition and proposed a joint formulation in the same probabilistic framework. The superior performance on comprehensive evaluations shows that the classic Bayesian face recognition is still highly competitive and shining, given modern low-level features and a training data with the moderate size.

References

1. Moghaddam, B., Jebara, T., Pentland, A.: Bayesian face recognition. Pattern Recognition **33** (2000) 1771–1782
2. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The feret evaluation methodology for face-recognition algorithms. PAMI **22** (2000) 1090–1104
3. Wang, X., Tang, X.: A unified framework for subspace face recognition. PAMI **26** (2004) 1222–1228
4. Wang, X., Tang, X.: Subspace analysis using random mixture models. In: CVPR. (2005)
5. Wang, X., Tang, X.: Bayesian face recognition using gabor features. (2003) 70–73
6. Li, Z., Tang, X.: Bayesian face recognition using support vector machine and face clustering. In: CVPR. (2004)
7. Weinberger, K.Q., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. Volume 10. (2005) 207–244
8. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: ICML. (2007)
9. Guillaumin, M., Verbeek, J.J., Schmid, C.: Is that you? metric learning approaches for face identification. In: ICCV. (2009)
10. Ying, Y., Li, P.: Distance metric learning with eigenvalue optimization. Journal of Machine Learning Research **13** (2012) 1–26
11. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV. (2009)
12. Taigman, Y., Wolf, L., Hassner, T.: Multiple one-shots for utilizing class label information. In: BMVC. (2009)
13. Yin, Q., Tang, X., Sun, J.: An associate-predict model for face recognition. In: CVPR. (2011)
14. Zhu, C., Wen, F., Sun, J.: A rank-order distance based clustering algorithm for face tagging. In: CVPR. (2011)
15. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E., Hanson, A.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. ECCV (2008)
16. Taigman, Y., Wolf, L.: Leveraging billions of faces to overcome performance barriers in unconstrained face recognition. Arxiv preprint arXiv:1108.1122 (2011)
17. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. PAMI **19** (1997) 711–720
18. Ioffe, S.: Probabilistic linear discriminant analysis. In: ECCV. (2006)
19. Prince, S., Li, P., Fu, Y., Mohammed, U., Elder, J.: Probabilistic models for inference about identity. PAMI **34** (2012) 144–157
20. Susskind, J., Memisevic, R., Hinton, G., Pollefeys, M.: Modeling the joint density of two images under a variety of transformations. In: CVPR. (2011)
21. Ramanan, D., Baker, S.: Local distance functions: A taxonomy, new algorithms, and an evaluation. In: ICCV. (2009)
22. Nguyen, H.V., Bai, L.: Cosine similarity metric learning for face verification. In: ACCV. (2010)
23. Liang, L., Xiao, R., Wen, F., Sun, J.: Face alignment via component-based discriminative search. In: ECCV. (2008)
24. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. PAMI **24** (2002) 971–987
25. Cao, Z., Yin, Q., Tang, X., Sun, J.: Face recognition with learning-based descriptor. In: CVPR. (2010)