Learning the Face Prior for Bayesian Face Recognition

Chaochao Lu and Xiaoou Tang

Department of Information Engineering,
The Chinese University of Hong Kong, China

Abstract. For the traditional Bayesian face recognition methods, a simple prior on face representation cannot cover large variations in facial poses, illuminations, expressions, aging, and occlusions in the wild. In this paper, we propose a new approach to learn the face prior for Bayesian face recognition. First, we extend Manifold Relevance Determination to learn the identity subspace for each individual automatically. Based on the structure of the learned identity subspaces, we then propose to estimate Gaussian mixture densities in the observation space with Gaussian process regression. During the training of our approach, the leave-set-out algorithm is also developed for overfitting avoidance. On extensive experimental evaluations, the learned face prior can improve the performance of the traditional Bayesian face and other related methods significantly. It is also proved that the simple Bayesian face method with the learned face prior can handle the complex intra-personal variations such as large poses and large occlusions. Experiments on the challenging LFW benchmark shows that our algorithm outperforms most of the state-of-art methods.

1 Introduction

Face recognition is an active research field in computer vision, and has been studied extensively [36,1,23,21,38,7,27,15,4,2,8,31]. It mainly consists of two subproblems: face verification (i.e., to verify whether a pair of face images are from the same person.) and face identification (i.e., to recognize the identity of a query face image given a gallery face set.). As the former is the foundation of the latter and has more applications, we focus on face verification in this paper.

Among the face verification methods, Bayesian face recognition [23] is a representative and successful one. It presents a probabilistic similarity measure based on the Bayesian belief that the difference $\Delta = x_1 - x_2$ of two faces x_1 and x_2 is characteristic of typical facial variations in appearance of an individual. It then formulates the face verification as a binary Bayesian decision problem. In other words, it classifies Δ as intra-personal variations Ω_I (i.e., the variations are from the same individual) or extra-personal variations Ω_E (i.e., the variations are from different individuals). Therefore, based on the MAP (Maximum a Posterior) rule, the similarity measure between x_1 and x_2 can be expressed by the logarithm likelihood ratio between $p(\Delta|\Omega_I)$ and $p(\Delta|\Omega_E)$, where both $p(\Delta|\Omega_I)$ and $p(\Delta|\Omega_E)$ are assumed to follow one multivariate Gaussian distribution [23].

D. Fleet et al. (Eds.): ECCV 2014, Part IV, LNCS 8692, pp. 119–134, 2014.

However, two limitations have restricted the performance of Bayesian face recognition. First, the above Bayesian face method, including several related methods [38,39,37], is based on the difference of a given face pair, which discards the discriminative information and reduce the separability [7]. Second, the distributions of $p(\Delta|\Omega_I)$ and $p(\Delta|\Omega_E)$ are oversimplified, assuming one multivariate Gaussian distribution can cover large variations in facial poses, illuminations, expressions, aging, occlusions, makeups and hair styles in the real world.

Recently, Chen et al. [7] proposed a joint formulation for Bayesian face, which has solved the first problem successfully, but the second problem still remains unsolved. In [19,27], a series of probabilistic models were developed to evaluate the probability that two faces have the same underlying identity cause. These parametric models are less flexible when dealing with complex data distributions. Therefore, it is difficult to capture the intrinsic features of the identity space by means of these existing Bayesian face methods.

To overcome the second problem in this paper, we propose a method to learn the two conditional distributions of $\{x_1, x_2\}$, denoted by $p(\{x_1, x_2\}|\Omega_I)$ and $p(\{x_1, x_2\}|\Omega_E)$. For brevity, we call the two conditional distributions as the *face prior*. Our method mainly consists of two steps.

In the first step, we exploit three properties of Manifold Relevance Determination (MRD) [9]: (1) It can learn a factorized latent variable representation of multiple observation spaces; (2) Each latent variable is either associated with a private space or a shared space; (3) It is a fully Bayesian model and allows estimation of both the dimensionality and the structure of the latent representation to be done automatically. We first extend MRD to learn an identity subspace for each individual automatically. As MRD is based on Gaussian Process latent variable models (GP-LVMs) [16], it is flexible enough to fit complex data. Then, we can obtain their corresponding latent representations z_1 and z_2 for x_1 and x_2 in the learned identity subspace. Therefore, two categories can be generated for training. One category includes K matched pairs, where each pair $\{z_1, z_2\}$ is from the same individual. The other category includes K mismatched pairs, where each pair is from different individuals.

In the second step, we propose to estimate Gaussian mixture densities for each category in the observed data space with Gaussian process regression (GPR) [30]. For each category, there is a clear one-to-one relationship between the latent input $[z_1, z_2]$ and the observed output $[x_1, x_2]$. We model this relationship with GPR, where the leave-set-out (LSO) technique is proposed for training in order to avoid overfitting. In fact, we interpret latent points as centers of a mixture of Gaussian distributions in the latent space that are projected forward by the Gaussian process to produce a high-dimensional Gaussian mixture in the observation space. Since the latent space only contains the identity information, the learned density can fully reflect the distribution of identities of face pairs $[x_1, x_2]$ in the observation space. The resulting distributions $p(\{x_1, x_2\}|\Omega_I)$ and $p(\{x_1, x_2\}|\Omega_E)$ can further improve the performance of Bayesian face recognition.

In summary, there are three contributions in this paper:

1) We introduce MRD and extend it to learn the identity subspace accurately, where the estimation of both the dimensionality and the structure of the subspace can be done automatically.
2) We propose to estimate Gaussian mixture densities with Gaussian process regression (GPR), which allows to estimate the densities in the high-dimensional observation space based on the structure of the low-dimensional latent space. Moreover, in order to avoid overfitting for training, the leave-set-out technique is also proposed.
3) We demonstrate that the learned face prior can improve the performance of Bayesian face recognition significantly, and the simple Bayesian face method with our face prior even outperforms the state-of-art methods.

2 Related Work

Our method is to learn the face prior for Bayesian face recognition. It consists of two steps: learn identity subspace and learn the distributions of identity. Therefore, we introduce some works of particular relevance to ours from the following two perspectives: learn subspace and learn the distributions of face images.

From the perspective of learning subspace, it has been extensively studied in recent face recognition [38,39,36,1,35,13,16]. The representative subspace methods are Principal Component Analysis (PCA) [36] and Linear Discriminant Analysis (LDA) [1]. The former produces the most expressive subspace for face representation, and the latter seeks the most discriminative subspace. Wang et al. [38] proposed a unified framework for subspace face recognition, where face difference is decomposed into three components: intrinsic difference, transformation difference, and noise. They only extracted the intrinsic difference for face recognition, and better performance can be achieved. In [39], a random mixture model was developed to handle complex intra-personal variations and the problem of high dimensions. As mentioned previously, most of these methods are based on the difference of a given face pair, which discards the discriminative information and reduce the separability. Besides, it is also unrealistic to accurately obtain the intra-personal subspace using the linear or simple parametric model in the complex real world. Besides, several probabilistic models, such as Probabilistic Principal Component Analysis (PPCA) [35], Probabilistic Linear Discriminant Analysis (PLDA) [13] and Gaussian Process Latent Variable Models (GP-LVMs) [16], were also proposed. However, these models assume that a single latent variable can represent general modalities, which is not realistic in the complex environment.

From the perspective of learning the distributions of face images, of particular relevance to our work is the Gaussian mixture model with GP-LVMs proposed by Nickisch et al. [25]. However, the problem in [25] is different from ours. In [25], in order to model the density for high dimensional data, GP-LVMs is firstly used to obtain a lower dimensional manifold that captures the main characteristics of

the data, and then the density of high-dimensional data can be estimated based on the low-dimensional manifold, but the hyperparameters of the model and the low-dimensional manifold need to be estimated simultaneously. In our method, the low-dimensional manifold (i.e., identity subspace) has been obtained from the first step, and only the hyperparameters need to be estimated. Thus GP-LVMs is not applicable for our problem. Further, as the low-dimensional manifold is fixed, the leave-out technique for overfitting avoidance is not suitable for our problem. A series of probabilistic models for inference about identity were also given in [19,27]. These parametric models assume that there exists a parametric function between the observation space and the latent space, so they are not flexible enough to learn a valid latent space in the complex real world. This also restricts their ability to learn the valid distribution for the identity.

3 Learning Identity Subspace

In this section, we first present how to extend MRD [9] to automatically learn the identity subspace for each individual, and then introduce the construction of the identity subspace. Finally, the construction of the training set for Bayesian face is presented.

3.1 Notation

We assume that the training set consists of N face images from M individuals, where the i-th individual has N_i ($N_i \geq 2$) D-dimensional face images, denoted by $X_i \in \mathbb{R}^{N_i \times D}$, and $N = N_1 + \cdots + N_M$. For each individual, we assume that X_i is partitioned into c subsets of the same size n_i, denoted by $X_i = \{X_i^1, \cdots, X_i^j, \cdots, X_i^c\}$, where $X_i^j \in \mathbb{R}^{n_i \times D}$. We further assume that the single latent identity subspace $Z_i \in \mathbb{R}^{n_i \times Q}$ ($Q \ll D$) exists for each individual, which gives a low-dimensional latent representation of the observed data through the mappings $F^{i,j} = \{f_d^{i,j}\}_{d=1}^D : Z_i \mapsto X_i^j$. In detail, we have $x_{nd}^{i,j} = f_d^{i,j}(z_n^i) + \epsilon_{nd}^{i,j}$, where $x_{nd}^{i,j}$ represents dimension d of point n in the observation space X_i^j, z_n^i represents point n in the latent space Z_i, and ϵ is the additive Gaussian noise.

3.2 The Extended Model of MRD

Although the proposed MRD in [9] only gave the analysis on the case of two views of data, it is easy to extend the model to the case of multiple views of data, as shown in Figure 1. For each observation space X_i^j, D latent functions $f_d^{i,j}$ are selected to be independent draws of a zero-mean Gaussian processes (GPs) with an automatic relevance determination (ARD) [30] covariance function of the form as follows,

$$k^{i,j}(z_a^i, z_b^i) = (\sigma^{i,j})^2 \exp\big(-\frac{1}{2}\sum_{q=1}^{Q} w_q^{i,j}(z_{aq}^i - z_{bq}^i)^2\big), \tag{1}$$

where we define the ARD weights as $\mathbf{w}^{i,j} = \{w_q^{i,j}\}_{q=1}^Q$ that can automatically infer the responsibility of each latent dimension for each observation space X_i^j. Thus, we can obtain the following likelihood,

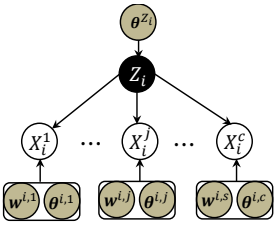

Fig. 1. The graphical model for multiple views of data in the extended model of MRD. In this figure, in order to emphasize the function of the ARD weights, $\mathbf{w}^{i,j}$ are separated from other hyperparameters $\boldsymbol{\theta}^{i,j}$ such as $\sigma^{i,j}$ and those in the addictive Gaussian noise. The ARD weights can encode the relevance of each dimension in the latent space Z_i for each observation space X_i^j. $\boldsymbol{\theta}^{Z_i}$ is the hyperparameters of prior knowledge about Z_i.

$$p(X_i^1, \cdots, X_i^c | Z_i, \boldsymbol{\theta}^{X_i}) = \prod_{j=1}^{c} \int p(X_i^j | F^{i,j}) p(F^{i,j} | Z_i, \mathbf{w}^{i,j}, \boldsymbol{\theta}^{i,j}) \mathrm{d}F^{i,j}, \quad (2)$$

where $\boldsymbol{\theta}^{X_i} = \{\mathbf{w}^{i,1}, \cdots, \mathbf{w}^{i,c}, \boldsymbol{\theta}^{i,1}, \cdots, \boldsymbol{\theta}^{i,c}\}$, and $p(F^{i,j} | Z_i, \mathbf{w}^{i,j}, \boldsymbol{\theta}^{i,j})$ can be modeled as a product of independent GPs parameterized by $k^{i,j}$. A fully Bayesian training procedure requires to maximize the joint marginal likelihood as follows,

$$p(X_i^1, \cdots, X_i^c | \boldsymbol{\theta}^{X_i}, \boldsymbol{\theta}^{Z_i}) = \int p(X_i^1, \cdots, X_i^c | Z_i, \boldsymbol{\theta}^{X_i}) p(Z_i | \boldsymbol{\theta}^{Z_i}) \mathrm{d}Z_i, \quad (3)$$

where $p(Z_i | \boldsymbol{\theta}^{Z_i})$ is a prior distribution placed on Z_i. We then use the approach proposed in [9] to obtain the final solution $\{Z_i, \boldsymbol{\theta}^{X_i}, \boldsymbol{\theta}^{Z_i}\}$.

3.3 The Construction of Identity Subspace

After the Bayesian training, we can acquire $\{Z_i, \boldsymbol{\theta}^{X_i}, \boldsymbol{\theta}^{Z_i}\}$ for each individual. Then, a segmentation of the latent space Z_i can be automatically determined as $Z_i = (Z_i^S, Z_i^1, \cdots, Z_i^j, \cdots, Z_i^c)$, where $Z_i^S \in \mathbb{R}^{n_i \times Q_S^i}$ $(Q_S^i \leq Q)$ is the latent space shared by $\{X_i^j\}_{j=1}^c$, and $Z_i^j \in \mathbb{R}^{n_i \times Q_j^i}$ $(Q_j^i \leq Q)$ is the private latent space for each X_i^j. Each dimension of Z_i^S, denoted by q, is selected from the set of dimensions $\{1, \cdots, Q\}$ with the constraint that $w_q^{i,1}, \cdots, w_q^{i,c} > \delta$, where δ is a threshold close to zero. Similarly, each dimension of Z_i^j is selected with the constraint that $w_q^{i,j} > \delta$ and $w_q^{i,1}, \cdots, w_q^{i,j-1}, w_q^{i,j+1}, \cdots, w_q^{i,c} < \delta$. Since Z_i^S only contains the information about the identity, we call it *identity subspace* for each individual.

Clearly, the model is independently trained for each individual. So the dimensions of their shared latent spaces may be different, meaning that the values of $\{Q_S^i\}_{i=1}^M$ are not consistent. To make each individual lie in the identity subspace with the same dimension Q_S, we let $Q_S = \min(Q_S^1, \cdots, Q_S^M)$. For $Q_S^i > Q_S$, we only select the dimensions with Q_S largest ARD weights.

3.4 The Construction of Training Set for Bayesian Face

Until now, each individual has two types of data: the identity subspace Z_i^S and the observation space X_i, where each z_n^i corresponds to the set $\{x_n^{i,j}\}_{j=1}^c$ through

the mapping set $F^{i,j}$. More precisely, for each individual, we can construct the following $n_i \times c$ correspondences between the identity subspace and the observation space,

$$\begin{bmatrix} \{z_1^i, x_1^{i,1}\} & \cdots & \{z_1^i, x_1^{i,j}\} & \cdots & \{z_1^i, x_1^{i,c}\} \\ \vdots & & \vdots & & \vdots \\ \{z_n^i, x_n^{i,1}\} & \cdots & \{z_n^i, x_n^{i,j}\} & \cdots & \{z_n^i, x_n^{i,c}\} \\ \vdots & & \vdots & & \vdots \\ \{z_{n_i}^i, x_{n_i}^{i,1}\} & \cdots & \{z_{n_i}^i, x_{n_i}^{i,j}\} & \cdots & \{z_{n_i}^i, x_{n_i}^{i,c}\} \end{bmatrix}. \tag{4}$$

Based on these correspondences from all individuals in the training set, two categories respectively consisting of K matched pairs and K mismatched pairs, denoted by Π_1 and Π_2, can be generated using the following criterion,

$$\pi^k = \{[z_a^{i_a}, z_b^{i_b}], [x_a^{i_a,j_a}, x_b^{i_b,j_b}]\}, \quad k = 1, \ldots, K \tag{5}$$

where $\pi^k \in \Pi_1$ when $i_a = i_b$ and $\pi^k \in \Pi_2$ when $i_a \neq i_b$. For convenience in the following sections, let $\pi^k = \{\mathbf{z}^k, \mathbf{x}^k\}$, where $\mathbf{z}^k = [z_a^{i_a}, z_b^{i_b}]^\top \in \mathbb{R}^{2Q_S}$ and $\mathbf{x}^k = [x_a^{i_a,j_a}, x_b^{i_b,j_b}]^\top \in \mathbb{R}^{2D}$. The two categories can be regarded as the training set for Bayesian face.

As mentioned above, we learn the identity subspace for each individual independently, thus the Bayesian training procedure can be conducted in parallel. Also, each individual generally does not contain too many images. Therefore, the time of Bayesian training is short, and the usage of memory can be controlled adaptively and reasonably.

4 Learning the Distributions of Identity

In this section, we propose to utilize GPR to estimate the density in the high-dimensional observation space based on the structure in the low-dimensional identity subspace. As we know, Gaussian mixture models (GMMs) are hard to fit in high dimensions while working well in low dimensions, as each component is either diagonal or has in the order of D^2 parameters [3]. Therefore, we first fit GMMs in the low-dimensional identity subspace, and then map it to the density in the high-dimensional observation space using GPR. Moreover, the leave-set-out technique is also proposed to avoid overfitting for training. In addition, we also present how to use the face prior for Bayesian face recognition.

4.1 Review of GPs and GPR

Here, we give a brief review of Gaussian processes (GPs) and GPR [30]. GPs are the extension of multivariate Gaussian distributions to infinite dimensionality. It is a probability distribution over functions, which is parameterized by a mean function $m(\cdot)$ and a covariance function $k(\cdot, \cdot)$. Without loss of generality, we let $m(\cdot) = 0$ and $k(\cdot, \cdot)$ be the ARD covariance function with the similar form as Equation (1),

$$\hat{k}(\mathbf{z}^a, \mathbf{z}^b) = \sigma_f^2 \exp\big(-\frac{1}{2}\sum_{q=1}^{2Q_S} w_q (z_q^a - z_q^b)^2\big) + \sigma_\epsilon^2 \delta(\mathbf{z}^a, \mathbf{z}^b), \tag{6}$$

where $\delta(\cdot,\cdot)$ is the Kronecker delta function, σ_f^2 and σ_ϵ^2 denote the signal and noise variances, respectively. For simplicity, these hyperparameters are collectively denoted by $\boldsymbol{\theta}^{\mathcal{K}} = \{w_1, \ldots, w_{2Q_S}, \sigma_f^2, \sigma_\epsilon^2\}$. Compared with Equation (1), the noise is folded into the covariance function for simplicity in the following. In GPR with vector-valued outputs, $2D$ independent GP priors with the same covariance and mean functions are placed on the latent functions $\mathbf{f} = \{f_i\}_{i=1}^{2D} : \mathcal{Z} \mapsto \mathcal{X}$. Given the training set $\{\mathbf{z}^k, \mathbf{x}^k\}_{k=1}^K$, if $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{z}, \boldsymbol{\Sigma}_\mathbf{z})$, then the distribution of $\mathbf{x}$ can be approximated by the following Gaussian distribution,

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma}_\mathbf{x}), \tag{7}$$

with $\boldsymbol{\mu}_\mathbf{x} = \mathbf{C}\bar{\mathbf{k}}$, and $\boldsymbol{\Sigma}_\mathbf{x} = \big(\bar{k} - \mathrm{Tr}(\mathbf{K}^{-1}\bar{\mathbf{K}})\big)\mathbf{I} + \mathbf{C}(\bar{\mathbf{K}} - \bar{\mathbf{k}}\bar{\mathbf{k}}^\top)\mathbf{C}^\top$, where $\mathbf{C} = [\mathbf{x}^1, \ldots, \mathbf{x}^K]\mathbf{K}^{-1}$, $\bar{\mathbf{k}} = \mathbb{E}[\mathbf{k}]$, $\bar{\mathbf{K}} = \mathbb{E}[\mathbf{k}\mathbf{k}^\top]$, $\mathbf{k} = [\hat{k}(\mathbf{z}^1, \mathbf{z}), \ldots, \hat{k}(\mathbf{z}^K, \mathbf{z})]^\top$, $\mathbf{K} = [\hat{k}(\mathbf{z}^a, \mathbf{z}^b)]_{a,b=1..K}$ and $\bar{k} = \hat{k}(\boldsymbol{\mu}_\mathbf{z}, \boldsymbol{\mu}_\mathbf{z})$. The two expectations can be evaluated in closed form [25,29].

4.2 Gaussian Mixture Modeling with GPR

According to the relationship between the distributions of $\mathbf{z}$ and $\mathbf{x}$, firstly, it is natural to build a GMM model on the latent identity subspace $\mathcal{Z} = \{\mathbf{z}^k\}_{k=1}^K$ as follows,

$$p(\mathbf{z}) = \sum_{l=1}^{L} \lambda_l \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\mathbf{z}^l, \boldsymbol{\Sigma}_\mathbf{z}^l), \tag{8}$$

where L is the number of components, $\{\lambda_l\}_{l=1}^L$ are the mixture weights satisfying the constraint that $\sum_{l=1}^L \lambda_l = 1$, and each mixture component of the GMM is a $2Q_S$-variate Gaussian density with the mean $\boldsymbol{\mu}_\mathbf{z}^i$ and the covariance $\boldsymbol{\Sigma}_\mathbf{z}^i$. These parameters are collectively represented by the notation, $\boldsymbol{\theta}^{\mathcal{G}} = \{\lambda_l, \boldsymbol{\mu}_\mathbf{z}^l, \boldsymbol{\Sigma}_\mathbf{z}^l\}_{l=1}^L$. We resort to the Expectation-Maximization (EM) algorithm to obtain an estimate of $\boldsymbol{\theta}^{\mathcal{G}}$. Secondly, each point $\mathbf{z}^k$ in the identity subspace is assigned to certain mixture component with the highest probability $\mathcal{N}(\mathbf{z}^k|\boldsymbol{\mu}_\mathbf{z}^l, \boldsymbol{\Sigma}_\mathbf{z}^l)$. In other words, each mixture component should contain a subset of points in the identity subspace, denoted by $\{\mathbf{z}^k\}_{k\in I_l}$, where I_l is the subset of indices of $\mathcal{Z}$ assigned to the l-th mixture component of the GMM. Thirdly, assuming that the parameters $\boldsymbol{\theta}^{\mathcal{K}}$ of the covariance function in Equation (6) have been estimated, we can utilize Equation (7) to calculate $\boldsymbol{\mu}_\mathbf{x}^l$ and $\boldsymbol{\Sigma}_\mathbf{x}^l$ based on $\{\mathbf{z}^k, \mathbf{x}^k\}_{k\in I_l}$ and $\{\mathbf{z}^k\}_{k\in I_l} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{z}^l, \boldsymbol{\Sigma}_\mathbf{z}^l)$, and then obtain $\{\mathbf{x}^k\}_{k\in I_l} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{x}^l, \boldsymbol{\Sigma}_\mathbf{x}^l)$. Therefore, we can finally acquire the distribution of identity in the observation space as follows,

$$p(\mathbf{x}) = \sum_{l=1}^{L} \lambda_l \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{x}^l, \boldsymbol{\Sigma}_\mathbf{x}^l). \tag{9}$$

4.3 The Leave-Set-Out Method

Now, the last question is how to estimate the parameters $\boldsymbol{\theta}^{\mathcal{K}}$ of the covariance function in Equation (6) on the training set $\{\mathbf{z}^k, \mathbf{x}^k\}_{k=1}^K$. Intuitively, we can attain $\boldsymbol{\theta}^{\mathcal{K}}$ by maximizing the following log likelihood of the data,

$$\mathcal{L}(\boldsymbol{\theta}^{\mathcal{K}}) = \sum_{k=1}^{K} \ln p(\mathbf{x}^k) = \sum_{k=1}^{K} \ln \sum_{l=1}^{L} \lambda_l \mathcal{N}(\mathbf{x}^k | \boldsymbol{\mu}_{\mathbf{x}}^l, \boldsymbol{\Sigma}_{\mathbf{x}}^l). \tag{10}$$

However, the above logarithm likelihood easily leads to overfitting on the training set. Inspired by the leave-out techniques in [40,25], for our specific problem, we propose the leave-set-out (LSO) method to prevent overfitting,

$$\mathcal{L}_{LSO}(\boldsymbol{\theta}^{\mathcal{K}}) = \sum_{l=1}^{L} \sum_{k \in I_l} \ln \sum_{l' \neq l} \lambda_{l'} \mathcal{N}(\mathbf{x}^k | \boldsymbol{\mu}_{\mathbf{x}}^{l'}, \boldsymbol{\Sigma}_{\mathbf{x}}^{l'}). \tag{11}$$

Compared with $\mathcal{L}(\boldsymbol{\theta}^{\mathcal{K}})$ in the objective (10), $\mathcal{L}_{LSO}(\boldsymbol{\theta}^{\mathcal{K}})$ enforces that the set of $\{\mathbf{x}^k\}_{k \in I_l}$ has the high density even though the set of the mixture components $\{\lambda_l \mathcal{N}(\mathbf{x}^k | \boldsymbol{\mu}_{\mathbf{x}}^l, \boldsymbol{\Sigma}_{\mathbf{x}}^l)\}_{k \in I_l}$ has been removed from the mixture, so we call it the *leave-set-out* method. Finally, we use the scaled conjugate gradients [24] to optimize $\mathcal{L}_{LSO}(\boldsymbol{\theta}^{\mathcal{K}})$ with respect to $\boldsymbol{\theta}^{\mathcal{K}}$.

4.4 Bayesian Face Recognition Using the Face Prior

When the probability (9) is obtained from Π_1, it describes the distribution of the identity information from the same individual in the observation space, thus we regard it as $p(\mathbf{x}|\Omega_I)$. Similarly, when the probability (9) is obtained from Π_2, it describes the distribution of the identity information from different individuals in the observation space, and we regard it as $p(\mathbf{x}|\Omega_E)$. At the testing step, given a pair of face images x_1 and x_2, so the similarity metric between them can be computed using the following logarithm likelihood ratio,

$$s(x_1, x_2) = \log \frac{p(\mathbf{x}|\Omega_I)}{p(\mathbf{x}|\Omega_E)}, \tag{12}$$

where $\mathbf{x} = [x_1, x_2]$. Since the above formulation is the traditional Bayesian face recognition based on the leaned face prior, for notational convenience in the following, we call it *the learned Bayesian*.

4.5 Discussion

It is worth noting that $\mathcal{L}_{LSO}$ proposed in this paper is different from the leave-out techniques such as $\mathcal{L}_{LOO}$ and $\mathcal{L}_{LPO}$ in [25]. There are four main differences as follows: (a) Only the hyperparameters need to be estimated in $\mathcal{L}_{LSO}$, whereas both $\mathcal{L}_{LOO}$ and $\mathcal{L}_{LPO}$ need to estimate the latent subspace and the hyperparameters; (b) Since we build a GMM in the latent identity subspace in advance,

and all points have been partitioned into different disjoint subsets, therefore, removing the mixture component is enough to avoid overfitting. However, this is not the case in $\mathcal{L}_{LOO}$ and $\mathcal{L}_{LPO}$, because the latent points are still unknown and need to be computed; (c) It is easy to leave out the set of points in $\mathcal{L}_{LSO}$, but it is hard in $\mathcal{L}_{LOO}$ and $\mathcal{L}_{LPO}$ as the number of points left out cannot be determined accurately; (d) In $\mathcal{L}_{LSO}$, a set of points with I_l shares the same mixture component $\mathcal{N}(\mathbf{x}^k|\boldsymbol{\mu}_{\mathbf{x}}^l, \boldsymbol{\Sigma}_{\mathbf{x}}^l)$ rather than each point has one unique Gaussian density in $\mathcal{L}_{LOO}$ and $\mathcal{L}_{LPO}$. Therefore, our method is much faster than the methods in [25] during the training procedure.

5 Experimental Results

In this section, we first introduce several datasets used in our experiments, and then analyze the validity of our approach. Next, we compare our approach with conventional Bayesian face. Finally, our approach is also compared with other competitive face verification methods in different tasks.

5.1 Datasets

In our experiments, the following five datasets are used for different tasks,

- **Multi-PIE** [11] This dataset contains 755,370 face images from 337 individuals under 15 view points and 20 illumination conditions in four recording sessions. Each individual has hundreds of face images.
- **Label Face in the Wild (LFW)** [12] This dataset contains 13,233 uncontrolled face images of 5,749 public figures collected from the Web with large variations in poses, expressions, illuminations, aging, hair styles and occlusions. Of these, 4069 people have just a single image, and only 95 people have more than 15 images in the dataset.
- **AR** [22] This dataset consists of over 4,000 color images from 126 people (70 males and 56 females). All images correspond to frontal view faces with different facial expressions, different illumination conditions and with different occlusions (people wearing sun-glasses or scarf). The number of image per person is 26.
- **PubFig** [15] This dataset is a large, real-world face dataset consisting of 58,797 images of 200 people collected from the Internet. Although the number of persons is small, every person has more than 200 images on average.
- **Wide and Deep Reference (WDRef)** [7] This dataset contains 99,773 images of 2995 people. Of them, 2065 people have more than 15 images, and over 1000 people have more 40 images. It is worth emphasizing that there is no overlap between this dataset and LFW.

To perform the fair comparison with the recent face verification methods, each face image is cropped and resized to 150×120 pixels with the eyes, nose, and mouth corners aligned, and then LBP feature [26] is extracted in each rectified holistic face (if not otherwise specified).

5.2 Parameter Setting

According to the descriptions in the preceding sections, our approach involves two types of parameters: the hyperparameters $\{\boldsymbol{\theta}^{\mathcal{G}}, \boldsymbol{\theta}^{\mathcal{K}}\}$ and the general parameters $\{c, L\}$. Since the hyperparameters can be automatically learned from the data, so we only need to focus on how to select the values of the general parameters. In fact, the parameter c controls the number of conditions influencing intra-personal variations, and the parameter L implies the complexity of the distributions of identity. As the two general parameters play a very important role in our approach, we give a detailed description about how to determine them.

Given the training set and the validation set, we then can determine the values of $\{c, L\}$ using the following two methods based on the characteristics of each dataset:

Method 1. For the datasets under controlled conditions (e.g., Multi-PIE and AR), we directly let c be the number of controlled conditions, and then tune L. Each time we tune L, our approach can be trained on the training set, and then tested on the validation set. Finally, the value of L that leads to the best performance on the validation set is determined.

Method 2. For the datasets under uncontrolled conditions (e.g., LFW, PubFig, and WDRef), we first fix c, and then tune L in the same method as **Method 1**. After the optimal L is determined, we fix L, and then tune c in the same method again. Thus we can obtain the final c and L.

5.3 Performance Analysis of the Proposed Approach

In this section, we conduct three experiments to analyze the validity of our approach. All the experiments are performed by using the training set (PubFig), validation set (the testing set in View 1 of LFW) and testing set (View 2 of LFW). In the training set, all 200 different individuals are used, and 200 images are randomly selected for each individual. In the testing set, we strictly follow the standard 10 fold cross validation experimental setting of LFW under the unrestricted protocol.

For the first experiment, we demonstrate the validity of our method for learning identity subspace by comparing PCA [36], LDA [1], PPCA [35], PLDA [13] with our extension of MRD. In detail, since our approach consists of two steps, so we can replace our extension of MRD with the above conventional subspace methods in the first step to learn the identity subspace, while both the construction of training set in Section 3.4 and the method of learning the distributions of identity in Section 4 are kept unchanged. In the experiment, for PCA and PPCA, the original 10,620 ($15 \times 12 \times 59$) dimensional LBP feature can be directly reduced to the best dimension. However, for LDA and PLDA, the original LBP feature can be first reduced to the best dimension by PCA, and then is further reduced to lower dimensional subspace. For our extension of MRD, the dimension of identity subspace can be determined automatically. We vary the the number of individuals in the training set from 50 to 200 to study the performance of our approach w.r.t. the training data size. Each time the training data size changes, the best c and L is estimated using **Method 2** in Section 5.2 for

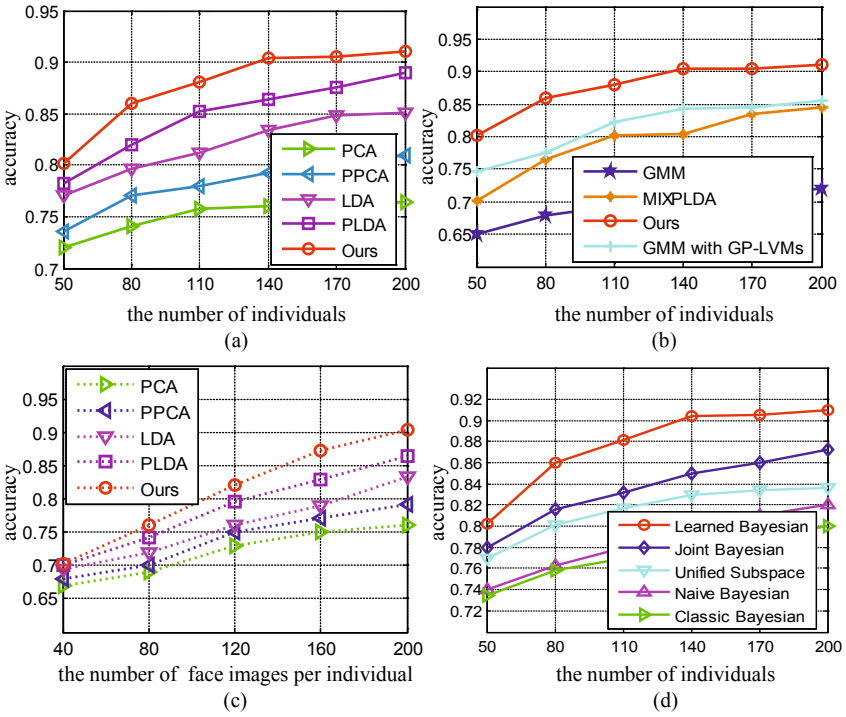

Fig. 2. Verification of the validity of our approach. (a) To verify the validity of learning identity subspace in our approach. (b) To verify the validity of learning the distributions of identity in our approach. (c) To verify the relationship between the number of images for each individual and the performance of our approach. (d) Comparison with other Bayesian face methods.

our approach, because PubFig is an uncontrolled dataset. Figure 2 (a) shows the performances of our approach with different subspace methods replaced in the first step, where the performance of our approach with the extension of MRD is better than others on various training data sizes. This has demonstrated the validity of our method for learning identity subspace.

For the second experiment, we prove the validity of our method for learning the distributions of identity. This step is to estimate the Gaussian mixture density in the observation space based on its corresponding known latent subspace. From the view of mixture models, we can compare our method with conventional GMMs, GMM with GP-LVMs [25] and Mixtures of PLDAs (MIXPLDA) [19]. For the fair comparison, the number of mixture components is set to the same L as ours for all methods. Similar to that in the first experiment, we also vary the number of individuals in the training set from 50 to 200 to study the performance of our approach. Each time we estimate the optimal c and L using **Method 2**. As shown in Figure 2 (b), our method for learning the distributions of identity outperforms other methods on the training set with different numbers of individuals.

For the third experiment, we analyze the relationship between the number of images for each individual and the performance of our approach. We use the same experiment setting as described in the first experiment. The number of individuals on the training set is fixed to 140, we then vary the number of face images per individual from 40 to 200 to study its influence on the performance of our approach. As shown in Figure 2 (c), the performance of our approach can be improved more rapidly than other methods with the increasing number of images per individual. That is because our method can capture the identity information more accurately when each individual contains more images. With the advent of the era of big data, it has become much easier to obtain many samples for each individual. Therefore our approach will be more widely used.

5.4 Comparison with Other Bayesian Face Methods

In this experiment, we verify that the Bayesian face with the learned face prior (the learned Bayesian face) outperforms the conventional Bayesian face [23]. Besides, we also compare unified subspace [38], naive Bayesian formulation [7], joint Bayesian formulation [7] with our learned Bayesian face. Here, the same experiment setting as described in Section 5.3 is used. The LBP feature is reduced by PCA to the best dimension for those methods. Obviously, the results in Figure 2 (d) shows that the learned face prior can improve the performance of Bayesian face recognition significantly.

5.5 Handling Large Poses

Face recognition with large pose variations is always a challenging problem. In this experiment, we demonstrate that our approach is also robust to large pose variations. Existing methods can be mainly divided into two categories: 2D methods and 3D methods (or their hybrids). Although 3D model based methods generally have higher precision than 2D methods, our approach is the 2D method, and therefore compared with several recent popular 2D pose robust methods: APEM [18], Eigen light-fields (ELF) [10], coupled bias-variance tradeoff (CBVT) [17], tied factor analysis (TFA) [28], Locally Linear Regression (LLR) [6], and multi-view discriminant analysis (MvDA) [14]. Of them, APEM, CBVT, TFA and MvDA are from authors' implementation, and the remaining is based on our own implementation. All methods are tested on the Multi-PIE dataset. As we only consider the pose variations in this experiment, so we choose a subset of individuals from MultiPIE, where each individual contains the images with all 15 poses, the neutral expression, and 6 similar illumination conditions (the indices of the selected illumination conditions are $\{07, 08, 09, 10, 11, 12\}$ in this experiment). Then, the subset is split into two mutually exclusive parts: 100 different individuals are used for testing, and the others are for training. At the training step, we let $c = 15$, meaning that all images of an individual is partitioned into 15 subsets, where each subset only contains the images with one pose. Then, L is estimated using **Method 1** on the training set and the validation set (View 1 of LFW), where 10,000 matched pairs and mismatched pairs are constructed respectively. At the testing step, to verify the performance

Table 1. Results (%) on the Multi-PIE dataset

Pose Pairs	APEM	ELF	CBVT	TFA	LLR	MvDA	Learned Bayesian
$\{0°, +60°\}$	65.3	77.4	86.7	89.1	85.4	86.4	**93.6**
$\{0°, +75°\}$	51.7	63.9	79.2	86.5	74.7	82.3	**91.2**
$\{0°, +90°\}$	40.1	38.9	70.1	82.4	64.2	73.6	**88.5**
$\{+15°, +75°\}$	60.2	75.1	81.6	86.5	82.3	75.4	**89.1**
$\{+15°, +90°\}$	45.8	55.2	75.2	81.2	78.6	79.3	**89.2**
$\{+30°, +90°\}$	41.2	57.3	73.2	84.4	79.1	77.2	**90.3**

of our approach on large poses, we split the testing set into different groups. Each group contains all images from one pose pair in the testing set. Similar to the protocol in LFW, all images in each group are also divided into 10 cross-validation sets and each set contains 300 intra-personal and extra-personal pairs. all the methods are tested on each group. Due to space limitation, we only present some results on the groups with over 45° pose differences. As shown in Table 2, our approach outperforms other methods on these groups. Further, we can observe that the performance of our approach becomes more noticeable with the increasing pose differences.

5.6 Handling Large Occlusions

In this experiment, we show that our approach can handle the face images with large occlusions. Our approach is compared with three representative methods: sparse representation classification (SRC) [41], the sparsity based algorithm using MRFs (SMRFs) [43], and Gabor-feature based SRC (GSRC) [42]. All methods are tested on the AR dataset. First, we chose a subset of AR dataset, where only the images with the neutral expression and the norm illumination are considered. Then, we partition the selected subset into two parts: 40 individuals are used for testing, and the remaining are used for training. During the training procedure, let c be the number of types of occlusions ($c = 3$ in this experiment, i.e., all images of each individual are split into three subsets: no wearing, wearing glasses, and wearing scarf), and then L is optimized using **Method 1** on the training set and the validation set (View 1 of LFW), where 400 matched pairs and mismatched pairs are constructed respectively. At the testing step, similar to the protocol in LFW, the testing images are divided into 10 cross-validation sets and each set contains 100 intra-personal and extra-personal pairs. As shown in Table 2, our approach is also robust to large occlusions, because our approach can accurately learn the identity subspace for each individual with occlusions.

Table 2. Results on the AR dataset

Method	SRC	SMRFs	GSRC	Learned Bayesian
Accuracy (%)	87.13	92.42	94.38	**96.23**

5.7 Comparison with the State-of-Art Methods

Finally, to compare with the state-of-art methods and better investigate our approach, we present our best verification result on the LFW benchmark with the outside training data (WDRef). LBP [26] and LE [5] features are extracted from these two datasets[1]. We combine the similar scores with a linear SVM classifier to make the final decision. In the experiment, we strictly follow the standard unrestricted protocol in LFW. First, to make better use of the strengths of our approach as indicated in the third experiment of Section 5.3, we choose a subset of WDRef with the individuals containing at least 30 images. Then, our approach is trained on WDRef and validated on the View 1 of LFW to estimate the optimal general parameters L and c. Finally, we test our approach on the View 2 of LFW under the standard unrestricted protocol. As shown in Figure 3, our approach, i.e., the learned Bayesian face, achieves **96.65**% accuracy. The previously published best Bayesian result on the LFW dataset (96.33%, unrestricted protocol) was achieved by the transfer learning algorithm [4] trained on the WDRef dataset based on the combined Joint Bayesian method [7] and the high-dimensional features [8], while our approach is trained on the same dataset using only the simple low-dimensional features. It is also shown that the accuracy of the simple Bayesian face method with our face prior can outperform most of the state-of-art methods [2,8,31,4,7], and is even comparable with the current best results [34,20,33,32].

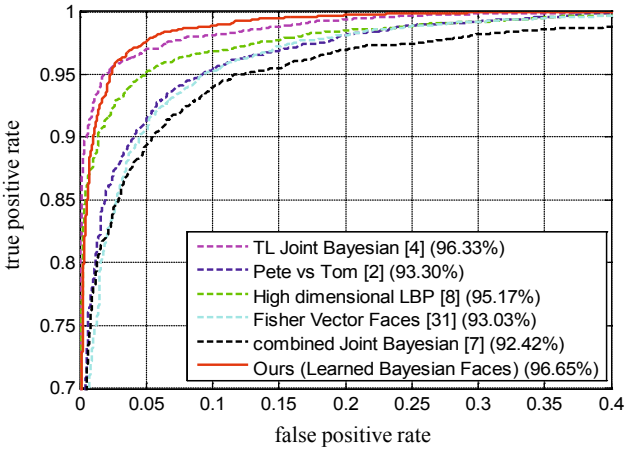

Fig. 3. Verification performance on LFW with the outside training data

6 Conclusions

In this paper, we have proposed a new approach to learn the face prior for the traditional Bayesian face recognition. Our approach consists of two steps. In

[1] These two kinds of extracted features of the LFW and WDRef datasets and annotations are provided by the authors [7], and can be downloaded from their project website.

the first step, MRD is extended to automatically learn the identity subspace for each individual. In the second step, GMM with GPR is proposed to estimate the density of identities in the observation space based on the structure of identity subspace. Moreover, we propose to use the leave-set-out technique to avoid overfitting. Extensive experiments shows that the learned face prior significantly improves the performance of the Bayesian face method, and the simple Bayesian face method with our face prior even outperforms most of the state-of-art methods.

References

1. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. TPAMI (1997)
2. Berg, T., Belhumeur, P.N.: Tom-vs-pete classifiers and identity-preserving alignment for face verification. In: BMVC (2012)
3. Bishop, C.M.: Pattern recognition and machine learning (2006)
4. Cao, X., Wipf, D., Wen, F., Duan, G., Sun, J.: A practical transfer learning algorithm for face verification. In: ICCV (2013)
5. Cao, Z., Yin, Q., Tang, X., Sun, J.: Face recognition with learning-based descriptor. In: CVPR (2010)
6. Chai, X., Shan, S., Chen, X., Gao, W.: Locally linear regression for pose-invariant face recognition. TIP (2007)
7. Chen, D., Cao, X., Wang, L., Wen, F., Sun, J.: Bayesian face revisited: A joint formulation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 566–579. Springer, Heidelberg (2012)
8. Chen, D., Cao, X., Wen, F., Sun, J.: Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In: CVPR (2013)
9. Damianou, A., Ek, C., Titsias, M.K., Lawrence, N.D.: Manifold relevance determination. In: ICML (2012)
10. Gross, R., Matthews, I., Baker, S.: Appearance-based face recognition and light-fields. TPAMI (2004)
11. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multipie. Image and Vision Computing (2010)
12. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., University of Massachusetts, Amherst (2007)
13. Ioffe, S.: Probabilistic linear discriminant analysis. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part IV. LNCS, vol. 3954, pp. 531–542. Springer, Heidelberg (2006)
14. Kan, M., Shan, S., Zhang, H., Lao, S., Chen, X.: Multi-view discriminant analysis. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 808–821. Springer, Heidelberg (2012)
15. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and Simile Classifiers for Face Verification. In: ICCV (2009)
16. Lawrence, N.D.: Gaussian process latent variable models for visualisation of high dimensional data. In: NIPS (2003)
17. Li, A., Shan, S., Gao, W.: Coupled bias–variance tradeoff for cross-pose face recognition. TIP (2012)
18. Li, H., Hua, G., Lin, Z., Brandt, J., Yang, J.: Probabilistic elastic matching for pose variant face verification. In: CVPR (2013)

19. Li, P., Fu, Y., Mohammed, U., Elder, J.H., Prince, S.J.: Probabilistic models for inference about identity. TPAMI (2012)
20. Lu, C., Tang, X.: Surpassing human-level face verification performance on lfw with gaussianface. arXiv preprint arXiv:1404.3840 (2014)
21. Lu, C., Zhao, D., Tang, X.: Face recognition using face patch networks. In: ICCV (2013)
22. Martinez, A.M.: The ar face database. CVC Technical Report (1998)
23. Moghaddam, B., Jebara, T., Pentland, A.: Bayesian face recognition. Pattern Recognition (2000)
24. Nabney, I.: Netlab: algorithms for pattern recognition. Springer (2002)
25. Nickisch, H., Rasmussen, C.E.: Gaussian mixture modeling with gaussian process latent variable models. In: Goesele, M., Roth, S., Kuijper, A., Schiele, B., Schindler, K. (eds.) DAGM 2010. LNCS, vol. 6376, pp. 272–282. Springer, Heidelberg (2010)
26. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. TPAMI (2002)
27. Prince, S.J., Elder, J.H.: Probabilistic linear discriminant analysis for inferences about identity. In: ICCV (2007)
28. Prince, S.J., Warrell, J., Elder, J.H., Felisberti, F.M.: Tied factor analysis for face recognition across large pose differences. TPAMI (2008)
29. Quinonero-Candela, J., Girard, A., Rasmussen, C.E.: Prediction at an Uncertain Input for Gaussian Processes and Relevance Vector Machines Application to Multiple-Step Ahead Time-Series Forecasting. IMM, Informatik og Matematisk Modelling, DTU (2003)
30. Rasmussen, C.E., Williams, C.K.: Gaussian processes for machine learning (2006)
31. Simonyan, K., Parkhi, O.M., Vedaldi, A., Zisserman, A.: Fisher vector faces in the wild. In: BMVC (2013)
32. Sun, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. arXiv preprint arXiv:1406.4773 (2014)
33. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: CVPR (2014)
34. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: CVPR (2014)
35. Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology) (1999)
36. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of cognitive neuroscience (1991)
37. Wang, X., Tang, X.: Bayesian face recognition using gabor features. In: ACM SIGMM Workshop on Biometrics Methods and Applications (2003)
38. Wang, X., Tang, X.: A unified framework for subspace face recognition. TPAMI (2004)
39. Wang, X., Tang, X.: Subspace analysis using random mixture models. In: CVPR (2005)
40. Wasserman, L.: All of nonparametric statistics. Springer (2006)
41. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. TPAMI (2009)
42. Yang, M., Zhang, L.: Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 448–461. Springer, Heidelberg (2010)
43. Zhou, Z., Wagner, A., Mobahi, H., Wright, J., Ma, Y.: Face recognition with contiguous occlusion using markov random fields. In: ICCV (2009)