Few Samples Face Recognition Based on Generative Score Space

Bin Wang[1*], Cungang Wang[2], Qian Zhang[1], Jifeng Huang[1]
[1] College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, 200234, China
[2]School of Computer Science, Liaocheng University, Liaocheng, 252000, China
[Email : wangbin527.2008@163.com]
*Corresponding author: Bin Wang

Received July 19, 2016; revised October 4, 2016; accepted October 31, 2016; published December 31, 2016

Abstract

Few samples face recognition has become a highly challenging task due to the limitation of available labeled samples. As two popular paradigms in face image representation, sparse component analysis is highly robust while parts-based paradigm is particularly flexible. In this paper, we propose a probabilistic generative model to incorporate the strengths of the two paradigms for face representation. This model finds a common spatial partition for given images and simultaneously learns a sparse component analysis model for each part of the partition. The two procedures are built into a probabilistic generative model. Then we derive the score function (i.e. feature mapping) from the generative score space. A similarity measure is defined over the derived score function for few samples face recognition. This model is driven by data and specifically good at representing face images. The derived generative score function and similarity measure encode information hidden in the data distribution. To validate the effectiveness of the proposed method, we perform few samples face recognition on two face datasets. The results show its advantages.

Keywords: Generative score space, face representation, few samples face recognition

1. Introduction

A number of approaches [1][2] have been proposed for human face representation towards recognition and synthesis. Among them, component analysis based approaches [1][3], e.g. two-dimensional principle component analysis (2DPCA) [1], are widely used in both recognition and analysis. Although various approaches, e.g., image descriptor based approaches [4][5] (such as Gabor and its variants [6]), have been proposed to extract the discrimination information of faces and show state-of-the-art performance, component analysis approaches are much more popular due to their abilities in both recognition and synthesis.

Component analysis approaches [7,8,9] are widely used for image representation, and achieve success in a variety of computer vision tasks [1][3], such as recognition, detection and visualization. Commonly, these approaches find a set of components such that they represent training data with the least error. As a specific class of component analysis, parts-based representation [10,11,12] is successful in a variety of vision tasks, especially good at representing human faces because faces are typically structured. These methods encourage the components to be some semantic parts of the training images, and represent each image using a combination of these parts. At the same time, recent researches have seen the success of sparse component analysis (SCA) [9,13] in vision problems. The basic idea of SCA is to learn a set of overcomplete components from the training data but activate as few as possible components to represent each data point, which has a solid physiologic foundation [9]. SCA is efficient in coding and robust to data variance. All the above methods model the observed data $\mathbf{x} \in R^D$ as the outputs of a random linear system $\mathbf{x} = W\mathbf{z} + \mu + \varepsilon$, where $W \in R^{D \times M}$ is a matrix composed of M components; $\mathbf{z} \in R^{M \times 1}$ is the vector of combination coefficients corresponding to M components; $\mu \in R^{D \times 1}$ is the vector of mean value and $\varepsilon \in R^{D \times 1}$ is the vector of random noise. Note that W and μ are determined to fit the training data.

Parts-based representation [10,11,12] encourages the components W to be sparse and expects the components to capture the semantic parts of given images. It represents an image in the part-by-part style, where each part is represented by a set of components like holistic methods (e.g. PCA [7]). Among them, multiple cause factor analysis (MCFA) [11] and structured sparse principle component analysis (SSPCA) [12] show attracting abilities, e.g., better factorization [11,12] than non-negative matrix factorization (NMF) [10]. MCFA learns a common spatial partition for given data, and models each part of the partition as a discrete state space or a linear space. This method is particularly good at representing aligned faces. However, it may result in non-continuous parts and its parts cannot benefit from the robustness of sparse component analysis. SSPCA is able to find convex areas from given images and could be straightforwardly scaled to several vision problems. For face representation, however, convex areas [12] are not usually flexible enough to capture the real face structures. Recently, [14] proposes a parts-based method (latent spaces with structured sparsity, LSSS) based on SSPCA, where the optimization is elegantly converted to two convex optimization problems. Also, [15] proposes a hierarchical parts-based model for human face parse, and [16] proposes a factored shapes and appearances (FSA) model for parts-based object recognition.

However, the above approaches didn't fully exploit the distribution information hidden in the data, which is demonstrated to be very useful in face recognition, especially when only a

few samples are available. Recently, a probabilistic branch of approaches, called generative score space, have received increasing attention. Generative score space is a class of principled approaches, which aim to integrate the abilities of discriminative models and probabilistic generative models for recognition. In these approaches, generative distribution can drive feature mappings over observed variables, hidden variables and model parameters while classification can be then performed in the derived feature spaces. The advantage of using generative score space to perform face recognition is that, score space is able to discover the information hidden in data by inferring hidden variables, which is additional and useful for few samples face recognition. Different from traditional generative score space approaches which utilize existing generative model (e.g. gaussian mixture model, hidden markov model), in this paper, we propose a new generative model to represent human face for few samples face recognition, i.e. generative score space based face recognition (GSSFR). Specifically, to integrate the flexibility of parts-based representation and the robustness of sparse component analysis, we built the two procedures into a probabilistic generative model and propose a parts-based sparse component analysis method to model human face. In the path from observed images to hidden variables, it first decomposes images to several parts and then learns a sparse component analysis model for each part. In the path from hidden variables to observed images, it generates an image pixel-by-pixel where, for each pixel, it first selects a part and then generates the pixel from the sparse component analysis model of the part. Then a feature mapping (i.e. score function) encoded hidden information is derived from the model. At last, a similarity measure is defined over the feature mapping to perform few samples face recognition. See **Fig. 1** for the illustration of our proposed approach. As shown in **Fig. 1**, the procedures of the proposed approach can be summarized as follows:

Step 1. We learn GSSFR model which incorporates the superiorities of both parts-based representation and sparse component analysis.

Step 2. We derive feature mapping for training set and test set respectively from the learned GSSFR via regularized inference.

Step 3. We define a similarity measure over the learned feature mappings. Face recognition can be then performed based on the defined similarity measure, i.e. encouraging the defined similarity to take a large value for a pair of images with the same label and to take a small value for a pair of images with distinct labels.

The advantages of our proposed approach for the task of few samples face recognition can be briefly summerized as follows:

(1) We propose a probabilistic generative model for face representation. The proposed model can learn continuous and flexible face parts.

(2) The derived score function from the proposed model is a function over observed variables, hidden variables and model parameters, which is informative for the few samples face recognition task.

(3) Our approach is able to exploit unlabeled data and works well with few labeled training data which is usually expensive to obtain.

The remainder of this paper is organized as follows. We first introduce related works in Section 2. Section 3 presents the generative score space based sparse component analysis model, and derives the feature mapping and similarity measure. Section 4 experimentally evaluates the model for few samples face recognition on two popular dataset. Section 5 draws a conclusion.

2. Preliminaries and Related Work

This section will briefly review some methods which are closely related to our approach, leaving other methods out.

Sparsity inducing priors. Different from the deterministic framework [13] that uses certain norms to induce sparsity, the probabilistic framework induces sparsity in terms of certain probabilistic priors [17,18][32,33], e.g. Laplace prior, Jeffrey prior and Inverse Gamma prior. The Laplace prior is the probabilistic implementation of l_1 norm, with the density function:

$$P(x;u,b)=\frac{1}{2b}\exp(-\frac{|x-u|}{b}) \tag{1}$$

where u is a location parameter and $b>0$ is a scale parameter. Letting $u=0$ and $b=1$, the objective function (i.e., log likelihood function) is $\|x\|_1$ which induces sparsity. The objective function of Jeffrey prior is $\log x$ that induces higher sparsity and constrains variables positive. Inverse Gamma prior allows to tune the sparsity degree through configuring its parameters.

We note that Multinomial prior also induces sparsity when the trail number is configured to $n=1$. Letting the variable be $\mathbf{r}=\{r_1,\cdots,r_k\}$, its distribution is:

$$P(\mathbf{r};n=1;\alpha)=\prod_{k=1}^{K}\alpha_k^{r_k} \tag{2}$$

where K is the number of probable events; r_k is the times of observing the k-th event in n trials and satisfies $\sum_k r_k=n$; α_k is the probability of observing the k-th event in a trial and satisfies $\sum_k \alpha_k=1$. For each sample drawn from $P(\mathbf{r};n=1;\alpha)$, only one element takes 1 and others take 0. For instance, when $K=4$, a possible sample is $\mathbf{r}=(0,1,0,0)^T$. This prior induces extreme sparsity.

Multiple Cause Factor Analysis (MCFA). For images, cause and factor correspond to image part and component respectively, where component shares the same definition with [13], i.e. linear basis. So we refer to them as part and component respectively. MCFA [11] models the image vector $\mathbf{x}\in\mathbf{R}^D$ as the output of a multiple cause model. To generate an image $\mathbf{x}$, for each pixel x_d, this model first selects a part (recorded by $r_{d.}$) and then generates the pixel using the linear model corresponding to the part, with the combination coeffcients $z_{k.}$. The joint distribution of this model is:

$$P(\mathbf{x},R,Z)=\prod_{d,k}N(x_d;\sum\nolimits_m w_{dkm}z_{km}+\mu_{dk},\sum\nolimits_{dk})^{r_{dk}}\prod_{k,m}N(z_{km};0,1)\prod_{d,k}\alpha_{dk}^{r_{dk}} \tag{3}$$

where d,k,m index pixel, part and component respectively; $Z=\{z_{km}\}_{km}$ is the coeffcient matrix for all $M\times K$ components; the vector $\mathbf{w}_{.km}=(w_{1km},\cdots,w_{Dkm})^T$ is the m-th component of the k-th part; μ_{dk} denotes the mean value and $\sum_{dk}$ denotes the variance of Gaussian

noise; $R=\{r_{dk}\}_{dk}$ is a set of binary indicators, where the binary indicator r_{dk} takes 1 if the d-th pixel belongs to the k-th part; $\alpha_{dk}=E[r_{dk}]\in[0,1]$ is the probability of the pixel d belonging to the part k.

Generally, there are three limitations for MCFA. First, MCFA does not exploit the distribution information hidden in the data, which is demonstrated to be very useful in few samples face recognition. Second, MCFA does not benefit from the advantages of sparse component analysis. Third, image parts learned by MCFA are usually discontinuous, which will potentially lead to the boundary-effect in synthesis. The reason accounting for the discontinuity is that the pixels are independently generated, without considering the dependence among neighbor pixels [11]. To overcome the three limitations, we will present an effcient model base on generative score space in Section 3.

3. Generative Score Space based Sparse Component Analysis

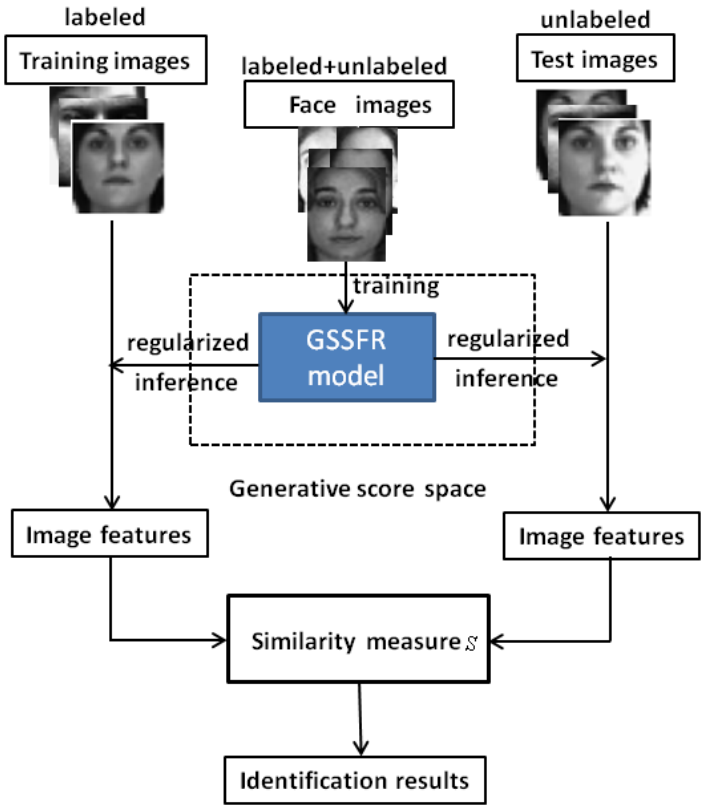


Fig. 1. The framework of the proposed approach.

In this section, we proceed to present the proposed model. Here we give a brief summary for the proposed method that comprises three main parts. First, train the generative model GSSFR using unlabeled data in an unsupervised style and derive the score function (feature mapping), as demonstrated in Sections 3.1, 3.2 and 3.3. Second, construct the similarity based on the trained model, as demonstrated in Section 3.4. Third, the similarity is embedded into a K-NN classifier for recognition. The comprehensive demonstration will be given in the experiments.

3.1 Model formulation

In this section, we first present the sparse component analysis model for each part and then glue the models for different parts together. Let D be the number of pixels of a face image, and $\mathbf{y}_k \in R^D$ be the k-th part of all K face parts. We model each part as a linear sparse component analysis model: $y_{dk} = \sum_{m=1}^{M} z_{km} w_{dkm}$, where M is the number of components; $\{\mathbf{w}_{.km}\}_{m=1}^{M}$ are the over-complete components for the k-th part; the coefficients $\{\mathrm{z}_{km}\}_{km}$ are encouraged to be sparse by the Laplace prior in Eq. (1).

The proposed model generates an image pixel-by-pixel. For the pixel x_d, it first selects a part from K ones, and then accordingly generates the pixel from the k-th sparse component analysis model. The selection is recorded by a random binary vector $\mathbf{r}_{d.} = (r_{d1}, \cdots, r_{dK})^T$ where $r_{dk} \in \{0,1\}$ follows Multinomial prior (Eq. (2)), and $r_{dk} = 1$ if the k-th part is selected. Then we have the model: $x_d = \sum_{k=1}^{K} r_{dk} y_{dk} + \mu_d + \varepsilon_d$, where y_{dk} is the d-th element of the k-th part defined above; μ_d denotes the mean value and $\varepsilon_d \sim N(0, \sigma_d^2)$ is the random noise; $\boldsymbol{\alpha}_{d.} = (\alpha_{d1}, \cdots, \alpha_{dK})^T$ is the parameter of Multinomial distribution. Then we have:

$$x_d = \sum_{k=1}^{K} r_{dk} \sum_{m=1}^{M} z_{km} w_{dkm} + \mu_d + \varepsilon_d \tag{4}$$

The conditional distribution over the observed variable $\mathbf{x}$ can be expressed as:

$$P(\mathbf{x} \mid R, Z) = \prod_d P(x_d \mid \mathbf{r}_{d.}, Z) = \prod_d N(x_d; \textstyle\sum_k r_{dk} \sum_m z_{km} w_{dkm} + \mu_d, \sigma_d^2) \tag{5}$$

The above conditional distribution is significantly different from that of MCFA (Eq. (3)). In MCFA, a pixel x_d is generated by a mixture of K independent Gaussian distributions, each of which corresponds to a part. When generating the pixel, one first selects a part and therefore a Gaussian from K ones. In the above model, a pixel x_d is generated by a single Gaussian distribution whose mean value is a function of the part index. When generating the pixel, one first selects a part and then determines the mean value of the Gaussian.

We also note that, in the above model, for a pixel x_d, K parts have K corresponding mean values but share only one variance parameter σ_d^2, which is different from MCFA where K parts have K variance parameters. This difference is important, since it makes the sampling distribution of z_{km} easy (see the sampling distribution in Eq. (9)).

3.2 Continuous and smooth partition

The proposed method seeks to partition faces into several parts. Such a part, e.g. mouth, is expected to be spatially continuous. Recall that the parts are inferred by selecting a part for each pixel, and the selection for the d-th pixel is recorded by the variable $\mathbf{r}_{d.}$ that follows the Multinomial prior. We refer to this prior as $P_1(\mathbf{r}_{d.})$. On the other hand, to learn

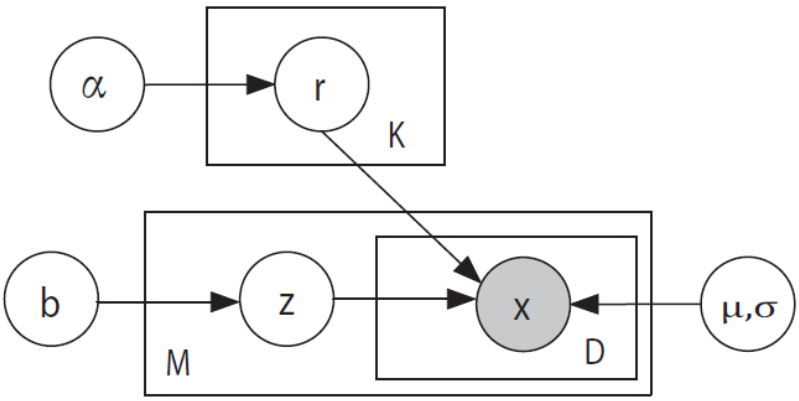


Fig. 2. Graphic illustration for our proposed approach.

continuous parts, we encourage that the selection for the pixel x_d (recorded by $\mathbf{r}_{d.}$) is continuous parts, we encourage that the selection for the pixel x_d (recorded by $\mathbf{r}_{d.}$) is consistent with the selections for its neighbors. We derive a continuity inducing prior over $\mathbf{r}_{.k} = (r_{1k}, \cdots, r_{Dk})$, and refer to this prior as $P_2(\mathbf{r}_{d.})$. The above two priors over $r_{d.}$ can be merged using the products of two experts [19] where each prior is an expert. That is $P(\mathbf{r}_{d.}) = \frac{1}{Z_d} P_1(\mathbf{r}_{d.}) \cdot P_2(\mathbf{r}_{d.})$, where Z_d is the partition function. Then the overall prior over $R = \{\mathbf{r}_{d.}\}$ is

$$P(R) = \prod_d P(\mathbf{r}_{d.}) = \prod_d \frac{1}{Z_d} P_1(\mathbf{r}_{d.}) \cdot P_2(\mathbf{r}_{d.}) \tag{6}$$

Considering $Z = \{z_{km}\}_{km}$ are encouraged to be sparse by the Laplace prior, we have

$$\begin{aligned} P(R,Z) &= P(R)P(Z) \\ &= \underbrace{\frac{1}{\prod_d Z_d} \prod_{d,k} \alpha_{dk}^{r_{dk}} \cdot \prod_{d,k} \exp\{-r_{dk}(1 - \frac{4}{|D_d|} \sum_{d' \in D_d} r_{d'k})\}}_{P(R)} \cdot \underbrace{\prod_{k,m} \frac{1}{2b} \exp(-\frac{|z_{km} - u|}{b})}_{P(Z)} \end{aligned} \tag{7}$$

where D_d is a set of neighbors of r_{dk}, and the number of the neighbors is set $|D_d| = 2$.

Note that the resulting prior is factorable according to the pixel d. This factorization is very important since it makes the learning and inference procedure tractable. Considering $P(\mathbf{x}|R,Z)$ (Eq.(5)), $P(R)$ (Eq.(7)), and $P(Z)$ (Eq.(7)), the joint distribution $P(\mathbf{x},R,Z) = P(\mathbf{x}|R,Z)P(R)P(Z)$ of the proposed model is

$$P(\mathbf{x},R,Z)=\underbrace{\prod_{d}N(x_d;\sum_k r_{dk}\sum_m z_{km}w_{dkm}+\mu_d,\sigma_d^2)}_{P(x|R,Z)}\cdot\underbrace{\prod_{k,m}\frac{1}{2b}\exp(-\frac{|z_{km}-u|}{b})}_{P(Z)}$$
$$\cdot\underbrace{\frac{1}{\prod_d Z_d}\prod_{d,k}\alpha_{dk}^{r_{dk}}\cdot\exp\{-r_{dk}(1-\sum_{d'\in D_d}r_{d'k})\}}_{P(R)} \tag{8}$$

where we set $b=1,u=0;\theta=(w_{dkm},\mu_d,\sigma_d,\alpha_{dk})$ is the parameter set to be learned. **Fig. 2** is the graphical illustration of this model. As shown in **Fig. 2**, r is parameterized by α and controls the generation of the part and z is parameterized by b and controls the generation of the sparse coefficient. r and z control the generation of image by means of a Gaussian distribution.

For face representation, the parameters and variables of our proposed approach can be intuitively interpreted. r_{dk} indicates whether the d-th pixel belongs to the k-th part, and $\alpha_{dk}=E[r_{dk}]$ is the prior probability that the d-th pixel belongs to the k-th face part. The parameter $\alpha_{.k}$ defines the shape of the k-th face part. $\mathbf{w}_{.km}$ is the m-th component of the k-th face part. $\mu.$ is the average face. Intuitively, if a random variable depends on a specific sample, it encodes the information related to the sample and is able to represent the sample. More specifically, for a sample, variables r_{dk} , z_{km} depend on the sample and are suffcient to reconstruct the sample.

3.3 Inference and learning

For the proposed model, it is difficult to estimate the posterior distribution of hidden variables using deterministic methods, such as variational inference [20]. As an alternative, we use Monto Carlo EM algorithm [21] to attack this problem. The method first draws the samples of hidden variables from the posterior sampling distribution using Gibbs sampling (E step) and then estimates model parameters using the drawn samples (M step).

E-step The Gibbs sampling distribution for r_{dk} can be formulated as:

$$P(r_{dk}|\mathbf{x},R_{-dk,}Z)=\frac{P(\mathbf{x},Z|R)P(r_{dk}|R_{-dk})P(R_{-dk})}{P(\mathbf{x},Z|R_{-dk})P(R_{-dk})}\propto P(\mathbf{x},Z|R)P(r_{dk}|R_{-dk})$$

where R_{-dk} denotes R but with r_{dk} omitted. Substituting Eq. (5) and Eq. (6) into the above formula, we have:

$$P(r_{dk}|\mathbf{x},R_{-dk,}Z)\propto N(x_d;\sum_k r_{dk}\sum_m z_{km}w_{dkm}+\mu_d,\sigma_d^2)\cdot\exp\left\{r_{dk}\left(\sum_{d'\in D_d}r_{d'k}-1\right)\right\}\alpha_{dk}^{r_{dk}} \tag{9}$$

Sampling from this distribution is very simple since r_{dk} is discrete. The sampling distribution for z_{km} can be similarly derived and takes the following form:

$$P(z_{km}\left|\mathbf{x},R,Z_{-km}\right.)\propto\begin{cases}N(z_{km};\mu_{zkm}^{+},\sigma_{zkm}^{2+}) & z_{km}\geq 0 \quad (P_{+})\\ N(z_{km};\mu_{zkm}^{-},\sigma_{zkm}^{2-}) & z_{km}\geq 0 \quad (P_{-})\end{cases} \tag{10}$$

where

$$\begin{aligned}&\mu_{zkm}^{+}=(-\sum_{d}r_{dk}w_{dkm}\varphi_{d}/\sigma_{d}^{2}-1)/\sigma_{zkm}^{2+},\ \mu_{zkm}^{-}=(-\sum_{d}r_{dk}w_{dkm}\varphi_{d}/\sigma_{d}^{2}+1)/\sigma_{zkm}^{2-}\\&\sigma_{zkm}^{2+}=\sigma_{zkm}^{2-}=1/[\sum_{d}(r_{dk}w_{dkm}/\sigma_{dk})^{2}],\ \varphi_{d}=\sum_{i\neq k}r_{di}\sum_{j\neq m}z_{ij}w_{dij}+\mu_{d}-x_{d}\end{aligned} \tag{11}$$

The resulting sampling distribution is discontinuous. However, we can effectively sample from it using a very simple yet effective strategy as follows: (1) sample from the distribution P_{+} in Eq.10 and output the non-negative samples ($z_{km}\geq 0$); (2) sample from the distribution P_{-} in Eq.11 and output the negative samples ($z_{km}\leq 0$); (3) combine the two set of samples as the output.

M-step Note that the above sampling procedures are performed for every sample $\mathbf{x}^{c}$. Let $\{R^{c,i}\}_{i=1}^{J}$ and $\{Z^{c,i}\}_{i=1}^{J}$ be the samples drawn from $P(R\left|\mathbf{x}^{c},Z\right.)$ and $P(Z\left|\mathbf{x}^{c},R\right.)$ respectively, then M step estimates model parameters by maximizing $\sum_{c=1}^{N}E\log P(\mathbf{x}^{c},R,Z)$ with respect to parameters.

Then we obtain the update rules for the parameters:

$$\begin{aligned}&\alpha_{dk}=\sum_{c}\left\langle r_{dk}^{c}\right\rangle/\sum_{c,k}\left\langle r_{dk}^{c}\right\rangle,\ w_{mkd}=\frac{\sum_{c}\left\langle r_{dk}^{c}z_{km}^{c}\right\rangle(x_{d}^{c}-\mu_{d}-\sum_{i\neq k,\,j\neq m}\left\langle r_{di}^{c}z_{ij}^{c}\right\rangle w_{dij})}{\sum_{c}\left\langle r_{dk}^{c}(z_{km}^{c})^{2}\right\rangle}\\&\mu_{d}=\frac{1}{N}\sum_{c}\left(x_{d}^{c}-\sum_{k,m}\left\langle r_{dk}^{c}z_{km}^{c}\right\rangle w_{dkm}\right),\ \sigma_{d}^{2}=\frac{1}{N}\sum_{c}\left\langle\left[\sum_{k,m}r_{dk}^{c}z_{km}^{c}w_{dkm}+\mu_{d}-x_{d}^{c}\right]^{2}\right\rangle\end{aligned} \tag{12}$$

where $\left\langle\cdot\right\rangle$ denotes the mean over samples, e.g., $\left\langle r_{dk}^{c}z_{km}^{c}\right\rangle=\frac{1}{J}\sum_{i}r_{dk}^{c,i}z_{km}^{c,i}$.

The overall algorithm is an iteration of the following two steps: (1) for each given training image $\mathbf{x}$, draw the samples of R and Z (Eq. (9)); (2) update $\alpha,\mathbf{w},\mu,\sigma$ using the drawn samples.

3.4 Score function derived from the proposed model

We use the proposed model for few sample face recognition, with a similarity measure. Note that the posteriors $P(Z^{c}\left|\mathbf{x}^{c}\right.)$ and $P(t^{c}\left|\mathbf{x}^{c}\right.)$ captures information associating with the sample $\mathbf{x}^{c}$. Although a number of methods [22,23] have been proposed to construct

and learn similarity measures, we define the similarity between the samples $\mathbf{x}^c$ and $\mathbf{x}^a$ as follows since it can naturally work on probabilistic generative models. We use the combination of projections $E_{P(Z^c|\mathbf{x}^c)}[Z^c]$ and $E_{P(R^c|\mathbf{x}^c)}[R^c]$ as the feature h^c to identify $\mathbf{x}^c$,

$$h^c = vec(\{E_{P(z_{km}^c|x^c,\theta)}[z_{km}^c], E_{P(r_{dk}^c|x^c,\theta)}[r_{dk}^c]\}_{k,m}) \tag{13}$$

Then the similarity is defined as the weighted inner product:

$$S_{GS}(\mathbf{x}^c, \mathbf{x}^a) = \mathbf{h}^{cT} X \mathbf{h}^a \tag{14}$$

where X is a diagonal matrix weighting face components and is determined by maximizing the inter-class distances. Both the distribution $P(Z^c, R^c|\mathbf{x}^c)$ and the projection $\mathbf{h}^c$ (Eq. (13)) are estimated from the samples of Z^c, R^c drawn by the Gibbs sampling. We refer to the similarity function as S_{GS} (inner product) throughout the remainder section.

4. Experimental Results

4.1 Few samples face recognition

Few samples face recognition is particular valuable in practice where usually only a few samples are available for each subject. We here introduce generative score space based sparse component analysis to attack the challenging problem. The points are three folds: (1) parts-based strategy is particularly good at representing face for its flexibility [24]; (2) the proposed method is able to learn the face representation from unlabeled face data unsupervisedly, which can derive score function (i.e. feature mapping) and similarity measure for the recognition task; (3) the derived feature mapping encodes information hidden in the human faces.

To validate our proposed approach can learn continuous and flexible face parts, we firstly learn face-parts priors. For face applications, such as face synthesis and smile learning, the learned face partition and resulting parts can be reused because all of faces present a comparable structure. Here we use the CBCL face database [27] to learn the face parts prior. For our methods and MCFA, the number of parts and components are set to be K=6 and $M = 200(< D)$ respectively. In fact, M in a wide range ($M \in [20, 240]$ empirically) works well and is able to learn a reasonable part prior. The learned parts of MCFA and our method are respectively shown in the top row and the bottom row of **Fig. 3**. The k-th column is the visualization of the part-defining parameter $\alpha_{\cdot k}$. The pixel value denotes the probability of the pixel belonging to the part. There is a remarkable difference between our model and MCFA: ours learns a soft segmentation (the parts are partially overlapped at the boundaries among parts) while MCFA learns a hard segmentation.

In the following section, we use the proposed model GSSFR for few samples face recognition. The proposed model will compare with the component analysis methods and other related methods. We use offline cross validation to choose the parameters, i.e.,

selecting the parameters such that the model trained from training set obtains the best performance on the test set. Then the chosen parameters are fixed for the rest experiments. To learn a set of overcomplete components, the number of components M in sparse component analysis is usually set to be a large value ($M > D$) [9]. However, we found that $M < D$ could produce satisfied results. So in the following experiments, for computational efficiency, we set $M < D$.

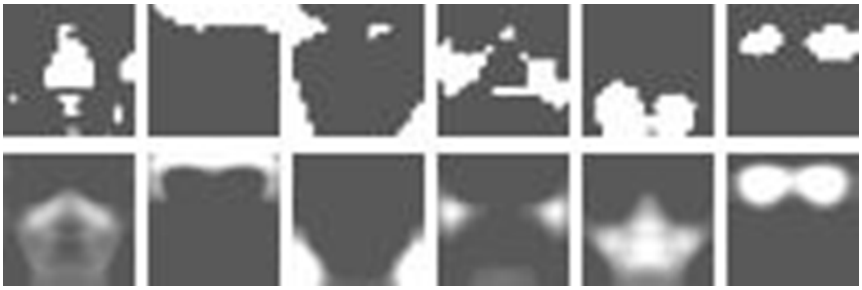


Fig. 3. Face parts learned by MCFA (top) and Ours (bottom). The pixel values denote the probabilities of pixels belonging to parts.

We evaluate the proposed method on two experiments: few samples face recognition on PIE database [25], and few samples face recognition on AR database [26]. The experimenting methods are summarized as follows:

Baseline. The baseline method is the nearest neighbor classification using the raw pixel, without utilizing the labeled data.

Eigenface [37]**.** The Eigenface method is based on linearly projecting the image space to a low dimensional feature space, and uses principal components analysis (PCA) for dimensionality reduction.

Fisherface [37]**.** The Fisherface method is based on Fisher's Linear Discriminant and produces well separated classes in a low-dimensional subspace, even under severe variation in lighting and facial expressions.

SSDA [28]. Semi-Supervised Discriminant Analysis using robust path-based similarity. It can utilize both labeled and unlabeled data to perform dimensionality reduction in the semi-supervised setting.

SGDA [29]. Semi-supervised Generalized Discriminant Analysis. It is an extension of generalized discriminant analysis (GDA) and utilizes unlabeled data to maximize an optimality criterion of GDA.

FE [30]. a Feature Encoding method. It models the distribution of data and derives features from the distribution.

GNMF [31]. Graph regularized Nonnegative Matrix Factorization. It seeks to find a compact representation, which uncovers the hidden semantics and simultaneously respects the intrinsic geometric structure.

DFD [4]. Discriminant Face Descriptor. It proposes a learning based discriminant face descriptor for face recognition .

MCFA [11]. Multiple Cause Factor Analysis. It is proposed for the unsupervised learning of parts-based representations of data.

SSPCA [12]. Structured Sparse Principle Component Analysis. It is an extension of sparse PCA , where the sparsity patterns of all dictionary elements are structured and constrained to belong to a pre-specified set of shapes.

KTPSL [36]. Kernel-based Transition Probability towards Similarity Learning for semi-supervised learning. It constructs similarities via comparing kernel least squares to variational least squares in the probabilistic framework.

PCSDA [38]. Pairwise Costs in Semisupervised Discriminant Analysis. It is a state-of-the-art semi-supervised approach for face recognition.

GSSFR. The proposed generatvie score space based face recognition approach, which presents a new generative model for face representation and derives feature mapping and similarity measure from the model for few samples face recognition.

For Eigenface, Fisherface, SSDA and SSPCA, we leverage the programs released by the authors or use the parameters suggested by authors. We implement the rest approaches according to the authors' suggestion. For the component analysis based methods, i.e., GNMF, MCFA, SSPCA and the proposed approach, we will select the parameters using offline cross validation and specify in the following experiments

4.1.1 Few samples face recognition on PIE database

The PIE face database [25] is used for evaluation of few samples face recognition for its high challenging. It contains 41, 368 face images which are captured from C = 68 individuals under illumination, varying lighting, and pose conditions. We choose the frontal pose (C27) for the recognition experiment, which is composed of about 49 face

Fig. 4. Examples of a subject with pose C27 in CMU PIE database.

images per individual [27]. Examples of a subject with pose C27 in CMU PIE database are shown in **Fig. 4**. All these images are cropped and normalized to 32 × 32 gray images. For each individual, we randomly select 30 images as the training dataset and the rest as the testing set. For each person in the training set, we randomly select $p \in \{2, 3, 4, 5\}$ images and label them, and remain the rest images unlabeled. The recognition phase assigns each unlabeled or test image to the nearest labeled sample via similarity function S_{GS}.

Table 1. Comparisons of recognition accuracy on PIE database. p is the number of labeled images randomly chosen from dataset in each round of experiment. The number of neighbors k used in K-NN classifier is determined by cross validataion, k = 1, 1, 2, 2 for p = 2, 3, 4, 5 respectively.

Method	$p = 2$		$p = 3$	
	Unlabeled	Test	Unlabeled	Test
Baseline	24.99±1.22	25.11±2.55	25.48±2.76	25.22±2.98
Eigenface[37]	21.91±2.46	20.98±2.93	21.98±3.25	22.00±1.92
Fisherface[37]	47.92±2.97	48.64±3.08	49.27±1.67	48.89±1.99
SSDA [28]	46.83±2.11	46.92±2.44	47.79±2.17	47.59±1.27
SGDA [29]	47.99±1.94	48.03±3.53	48.52±1.72	49.06±1.55
FE [30]	47.66±2.32	47.87±2.54	48.28±2.55	49.46±2.06
GNMF [31]	49.32±2.77	49.09±2.97	50.99±2.38	51.01±2.37
DFD [4]	29.55±1.76	28.89±2.87	32.28±2.56	33.99±2.09
MCFA [11]	35.07±2.86	34.98±1.24	39.21±3.16	40.12±3.09
KTPSL [36]	50.32±1.02	49.65±2.43	**56.15±1.09**	49.87±1.21
PCSDA [38]	50.96±1.56	50.78±2.13	54.72±2.43	55.66±1.13
SSPCA [12]	41.04±2.76	42.96±2.98	44.17±3.14	45.86±3.43
GSSFR	**51.84±2.00**	**51.56±2.43**	55.78±2.04	**57.16±1.89**
Method	$p = 4$		$p = 5$	
	Unlabeled	Test	Unlabeled	Test
Baseline	25.99±2.21	25.91±1.56	26.18±3.76	26.52±1.99
Eigenface	22.21±1.46	22.78±1.94	23.58±2.25	23.94±1.76
Fisherface	46.82±1.97	47.68±2.13	49.65±2.67	50.09±2.43
SSDA [28]	47.89±2.41	47.92±2.64	48.79±3.17	49.23±2.27
SGDA [29]	48.97±2.94	49.00±2.53	49.65±2.72	50.06±2.55
FE [30]	49.66±1.32	48.47±1.52	48.98±1.55	49.96±2.26
GNMF [31]	51.32±2.21	52.09±1.87	53.59±2.68	53.91±1.37
DFD [4]	34.55±2.76	38.89±3.87	39.58±1.56	40.19±3.092
MCFA [11]	42.05±1.82	44.98±2.24	45.21±2.16	47.12±2.19
KTPSL [36]	56.76±3.21	53.32±1.54	57.87±2.09	60.09±2.98
PCSDA [38]	58.65±1.29	59.91±1.08	61.38±2.03	66.28±1.71
SSPCA [12]	45.94±1.76	46.86±3.98	47.17±2.14	49.96±1.44
GSSFR	**59.94±2.10**	**61.16±1.42**	**62.78±2.04**	**67.25±2.39**

The proposed method will compare with the state-of-the-art methods. In each test round, we train the proposed model using all training images. Then, for each labeled training image and each test image, we extract the feature mapping (i.e. score space) using Eq. (13) and compute the similarity using Eq. (14). Having the face feature mapping and similarity, we use k-nearest neighbor classifiers for face recognition. We set k = 1, 1, 2, 2 for p = 2, 3, 4, 5 respectively. Note that, the number of neighbor k should be smaller than the number of labeled images p. In this experiment, k is determined by cross validation over a small test set. We test the approaches for 30 rounds and report the average results. We respectively set K = 6 and M = 200 according to cross validation for MCFA and SSPCA. The parameters of the proposed model is configured to K = 6 and M = 20 ($< N_S$) using offline cross validation.

The average recognition accuracies are reported in **Table 1**. We find that, compared with Baseline and Eigenface, all the other approaches obtain significant improvements. In particular, the performance of Fisherface is consistently superior to that of Eigenface. The reason is that Fisherface is insensitive to gross variation in lighting direction and facial expression. Meanwhile, our proposed approach GSSFR outperforms parts-based methods GNMF, MCFA and SSPCA due to the consideration of probabilistic modeling of image distribution. As an unsupervised method which only exploits unlabeled data, FE shows competitive performance with semi-supervised methods (SSDA, SGDA, DFD, KTPSL, PCSDA) exploiting both labeled and unlabeled data. Overall, our proposed approach GSSFR with the similarity function S_{GS}, as shown in **Table 1**, achieves the best performance among these compared approaches over both unlabeled dataset(>1.0%) and test dataset(>1.5%) in most cases . Intuitively, the ability of our method to utilize unlabeled data comes from the probabilistic generative model which encodes the distribution information or manifold structure, and the derived feature mapping is a function over observed variables, hidden variables and model parameters, which is informative for few samples face recognition task.

4.1.2 Few samples face recognition on AR database

In this section, AR dataset [26] is applied to few samples face recognition. To compare with [28] fairly, we follow its experimental setting where 100 persons (50 men and 50 women) are selected from all 126 persons (70 men and 56 women) for this experiment, just as are done in [28]. Totally, we have 2600 images (26 images are taken from each person with frontal view) under different expressions, illuminations and occlusions. All images are converted to gray images and normalized to 33 × 24. For each person, 13 images are randomly selected to train the model and the rest for testing. Among the 13 training images, we randomly select $p \in \{2, 3, 4, 5\}$ images and give them labels. For MCFA and SSPCA, the number of parts is set to K = 6, and the number of components is set to be M = 200 using cross validation. **Fig. 5** presents some examples of eight subjects from AR database.

Similar with the previous experiment, the number of the components is set to be M = 20. In each test round, we train the proposed model using all training images. Then, for each labeled training images and each test images, we extract the sore function using Eq. (10) and compute the similarity using Eq. (11). Having the face feature mapping and similarity, we use k-nearest neighbor classifiers and set k = 1, 1, 2, 2 for p = 2, 3, 4, 5 respectively for face recognition. We perform each experiment for 20 rounds and report the

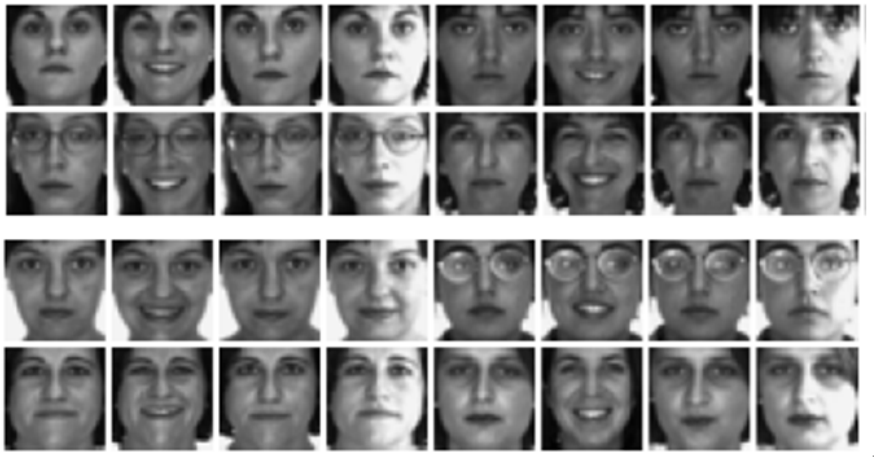


Fig. 5. Examples of eight subjets from AR database.

average results in **Table 2**. As shown in **Table 2**, Fisherface again gain a significant improvement over Eigenface. Similar to PIE dataset, semi-supervised learning approaches SSDA, SGDA, DFD, KTPSL, PCSDA show competitive performance. Unsupervised learning approach FE again obtains convincing results, which validates its ability to exploit unlabeled data. Our proposed method GSSFR outperforms other compared methods in most cases as good as on PIE dataset, which further validates its robustness to different databases. The reason accounting for this superiority is that the proposed approach can exploit information hidden in the data, which encodes high-level information especially useful in few sample face recognition.

4.2 Face recognition under random block occlusion

To further validate the effectiveness of our proposed approach on face representation, we design another experiment to perform face recognition under random block occlusion. The experiment is performed on face images chosen from AR dataset. Specifically, we choose images with two expressions: neutral and smile. We crop all the face images and normalize them to 32×32 gray images. 50 subjects are randomly chosen from 100 subjects to form training dataset. Afterwards, we make occlusions to the training set by means of setting the randomly chosen rectangles to be black. The size of the occlusion rectangles varies from 5×5 to 12×12. The compared methods include: Martinez's [34], which presents a probabilistic approach that is able to compensate for imprecisely localized, partially occluded and expression variant faces even when only one single training sample per class is available to the face recognition system; NRBM [35], which produces not only controllable decomposition of data into interpretable parts but also offers a way to estimate the intrinsic nonlinear dimensionality of data. For our method, we set $M=6$, $K=5$.

Table 2. Comparisons of recognition accuracy on AR database. p is the number of labeled images randomly chosen from dataset in each round of experiment. The number of neighbors k used in K-NN classifier is determined by cross validataion, k = 1, 1, 2, 2 for p = 2, 3, 4, 5 respectively.

Method	*p* = 2		*p* = 3	
	Unlabeled	Test	Unlabeled	Test
Baseline	14.35±1.19	14.83±1.20	18.90±1.35	19.20±1.44
Eigenface[37]	13.55±1.16	14.16±1.13	17.98±1.25	18.45±1.22
Fisherface[37]	47.41±1.92	48.12±1.73	52.29±2.65	52.96±2.29
SSDA [28]	57.32±3.85	58.16±3.63	71.80±2.36	71.51±2.17
SGDA [29]	59.04±3.64	58.92±3.45	71.07±2.71	72.76±2.53
FE [30]	56.67±2.93	56.89±2.86	68.82±2.54	68.56±2.36
GNMF [31]	58.36±2.77	58.64±2.61	70.72±2.38	70.51±2.27
DFD [4]	44.40±2.76	43.73±2.65	43.58±3.03	43.29±2.92
MCFA [11]	33.71±2.47	34.16±2.28	40.16±2.16	40.63±2.09
KTPSL [36]	58.61±1.11	45.98±1.52	71.32±1.97	70.87±2.21
PCSDA [38]	59.23±1.15	**59.78±1.09**	71.83±1.24	72.89±1.45
SSPCA [12]	39.63±2.69	40.07±2.42	47.15±2.10	47.71±2.03
GSSFR	**59.58±1.21**	59.61±1.55	**72.98±1.24**	**73.12±3.09**

Method	*p* = 4		*p* = 5	
	Unlabeled	Test	Unlabeled	Test
Baseline	23.16±1.06	23.05±1.19	26.62±1.30	26.46±1.36
Eigenface[37]	22.32±1.05	22.31±1.18	25.76±1.29	25.70±1.27
Fisherface[37]	69.79±1.34	69.86±1.52	67.91±1.88	67.50±1.50
SSDA [28]	70.13±2.13	72.54±2.49	75.71±1.42	76.55±1.23
SGDA [29]	73.22±2.26	74.53±2.55	76.27±1.29	77.12±1.55
FE [30]	71.24±1.86	72.12±2.55	71.58±2.31	73.15±2.45
GNMF [31]	72.26±2.34	73.33±2.41	74.15±2.33	75.78±2.38
DFD [4]	58.58±1.78	58.86±1.92	69.61±1.40	69.96±1.43
MCFA [11]	45.67±1.21	45.83±1.32	50.16±1.16	49.81±1.24
KTPSL [36]	70.32±1.71	76.55±2.03	77.56±1.88	75.89±1.42
PCSDA [38]	75.20±1.27	74.87±1.09	78.76±1.24	79.18±1.22
SSPCA [12]	53.07±1.18	53.28±1.27	58.37±1.12	58.82±1.19
GSSFR	**77.18±1.29**	**78.22±2.37**	**80.23±1.56**	**79.92±2.21**

Table 3. Comparison of recognition accuracy on face images from AR dataset with neural expression under random block occlusion

Method	Occluded region							
	5×5	6×6	7×7	8×8	9×9	10×10	11×11	12×12
Martinez's[34]	85.91	82.50	81.16	77.01	75.12	74.09	72.52	69.29
NRBM[35]	85.33	84.32	81.95	80.43	75.92	72.64	70.76	66.52
GSSFR	**86.54**	**85.72**	**83.55**	**80.87**	**77.21**	**76.54**	**73.48**	**71.27**

Table 4. Comparison of recognition accuracy on face images from AR dataset with smile expression under random block occlusion

Method	Occluded region							
	5×5	6×6	7×7	8×8	9×9	10×10	11×11	12×12
Martinez's[34]	76.10	74.37	67.28	65.09	63.51	61.93	60.04	58.73
NRBM[35]	76.65	75.23	74.81	**72.36**	66.72	61.27	57.17	58.15
GSSFR	**79.56**	**79.46**	**76.74**	72.15	**69.88**	**65.42**	**62.67**	**60.32**

The experimental results are shown in **Table 3** and **Table 4**. We find that, in most cases, our approach consistently outperforms Martinez's method and NRBM over the two test face expressions. With the increase of occlusion size, both Martinez's method and NRBM are no longer robust, especially when the occlusion size is large. Meanwhile, our proposed approach shows much more convincing performance since the parts-based sparse representation is robust to noise. In addition, we found that, training GSSFR model over the whole dataset could reach a higher performance, where the reason is that it could get a better-fitted model with lower training error. Of course, it can lead to more precise similarity measure. Although this route is discarded in our experiment because of unfairness in comparison, it is still valuable in the future work, e.g. exploiting more unlabeled data from other dataset.

4.3 Analysis of the experimental results of face recognition

In Section 4.1, we perform few samples face recognition on two popular face datasets. We compare the proposed approach with the state-of-the-art approaches. As shown in **Table 1** and **Table 2**, our approach presents convincing results. The reason accounting for its excellent performance is that, our approach incorporates the flexibility of parts-based representation and

the robustness of sparse representation scheme. We can learn more flexible parts. The derived similarity measure via regularized inference is a function over model parameter, hidden variable and observed variable, which encode high-level and more discriminative information useful in few samples face recognition. To further validate the effectiveness of the proposed approach on face recognition with random block occlusion, we perform another experiment in Section 4.2. We consider two kinds of face expressions: neural and smile. As shown in **Table 3** and **Table 4**, our approach is robust to facial expressions, even when the face image is occluded. Especially, our approach is much more robust for occlusion with large size than the other compared methods. The reason for this superiority is the robustness to noise of sparse representation.

5. Conclusion

In this paper, we present a probabilistic generative model to learn parts-based sparse component analysis for face representation, where the components of each part are treated as a group and are regularized to be sparse in spatial domain. It incorporates the flexibility of parts-based scheme and robustness of sparse component analysis. To perform few samples face recognition, we derive the generative score space (i.e. feature mapping) from the proposed model and a similarity measure is defined over the derived score space. Also, our proposed approach can exploit the data distribution, which is well adapted to data, and the derived feature mapping and similarity measure encode information hidden in the observed data and model parameters. The proposed model reaches convergence within 20 (50) iterations, while MCFA and SSPCA require about 100 iterations to reach the convergence. The convincing experimental results demonstrate the effectiveness of our proposed generative score space based approach for the few samples face recognition task. However, this method can further benefit from the exploiting of large dataset. Moreover, the computational efficiency is a main limitation of the training procedure, when the method scales to larger dataset. These works will leave in the future.

Acknowledgement

This work is supported by National Natural Science Foundation of China (Grant No.61503251), General Scientific Research Project of Shanghai Normal University(Grant No.SK201510) and Key Project of Scientific Research and Innovation of Shanghai Municipal Education Commission(Grant No. 1422125)

References

[1] Z. Cui, HongChang, ShiguangShan, BingpengMac, and XilinChen, “Joint sparse representation for video-based face recognition,” *Neurocomputing*, vol. 135, pp. 306–312, July, 2014. Article (CrossRef Link).

[2] H. Shim, “Probabilistic approach to realistic face synthesis with a single uncalibrated image,” *IEEE Trans. Image Processing*, vol. 21, no. 8, pp. 3784–3793, 2012. Article (CrossRef Link).

[3] Hu H, Klare B F, and Bonnen K, et al, “Matching Composite Sketches to Face Photos: A Component-Based Approach,” *IEEE Transactions on Information Forensics & Security*, vol. 8, no. 1, pp. 191-204, 2013. Article (CrossRef Link).

[4] Z. Lei, and S. Li, “Learning discriminant face descriptor for face recognition,” in *Proc. of Asian Conference on Computer Vision*, vol. 7725, no. 2, pp. 748-759, 2012. Article (CrossRef Link).

[5] N.-S. Vu, and A. Caplier, "Enhanced patterns of oriented edge magnitudes for face recognition and image match-ing," *IEEE Trans. on Image Processing*, vol. 21, no. 3, pp. 1352–1365, 2012. Article (CrossRef Link).
[6] Xie S, Shan S, and Chen X, et al, "Fusing local patterns of Gabor magnitude and phase for face recognition," *IEEE Transactions on Image Processing*, vol. 19, no. 5, pp. 1349-1361, 2010. Article (CrossRef Link).
[7] Jolliffe I T, "Principal Component Analysis," *Springer Berlin*, vol. 87, no. 100, pp. 41-64, 2015. Article (CrossRef Link).
[8] Westlake C, Mountain R W, and Paton T A L, "Factored Shapes and Appearances for Parts-based Object Understanding," in *Proc. of IEE - Part I: General*, vol. 101, no. 132, pp. 367-368, 2011. Article (CrossRef Link).
[9] B. Olshausen, and D. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997. Article (CrossRef Link).
[10] D. Lee, and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999. Article (CrossRef Link).
[11] D. Ross, and R. Zemel, "Learning parts-based representations of data," *The Journal of Machine Learning Research* , vol. 7, no. 3, pp. 2369-2397, 2006. Article (CrossRef Link).
[12] R. Jenatton, G. Obozinski, and F. Bach, "Structured sparse principal component analysis," in *Proc. of AISTATS*, vol. 9, no. 2, 131-160, 2009. Article (CrossRef Link).
[13] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265-286, 2006. Article (CrossRef Link).
[14] T. D. Y. Jia, and M Salzmann, "Factorized latent spaces with structured sparsity," in *Proc. of Neural Information Processing Systems*, pp. 982-990, 2011. Article (CrossRef Link).
[15] P. Luo, X. Wang, and X. Tang, "Hierarchical face parsing via deep learning," in *Proc. of Computer Vision and Pattern Recognition*, vol. 157, no. 10, pp. 2480-2487, 2012. Article (CrossRef Link).
[16] S. Eslami and C. Williams, "Factored shapes and appearances for parts-based object understanding," in *Proc. of British Machine Vision Conference*, vol. 101, no. 132, pp. 367-368, 2011. Article (CrossRef Link).
[17] Eisenstein J, Ahmed A, and Xing E P, "Sparse Additive Generative Models of Text," in *Proc. of International Conference on Machine Learning*, pp. 1041-1048, 2011. Article (CrossRef Link).
[18] C. Archambeau and F. Bach, "Sparse probabilistic projections," in *Proc. of Neural Information Processing Systems*, pp. 73-80, 2008. Article (CrossRef Link).
[19] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002. Article (CrossRef Link).
[20] Beal M J, "Variational algorithms for approximate Bayesian inference," *University College London*, 2015. Article (CrossRef Link).
[21] Levine, Richard A, and Casella, et al, "Implementations of the Monte Carlo EM Algorithm," *Journal of Computational & Graphical Statistics*, vol. 10, no. 10, pp. 422-439, 2001. Article (CrossRef Link).
[22] J. Li, J. Pan, and S. Chu, "Kernel class-wise locality preserving projection," *Information Sciences*, vol. 178, no. 7, pp. 1825–1835, 2008. Article (CrossRef Link).
[23] J. Li, J. Pan, and S. Chu, "Kernel self-optimized locality preserving discriminant analysis for feature extraction and recognition," *Neurocomputing*, vol. 74, no. 17, pp. 3019–3027, 2011. Article (CrossRef Link).
[24] Wright J, Yang A Y, and Ganesh A, et al. "Robust Face Recognition via Adaptive Sparse Representation," *Cybernetics IEEE Transactions on*, vol. 44, no. 12, pp. 2368 – 2378, 2014. Article (CrossRef Link).

[25] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database," in *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, Vol. 4, 2002. Article (CrossRef Link).

[26] A. Martinez and R. Benavente, "The AR face database," *CVC Technical Report*. Article (CrossRef Link).

[27] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proc. of International Conference on Computer Vision*, pp. 1–7, 2007. Article (CrossRef Link).

[28] Y. Zhang and D. Yeung, "Semi-supervised discriminant analysis using robust path-based similarity," in *Proc. of Computer Vision and Pattern Recognition*, pp. 1–8, 2008. Article (CrossRef Link).

[29] Y. Zhang and D.-Y. Yeung, "Semisupervised generalized discriminant analysis," *IEEE Trans. on Neural Net-works*, vol. 22, no. 8, pp. 1207–1217, 2011. Article (CrossRef Link).

[30] Chatfield K, Lempitsky V, and Vedaldi A, et al., "The devil is in the details: an evaluation of recent feature encoding methods," in *Proc. of British Machine Vision Conference*, vol. 76, pp. 76.1-76.12, 2011. Article (CrossRef Link).

[31] D. Cai, X. He, J. Han, and T. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548 –1560, 2011. Article (CrossRef Link).

[32] Jun B and Kim D, "Robust face detection using local gradient patterns and evidence accumulation," *Pattern Recognition*, vol. 45, no. 9, pp. 3304-3316, 2012. Article (CrossRef Link).

[33] Wen Y, Liu W, and Yang M, et al., "Regularized Robust Coding for Face Recognition," *Computer Science*, vol. 178, pp. 11-24, 2015. Article (CrossRef Link).

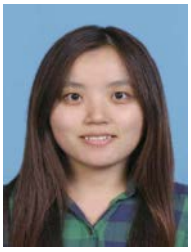
[34] Martinez A M, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 24, no. 6, pp. 748-763, 2010. Article (CrossRef Link).

[35] Tu D N, Tran T, and Phung D, et al. "Learning Parts-based Representations with Nonnegative Restricted Boltzmann Machine," in *Proc. of Asian Conference on Machine Learning*, pp. 133-148, 2013. Article (CrossRef Link).

[36] Kobayashi T, "Kernel-based transition probability toward similarity measure for semi-supervised learning," *Pattern Recognition*, vol. 47, no. 5, pp. 1994-2010, 2014. Article (CrossRef Link).

[37] Belhumeur P N, Hespanha J P, and Kriegman D J, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 19, no. 7, pp. 711-720, 1996. Article (CrossRef Link).

[38] Wan J, Yang M, and Gao Y, et al, "Pairwise Costs in Semisupervised Discriminant Analysis for Face Recognition," *IEEE Transactions on Information Forensics & Security*, vol. 10, no. 9, pp. 1569-1580, 2014. Article (CrossRef Link).



Bin Wang received her PhD degree from Department of Automation, Shanghai Jiao Tong University, Shanghai, China. She is now the lecturer of Shanghai normal University, China. Her research interests include computer vision, machine learning, image processing, multimedia analysis.

Cungang Wang received the B.E. degree in educational technology from Liaocheng University, Liaocheng, China, in 2000, and the M.S. degree in computer science and technology from Ocean University of China, Qingdao, China, in 2006. He worked as a leturer in the School of Computer Science at Liaocheng University since 2008. His research interests include machine learning, image processing, pattern recognition.

Qian Zhang is now the lecturer of Shanghai normal University, China. She received her P.HD. from Shanghai University in China. Her research interest fields include video processing

Jifeng Huang received the Ph.D. degree from East China University of Science and Technology, Shanghai, China, in 2005. He has been a professor in Shanghai Normal University from 2007. His research interests are in image processing and computer vision, especially in image processing.