Collaborative Random Faces-Guided Encoders for Pose-Invariant Face Representation Learning

Ming Shao, *Member, IEEE*, Yizhe Zhang, *Student Member, IEEE*, and Yun Fu, *Senior Member, IEEE*

***Abstract*— Learning discriminant face representation for pose-invariant face recognition has been identified as a critical issue in visual learning systems. The challenge lies in the drastic changes of facial appearances between the test face and the registered face. To that end, we propose a high-level feature learning framework called "collaborative random faces (RFs)-guided encoders" toward this problem. The contributions of this paper are three fold. First, we propose a novel supervised autoencoder that is able to capture the high-level identity feature despite of pose variations. Second, we enrich the identity features by replacing the target values of conventional autoencoders with random signals (RFs in this paper), which are unique for each subject under different poses. Third, we further improve the performance of the framework by incorporating deep convolutional neural network facial descriptors and linking discriminative identity features from different RFs for the augmented identity features. Finally, we conduct face identification experiments on Multi-PIE database, and face verification experiments on labeled faces in the wild and YouTube Face databases, where face recognition rate and verification accuracy with Receiver Operating Characteristic curves are rendered. In addition, discussions of model parameters and connections with the existing methods are provided. These experiments demonstrate that our learning system works fairly well on handling pose variations.**

***Index Terms*— Collaborative encoders, discriminant feature learning, face representation learning, pose-invariant feature, random faces (RFs).**

I. Introduction

Learning discriminant face representation for face recognitions has long been discussed in learning communities, under either controlled lab environment, or unrestricted environment [1]. Most of these applications run on 2-D facial images and acquire facial descriptors through certain learning systems [2]–[4]. By nature, such learning systems cannot avoid impact factors, such as expression, illumination, and pose. Among them, pose is extremely challenging as facial appearance is completely different between two poses, let alone other complex ones, e.g., nonrigid motions [5] and aging [6]. Although 3-D face recognition algorithms work well against pose variations, factors, such as expression, expensiveness, and slow data acquisition process, still limit their broad use [7].

To compensate for the pose and rigid/nonrigid motions, face alignment is first conducted as the standard preprocessing [1]. There are two kinds of face alignments: appearance level and feature level. Appearance level alignment usually warps the face to the frontal or designated pose based on the detected fiducial points, with or without 3-D models [8], [9]. Consequently, the learning systems only need to compare faces under the same pose. Differently, feature level alignment manages to explore a discriminative identity feature space, regardless of pose variations. A common strategy of feature level alignment is to project data to some data-driven latent feature space [10]–[13].

In this paper, following the line of feature level alignment, we propose a new pose-invariant discriminant identity feature and its learning system, inspired by the followed observations.

1) Facial features under different poses are transferable, in either linear or nonlinear way [10], [13]. For instance, we are able to map the side-view facial feature to the front-view by a linear transform function [10].
2) Faces have common structures although their identities are different. Thus, discriminative features should be able to model both common facial attributes and private ones.
3) Identity is unique for each individual; however, identity feature could be arbitrary vector mapped to the corrected identity as long as it is distinct in the feature space.

Inspired by the above-mentioned points, in this paper, we develop a new approach called "**R**andom **F**aces-guided **S**parse **M**any-to-one **E**ncoder" (RF-SME). The entire framework is shown in Fig. 1. First, a Single-hidden-layer Neural Network (SNN) is built to guide the identity feature learning. Essentially, it maps facial features under different poses to the unique one (many-to-one), i.e., frontal pose. Second, we enrich the identity feature and improve its discriminative power by replacing frontal faces with multiple random faces (RFs) for autoencoders. Thus, we can augment features by stacking hidden layers of multiple autoencoders and encode both common and private attributes. In addition, we incorporate within/between class constraint to RF-SME, and synchronize

Manuscript received March 24, 2016; revised July 1, 2016, October 21, 2016, and December 8, 2016; accepted December 11, 2016. Date of publication February 1, 2017; date of current version March 15, 2018. This work was supported in part by NSF IIS under Award 1651902, in part by NSF CNS under Award 1314484, in part by ONR under Award N00014-12-1-1028, in part by ONR Young Investigator under Award N00014-14-1-0484, and in part by U.S. Army Research Office Young Investigator under Award W911NF-14-1-0218.

M. Shao is with the Department of Computer and Information Science, University of Massachusetts Dartmouth, Dartmouth, MA 02747 USA (e-mail: mshao@umassd.edu).

Y. Zhang is with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: yizhe.zhang.190@nd.edu).

Y. Fu is with the Department of Electrical and Computer Engineering, College of the Computer and Information Science, Northeastern University, Boston, MA 02115 USA (e-mail: yunfu@ece.neu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2017.2648122

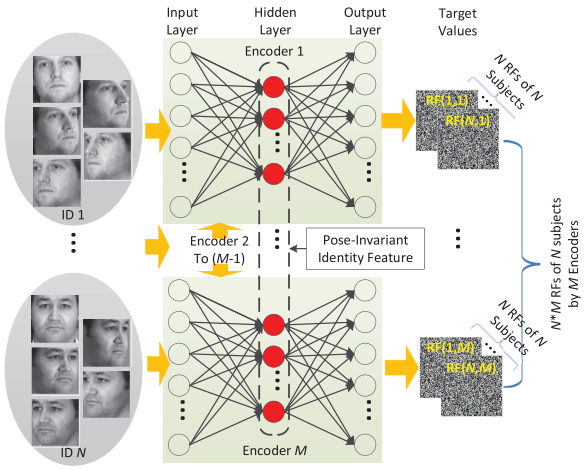


Fig. 1. Framework of "RFs-guided sparse many-to-one encoder." There are multiple faces under different poses for each "ID." They are the inputs of the single-hidden-layer neural network, while RFs are the target values of this network. Assume that we train M encoders and thus need M RFs for each unique ID. The concatenated nodes in hidden layers compose the pose-invariant feature (red nodes in the dashed area).

multiple autoencoders by enforcing the pairwise distances of the same subjects from different encoders to be close. We call it "**C**ollaborative RF-SME" (CRF-SME). Furthermore, we incorporate the powerful deep facial descriptors VGG-Face [57], and propose a new deep model "VGG-RF/CRF-SME." The contributions can be summarized as follows.

1) We design a novel "broad" representation learning framework using autoencoders as the building blocks for high-level pose-invariant features.
2) We propose to use random signals instead of frontal faces as the target values for feature augment, which significantly boost the performance.
3) Multiple autoencoders are synchronized to retain discriminant knowledge across different autoencoders, and deep convolutional neural network (CNN) facial descriptors are incorporated to boost the overall system performance.

This paper is an extension of our previous conference work [14], which details a competitive learning system for face representation learning. Essentially, we improve the RF-SME model by CRF-SME and VGG-RF/CRF-SME [by 7% on Multi-PIE and 14% on YouTube Face (YTF)], and conduct more experiments for demonstration. Specifically, we introduce discriminant analysis to the hidden layer and encourage the consistence between features guided by different RFs. In addition, we evaluate our model and competitive methods on YTF database. New results and illustrations about evaluations and model analysis can be found in Figs. 4, 5, 10, 11, 13, and 14, and Tables I, III, V, and VI.

II. Related Work

Holistic feature learning for face representation considers the whole face as the input and manipulate on all facial components without distinction [15]–[17]. However, they usually suffer from illumination or pose variations given limited training data. Besides, image filters, such as Gabor [18] and bioinspired features [19], are applied on face for robust features. On the other hand, local descriptors [20]–[22] collect the hand-craft codes from local patches and assemble them to formulate the final representation of each facial image. Recent advances in feature learning prefer statistical learning over the hand-craft fashion for more discriminative feature [23]–[25]. Our method follows this trend, but carries the semantics "many-to-one" that features the pose-invariant representations.

In addition, a group of pose specified 2-D face recognition algorithms have been proposed recently [5]. A straightforward way is to store or expand the poses of the registered faces to exhaustively cover all possible views of test faces [26], [27]. On high-level, features from different views are transformed to the ideal one, e.g., Tied Factor Analysis [10]. Multiview Discriminant Analysis (MvDA) [11] considers the discriminant information in an explicit way by which multiple view-specific transforms can be jointly learned. In addition, facial patches are explicitly considered in progressive face warping [28] with maximal intracorrelation [29] and probabilistic pose matching [30]. Recently, coupled latent space discriminative analysis (CLSDA) [12] considers latent space for face recognition under multiple poses. Besides, there are a large group of methods targeting at pose issues for faces in the wild [31]–[35]. Different from those using linear transforms, ours arguments the identity features through multiple nonlinear mappings for better performance.

The 3-D face model is able to simulate facial appearance from intended viewpoints [5]. Typical works include Morphable Face [8] and 3-D Generic Elastic Model [36]. Pose Normalization [37] invents a new matching scheme for each reference and test image, i.e., it renders a new virtual frontal face. Similarly, morphable displacement field (MDF) [38] considers virtual faces to match the gallery both globally and locally. Different from appearance level matching, pose adaptive filter [39] attempts to adapt filters in a fast manner according to the pose of the input images. Recently, a High-fidelity Pose and Expression Normalization method (HPEN) with the 3-D Morphable Model has been proposed and achieved the state-of-the-art performance on Multi-PIE data set [40]. In brief, the above-mentioned methods require an automatic 3-D face model fitting process given a 2-D facial image which may, however, easily be affected by illumination and expression changes.

TABLE I
Notations and Descriptions

Variable	Description
$x_{i,j}$	facial features of i-th subject, under j-th pose
$\hat{x}_i$	frontal face of the i-th subject
$W_{1/2}$	wight matrix of the first/second layer of autoencoder
$b_{1/2}$	bias term of the first/second layer of autoencoder
λ	weight decay parameter
y_i	output of the first layer of autoencoder
z_i	output of the second layer of autoencoder
a_i	hidden layer representation
r_i	random faces for the i-th subject
$h(\cdot)$	non-linear activation function
$t(\cdot)$	hypothesis of the autoencoder
D	input feature dimension
d	hidden layer feature dimension

Notably, face representation learning has attracted lots of attention from learning communities, and many dedicated learning systems have been developed, e.g., manifold/subspace learning for weakly labeled images and aging issues [6], [41], metric learning for human reidentification and face verification [42], sparse and representation learning [2], [43], and low-rank discriminative feature learning [3], [4].

Recently, deep learning [23], [24], [44] has shown superior performance on benchmark tests [45], [46]. Typical approaches for face recognition include deep belief networks [47], deep metric learning [48], hybrid deep models [49], face identity preserving (FIP) feature [50], Stack Progressive AutoEncoders [13], deep and low-rank modeling [51], [52], DeepID [53]–[55], DeepFace [56], VGG-Face [57], and FaceNet [58]. Most of them rely on CNNs to learn discriminant deep features against varied impact factors for faces in the wild. Most recently, a deep representation learning framework with target coding has been brought to our attention as it is similar to RFs procedure [59]. Different from it and other deep CNN-based face recognition framework, our method assembles multiple autoencoders under consistent constraints to construct a "broad" structure for high-level pose-invariant features, which could be a useful complement for existing facial descriptors.

III. Pose-Invariant Facial Feature Learning

We summarize variables and their descriptions in Table I.

A. Sparse Many-to-One Encoder

SME, built on an SNN, improves the traditional SNN learning scheme. In SME, we formulate a many-to-one mapping that enables SNN to extract identity preserved pose-invariant features. Specifically, "many" here means that the inputs of the SME are faces under different poses, while "one" means that the target values are frontal faces of the same identity as the inputs. We can see that SME encourages the outputs of the SNN being close to the frontal face of the same identity, in spite of poses of inputs. Next, we mathematically detail this procedure.

Assume that we have I different subjects, each of which is under J different poses. We denote $x_{i,j} \in \mathbb{R}^D$ as the input feature of the ith subject under the jth pose. In a typical neural network, the neuron in the hidden layer is essentially the weighted input plus a bias followed by a nonlinear activation. In our model, the neuron vector can be interpreted as a pose-invariant high-level representation. Formally, this "input→hidden layer" transform $f_1(\cdot)$ can be written in

$$a_{i,j} = f_1(x_{i,j}) = h(W_1 x_{i,j} + b_1) \tag{1}$$

where $W_1 \in \mathbb{R}^{d\times D}(d < D)$ is the weight matrix, $b_1 \in \mathbb{R}^d$ is the bias vector, $h(x) = (1+e^{-x})^{-1}$ is the nonlinear activation function, and $a_{i,j} \in \mathbb{R}^d$ is the hidden vector.

In conventional autoencoders, the output is the reconstruction of the input signal by linearly weighting the hidden neuron vector followed by another nonlinear activation. If we use $f_2(\cdot)$ for the "hidden layer→ouput" transform, then we have

$$t(x_{i,j}) = f_2(a_{i,j}) = h(W_2 a_{i,j} + b_2) \tag{2}$$

where $W_2 \in \mathbb{R}^{D\times d}$ is the weight matrix, $b_2 \in \mathbb{R}^D$ is the bias, and $t(x_{i,j})$ is the hypothesis generated by the autoencoder.

In traditional SNN or autoencoder, the target values are either meaningful labels, e.g., identity, object category, or identical values of the inputs. The objective function encourages hypothesis outputs to be as close to target values as possible. Different from them, we enforce the target values being close to the corresponding frontal facial features, i.e., $t(x_{i,j}) \approx \hat{x}_i$ where $\hat{x}_i$ represents the ith subject's frontal pose feature. As the output layer is generated by a feed-forward process using hidden layer as the input, the hidden layer can be seen as a basis as well as pose-invariant representation for the input.

By considering all training images, we formulate the objective function of the proposed SME as

$$\min_{W_1,b_1,W_2,b_2} \frac{1}{2N}\sum_{i,j} \|t(x_{i,j}) - \hat{x}_i\|_2^2 \tag{3}$$

where $N = I \times J$ is the number of training samples. Typically, this unconstrained optimization can be solved by gradient descent algorithms given hypothesis $t(\cdot)$.

However, the flexible nature of the proposed SME will easily make the model in (3) overfit. A typical solution is introducing a regularization term to the weight W. Inspired by the comparative study of regularizers for regression problem [60], we penalize the complexity of W_1 and W_2 via matrix l_1 norm and pursuit sparsity. Reasons are twofold. First, features are not equally important, especially for structure datalike face. Sparsity is able to select the most discriminative feature by zeroing out irrelevant factors. Second, it can avoid overfitting. Adding l_1 regularizers to (3), we have a new unconstrained optimization problem

$$\min_{\substack{W_1,W_2,\\ b_1,b_2}} \frac{1}{2N}\sum_{i,j} \|t(x_{i,j}) - \hat{x}_i\|_2^2 + \sum_{k=1}^{2} \lambda_k \|W_k\|_1 \tag{4}$$

where $\|W_k\|_1 = \sum_{ij} |[W_k]_{i,j}|$ is the sum of absolute values of elements in the matrix W_k, and λ_k is the weight decay parameter that suppresses the element values in W_k. In practice, this unconstrained optimization problem can be solved via Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (L-BFGS) optimizer [61], [62] that is able to approach large data with limited memory. More details about the optimizer and solutions can be found in [61] and the off-the-shelf package.[1]

After W_1, W_2, b_1, and b_2 are learned, we collect the hidden layer vector $a_{i,j}$ as the pose-invariant high-level feature for input $x_{i,j}$, and efficient classifiers, e.g., nearest neighbor classifier, can be used for final recognition task.

B. Random Faces

In a single SME model, we penalize the difference between the frontal facial feature of each subject and the output layer, and encourage $t(x_{i,j}) = \hat{x}_i$. This guides the output to approximate the frontal face regardless of the poses of input faces. However, the many-to-one mapping can only build a

[1] http://www.di.ens.fr/ mschmidt/Software/minFunc.html

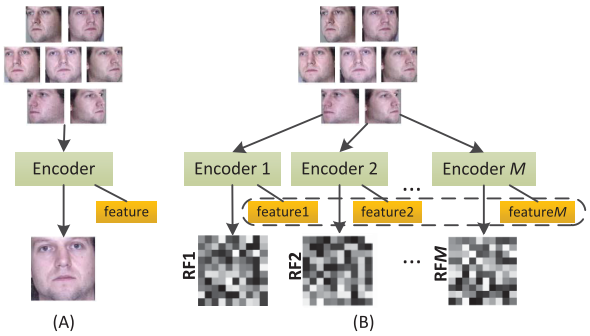


Fig. 2. Illustrations of RFs. In (a), we use a single frontal face as the target value, M RFs in (b) simulate the common and private attributes between different individuals by randomness. Therefore, feature $1-M$ in hidden layer would be discriminative representations for identity.

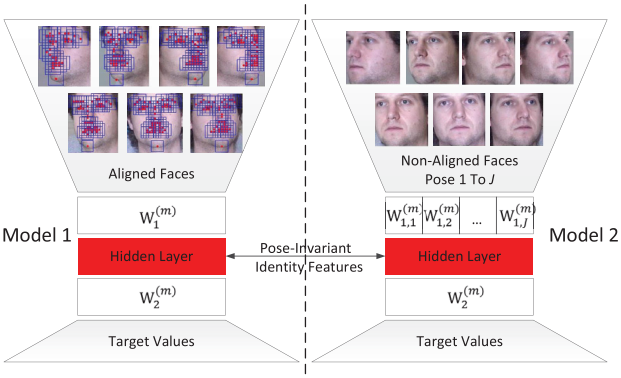


Fig. 3. Left: model-1 with a single $\{W_1^{(m)}, b_1^{(m)}\}$ pair learned by full-aligned faces. Right: model-2 with multiple $\{W_{1,j}^{(m)}, b_{1,j}^{(m)}\}, 1 \le j \le J$ pairs learned by nonaligned faces.

single model corresponding to the single frontal face. We are motivated to enrich the identity features by the following.

First, on the abstract level, the frontal facial feature of each subject used for the target value is simply a representation. Any unique matrix can work, not necessarily the frontal facial feature. Second, we can encode the identity information into many different "virtual" frontal faces rather than a "real" frontal face, by including both common and private attributes.

To that end, we propose to use random signals or RFs to replace the single frontal face as the target value in SME to augment the identity features, which is shown in Fig. 2. For the ith subject, we generate M RFs $r_i^{(m)} \in \mathbb{R}^D$, $1 \le m \le M$, where each single element in the vector $r_i^{(m)}$ is 0∼1 uniform distributed (independent and identically distributed). Thus, for two different RFs, they have extremely low probability being identical, which are ideal replacement for frontal faces. Notably, the randomized vector $r_i^{(m)}$ is not even close to faces in terms of appearance [Fig. 2(b) bottom part], but they function similar to frontal facial features in training the SME. For each element in the RFs, it is randomly encoded by either common attributes or private attributes, depending on its similarities to RFs for other identities. Then, the original reconstruction loss function in (3) becomes

$$\sum_{i,j} \frac{1}{2N} \left\| t^{(m)}(x_{i,j}) - r_i^{(m)} \right\|_2^2, \quad 1 \le m \le M. \tag{5}$$

Solving M such problems, we are able to obtain M groups of model parameters $\{W_k^{(m)}, b_k^{(m)}\}$. By stacking all the identity features vertically, $[a_{i,j}^{(1)}; a_{i,j}^{(2)}; \ldots; a_{i,j}^{(M)}]$, we obtain the RFs guided pose-invariant identity feature.

C. Full-Aligned Versus Nonaligned Faces

If faces are *aligned* already, then one encoder is able to map features from different poses to a unique one; otherwise, J different encoders will be learned for J different poses. This brings out two corresponding models (Fig. 3).

1) Full-Aligned Faces: For faces under different poses, if we select dense facial landmarks and extract features from the neighborhood of landmarks, i.e., small local patches on face, and then, the features have been aligned already. Still, we need frontal facial features or RFs to be the target values. We name it `Model-1` in this paper, which needs a single pair of $\{W_1^{(m)}, b_1^{(m)}\}$ to produce pose-invariant identity feature. Nonetheless, finding landmarks could be time-consuming and introduce many incorrect correspondences as well.

2) Nonaligned Faces: We may skip the face alignment step and learn different RF-SMEs for different poses. This essentially breaks down many-to-one mapping to a few one-to-one mappings. For example, inputs for training a specific pair of $W_{1,j}^{(m)}$ and $b_{1,j}^{(m)}$ will be facial features under the jth pose, and outputs are either features under frontal pose or RFs. For the test under arbitrary pose, we need to approximately estimate its pose first, and then encode it by an appropriate $\{W_{1,j}^{(m)}, b_{1,j}^{(m)}\}$. We call this `Model-2`.

D. Deep CNN Boosted RF-SME

While RF-SME has been proved powerful in our previous work [14], there are still a few issues that should be addressed. First, Model-1 relies on the alignment, and local correspondences to provide discriminant facial features. This step not only takes additional time, but also may introduce incorrect correspondences. Second, Model-2 usually trains more than one encoders depending on the poses/views in the problem. Therefore, it requires training samples under typical or representative views/poses. This may not be satisfied in a few real-world applications.

Recently, deep learning [23], [24], especially deep CNN has been widely and successfully applied to face recognition problem [40], [50], [53]–[58]. The learned deep facial descriptors are robust against pose, lighting, expressions, and generalized well to unseen face images. Therefore, deep CNN could work well, despite of missing correspondence or pose information in real-world applications, which offers a powerful substitution for Model-1 or Model-2.

Given a large amount of training data, e.g., CASIA-WebFace [63], we could learn the deep discriminant facial descriptors by following the well-established deep structure. In this paper, we use the open source deep CNN model released in [57] to build our deep model, which we call "VGG-RF/CRF-SME." The whole procedure is shown in Fig. 4. First, the nonaligned training data are fed to the pretrained deep CNN model. Second, we replace the last fully connected

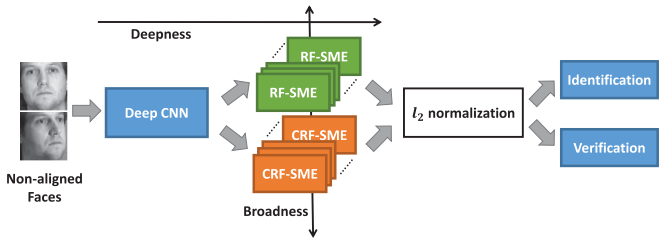


Fig. 4. Illustration of the proposed VGG-RF/CRF-SME. In this model, neither dense correspondence nor pose information is required. Raw images are first fed into deep CNN, and then passed to RF-SMEs or CRF-SMEs to extract pose-invariant features, followed by l_2 normalization. Finally, the normalized features are used for face problems.

layer with RF/CRF-SME and, then, freeze all the previous layers. Then, we can take advantage of deep structure for better pose-invariant feature learning. The learned features in the hidden layer before activation from different SMEs will be concatenated and, then, normalized for different face problems.

In brief, the VGG-RF/CRF-SME can be seen as an interesting practice with both broadness and deepness in the learning model, as shown in Fig. 4. With deepness, we could take advantage of the well-established deep CNN for robust facial descriptors learned from millions of face images, while with broadness, we could fulfill our many-to-one strategy and extract high-level pose-invariant features. The importance of both of them is demonstrated in the experiment sections.

IV. Collaborative Random Faces-Guided Encoders

In the RF-SME model, faces of each subject are enforced to map to certain RFs, but RFs for the same subject are generated independently. In other words, M RF-SMEs have weak connections as no constraint is imposed among them. In addition, within each SME, no explicit discriminant criterion is considered. To that end, we propose a novel CRF-SME model.

Recall that the originally proposed loss function of the mth RF-SME can be written as

$$E_1^{(m)} = \frac{1}{2N}\sum_{i=1}^{N}\left\|t^{(m)}(x_i) - r_i^{(m)}\right\|_2^2 + \sum_{k=1}^{2}\lambda_k\left\|W_k^{(m)}\right\|_1. \quad (6)$$

Note for the facility of later deductions, we skip the pose index j and follow the `Model-1` setting. This means r_i will be the paired RFs for a specific x_i. In addition, we introduce two intermediate variables

$$\begin{cases} y_i^{(m)} = W_1^{(m)}x_i + b_1^{(m)} \\ z_i^{(m)} = W_2^{(m)}a_i^{(m)} + b_2^{(m)}. \end{cases} \quad (7)$$

It can be checked that $a_i^{(m)} = h(y_i^{(m)})$, and $t^{(m)}(x_i) = h(z_i^{(m)})$. We will keep these notations through this paper.

A. Fisher Criterion and Feature Consistency

For face identification problem, a common criterion used for discriminative feature learning is to make samples of the same class as condense as possible while keep those from different classes as separated as possible, i.e., Fisher Criterion [15]. In the proposed RF-SME framework, this corresponds to enforcing the hidden units of each input to have such relations.

First, for within-class constraint of Fisher Criterion of the mth SME, we have

$$E_2^{(m)} = \frac{1}{2N}\sum_{i=1}^{N}\left\|a_i^{(m)} - \frac{1}{n_{y_i}}\sum_{y_j=y_i}a_j^{(m)}\right\|_2^2 \quad (8)$$

where y_i is the label of x_i, n_{y_i} is the number of samples in class y_i. Intuitively, $E_2^{(m)}$ measures the compactness of each class via the sum of square distance between each sample and its class center.

Second, for between-class constraint of Fisher Criterion of the mth RF-SME, we have

$$E_3^{(m)} = \frac{1}{2C}\sum_{c=1}^{C}\left\|\frac{1}{n_c}\sum_{y_i=c}a_i^{(m)} - \frac{1}{N}\sum_{j=1}^{N}a_j^{(m)}\right\|_2^2 \quad (9)$$

where C is the number of classes. Differently, $E_3^{(m)}$ should be as large as possible to make data separable.

On the other hand, we need to intentionally couple features from different RF-SMEs. Although each RF-SME is formulated and solved independently, the features generated by different RF-SMEs should keep the consistency, meaning pairwise relations from one RF-SME should be similar to those from another RF-SME. Suppose $d_{ij}^{(m)}$ and $d_{ij}^{(n)}$ are pairwise distances between features a_i and a_j generated by the mth and the nth RF-SME, then naturally, the following loss function should be minimized:

$$E_4^{(m)} = \frac{1}{2}\sum_{\substack{n \\ m<n}}\sum_{\substack{i,j \\ i\neq j}}\left(d_{ij}^{(m)} - d_{ij}^{(n)}\right)^2$$

where

$$d_{ij}^{(m)} - d_{ij}^{(n)} = \left(\left\|a_i^{(m)} - a_j^{(m)}\right\|_2 - \left\|a_i^{(n)} - a_j^{(n)}\right\|_2\right). \quad (10)$$

Therefore, we are ensured that the pairwise relations keep consistent across different features from different RF-SMEs.

Incorporating both the discriminative and pairwise distance consistency constraints of multiple RF-SMEs, we can build a new CRF-SME by minimizing the following loss function:

$$E = \sum_{m}\left(E_1^{(m)} + \omega_1\left(E_2^{(m)} - E_3^{(m)}\right) + \omega_2 E_4^{(m)}\right) \quad (11)$$

where ω_1 and ω_2 are two balancing parameters. Fig. 5 summarizes the loss functions of CRF-SME discussed earlier.

B. Solutions

The problem proposed in (11) is not convex, and to the best of our knowledge, it does not have a closed-form solution due to the sum operation over m. However, we could break down the entire problem into M subproblems and iteratively solve them one at a time.

It is easy to check that the loss functions of $E_2^{(m)}$, $E_3^{(m)}$, and $E_4^{(m)}$ in (11) are smooth and have a second-order derivative.

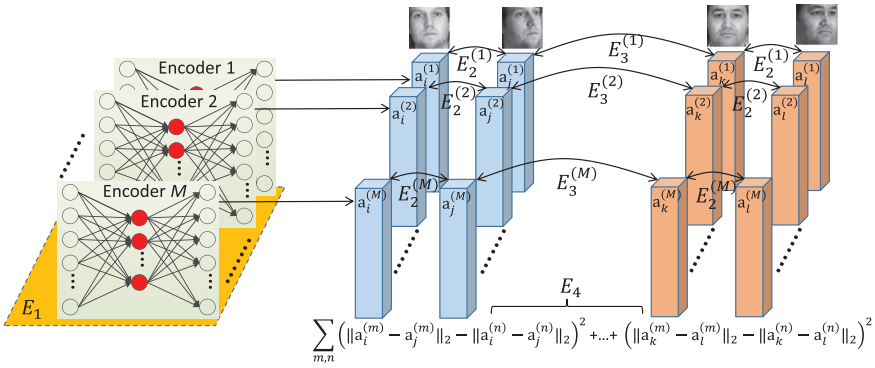


Fig. 5. Illustration of the CRF-SME model. We use the same color for feature vectors from the same subject. CRF-SME aims to maximize the distance between blue data and red data by maximizing $\sum_m E_3^{(m)}$, $1 \le m \le M$, while keep red (blue) data itself condense by minimizing $\sum_m E_2^{(m)}$. In addition, we maintain the pairwise relations of two features from two different SMEs formulated by E_4 to guarantee the data consistency across multiple SMEs.

Still, we are able to use L-BFGS optimizer for the solutions of this unconstrained optimization problem. Similar to quasi-Newton algorithms, we have the following updating rules at step t using L-BFGS optimizer:

$$\begin{cases} W_{t+1} = W_t - \alpha_t U_t \dfrac{\partial E}{\partial W}\Big|_{W_t} \\ b_{t+1} = b_t - \beta_t V_t \dfrac{\partial E}{\partial b}\Big|_{b_t} \end{cases} \tag{12}$$

for all $\{W_k^{(m)}, b_k^{(m)}\}$ pairs, where α_t and β_t are the learning rates, and U_t and V_t are the approximations for the inverse Hessian matrices of E with respect to W and b, respectively. For more details and discussions of α_t, β_t, U_t, and V_t, which are beyond the scope of this paper, readers can refer to work [61]. Here, we concentrate more on the derivatives of E with respect to W and b as they are affected by the Fisher Criterion and feature consistency constraints introduced in (8)–(10), and closely related to both learning rate $\{\alpha_t, \beta_t\}$, and inverse Hessian matrices approximations $\{U_t, V_t\}$.

The derivatives of $E_1^{(m)}$ have two different formulations depending on the layer k of $W_k^{(m)}$ and $b_k^{(m)}$

$$\begin{cases} \dfrac{\partial E_1^{(m)}}{\partial W_1^{(m)}} = \sum_i \delta_{i,1}^{(m)} \left(x_i^{(m)}\right)^\top + \lambda_1 \Theta_{W_1}^{(m)} \\ \dfrac{\partial E_1^{(m)}}{\partial W_2^{(m)}} = \sum_i \delta_{i,2}^{(m)} \left(a_i^{(m)}\right)^\top + \lambda_2 \Theta_{W_2}^{(m)} \end{cases} \tag{13}$$

and $(\partial E_1^{(m)}/\partial b_k^{(m)}) = \sum_i \delta_{i,k}^{(m)}$, where

$$\begin{cases} \delta_{i,1}^{(m)} = \left(W_2^{(m)}\right)^\top \delta_{i,2}^{(m)} \otimes h'\left(y_i^{(m)}\right) \\ \delta_{i,2}^{(m)} = \left(t^{(m)}(x_i) - r_i^{(m)}\right) \otimes h'\left(z_i^{(m)}\right) \end{cases} \tag{14}$$

$\otimes$ is the elementwise multiplication and $\Theta_{W_k}^{(m)}$ is an indicator matrix with element 1 for positive element and -1 for negative element in $W_k^{(m)}$.

For $E_2^{(m)}$, $E_3^{(m)}$, and $E_4^{(m)}$, since they are related to the constraints in the hidden layer, the derivatives are only computed in the hidden layer. For $E_2^{(m)}$, we have

$$\frac{\partial E_2^{(m)}}{\partial b_1^{(m)}} = \frac{1}{N}\sum_i \left(a_i^{(m)} \otimes h'\left(y_i^{(m)}\right) - \frac{1}{n_{y_i}} \sum_{y_j = y_i} a_j^{(m)} \otimes h'\left(y_j^{(m)}\right) \right) \tag{15}$$

and

$$\frac{\partial E_2^{(m)}}{\partial W_1^{(m)}} = \frac{1}{N}\sum_i \left(a_i^{(m)} \otimes h'\left(y_i^{(m)}\right) x_i^\top - \frac{1}{n_{y_i}} \sum_{y_j = y_i} a_j^{(m)} \otimes h'\left(y_j^{(m)}\right) x_j^\top \right). \tag{16}$$

Similarly, for $E_3^{(m)}$, we have

$$\frac{\partial E_3^{(m)}}{\partial b_1^{(m)}} = \frac{1}{C}\sum_{c=1}^{C} \left(\bar{a}_c^{(m)} \otimes h'\left(\bar{y}_c^{(m)}\right) - \bar{a}^{(m)} \otimes h'\left(\bar{y}^{(m)}\right) \right) \tag{17}$$

where $\bar{a}_c^{(m)}$ is the mean of hidden layer vectors from class c, $\bar{y}_c^{(m)}$ is the mean of vectors $y_i^{(m)}$ from class c, $\bar{a}^{(m)}$ is the mean of all hidden vectors, and $\bar{y}^{(m)}$ is the mean of all $y_i^{(m)}$ vectors. Then, the derivative with respect to W_1 can be written as

$$\frac{\partial E_3^{(m)}}{\partial W_1^{(m)}} = \frac{1}{C}\sum_{c=1}^{C} \left(\bar{a}_c^{(m)} \otimes h'\left(\bar{y}_c^{(m)}\right) \bar{x}_c^\top - \bar{a}^{(m)} \otimes h'\left(\bar{y}^{(m)}\right) \bar{x}^\top \right). \tag{18}$$

For $E_4^{(m)}$, we have a slightly complex derivative since we have two intermediate derivative corresponding to $a_i^{(m)}$ and $a_j^{(m)}$, respectively

$$\frac{\partial E_4^{(m)}}{\partial W_1^{(m)}} = \sum_{\substack{i,j \\ m<n}} \left(\nabla d_{ij}^{(m)} \otimes h'\left(y_i^{(m)}\right) x_i^\top - \nabla d_{ij}^{(m)} \otimes h'\left(y_j^{(m)}\right) x_j^\top \right) \tag{19}$$

Algorithm 1 Gradient Descent Algorithm for CRF-SME

Input: Training samples $x_1, x_2, \ldots, x_N$, random faces $r_1^{(1)}, \ldots, r_N^{(1)}, \ldots, r_1^{(M)}, \ldots, r_N^{(M)}$, and model parameters $\lambda_1, \lambda_2, \omega_1, \omega_2$.

Output: Weights matrices and bias vectors $W_1^{(m)}, b_1^{(m)}, W_2^{(m)}, b_2^{(m)} (1 \le m \le M)$.

Initialization: random initialization for $W_1^{(m)}, b_1^{(m)}, W_2^{(m)}, b_2^{(m)} (1 \le m \le M)$ with the purpose of symmetry breaking.

repeat

  Feed-forward process, compute $a_i^{(m)}$ and $t^{(m)}(x_i), (1 \le i \le N, 1 \le m \le M)$ by Eq. (1), (2).

  for $m = 1$ **to** M **do**

    Back-propagation process, compute the gradient of E w.r.t. $W_1^{(m)}, b_1^{(m)}, W_2^{(m)}, b_2^{(m)}$ by Eq. (13)-(20).

    Update $W_1^{(m)}, b_1^{(m)}, W_2^{(m)}, b_2^{(m)}$ by Eq. (12).

  end

  Compute total loss E_t by Eq. (11)

  $t = t + 1$

until The variation of the objective function between two consecutive iterations is smaller than predefined threshold, i.e., $|E_t - E_{t+1}| < \varepsilon$;

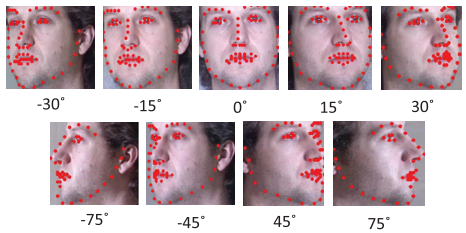


Fig. 6. Samples of landmark localization and across pose alignment on Multi-PIE database. In the second row, note that some landmarks are invisible. Identity features under full-aligned setting are extracted from local patches defined by the red landmarks.

where $\nabla d_{ij}^{(m)} = (a_i^{(m)} - a_j^{(m)}) - 2(a_i^{(n)} - a_j^{(n)})$ and

$$\frac{\partial E_4^{(m)}}{\partial b_1^{(m)}} = \sum_{\substack{i,j \\ m<n}} \left(\nabla d_{ij}^{(m)} \otimes h'\left(y_i^{(m)}\right) - \nabla d_{ij}^{(m)} \otimes h'\left(y_j^{(m)}\right)\right). \quad (20)$$

Finally, we keep updating $\{W_k^{(m)}, b_k^{(m)}\}$ until the change of total loss is less than a predefined threshold ε, i.e., $|E_{t+1} - E_t| < \varepsilon$. We summarize the whole procedure in Algorithm 1.

V. Experiments

A. Face Identification Results

We consider using Multi-PIE [64] database for face identification, and compare the proposed models with others. In full-aligned setting, we use the face alignment model proposed in [65] for landmark localization,[2] as shown in Fig. 6. In nonaligned setting, faces are manually cropped based on the `boundary` of the face, and then resized to the size of 128×128.

1) Pose Range: It can be learned from Fig. 6 that some face landmarks become invisible if the pose angle goes beyond 45°. Therefore, in full-aligned setting, only pose angles in $[-45°, 45°]$ are considered, while in nonaligned setting, all poses in $[-75°, 75°]$ are included.

2) Model Parameters: In full-aligned setting, no pose information is needed in both training and test phases, and we have two pairs of parameters: $\{W_1^{(m)}, b_1^{(m)}\}$ and $\{W_2^{(m)}, b_2^{(m)}\}$ learned from the proposed SME. In nonaligned setting, however, we need to learn a pair of parameters $\{W_{1,j}^{(m)}, b_{1,j}^{(m)}\}$ for each pose using either holistic or local features. To obtain approximate pose in the nonaligned setting, we use three key points: two eyes and nose tip detected by [65] to estimate the pose of test faces. And, therefore, we could choose appropriate model from the set of $\{W_{1,j}^{(m)}, b_{1,j}^{(m)}\}$.

In our evaluations, both ω_1 and ω_2 are set to 0.001, and λ_1 and λ_2 are set to 0.0001. In most of the cases, Algorithm 1 will converge in 100 iterations.[3] Note that the data consistency constraint is time-consuming since it will look through all possible pairs from different RF-SMEs. To alleviate the computational burden and meanwhile keep the comparable performance, we randomly sample a few pairs from all possible combinations of RF-SMEs and feature vectors. For the initialization of $\{W_k^{(m)}, b_k^{(m)}\}$, we refer to the approach suggested in [68], which empirically shows better performance. For more details about model parameters, please refer to Section V-D.

3) Notations: We use Model-1/2 + SME to indicate the SME model with corresponding frontal face feature as the output, and use Model-1/2 + RF/CRF + SME(·) to indicate the RF-SME, followed by the number of RFs in the bracket in Tables II and III. In addition, the deep CNN boosted RF/CRF models are represented by VGG-RF/CRF-SME. Both the results in our previous work [14] and in this paper are highlighted throughout the experiments with **bold fonts**.

4) Running Environment: We experiment on an Intel i7 desktop with 16-GB memory. All codes are written in MATLAB and optimized by parallel computing toolbox. Furthermore, the unconstrained optimization problem in our model is solved by L-BFGS optimizer. For deep CNN facial descriptors, we adopt the VGG-Face model [57] and take fc7 as the facial descriptors.

5) Full-Aligned Faces: In full-aligned setting, faces of all subjects under neutral expressions and illumination from session 1 to session 4 are selected. Also, we extract facial features from local patches defined by the detected landmarks, and in total, there are 52 different 20×20 patches which yields a 20 800-D feature vector to represent each face. We then use Whitening Principal Component Analysis (PCA) to further reduce it to 400-D. Unless otherwise stated, we set hidden layer size to 100, and the output layer size to 40. We choose the last 88 subjects' face images as the training data, and the rest from 249 subjects' as test data. Finally, we use the nearest neighbor classifier to predict the test image's identity. Note that we have two different registration methods here.

[2]It should be noted that to avoid trivial hand labeling work and meanwhile obtain high-quality landmarks for this setting, we use [65], which was trained by 900 Multi-PIE images. These landmarks, which should be provided before the training, however, will not affect our model training and its effectiveness.

[3]Note this is different from "epoch" used in many deep learning models.

TABLE II

FULL-ALIGNED FACES-BASED IDENTIFICATION RESULTS UNDER TWO DIFFERENT SETTINGS ON MULTI-PIE. THE 3-DGEM MODEL IS LEARNED ON USF HUMAN-ID DATABASE [66], INCLUDING 94 DIFFERENT 3-D FACE MODELS. NOTE 3-DGEM DID NOT CONSIDER EYEGLASSES IN MODEL LEARNING. IN THE RESULTS, "GLASSES" INDICATE THE ACCURACY OF THE ORIGINAL TEST SET (249 INDIVIDUALS) WITH EYEGLASSES ON FACE, WHILE "NO-GLASSES" INDICATE THE ACCURACY OF A SUBSET (158 INDIVIDUALS) OF THE ORIGINAL TEST SET WITHOUT EYEGLASSES. DURING THE TRAINING, WE TAKE EYEGLASSES INTO ACCOUNT, AND THE ACCURACY OF OUR MODEL IS EVALUATED ON THE ORIGINAL 249 PEOPLE'S TEST SET

Method/Degree		$-45°$	$-30°$	$-15°$	$0°$	$+15°$	$+30°$	$+45°$	Avg.
Setting-1	3DGEM+Glasses [36]	65.0%	86.7%	97.6%	N/A	93.2%	83.5%	65.0%	81.8%
	3DGEM+No-Glasses [36]	78.3%	92.2%	97.4%	N/A	93.5%	87.0%	83.1%	88.6%
	Raw-pixel	38.2%	55.8%	77.5%	N/A	57.0%	50.6%	43.4%	53.8%
	LBP [21]	82.3%	99.6%	100.0%	N/A	100.0%	98.9%	76.7%	92.9%
	HOG [22]	64.7%	94.8%	100.0%	N/A	100.0%	94.0%	64.7%	86.3%
	LDA [15]	92.4%	98.8%	98.8%	N/A	98.8%	96.8%	94.0%	96.6%
	Model-1+SME [14]	81.5%	93.2%	98.4%	N/A	96.8%	92.4%	88.8%	91.8%
	Model-1+RF-SME(20) [14]	96.8%	100.0%	100.0%	N/A	100.0%	100.0%	96.4%	98.8%
	Model-1+CRF-SME(20) (Ours)	99.0%	100.0%	100.0%	N/A	100.0%	100.0%	98.6%	99.7%
Setting-2	Raw-pixel	35.0%	46.8%	46.5%	42.6%	43.6%	35.2%	32.3%	40.3%
	HOG [22]	56.6%	69.0%	73.8%	72.1%	69.9%	63.9%	50.5%	65.1%
	LBP [21]	61.6%	75.9%	85.4%	89.2%	86.2%	74.9%	58.5%	76.0%
	LDA [15]	90.1%	95.6%	95.2%	95.0%	94.4%	93.6%	89.5%	93.4%
	Model-1+RF-SME(20) [14]	90.9%	97.7%	98.1%	98.5%	98.2%	95.7%	87.2%	95.2%
	Model-1+CRF-SME(20) (Ours)	91.6%	98.3%	98.5%	98.9%	98.7%	96.4%	88.1%	95.8%

TABLE III

NONALIGNED FACES-BASED IDENTIFICATION RESULTS UNDER THREE DIFFERENT SETTINGS ON MULTI-PIE

Method/Degree		$-75°$	$-45°$	$-30°$	$-15°$	$+15°$	$+30°$	$+45°$	$+75°$	Avg.
Setting-1	Raw-pixel	11.0%	10.0%	17.0%	36.0%	46.0%	21.0%	13.0%	11.0%	20.6%
	LBP [21]	4.0%	12.0%	24.0%	61.0%	57.0%	21.0%	13.0%	6.0%	24.8%
	HOG [22]	4.0%	10.0%	17.0%	71.0%	65.0%	18.0%	13.0%	6.0%	25.5%
	MvDA [11]	29.0%	55.0%	64.0%	70.0%	74.0%	62.0%	58.0%	43.0%	56.9%
	VGG-Face [57]	83.0%	90.0%	99.0%	100%	100%	99.0%	93.0%	84.0%	93.5%
	Model-2+SME [14]	57.0%	75.0%	79.0%	94.0%	92.0%	84.0%	78.0%	61.0%	77.5%
	Model-2+RF-SME(20) [14]	79.0%	88.0%	92.0%	97.0%	98.0%	96.0%	91.0%	80.0%	90.1%
	Model-1+VGG-CRF-SME(20) (Ours)	87.0%	95.0%	100%	100%	100%	100%	99.0%	89.0%	96.3%
Setting-2	CLSDA [12]	42.2%	84.4%	96.6%	99.2%	99.2%	96.2%	89.0%	47.7%	81.8%
	VGG-Face [57]	59.5%	86.1%	97.1%	100%	100%	99.6%	94.1%	62.0%	87.3%
	Model-2+RF-SME(20) [14]	50.6%	87.3%	97.9%	99.2%	99.2%	97.4%	91.9%	54.8%	84.8%
	Model-1+VGG-CRF-SME(20) (Ours)	63.3%	90.7%	98.3%	100%	100%	100%	95.4%	67.1%	89.4%
Setting-3	LE [67]	N/A	86.9%	95.5%	99.9%	99.7%	95.5%	81.8%	N/A	93.2%
	MDF [38]	N/A	84.7%	95.0%	99.3%	99.0%	92.9%	85.2%	N/A	92.7%
	SPAE [13]	N/A	84.9%	92.6%	96.3%	95.7%	94.3%	84.4%	N/A	91.4%
	VGG-Face [57]	N/A	89.1%	97.8%	100%	100%	100%	92.7%	N/A	96.6%
	FIP [50]	N/A	95.6%	98.5%	100%	99.3%	98.5%	97.8%	N/A	98.3%
	HPEN+LDA [40]	N/A	97.4%	99.5%	99.5%	99.7%	99.0%	96.7%	N/A	98.6%
	Model-2+RF-SME(20) [14]	N/A	92.3%	98.7%	99.7%	99.3%	98.3%	91.7%	N/A	96.7%
	Model-1+VGG-CRF-SME(20) (Ours)	N/A	99.3%	100%	100%	100%	100%	100%	N/A	99.9%

Setting-1 adopts frontal face from each subject, while **Setting-2** randomly chooses an image under arbitrary pose from each subject as the reference. Both are reported in Table II. Note that the results for **Setting-2** are averaged after 20 trials.

Besides, we render the virtual frontal faces generated by single sparse many-to-one encoder with frontal face as the output. Specifically, in full-aligned setting, we use the hypothesis output $t(x) = h(W_2a + b_2)$, where a is the pose-invariant feature, to recover the virtual frontal face of input x. We illustrate these virtual frontal faces in Fig. 7. These results show that our SME can map features in different formats to frontal face feature space regardless of pose, which also demonstrates the property of many-to-one of SME.

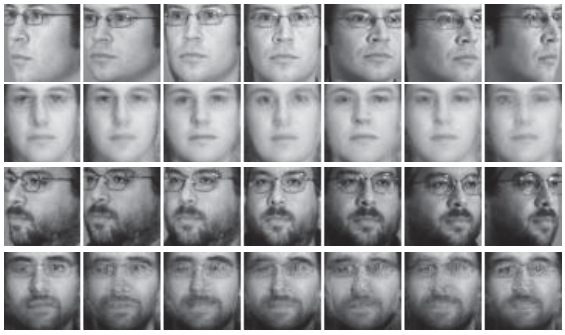


Fig. 7. Virtual frontal faces on Multi-PIE database (by Model-1). Odd rows: test faces. Even rows: virtual front faces by Model-1.

From Table II, we learn that the majority of them performs well due to face alignment. Classical facial features, such as Local Binary Patterns (LBP) and Linear Discriminant Analysis (LDA) work pretty well given only 2-D images, and most of the 2-D face recognition algorithms perform better than the 3-D method proposed in [36]. Still, we believe that accurate alignment plays a big role in the improvement. In general, we can see that our new method is superior to competitors

with face being aligned. On the other hand, we witness a performance drop in **Setting-2** of Table II, compared with the same method in **Setting-1**. These can be easily observed from the local descriptors-based methods, e.g., LBP and Histogram of Oriented Gradients. However, our methods still keep very competitive performance. Besides, it can be found that Model-1 with RFs (RF or CRF) performs better than a single SME, i.e., Model-1, which demonstrates the importance of RFs. Finally, the new model proposed in this paper consistently performs better than our previous conference work [14].

6) Nonaligned Faces: This setting has faces manually cropped to 128×128 according its boundary, and no landmarks are required. Therefore, no faces are aligned, and Model-2 is deployed to learn separated $\{W_{1,j}^{(m)}, b_{1,j}^{(m)}\}$ for different poses. Notably, the target values are set to a 2500-D random vector given raw facial images as inputs. There are three settings in nonaligned faces-based experiments. **Setting-1** takes the last 237 subjects' facial images in Multi-PIE as the training data, and the rest 100 subjects' facial images as the test data. **Setting-2** follows the configurations in [12] where the first 100 subjects' images are taken for model training and the rest for testing. In **Setting-3**, we follow the configuration in [13] and take the first 200 subjects' images from four sessions as training samples, and use the rest images for testing purpose.

From the Setting-1 results in Table III, we observe performance degradation from almost all the methods, except for ours. As our methods do not need alignment and dense landmarks, we also have profile $(-75°, 75°)$ in the evaluation. Instead of LDA [15], we compare with the recent state-of-the-art MvDA [11] (multiview LDA) Table III. In addition, in Setting-2 of Table III, we compare with CLSDA [12]. In both settings, our method works better even for 45° or 75° poses. This demonstrates that the proposed Model-2 can extract pose-invariant features without alignment or dense landmarks. Finally, we also compare with the recently published pose-invariant feature learning methods, e.g., LE [67], MDF [38], HPEN [40], and deep neural networks-based pose-invariant feature learning methods, e.g., FIP [50] and Stacked Progressive Auto-Encoders [11] in Setting-3. Results demonstrate the newly proposed CRF-SME, especially, the deep version, i.e., VGG-CRF-SME performs fairly well compared with these state-of-the-art works.

Note that we adopt Model-1 for deep facial descriptors, i.e., Model-1+VGG-CRF-SME in nonaligned setting in Table III as we believe that the deep facial descriptors can partially solve the pose issue. We can see that the deep facial descriptor VGG-Face performs well in all three settings, but not the best among the other competitors. However, with Model-1, the newly proposed VGG-CRF-SME performs best compared with the state-of-the-art works. Additionally, our new model in this paper performs consistently better than our conference work [14] in Table III. Notably, the proposed VGG-CRF-SME outperforms our previous conference work by 7% in Setting-1.

B. Verification Results on LFW

"Labeled Faces in the Wild" (LFW) [45] is a benchmark that evaluates face verification algorithms through real-world public figures facial images (Fig. 8). This database collects over 13 000 images of faces from the Internet and 1680 individuals with at least two face images. As our model requires the identity information of the training data, we follow the "unrestricted" protocol in this experiment. We use LBP and High-dimensional LBP (HLBP) features provided by [35] to evaluate our method. Whitening PCA is then applied to reduce the dimensionality of both features to 400. We run our evaluations following the tenfold cross validation protocol and pick individuals who have at least five facial images for training. The size of hidden layer is 100 and that of output layer is 40.

Fig. 8. Sample faces of two different people from LFW database, subject to pose, illumination, and expression variations.

TABLE IV
VERIFICATION ACCURACIES (MEAN + STD) ON LFW DATABASE USING UNRESTRICTED PROTOCOL

Method	Accuracy($\mu \pm S_E$)	External Data
LBP+Multiple One-Shot [69]	0.8517 ± 0.0061	No
LBP+PLDA [70]	0.8733 ± 0.0055	No
Combined Joint Bayesian [71]	0.9090 ± 0.0148	No
Fisher Vector Faces [72]	0.9303 ± 0.0105	No
HLBP+Joint Bayesian [35]	0.9318 ± 0.0107	No
LBP+RF-SME(20) [14]	0.8775 ± 0.0060	No
LBP+CRF-SME(20) (Ours)	0.8873 ± 0.0091	No
HLBP+RF-SME(20) [14]	0.9230 ± 0.0083	No
HLBP+CRF-SME(20) (Ours)	0.9328 ± 0.0106	No
HLBP+HPEN+3DMM [40]	0.9487 ± 0.0038	3D model
ConvNet-RBM [49]	0.9252 ± 0.0038	87K images
VGG-Face [57]	0.9684 ± 0.0017	2.6M images
DeepFace [56]	0.9725 ± 0.0081	4.4M images
VGG-CRF-SME(20) (Ours)	0.9793 ± 0.0019	2.6M images
DeepID2 [53], [54]	0.9915 ± 0.0013	
FaceNet + Alignment [58]	0.9963 ± 0.0009	200M images

During the training, we follow the pose grouping approach suggested in [69] to centralize faces according to their poses. Afterward, Model-1 is adopted to learn the pose-invariant discriminant features, and both training and test data are projected into the new feature space. Note that we normalize each feature vector to have unit length. To conduct face verification, we compute the similarity scores of each pair of faces from an RF/CRF-SME. All such scores from different RF/CRF-SMEs formulate a new feature vector that will be used in similarity computing. For VGG-CRF-SME, we use the cosine similarity for the verification purpose. The average verification rates over ten trials and Receiver Operating Characteristic (ROC) curves of both our methods and competitors are shown in Table IV and Fig. 9.

As images in LFW (Fig. 8) have arbitrary poses, expressions, and illuminations, it is quite challenging for both verification algorithms and human being. As our model is

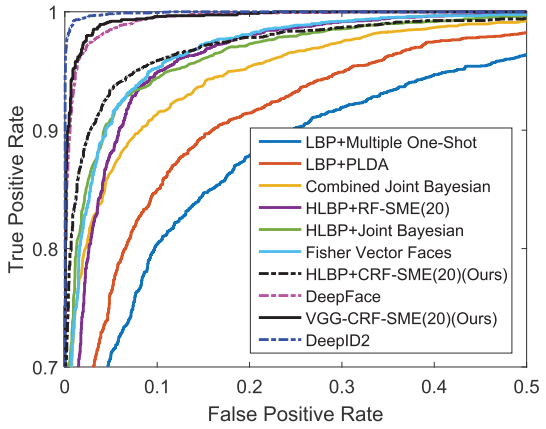


Fig. 9. ROC curves of face verification on LFW database using unrestricted protocol.

Fig. 10. Sample faces of four different people from YTF database subject to pose and expression variations.

compatible with both shallow and deep structure, we show the results of both with and without external training data. First, for shallow structure, it is clear that CRF-SME performs better than RF-SME on both LBP and HLBP features due to the Fisher Criterion and proposed data consistency constraint. In addition, we can see that the proposed CRF-SME model with HLBP feature performs better than state-of-the-art methods in Table IV. There are two reasons for the improvement: first, HLBP with dense landmarks provides robust facial features; second, our CRF-SME can take good advantage of these features (see our improvements over results in [35, Table IV]. Second, when external data are allowed, our VGG-CRF-SME model works comparably with those state-of-the-art deep models, e.g., DeepFace, and better than VGG-Face, as shown in the bottom part of Table IV. It should also be noted that the structure of our deep model is relatively simple, i.e., one deep CNN + 20 parallel SNNs, compared with 200 networks in DeepID2.

C. Verification Results on YTF

YTFs database [46] is a data set collected from YouTube videos targeting at the problem of unconstrained face verification in videos. The database collects 1595 people's 3425 videos. Like LFW database, most of them are public figures. In this database, each subject has an average of 2.15 videos and video clips have an average length of 181.3 frames. Key frames of four people are shown in Fig. 10.

Similar to LFW benchmark, YTF provides ten splits of video pairs for face verification purpose. Specifically, it randomly collects 5000 video pairs from the database. Half of these pairs include videos of the same person, while the rest half video pairs have different identities. In the benchmark

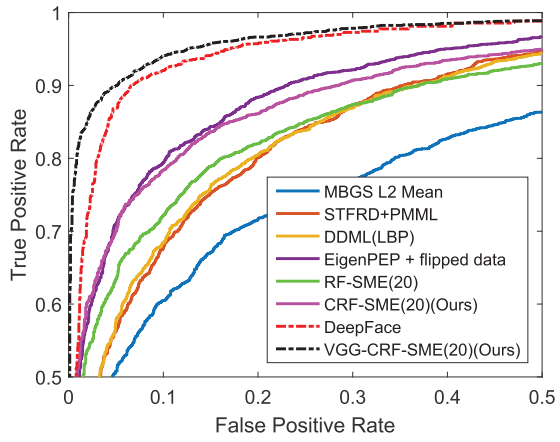


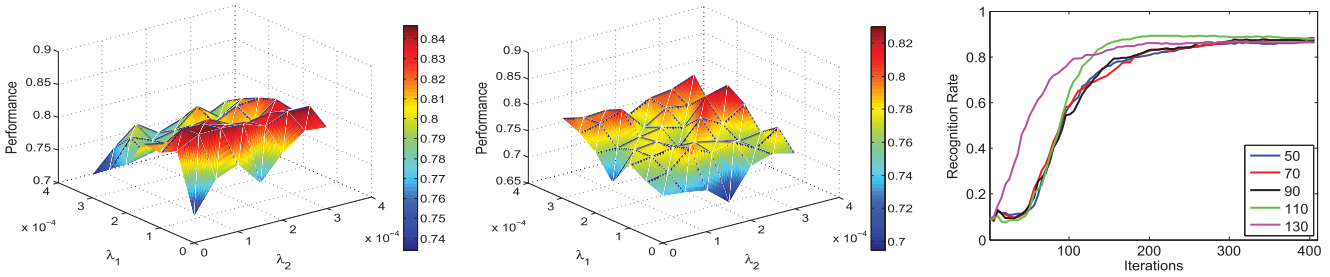
Fig. 11. ROC curves of face verification on YTF database.

TABLE V
VERIFICATION ACCURACIES (MEAN + STD) ON YTF DATABASE

Method	Accuracy($\mu \pm S_E$)	External Data
MBGS L2 mean (LBP) [46]	0.764 ± 0.018	No
STFRD+PMML [31]	0.795 ± 0.025	No
DDML (LBP) [48]	0.813 ± 0.016	No
DDML (Combined) [48]	0.823 ± 0.015	No
EigenPEP [73]	0.824 ± 0.017	No
EigenPEP+Flipped Data [73]	0.848 ± 0.014	No
LBP+RF-SME(20) [14]	0.818 ± 0.023	No
LBP+CRF-SME(20) (Ours)	0.839 ± 0.019	No
DeepFace [56]	0.914 ± 0.011	4.4M images
VGG-Face [57]	0.919 ± 0.015	2.6M images
DeepID2+ [55]	0.932 ± 0.020	
VGG-CRF-SME(20) (Ours)	0.938 ± 0.013	2.6M images
FaceNet + Alignment [58]	0.951 ± 0.004	200M images

test, each single split has 250 "same" and 250 "not-same" video pairs. We adopt the image restricted protocol and tenfold cross validation for evaluation. We use the LBP feature descriptors provided by [46] as our low-level features, and project the original LBP feature into a 400-D feature space before processed by RF/CRF-SME. We randomly select 40 images from each folder included in the training pairs and consider them as the inputs for each RF/CRF-SME. Other settings follow those in LFW and both accuracies and ROC curves are reported in Table V and Fig. 11, respectively.

From the experimental results, we can observe that YTF database is more challenging, due to the low-resolution images, and extreme lighting or facial expressions, as shown in Fig. 10. When there is no external data, compared with other state-of-the-art methods, our methods perform well. It should be noted that in [73], more training samples are created by flipped frames, which is not used in our methods. Without such process, the performance of EigenPEP drops down to 0.824 ± 0.017, which is inferior to ours. Given external training data, our method performs better than most of the deep models in Table V. The improvement of our method over VGG-Face proves that the proposed CRF-SME is useful to extract high-level pose-invariant face features. Notably, the proposed VGG-CRF-SME outperforms our previous conference work by 14%.



(a) l_1 norm (b) l_2 norm (c) Hidden layer size on Multi-PIE.

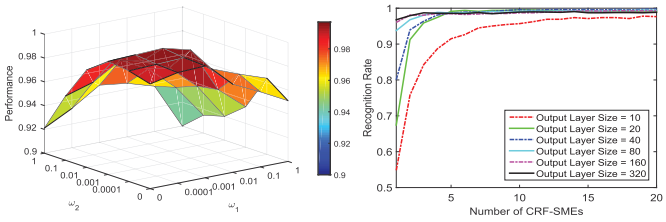
Fig. 12. Identification rates of l_1 (ours) and l_2 norms for a single SME over different values of λ_1 and λ_2 on Multi-PIE database. (a) l_1 norm. (b) l_2 norm. (c) Impacts of the hidden layer size.

D. Model Selection and Parameters Discussion

1) Weight Decay Parameters: We show the impacts of model parameters λ_1 and λ_2 and hidden layer size on model's performance. First, we show that how λ_1 and λ_2 will affect the performance in Fig. 12. Results are collected by Model-1 with single SME on Multi-PIE database following the Setting-1 of full-aligned faces. Clearly, different λ_1 and λ_2 values render different results, and their impacts on l_1 (our method) and l_2 matrix regularizers are shown in Fig. 12(a) and (b), respectively. In Fig. 12(a) and (b), the improvement by regularizers over the conventional methods ($\lambda_1 = \lambda_2 = 0$) is significant. Besides, l_1 norm performs slightly better than l_2 norm in our model. Although this does not mean to prove that l_1 matrix norm is superior in selecting discriminative features, it supports the claim l_1 regularizer empirically works well [60].

2) Hidden Layer Size: Hidden layer size of RF/CRF-SME is also of great importance. Intuitively, a large hidden layer size is easy to keep the intrinsic features, but would take longer time for training, while a small hidden layer size may suffer from inferior performance due to weak representation capability. To make this clear, we show the impact of layer size on identification task in Fig. 12(c). In this experiment, we use a single SME and follow the Setting-1 of full-aligned faces, and, therefore, are able to concentrate only on the size of hidden layer and performance. L-BFGS algorithm runs 400 iterations with different layer sizes, ranging from 50 to 140. In this experiment, layer sizes that are shown have different convergence speeds and larger hidden layer sizes converge faster. This can also compensate for the longer running time caused by the larger layer size.

3) Balancing Parameters: Recall in the loss function (11), we have two balancing parameters ω_1 and ω_2 for the terms $E_2^{(m)} - E_3^{(m)}$ and $E_4^{(m)}$, respectively, which can control the impacts of the two terms. A larger balancing parameter will suppress the corresponding loss function more, and vice versa. When both of them are set to zeros, CRF-SME model degenerates to RF-SME with only the first term in (12). To better illustrate their impacts, we show different performances of our model given different ω_1 and ω_2 values in Fig. 13(a). As we can see the range 0.001–0.0001 works well for both of them. A larger value of ω_1 will punish more on $E_2^{(m)} - E_3^{(m)}$, and may lead to overfitting, while a larger value of ω_2 will enforce different CRF-SMEs to be identical. Both of them lead to poor performance, as shown in Fig. 13(a).

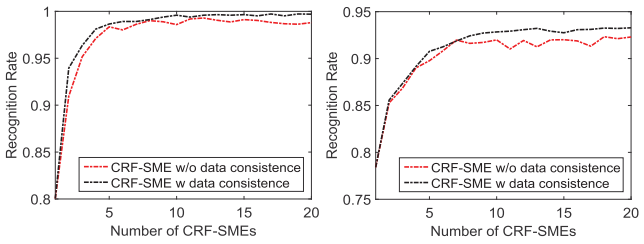


(a) Impacts of ω_1 and ω_2 (b) Output layer sizes

Fig. 13. Impacts of ω_1 and ω_2 on the performance (left) and output layer sizes of the CRF-SME model on Multi-PIE (full-aligned Setting-1) over different number of SMEs (right).

4) Output Layer Size: In addition, we discuss the sizes of the RFs in the output layer. Intuitively, small sizes of the RFs have weak capability on representation, but it promises to keeps good consistency between RFs. On the other hand, large sizes have better expressiveness, but may suffer from the weak consistency of identity features. To quantitatively measure the impacts of sizes of output layers on the system performance, we conduct another experiment on Multi-PIE database by fixing the size of the hidden layer at 100 and varying the sizes of output layer gradually from 10 to 320. In Fig. 13(b), we can see that the performance is affected in two ways. First, larger output sizes will yield better performance for the single SME which can be observed when the value of x-axis is equal to 1. In the meanwhile, larger number of CRF-SMEs results in better performance, but the improvements vary depending on the sizes of output layer. Clearly, small sizes of the output layer enjoy a significant improvement, while the larger sizes (e.g., 320) achieve a slight improvement and subject to a decrease at the later stage. Therefore, we can conclude that the moderate size of output layer is a key for better performance as well as data consistency.

5) Consistency: We also evaluate the importance of data consistency of our CRF-SME framework, and results are shown in Fig. 14. In this experiment, hidden layer and output layer sizes are set to 100 and 40, respectively, for both databases, and full-aligned faces with Setting-1 are used for Multi-PIE. From the results on two databases, we can observe that data consistency constraint helps to keep boosting the performance when the number of CRF-SME increases. Without such constraint, RF-SME suffers from unstable performance when the number of encoders is increasing. There



(a) Multi-PIE (b) LFW

Fig. 14. Illustration of data consistency. We show the importance of data consistency on Multi-PIE and LFW databases.

TABLE VI
TRAINING TIME COMPARISONS ON LFW DATABASE

Method	HLBP+JB	DDML	FV Faces	RF-SME	CRF-SME
Time (s)	6358.4	107.3	237.2	31.8	96.4

are some points where more RF-SMEs even ruin the system performance; however, CRF-SME with data consistency nearly increases the system performance monotonously.

6) Running Time: Finally, we compare the running time with competitive methods that do not relay on large-scale training data in Table VI. Specifically, we compare with HLBP+Joint Bayesian [35], Discriminative Deep Metric Learning [48], and Fisher Vector Faces [72]. From the results, we can see that the proposed RF-SME is very efficient, using the shortest training time. Since we add consistency constraint in CRF-SME, it needs more running time to get converged.

VI. CONCLUSION

In this paper, a novel collaborative face identity feature learning system robust to arbitrary poses had been proposed. First, a sparse many-to-one encoder was designed to mitigate negative factors incurred by arbitrary poses. Second, we invented a new target signal, i.e., RFs to enrich the pose-invariant features, which can be further boosted by the deep CNN facial descriptors. Since these encoders did not explicitly align themselves, we introduced a new learning criterion to hidden layers to enforce the data consistency. We evaluated our methods on both Multi-PIE and real-world face databases including variety of negative factors. Sufficient experiments validated the effectiveness and advantage of our learning system over state-of-the-art works.

REFERENCES

[1] A. K. Jain and S. Z. Li, *Handbook of Face Recognition*. New York, NY,USA: Springer-Verlag, 2005.
[2] C. Peng, X. Gao, N. Wang, D. Tao, X. Li, and J. Li, "Multiple representations-based face sketch–photo synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2201–2215, Nov. 2016.
[3] P. Zhou, Z. Lin, and C. Zhang, "Integrated low-rank-based discriminative feature learning for recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 5, pp. 1080–1093, May 2016.
[4] S. Li and Y. Fu, "Learning robust and discriminative subspace with low-rank constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2160–2173, Nov. 2016.
[5] X. Zhang and Y. Gao, "Face recognition across pose: A review," *Pattern Recognit.*, vol. 42, no. 11, pp. 2876–2896, Nov. 2009.
[6] D. Bouchaffra, "Nonlinear topological component analysis: Application to age-invariant face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 7, pp. 1375–1387, Jul. 2015.
[7] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition," *Comput. Vis. Image Understand.*, vol. 101, no. 1, pp. 1–15, Jan. 2006.
[8] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1063–1074, Sep. 2003.
[9] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.
[10] S. J. D. Prince, J. Warrell, J. H. Elder, and F. M. Felisberti, "Tied factor analysis for face recognition across large pose differences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 970–984, Jun. 2008.
[11] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 808–821.
[12] A. Sharma, M. A. Haj, J. Choi, L. S. Davis, and D. W. Jacobs, "Robust pose invariant face recognition using coupled latent space discriminant analysis," *Comput. Vis. Image Understand.*, vol. 116, no. 11, pp. 1095–1110, Nov. 2012.
[13] M. Kan, S. Shan, H. Chang, and X. Chen, "Stacked progressive auto-encoders (SPAE) for face recognition across poses," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1883–1890.
[14] Y. Zhang, M. Shao, E. K. Wong, and Y. Fu, "Random faces guided sparse many-to-one encoder for pose-invariant face recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2416–2423.
[15] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
[16] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2004.
[17] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
[18] C. Liu, "Gabor-based kernel PCA with fractional power polynomial models for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 572–581, May 2004.
[19] E. Meyers and L. Wolf, "Using biologically inspired features for face processing," *Int. J. Comput. Vis.*, vol. 76, no. 1, pp. 93–104, 2008.
[20] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[21] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, no. 1, pp. 886–893.
[23] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[24] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2007.
[25] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2011, pp. 215–223.
[26] D. J. Beymer, "Face recognition under varying pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 1994, pp. 756–761.
[27] R. Singh, M. Vatsa, A. Ross, and A. Noore, "A mosaicing scheme for pose-invariant face recognition," *IEEE Trans. Syst., Man, Cybern. B*, vol. 37, no. 5, pp. 1212–1225, Oct. 2007.
[28] A. B. Ashraf, S. Lucey, and T. Chen, "Learning patch correspondences for improved viewpoint invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[29] A. Li, S. Shan, X. Chen, and W. Gao, "Maximizing intra-individual correlations for face recognition across pose differences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 605–611.
[30] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, "Probabilistic elastic matching for pose variant face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3499–3506.

[31] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen, "Fusing robust face region descriptors via multiple metric learning for face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3554–3561.
[32] Q. Cao, Y. Ying, and P. Li, "Similarity metric learning for face recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2408–2415.
[33] X. Cao, D. Wipf, F. Wen, G. Duan, and J. Sun, "A practical transfer learning algorithm for face verification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3208–3215.
[34] Q. Yin, X. Tang, and J. Sun, "An associate-predict model for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 497–504.
[35] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3025–3032.
[36] U. Prabhu, J. Heo, and M. Savvides, "Unconstrained pose-invariant face recognition using 3D generic elastic models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 1952–1961, Oct. 2011.
[37] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. V. Rohith, "Fully automatic pose-invariant face recognition via 3d pose normalization," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 937–944.
[38] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan, "Morphable displacement field based image matching for face recognition across pose," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 102–115.
[39] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3539–3545.
[40] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 787–796.
[41] S. Xiao, D. Xu, and J. Wu, "Automatic face naming by learning discriminative affinity matrices from weakly labeled images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2440–2452, Oct. 2015.
[42] X. Xu, W. Li, and D. Xu, "Distance metric learning using privileged information for face verification and person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 12, pp. 3150–3162, Dec. 2015.
[43] R. He, W.-S. Zheng, B.-G. Hu, and X.-W. Kong, "Two-stage nonnegative sparse representation for large-scale face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 1, pp. 35–46, Jan. 2013.
[44] J. Tang, C. Deng, and G.-B. Huang, "Extreme learning machine for multilayer perceptron," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 809–821, Apr. 2016.
[45] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," College Inf. Comput. Sci. Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.
[46] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 529–534.
[47] G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2518–2525.
[48] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1875–1882.
[49] Y. Sun, X. Wang, and X. Tang, "Hybrid deep learning for face verification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1489–1496.
[50] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 113–120.
[51] M. Shao, Z. Ding, and Y. Fu, "Sparse low-rank fusion based deep features for missing modality face recognition," in *Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, vol. 1. May 2015, pp. 1–6.
[52] Z. Ding, M. Shao, and Y. Fu, "Deep robust encoder through locality preserving low-rank dictionary," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 567–582.
[53] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1891–1898.
[54] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2014, pp. 1988–1996.
[55] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2892–2900.
[56] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.
[57] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, vol. 1. Sep. 2015, p. 6.
[58] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.
[59] S. Yang, P. Luo, C. C. Loy, K. W. Shum, and X. Tang, "Deep representation learning with target coding," in *Proc. AAAI Conf. Artif. Intell.*, Jan. 2015, pp. 3848–3854.
[60] A. Y. Ng, "Feature selection, L_1 vs. L_2 regularization, and rotational invariance," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2004, p. 48.
[61] J. Nocedal, "Updating quasi-Newton matrices with limited storage," *Math. Comput.*, vol. 35, no. 151, pp. 773–782, 1980.
[62] J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, Q. V. Le, and A. Y. Ng, "On optimization methods for deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 265–272.
[63] D. Yi, Z. Lei, S. Liao, and S. Z. Li. (2014). "Learning face representation from scratch." [Online]. Available: https://arxiv.org/abs/1411.7923
[64] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.
[65] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2879–2886.
[66] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanID gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, Feb. 2005.
[67] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2707–2714.
[68] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, May 2010, pp. 249–256.
[69] Y. Taigman, L. Wolf, and T. Hassner, "Multiple one-shots for utilizing class label information," in *Proc. Brit. Mach. Vis. Conf.*, vol. 2. 2009, pp. 1–12.
[70] P. Li, Y. Fu, U. Mohammed, J. H. Elder, and S. J. D. Prince, "Probabilistic models for inference about identity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 144–157, Jan. 2012.
[71] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 566–579.
[72] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," in *Proc. Brit. Mach. Vis. Conf.*, vol. 1. 2013, pp. 1–12.
[73] H. Li, G. Hua, X. Shen, Z. Lin, and J. Brandt, "Eigen-PEP for video face recognition," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 17–33.

Ming Shao (S'11–M'16) received the B.E. degree in computer science, the B.S. degree in applied mathematics, and the M.E. degree in computer science from Beihang University, Beijing, China, in 2006, 2007, and 2010, respectively, and the Ph.D. degree in computer engineering from Northeastern University, Boston, MA, USA, in 2016.

He has been a Tenure-Track Assistant Professor with the College of Engineering, University of Massachusetts Dartmouth, Dartmouth, MA, USA, since 2016. His current research interests include sparse modeling, low-rank matrix analysis, deep learning, and applied machine learning on social media analytics.

Dr. Shao was a recipient of the Presidential Fellowship of the State University of New York at Buffalo from 2010 to 2012.

Yizhe Zhang (S'15) received the B.S. degree in computer science from Hohai University, Nanjing, China, in 2009, and the M.S. degree in computer science and engineering from the Polytechnic Institute of NYU, New York City, NY, USA, in 2012. He is currently pursuing the Ph.D. degree in computer science and engineering with the University of Notre Dame, Notre Dame, IN, USA.

His current research interests include medical image analysis and deep learning.

Yun Fu (S'07–M'08–SM'11) received the B.Eng. degree in information engineering and the M.Eng. degree in pattern recognition and intelligence systems from Xian Jiaotong University, China, respectively, and the M.S. degree in statistics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, respectively.

He has been an interdisciplinary Faculty Member with the College of Engineering and the College of Computer and Information Science, Northeastern University, Boston, MA, USA, since 2012. He has extensive publications in leading journals, books/book chapters, and international conferences/workshops. His current research interests include machine learning, computational intelligence, big data mining, computer vision, pattern recognition, and cyber-physical systems.

Dr. Fu is a fellow of IAPR, a Lifetime Senior Member of ACM and SPIE, a Lifetime Member of AAAI, OSA, and Institute of Mathematical Statistics, a member of Global Young Academy and INNS and a Beckman Graduate Fellow from 2007 to 2008. He serves as an Associate Editor, the Chair, a PC Member, and a Reviewer of many top journals and international conferences/workshops. He received seven Prestigious Young Investigator Awards from NAE, ONR, ARO, IEEE, INNS, UIUC, and Grainger Foundation; seven Best Paper Awards from IEEE, IAPR, SPIE, and SIAM; and three major Industrial Research Awards from Google, Samsung, and Adobe. He is currently an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEANING SYSTEMS.