Random Faces Guided Sparse Many-to-One Encoder for Pose-Invariant Face Recognition

Yizhe Zhang[1*] Ming Shao[2*] Edward K. Wong[1] Yun Fu[23]
[1]Department of Computer Science and Engineering, Polytechnic Institute of NYU, NY, USA
[2]College of Computer and Information Science, Northeastern University, MA, USA
[3]Department of Electrical and Computer Engineering, Northeastern University, MA, USA
zhangyizhe1987@gmail.com, mingshao@ccs.neu.edu, wong@poly.edu, yunfu@ece.neu.edu

Abstract

One of the most challenging task in face recognition is to identify people with varied poses. Namely, the test faces have significantly different poses compared with the registered faces. In this paper, we propose a high-level feature learning scheme to extract pose-invariant identity feature for face recognition. First, we build a single-hidden-layer neural network with sparse constraint, to extract pose-invariant feature in a supervised fashion. Second, we further enhance the discriminative capability of the proposed feature by using multiple random faces as the target values for multiple encoders. By enforcing the target values to be unique for input faces over different poses, the learned high-level feature that is represented by the neurons in the hidden layer is pose free and only relevant to the identity information. Finally, we conduct face identification on CMU Multi-PIE, and verification on Labeled Faces in the Wild (LFW) databases, where identification rank-1 accuracy and face verification accuracy with ROC curve are reported. These experiments demonstrate that our model is superior to other state-of-the-art approaches on handling pose variations.

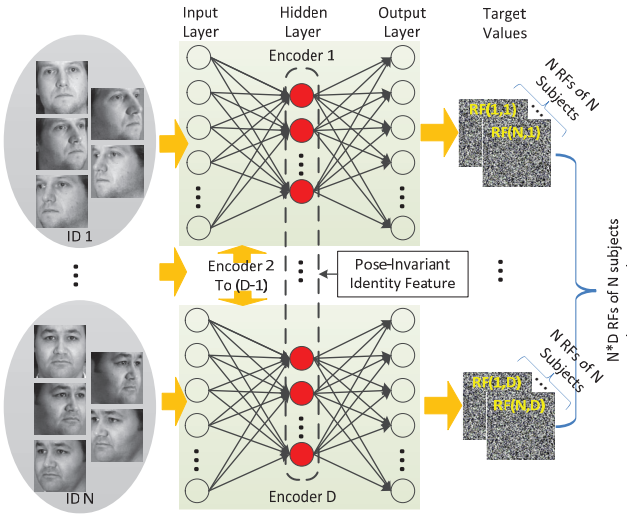

Figure 1. Framework of random faces guided sparse many-to-one encoder. Each unique "ID" has many facial images in different poses. We feed them into the single-hidden-layer neural network, i.e., the encoder, and set the target values to be random faces (RF). We design D encoders and therefore have D random faces for each ID. The concatenated nodes in hidden layers compose the high-level pose-invariant feature (red nodes in the dash area).

1. Introduction

Human facial images play important roles in security issues and social media analytics, where many real-world applications have been successfully developed during the past decades, e.g., face identification and verification, facial expression recognition, facial illumination simulation and removing, aging simulation and age estimation, under either controlled lab environment, or unrestricted environment. However, in both environments, pose is one of the most critical problems since faces in 2D images with different poses are significantly different from each other even though they are of the same identity.

For most of the state-of-the-art face recognition algorithms, finding correspondence or face alignment is the first yet the most essential step because all experiments based on comparisons between registered and test faces need either pixel-wise or semantic level alignment. To address the pose variation, two kinds of alignments are typically used in applications, i.e., appearance level, feature level. Apparently, appearance level alignment explicitly transforms the test face to the pose of the registered face, and then conducts any suitable face recognition algorithms; feature level alignment works in the feature space by projecting all faces to the discriminative identity feature space, regardless of the pose.

*indicates equal contributions.

Following the line of feature level alignment, in this paper, we aim to learn a high-level pose-invariant and discriminative identity feature. The benefits are twofold. First, this high-level pose free feature reduces the impact of diverse poses in the feature space. Second, the feature encodes both common and private attributes of faces, thus mitigating the over-fitting and bad extrapolation.

Our feature learning scheme is based on the following observations: (1) Facial features from different views are transferable, by either linear or non-linear methods. For example, we can project side-view facial feature to front-view facial feature, by a transform function. (2) Faces share similar structures even though their identities are different. Therefore, good facial feature should keep its common attributes as well as private ones. Absence of either one leads to either over-fitting or weak discriminant. (3) Identity is a unique label for each subject, but identity feature could be any vector, with arbitrary length. For example, we use "1" to label the first subject, but its identity feature could be either vector x_1 or x_2, or concatenated vector $[x_1; x_2]$ as long as they are not identical with other subjects' identity feature. And complex identity feature allows us to encode both its private and common attributes.

Based on above observations, we propose a novel approach called "random faces guided sparse many-to-one encoder" (RF-SME) in this paper, which is outlined in Figure 1:

First, we build a single-hidden-layer neural network (S-NN) with sparse constraint that can map faces in different poses to the unique one (many-to-one), i.e., frontal face, which guides the supervised feature learning in the hidden layer. Since the output of this S-NN only relies on the value in the hidden layer, neurons in the hidden layer are potentially good representations for pose free identity feature.

Second, we enhance the discriminative power of the proposed identity feature by assigning random faces to the target values of S-NN. In fact, what we need for target values in S-NN is nothing but an identity representation. Introducing multiple random faces allows us to learn multiple encoders which randomly encode private or common attributes to the identity feature.

Third, we demonstrate the effectiveness of the proposed method by facial images over different poses captured in the controlled environment (Multi-PIE) and facial images in the real-world (LFW) over different poses, mixed with other impact factors, such as illuminations, expressions.

1.1. Related Work

There are two lines in the related work: (1) face feature representation, (2) pose-invariant face recognition, which are highly related to the proposed model in this paper.

In general, face feature representation contains two categories, namely, holistic feature, and local descriptor. Holistic feature uses the entire face region as the input, followed by certain operations, e.g., linear projection [28, 3, 10], to extract discriminative features. On the other hand, local descriptor [17, 21, 8] relies on the hand-craft coding in the local patch and assembles these features to form the final representation of the entire face. A new trend in feature learning recently proposes to use statistical learning for more discriminative and compatible feature [7, 4, 24]. Most of these feature extraction processes are described as an appearance honestly reflected low-level image pre-treatment. Our method is also based on the statistical learning model, but bears the semantic meaning "many-to-one" as well that characterizes the high-level pose free feature.

Other than general face feature representation, there are also a group of pose specified face recognition algorithms. Tied Factor Analysis [23] is a probabilistic approach towards pose-invariant face recognition. The core idea of this method is to compute identity feature regardless of poses through a group of angle specified linear functions. In addition to linear transform, Multi-view Discriminant Analysis (MvDA) [13] explicitly considers the discriminant information and jointly learns multiple view-specific transforms by optimizing a generalized Rayleigh quotient for object recognition. Recently, Coupled Latent Space Discriminative Analysis (CLSDA) [26] has been proposed to tackle the multiple pose face recognition. The model integrates Partial Least Squares (PLS) [29], Bilinear Model (BLM) and Canonical Correlational Analysis (CCA) [11] into one framework, and considers small pose errors in the latent space, therefore enhancing the performance. Different from theirs, our approach generates the identity feature directly through a non-linear mapping and this identity feature can be expanded for the purpose of discriminant.

Researchers also adopt local patch based methods to tackle the pose problem [1, 14]. In [1], authors present an alignment strategy called "stack flow" that discovers viewpoint induced spatial deformities undergone by a face on the local patch level. They learn the relationship of face images between every two adjacent angle bin to form an incremental wrapping knowledge. By this knowledge, virtual frontal faces can be generated from non-frontal faces through one or multiple times of face wrapping, and recognition can be done on the same frontal pose images by off-the-shelf approaches. [14] further develops the former local patch based approach by considering maximizing intra-individual correlations. Compared with the previous method, it is more stable and compact, and reasonably increases the performance.

3D face model has been proposed for pose-invariant face recognition [2, 22, 16]. Pose Normalization [2] creates a novel match scheme that for each gallery and probe image, it generates a virtual frontal face, and the similarity between probe and gallery images could be evaluated on the same frontal pose condition. 3D Generic Elastic Mod-

el [22] learns a 3D generic elastic model from 3D face images. With 3D models, they synthesize a group of virtual face images in different poses for each gallery image in frontal pose. The recognition process first estimates the pose angle of the probe face image, and then performs face matching with virtual gallery face images of the same pose. Morphable Displacement Field (MDF) [16] also considers generating virtual faces to match the gallery. By using a convex combination of a number of template displacement fields, MDF guarantees both global conformity and local consistency. In brief, above methods heavily rely on automatically and robustly fitting a 3D face model to a 2D input image, which is easily affected by factors such as illumination and expression.

2. Pose-Invariant Feature Learning

We detail the proposed framework RF-SME here, which comprises two components, namely, "sparse many-to-one encoder" (SME) and "random faces" (RF). Sparse many-to-one encoder takes responsibility for mapping different poses to the frontal face, therefore yielding a high-level pose free feature in the hidden layer contained in the S-NN. On the other hand, random faces provide many options for the output of S-NN, and artificially produce many random shared structures between two identities. Consequently, it reduces the over-fitting and enhances the discriminative power of the model as well. The entire framework is illustrated in Figure 1.

2.1. Sparse Many-to-One Encoder

The proposed "sparse many-to-one encoder" (SME) is based on a single-hidden-layer neural network (S-NN) (Figure 1). Different from the traditional S-NN learning scheme, we use this structure to extract discriminative features by following a many-to-one mapping. Specifically, in our problem, the input of the SME is training facial images over different poses (many), while the target values are facial images of the same identity as the input but with frontal pose (one). The basic idea of this encoder is that regardless of the input pose, we encourage the output of this single-hidden-layer neural network to be close to the frontal pose facial image of the same identity. We detail this idea in the following part.

Suppose there are I subjects, each of which has J different poses. We use $x_i^j \in \mathbb{R}^n$ to index the input feature of i-th person's j-th pose. We first centralize each feature by the mean feature of a specific pose over all subjects, namely,

$$x_i^j = x_i^j - x^j, \quad \text{where } x^j = \frac{1}{I}\sum_{i=1}^{I} x_i^j. \tag{1}$$

In the feed-forward neural network, the element in the hidden layer is essentially the output of a weight function followed by an activation. In our model, it functions as a pose-invariant high-level feature representation, given the assumption that images of the same subject over different poses share the same high-level feature representation $\tau_i \in \mathbb{R}^m (m \ll n)$.

Suppose the transformation function for "input→hidden layer" is f_1, then the above feed-forward process can be expressed in:

$$\tau_i = f_1(x_i^j). \tag{2}$$

Note we ignore superscript j in τ since it is pose free. Similarly, we construct another transform function f_2 for "hidden layer→ouput", which maps the pose-invariant high-level feature to the hypothesis output h,

$$h(x_i^j) = f_2(\tau_i) = f_2(f_1(x_i^j)), \tag{3}$$

where both f_1 and f_2 can be expressed as a weight function followed by a non-linear activation function, namely,

$$\begin{cases} f_1(x) = \sigma(W_1 x + b_1), \\ f_2(x) = \sigma(W_2 x + b_2), \end{cases} \tag{4}$$

where $W_1 \in \mathbb{R}^{m\times n}$, $W_2 \in \mathbb{R}^{k\times m}$, $b_1 \in \mathbb{R}^m$, $b_2 \in \mathbb{R}^k$, and σ is the sigmoid function which has the form $\sigma(x) = (1+e^{-x})^{-1}$.

In traditional S-NN model, the target values are ground truth labels of training data. The objective function of N-N encourages hypothesis output to be close to these labels. However, in our model, we intentionally set the target values as the frontal pose facial images, i.e., $h(x_i^j) \approx x_i^1$ (frontal pose face is indexed as $j = 1$). Since the neurons in the hidden layer are basis for the output layer, our configuration of the target values enforces that the hidden layer has to be a pose-invariant high-level representation for the input.

We formulate objective function of the proposed encoder as:

$$\min_{W_1,b_1,W_2,b_2} \frac{1}{2N}\sum_{i,j} \|x_i^1 - h(x_i^j)\|_2^2, \tag{5}$$

where $N = I \times J$ is the total number of training images. This is a typical one-half square-loss function, that given the formulation of hypothesis h, can be solved via unconstraint optimization method.

However, due to the high flexibility of the neural network, the model in Eq. (5) easily gets over-fitting. A regularization term is often used on the weight W to overcome this issue. In this paper, inspired by the comparisons on regularization terms for regression problem [18], we impose l_1 norm on the weights W_1 and W_2 to promote them to be sparse. Reasons for sparsity are twofold. First, not all features are equally important, especially for faces that have

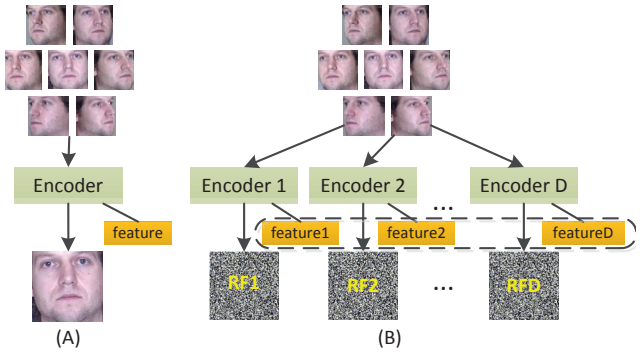

Figure 2. Learning feature with random faces. Compared with using a single frontal face as the target value in (A), random faces in (B) simulate the overlap facial parts between different individuals by randomness. The feature generated by hidden layer may contain more discriminative identity information.

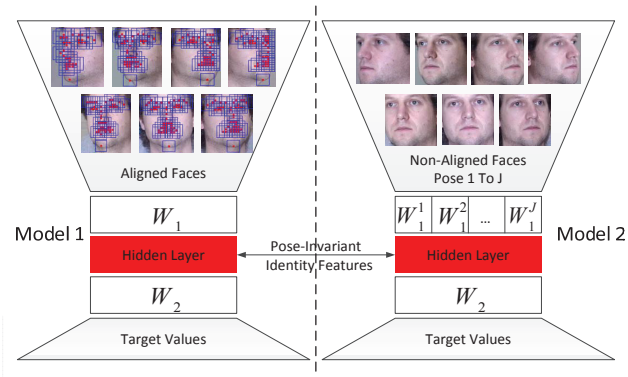

Figure 3. Left: using full-aligned faces for model-1 by learning a single W_1; Right: using non-aligned faces for model-2 by learning multiple W_1s.

significant structure, and being sparse can select the most critical feature. Second, it can avoid over-fitting. By adding l_1 norm regularization terms to Eq. (5), we have the following unconstraint optimization problem:

$$\min_{W_1,W_2,b_1,b_2} \frac{1}{2N}\sum_{i,j}\|x_i^1 - h(x_i^j)\|_2^2 + \lambda_1\|W_1\|_1 + \lambda_2\|W_2\|_1,$$

where $\|W\|_1 = \sum_{ij}|[W]_{i,j}|$ is the sum of absolute value of each element in matrix W. In practice, we solve this unconstraint optimization problem using L-BFGS optimizor [20] which enables to address large-scale data with limited memory. Details of the solution is trivial, and can be found in many related works[1].

After learning model parameters W_1, W_2, b_1, and b_2, we obtain the hidden layer output τ_i for each test x_i^j as a pose-invariant high-level feature, and any classifier can be used to do the recognition task.

2.2. Random Faces

In the previous model, we set the target value y_i^j as the frontal facial image of each subject, and encourage $y_i^j = h_i^j$. This produces output that approximates the frontal face regardless of input. Therefore, the hidden layer output can represent the pose-invariant high-level feature.

On abstract level, the frontal face for each subject in the proposed encoder model is only a representation. Therefore, any unique matrix can work as this representative during the training phase, not necessarily the frontal face of the input image. In addition, if the training and the test have no overlapped identification, the model should have strong extrapolation capability. Apparently, enforcing a specific target value for each subject will prevent the model from extrapolating. In fact, faces are not totally different, because they share similar structures. In the next section, we enhance the extrapolation capability of our model by assigning more than one target values to each input image.

We explain how to generate multiple target values for each input. For each subject i, we generate D random faces $y_i^d \in \mathbb{R}^n$, $1 \le d \le D$, where each single pixel is i.i.d., and follows 0∼1 uniform distribution. Apparently, these "random faces" are not even faces in terms of appearance, but they play the same roles of frontal faces as the representatives in training the encoder. For each input x_i (we omit pose index for simplicity), we train D different encoders and consequently, there are D outputs from the hidden layers, i.e., $[\tau_i^1, \tau_i^2, \ldots, \tau_i^D]$. We pile all this vectors vertically, and hence obtain the final pose-invariant, well-extrapolated high-level feature. This process is illustrated in Figure 2.

2.3. Aligned vs. Non-Aligned Face

In this part, we introduce two models corresponding to two different face alignment strategies, which are shown in Figure 3. As mentioned before, face alignment is the most important pre-processing step before feature extraction. If for each input face with arbitrary pose, we select dense correspondences (facial landmarks), and extract features from local patches defined by these correspondences, then the feature has already been aligned. Still, we need frontal faces to guide the hypothesis outputs. We call this `model-1` that we only use one pair of $\{W_1, b_1\}$ to generate the pose-invariant feature without considering pose information of the input, in either training or test phase. However, dense alignment is often time-consuming and generates many incorrect correspondences.

We may save time and avoid misalignment by skipping the face alignment, which motivates us to learn more than one weight functions and biases corresponding to each pose. For any test input with pose j, we do not need to align it to the frontal pose, rather we find its pose-invariant feature

[1] http://www.stanford.edu/class/cs294a/sparseAutoencoder.pdf

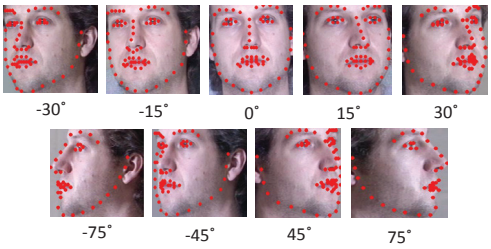

Figure 4. Samples of landmark localization and across pose alignment. Note that some landmarks disappear in the profile images in the second row. Feature of full-aligned setting is extracted based on these red landmarks and the areas they defined.

through corresponding W_1^j and b_1^j. We call this `model-2` in this paper.

3. Experiments

3.1. Face Identification Configuration

In this section, we use Multi-PIE [9] database to test the proposed models on face identification. For full-aligned experiments, we use the state-of-the-art face alignment model in [30] to do landmark localization, as Figure 4 shown. For non-aligned experiments, faces are manually cropped and resized to 128×128, based on the `boundary` of the face, rather than landmarks on the face.

Pose Range. From Figure 4 we can see that when the pose angle goes beyond $45°$, some face landmarks will disappear. As a result, for aligned faces, we only work on images of pose angle in $[-45°, 45°]$; while for non-aligned faces, we can include all poses in $[-75°, 75°]$.

Model Parameters. For full-aligned faces, in either training or test phase, we do not differentiate poses, and only two pairs of model parameters, i.e., $\{W_1, b_1\}$ and $\{W_2, b_2\}$ are learned for input and output layer of S-NN. For the non-aligned face images, we directly use either holistic or local feature to describe the input face, and for each pose j we train a pair of parameters $\{W_1^j, b_1^j\}$. In addition, many methods can be used to approximately estimate the pose for our non-aligned model. For example, in [5], people use multi-class SVM to coarsely group poses based on their appearance, which significantly improves the final performance. Different form theirs, in this paper, we use pose estimation model proposed in [30] to infer the pose for input parameter pair $\{W_1^j, b_1^j\}$.

We set the dimension of the feature space, i.e., size of the hidden layer to be approximately half of the number of individuals in the training set. For the sparse many-to-one encoder, we set the output to be the input's corresponding frontal face feature. For learning with random faces, we also set an encoder's output to be a 2500 dimensional random vector, if we use raw images as inputs. We set the sparse constraint parameter λ_1 or λ_2 to be 0.0001, and use L-BFGS optimizer to train the model with 400 iterations. For feature learning with random faces, we set the number of encoders to be 20, i.e., RF(20) in Table 1, 2, and 3

Figure 5. Samples of two different people from LFW database, including pose, illumination, and expression variations.

Running Environment. All experiments are run on an Intel i7 2600k computer with 16GB memory, implemented by Matlab 2012b and optimized by parallel CPU computing techniques (parfor function). In addition, we use L-BFGS[2] to solve the unconstraint optimization problem in the proposed model. All comparison methods, e.g., LBP [21], HOG [8], LDA [3] are tuned to be optimized for the test purpose based on cross-validation on the training set.

3.2. Face Identification Results

In face identification, we predict each probe image's identity by nearest-neighbor classifier. There are two registration settings.

`Setting-1` registers each individual's frontal face ($0°$) as the gallery. The rank-1 recognition rates for setting-1 are reported in Table 1. `Setting-2` randomly picks up an image from one individual as the gallery, and this image is not necessarily to be the frontal one. We repeat this 20 times and average identification results are reported in Table 3.

Full-Aligned Face. For full-aligned faces, we use landmarks [30] to locate local patches and extract features therein. We extract 20×20 local patches centered at 52 different landmarks. After that, we assemble all these local patches to formulate the complete face feature. So the dimension of the feature for each face is $20 \times 20 \times 52$. We use the first time appearance of all 337 people's face images from the session 1 to session 4 with neutral expression and illumination to build our data set. Specifically, we use the last 88 individuals' facial images to build the training set while the first 249 individuals' images to build the test set.

In addition, we also report the virtual frontal faces generated by model-1 (without random faces). That is, for full-aligned setting, we use the pose-invariant feature of the test sample to compute hypothesis output $h = \sigma(W_2\tau + b_2)$. We illustrate these virtual faces in Figure 6.

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

Method/Degree	-75°	-45°	-30°	-15°	$+15^\circ$	$+30^\circ$	$+45^\circ$	$+75^\circ$	Avg.
3DGEM+Glasses [22]	N/A	65.0%	86.7%	97.6%	93.2%	83.5%	65.0%	N/A	81.8%
3DGEM+No-Glasses [22]	N/A	78.3%	92.2%	97.4%	93.5%	87.0%	83.1%	N/A	88.6%
FA (Full-Aligned)	N/A	38.2%	55.8%	77.5%	57.0%	50.6%	43.4%	N/A	53.8%
FA+LBP	N/A	82.3%	99.6%	100.0%	100.0%	98.9%	76.7%	N/A	92.9%
FA+HOG	N/A	64.7%	94.8%	100.0%	100.0%	94.0%	64.7%	N/A	86.3%
FA+LDA	N/A	92.4%	98.8%	98.8%	98.8%	96.8%	94.0%	N/A	96.6%
FA+Model-1(Ours)	N/A	81.5%	93.2%	98.4%	96.8%	92.4%	88.8%	N/A	91.8%
FA+Model-1+RF(20)(Ours)	N/A	96.8%	100.0%	100.0%	100.0%	100.0%	96.4%	N/A	98.8%
NA (Non-Aligned)	11.0%	10.0%	17.0%	36.0%	46.0%	21.0%	13.0%	11.0%	20.6%
NA+LBP	4.0%	12.0%	24.0%	61.0%	57.0%	21.0%	13.0%	6.0%	24.8%
NA+HOG	4.0%	10.0%	17.0%	71.0%	65.0%	18.0%	13.0%	6.0%	25.5%
NA+MvDA [13]	29.0%	55.0%	64.0%	70.0%	74.0%	62.0%	58.0%	43.0%	56.9%
NA+Model-2(Ours)	57.0%	75.0%	79.0%	94.0%	92.0%	84.0%	78.0%	61.0%	77.5%
NA+Model-2+RF(20)(Ours)	79.0%	88.0%	92.0%	97.0%	98.0%	96.0%	91.0%	80.0%	90.1%

Table 1. Results from gallery `setting-1`. In 3DGEM, the model is trained by USF Human-ID database [25] which contains 94 people's 3D face images. In 3DGEM, they did not consider eyeglasses in training. There are two results reported. "Glasses" means the face recognition rate on the original testing set (249 individuals) which includes eyeglasses, while "No-Glasses" means the results on a subset (158 individuals) of the original testing set where there is no eyeglasses. We take eyeglasses into account during training, and the testing result of our model here is on the original 249 people's testing set.

Method/degree	-75°	-45°	-30°	-15°	$+15^\circ$	$+30^\circ$	$+45^\circ$	$+75^\circ$	Avg.
NA+CLSDA [26]	42.2%	84.4%	96.6%	99.2%	99.2%	96.2%	89.0%	47.7%	81.8%
NA+Model-2+RF(20)(Ours)	50.6%	87.3%	97.9%	99.2%	99.2%	97.4%	91.9%	54.8%	84.8%

Table 2. Identification results compared with the method in [26]. Results of ours is different from that in Table 1 since we follow the same setting in [26] in this experiment.

From the setting-1 results in Table 1, we can see that most of methods perform well thanks to the alignment. Both LBP and LDA achieve very high accuracy, even though only 2D images are utilized. More surprisingly, most of the 2D images based method are superior to the 3D model based method proposed in [22]. We believe this is mainly due to the accurate face alignment by [30] in the preprocessing step. This setting proves that our method is superior to others with full-aligned faces. From the setting-2 results in Table 3, we can see most methods perform worse compared with setting-1, especially for those using local descriptor, e.g., HOG, LBP, but ours is not affected so much by the new setting. This indicates that with different registered faces, our method still performs very well.

Non-Aligned Face. In this experiment, faces are manually cropped based on the boundary of faces, which do not rely on any landmarks, and resized to 128×128. Since all faces are not aligned cross poses, we use model-2 to implement non-aligned face identification where we learn separated $\{W_1^j, b_1^j\}$ for different poses. Apparently, this task is very challenging, and therefore we expand the training set and use the last 237 individuals' facial images in Multi-PIE as the training set, and the first 100 individuals' facial images as the test set.

From the bottom part of Table 1, we can see that all methods degrade significantly, except ours. Note that since we do not need cross pose alignment, we include profile $(-75^\circ, 75^\circ)$ as well. In addition, we use most recent state-of-the-art method MvDA [13] instead of LDA [3] in this table due to its advantage in multi-view learning. Specifically, we compare with the state-of-the-art method [26] which aims at maximizing correlations between different poses under the same setting in Table 2. In both tables, we find that the proposed model works especially well when the pose variation is large, i.e., $> 45^\circ$. Results conclude that even without face-alignment, our method can still extract pose-invariant features well.

Figure 6. Virtual frontal faces generated by model-1. Odd rows: test faces; Even rows: virtual front faces by model-1.

Method/Degree	-45°	-30°	-15°	0°	$+15^\circ$	$+30^\circ$	$+45^\circ$	Avg.
FA (Full-Aligned)	35.0%	46.8%	46.5%	42.6%	43.6%	35.2%	32.3%	40.3%
FA+HOG	56.6%	69.0%	73.8%	72.1%	69.9%	63.9%	50.5%	65.1%
FA+LBP	61.6%	75.9%	85.4%	89.2%	86.2%	74.9%	58.5%	76.0%
FA+LDA	90.1%	95.6%	95.2%	95.0%	94.4%	93.6%	89.5%	93.4%
FA+Model-1+RF(20)(Ours)	90.9%	97.7%	98.1%	98.5%	98.2%	95.7%	87.2%	95.2%

Table 3. Results from gallery `setting-2` for full-aligned facial image.

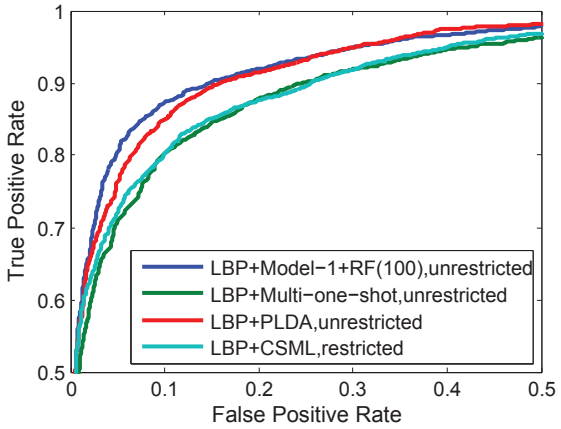

Figure 7. ROC curves of face verification on LFW database.

LBP+CSML, restricted [19]	0.8557 ± 0.0052
LBP+PLDA, unrestricted [15]	0.8733 ± 0.0055
LBP+multi-one-shot, unrestricted [27]	0.8517 ± 0.0061
LBP+Model-1+RF(100), unrestricted	0.8850 ± 0.0058

Table 4. Face verification accuracy by 10-fold cross validation on LFW dataset.

3.3. Face Verification in the Wild

"Labeled Faces in the Wild" (LFW) [12] is a benchmark database for evaluating face verification algorithm on "wild" real-world images. This dataset contains 13,000 images of faces collected from the Internet, and 1680 individuals with at least two face images. Since our feature learning scheme relies on the identity of the training set, we follow the unrestricted setting of the LFW. All images are cropped and resized into size of 150×120, and then LBP feature is extracted for each of them. We pick 1 fold apart for testing, and used the rest 9 folds' individuals who have at least 2 facial images for training. In this verification experiment, we run model-1 with 100 random faces, and the size of hidden layer is 10. Both λ_1 and λ_2 are set to 0.0001. For computational efficiency, we reduce the dimensionality of data by PCA.

We followed the method used in multi-one-shot [27] to centralize faces according to their poses, which is formulated in Eq. (1). Then, we adopt model-1 to learn the pose-invariant feature. After that, both training and test data are projected into pose-invariant feature space. Each feature vector is normalized to be of unit length. Then we combine two facial features in each pair by an element-wise multiplication, and use the pairs in the 9 folds, to train a linear SVM classifier [6]. The optimized penalty parameter C is learned through cross-validation on the training set. At last, we use face feature pairs in the test fold for face verification. We repeat the above process for 10 times, each time with different testing and training folds. The overall verification performance is stated in Table 4 and Figure 7.

From results we can see that LFW is very challenging since all the faces are from real-world with arbitrary poses, expressions as well as illuminations, as shown in Figure 5. The proposed model using LBP feature is still superior to state-of-the-art methods.

3.4. Model Selection

In this part, we show how the model parameters λ_1, λ_2 and size of hidden layer m affect the performance of the proposed model. We first show effects of different λ_1 and λ_2. In Figure 8, we run model-1 on Multi-PIE database with setting-1. The hidden layer size is fixed at 50. We can see that with different λ_1 and λ_2 the performance changes differently. Note that results in the left subfigure of Figure 8 is l_1 norm based regularization (our method) while the results in the middle subfigure are from l_2 regularization. In both subfigures, the improvement by regularization terms over that without regularization (at the original point) is significant. Besides, the best performance of l_1 norm is slightly better than that of l_2 norm. Although this does not demonstrate that l_1 norm is better than l_2 on selecting the most useful facial structure, it is consistent with the fact that l_1 regularization empirically works well [18].

In addition, we also show the impact of code size on identification in the right subfigure of Figure 8. We run L-BFGS algorithm 400 iterations with different code sizes, ranging from 50 to 140. Under this setting, we can see different code sizes have different convergence speeds, but their final performances are similar.

4. Conclusion

In this paper, we presented a novel many-to-one high-level face feature learning approach for extracting pose-invariant and discriminative identity feature from 2D facial

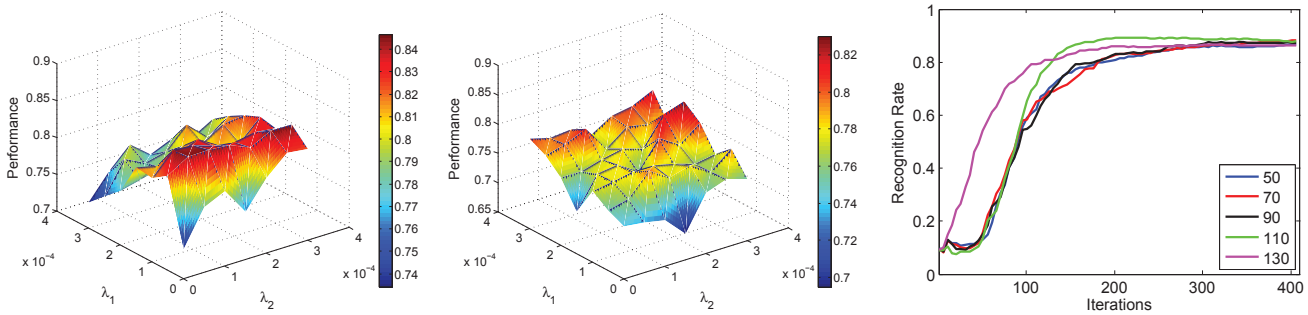

Figure 8. The recognition performance of l_1 (our model) and l_2 norm for regularized neural network over different parameters λ_1 and λ_2. Left: l_1 norm; Middle: l_2 norm; Right: Impact of the coder size.

images. First, we designed an l_1 norm regularized many-to-one encoder to remove the impact introduced by diverse poses from feature learning process. Second, we enhanced the discriminant of the pose free feature by setting multiple random faces as the target values of our encoders. Appealing results on both lab database, i.e., Multi-PIE, and real-world database, i.e., LFW demonstrate the effectiveness and superiority of our method.

5. Acknowledgement

This research is supported in part by the NSF CNS award 1314484, Office of Naval Research award N00014-12-1-0125 and N00014-12-1-1028, Air Force Office of Scientific Research award FA9550-12-1-0201, and U.S. Army Research Office grant W911NF-13-1-0160.

References

[1] A. B. Ashraf, S. Lucey, and T. Chen. Learning patch correspondences for improved viewpoint invariant face recognition. In *CVPR*, pages 1–8. IEEE, 2008. 2

[2] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith. Fully automatic pose-invariant face recognition via 3d pose normalization. In *ICCV*, pages 937–944. IEEE, 2011. 2

[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE TPAMI*, 19(7):711–720, 1997. 2, 5, 6

[4] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *NIPS*. MIT Press, 2007. 2

[5] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *CVPR*, pages 2707–2714. IEEE, 2010. 5

[6] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 7

[7] A. Coates, A. Y. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, pages 215–223, 2011. 2

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. IEEE, 2005. 2, 5

[9] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *IVC*, 28(5):807–813, 2010. 5

[10] X. He and P. Niyogi. Locality preserving projections. In *NIPS*. MIT Press, 2004. 2

[11] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. 2

[12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 7

[13] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. In *ECCV*, pages 808–821. Springer, 2012. 2, 6

[14] A. Li, S. Shan, X. Chen, and W. Gao. Maximizing intra-individual correlations for face recognition across pose differences. In *CVPR*, pages 605–611. IEEE, 2009. 2

[15] P. Li, Y. Fu, U. Mohammed, J. H. Elder, and S. J. Prince. Probabilistic models for inference about identity. *IEEE TPAMI*, 34(1):144–157, 2012. 7

[16] 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 *ECCV*, pages 102–115. Springer, 2012. 2, 3

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2

[18] A. Y. Ng. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *ICML*. ACM, 2004. 3, 7

[19] H. V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. In *ACCV*, pages 709–720. Springer, 2011. 7

[20] J. Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782, 1980. 4

[21] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE TPAMI*, 24(7):971–987, 2002. 2, 5

[22] U. Prabhu, J. Heo, and M. Savvides. Unconstrained pose-invariant face recognition using 3d generic elastic models. *IEEE TPAMI*, 33(10):1952–1961, 2011. 2, 3, 6

[23] S. J. Prince, J. Warrell, J. Elder, and F. M. Felisberti. Tied factor analysis for face recognition across large pose differences. *IEEE TPAMI*, 30(6):970–984, 2008. 2

[24] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *ICML*, pages 759–766. ACM, 2007. 2

[25] 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 TPAMI*, 27(2):162–177, 2005. 6

[26] 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. *CVIU*, 2012. 2, 6

[27] Y. Taigman, L. Wolf, and T. Hassner. Multiple one-shots for utilizing class label information. In *BMVC*, volume 2, pages 1–12, 2009. 7

[28] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *CVPR*, pages 586–591. IEEE, 1991. 2

[29] H. Wold. Partial least squares. *Encyclopedia of statistical sciences*, 1985. 2

[30] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886. IEEE, 2012. 5, 6