PAPER

Joint Multi-Patch and Multi-Task CNNs for Robust Face Recognition

Yanfei LIU[†], **Junhua CHEN**[††a)], ***Nonmembers*, *and* Yu QIU**[†††], ***Member***

SUMMARY In this paper, we present a joint multi-patch and multi-task convolutional neural networks (JMM-CNNs) framework to learn more descriptive and robust face representation for face recognition. In the proposed JMM-CNNs, a set of multi-patch CNNs and a feature fusion network are constructed to learn and fuse global and local facial features, then a multi-task learning algorithm, including face recognition task and pose estimation task, is operated on the fused feature to obtain a pose-invariant face representation for the face recognition task. To further enhance the pose insensitiveness of the learned face representation, we also introduce a similarity regularization term on features of the two tasks to propose a regularization loss. Moreover, a simple but effective patch sampling strategy is applied to make the JMM-CNNs have an end-to-end network architecture. Experiments on Multi-PIE dataset demonstrate the effectiveness of the proposed method, and we achieve a competitive performance compared with state-of-the-art methods on Labeled Face in the Wild (LFW), YouTube Faces (YTF) and MegaFace Challenge.

key words: *CNN, multi-task learning, face representation learning, face recognition*

1. Introduction

As the development of deep learning methods and accumulation of available training data of facial images, unconstrained face recognition, which has extensive application prospect in the field of security, authentication, and human-computer interaction, etc., has achieved remarkable results in recent years. Accurate face recognition depends on good face representations, which should be discriminative to the inter-person variations but retains robust to intra-person ones. Conventional face representation learning methods are usually based on artificial feature descriptors, such as histogram of oriented gradient (HoG) [1] and Local Binary Pattern (LBP) [2]–[4] etc. However, the representation composed by handcrafted feature descriptors is too shallow to distinguish the complex nonlinear facial appearance variances. Recently, deep learning based face recognition algorithms such as deep belief network (DBN) [5], stacked auto-encoder (SA) [6], [7] and convolutional neural networks (CNN) [8]–[13], etc. have drawn a lot of attention due to its superior performance on face-related tasks. Since deep model has deep architecture and powerful learning ability, it has achieved impressive performance in face representation learning and face recognition.

Although the deep learning based face representations have enabled great breakthrough in face recognition, there are still great challenges for unconstrained face recognition due to the existence of intra-person variations such as large pose variation, illumination variance, expression difference, and occlusion etc. in actual application environments. Particularly, pose variation is considered as the most challenging one among other non-identity variations, which will lead to severe decline on performance of face recognition.

To address the pose variation problem, previous works have proposed to simultaneously learn face recognition task and other face related tasks (such as facial attribute prediction [14], pose estimation [15], and face image synthesis [16], etc.) with multi-task learning. There are also some literatures that adopt multi-patch CNNs to obtain robust face representation by local feature fusion [8], [9], [17], [18]. It is intuitively that local features are not only important to face recognition task but also or even more important to other tasks like pose estimation. For example, in the circumstance of large pose variation, one of the eyes or half of the nose will be very different with that in the frontal face. That is to say, local features fusion is able to enhance the multi-task learning of face recognition and other tasks, and the enhanced other tasks can potentially further promote the performance of face recognition task by multi-task learning. However, to the best of our knowledge, no literature incorporates local feature fusion with multi-task learning together for pose-invariant face recognition. To this end, we combine multi-patch CNNs with multi-task learning (face recognition task and pose estimation task) in one framework and propose joint multi-patch and multi-task convolutional neural networks (JMM-CNNs) to learn more robust pose-invariant face representation for face recognition, as shown in Fig. 1. Under this framework, the global and local features are extracted by a set of multi-patch CNNs and then fused as a shared feature. Two fully connected layers are connected to the shared feature to perform classification of each task. Extensive experiments on one constrained face datasets: MultiPIE [19] and three unconstrained datasets: Labeled Faces in the Wild (LFW) [20], YouTube Faces (YTF) [21] and MegaFace Challenge [22], indicate that superior performance is achieved with the proposed JMM-CNNs framework.

To further enhance the insensitiveness of the extracted

Manuscript received March 18, 2020.
Manuscript revised May 25, 2020.
Manuscript publicized July 2, 2020.

[†] The author is with the Chongqing University of Technology, Chongqing, 400054, China.

[††] The author is with the Chongqing University of Posts and Telecommunications, Chongqing, 400065, China.

[†††] The author is with the Chongqing Industry Polytechnic College, Chongqing, 401120, China.

a) E-mail: chenjh@cqupt.edu.cn

DOI: 10.1587/transinf.2020EDP7059

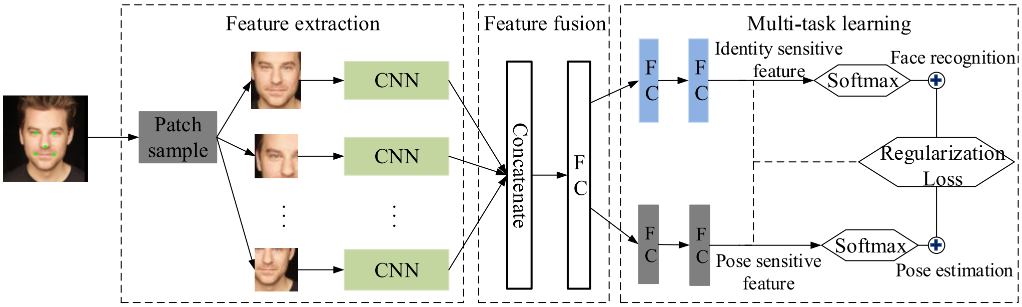


Fig. 1 Flowchart of the proposed JMM-CNNs framework. JMM-CNNs is essentially composed of three steps: global and local feature extracting using a set of CNNs, feature fusion by concatenating and a fully connected layer, and multi-task learning by fully connected layers and softmax plus regularization loss. The representations extracted from the last fully connected layer of the face recognition task, i.e. identity sensitive features, are used as the final feature for face recognition.

identity feature to pose variation, we introduce a similarity regularization term on features of the two tasks to form a regularized joint loss. The loss makes features sensitive to face recognition task be insensitive to pose estimation task and the features sensitive to pose variation task be insensitive to face recognition task.

For multi-patch CNNs based face recognition, the image patch sampling strategy will affect the performance of face recognition. Different from traditional methods that either randomly cropped patches [9] or uniformly sample a small number of patches [17] with the help of 3D model, we also uniformly crop the face image into a small number of image patches only depending on locations of the sparse facial landmarks for effectiveness of the method. There are two reasons why we adopt this patch sampling strategy. For one reason, sparse landmarks is more reliable than dense landmarks. Moreover, [17] has verified that sampling a small number of image patches uniformly in the semantic meaning can even outperform DeepID2 [9] which randomly cropped a large number of patches (25 patches). For another reason, it makes us achieve an end-to-end architecture. Existing works related to multiple patches usually sample patches separately and then take multiple patches as inputs of different network branches, because either there are so many patches need to be cropped or the patches is sampled with the help of other model. In contrast, the proposed patch sampling strategy is easy to implement and can be embedded in the whole framework as a layer of the network. Using this patch sampling strategy, the proposed framework has an end-to-end fashion that takes a holistic image with a set of facial landmarks as input and samples local patches with the patch sample block, then sends them to different network branches.

The main contributions can be concluded as follows: (1) We propose a joint multi-patch and multi-task convolutional neural networks (JMM-CNNs) framework to learn informative and pose-invariant facial feature representation by local facial information fusing and multi-task learning. (2) We introduce a similarity regularization term on features of the two tasks to further enhance the insensitiveness to pose variation of the extracted feature. (3) A simple but effective patch sampling strategy is used to make the proposed JMM-CNNs have an end-to-end architecture.

2. Related Work

2.1 Face Representation Learning

Existing face image representation learning methods can be divided into two categories: artificial feature learning methods and deep learning facial feature learning methods. The face representation learning methods before 2013 were mainly artificial descriptors or learning-based local descriptors. Therefore, the corresponding pipeline of face recognition at that time usually included artificial descriptor or learning-based local descriptor and distance metric learning. Artificial descriptors consist of scale-invariant feature transform (SIFT) [23], LBP and HoG, etc. These descriptors can be further improved and combined to obtain better performance. For example, Zhang etc. fused Gabor and LBP feature by extracting LBP feature on Gabor amplitude image and obtain facial feature with excellent illumination robustness [24]. Learning-based local descriptor learns semantic representation using artificial descriptor. The attribute and similarity classifier [25] proposed by Neeraj Kumar and the Tom-vs-Peter classifier [26] proposed by Thomas Berg are typical learning-based local descriptors.

In recent years, face representation learning methods based on deep learning have gradually become the research hotspot. Compared with artificial feature extraction methods, deep learning based methods can obtain more effective feature by hierarchical nonlinear mapping due to its deep architecture and powerful learning ability. The typical deep learning based facial feature extraction methods are DeepFace [13] proposed by Facebook, DeepID series [8]–[11], and FaceNet [12] proposed by Google. DeepFace adopted 3D method for face alignment and Siamese network architecture consisting of 2 normal convolutional layers, 3 local convolutional layers without weight sharing and 2 fully connected layers for facial feature extraction. DeepID extracted facial feature using the improved structure of CNN, which enhanced feature description ability by fusing the outputs of convolutional layer and its previous pooling layer. FaceNet mainly used a multi-branch local network topology named as GoogLeNet, in which the inception model [27] simultaneously combines multi-scale features and significantly reduces the number of training parameters by 1*1 convolutional kernel. FaceNet reached the best average classification accuracy result of 99.63% on LFW dataset, which mainly declared the termination of eight years performance competition on LFW dataset.

2.2 Multi-Patch CNN Based Face Recognition

Most of face recognition methods extract the global face representation from the holistic face image, which tend to cause recognition error in local variation conditions. To handle this problem, strategies that fuse local features using multiple CNNs were proposed. For example, DeepID [8] extracted facial representations from RGB image, gray-level image and gradient map and fused them in score level. DeepID2 [9] proposed to extract deep features from 25 image patches cropped with various scales and positions. [17] used a set of CNNs to extract a multimodal deep face representation by fusing features extracted from holistic image, 3D pose normalized holistic image, and image patches. These works verified that an ensemble of multiple networks corresponding to different image patches can improve the performance of face recognition.

2.3 Multi-Task Learning for Face Recognition

Multi-task learning (MTL) has been widely studied in computer vision and machine learning [28]–[30]. We focus our review on multi-task learning for facial feature extraction and face related tasks. [31] proposed a MTL network named as HyperFace for face detection landmarks localization, pose and gender estimation by fusing the intermediate layers of CNN for improved feature extraction. An all-in-one CNN framework [32] was proposed to realize more face related tasks including face detection, face alignment, pose estimation, gender recognition, smile detection, age estimation and face recognition simultaneously. Xi Yin et al. [15] proposed a multi-task CNN with pose estimation, illumination, and expression as the side task of the main task of face recognition. The proposed method falls under the multi-patch CNN method and the multi-task learning approach with CNNs for face recognition, but with several differences compared with existing methods. First, to the best of our knowledge, we are the first to combine multi-patch CNNs and multi-task learning into a uniform framework for facial feature extraction. Experiment results verify the effectiveness of the proposed combination method of multi-patch and multi-mask CNNs. Furthermore, different from existing multi-task CNN based methods, a similarity regularization term is introduced in loss function to further enhance pose robustness of the extracted features. Additionally, different from the patch sampling strategies of existing multi-patch CNN methods, we simply sample patches depending on sparse facial landmarks, which makes the proposed model have an end-to-end architecture.

3. The Proposed Method

In this section, we provide a detailed overview of the proposed face recognition method based on JMM-CNNs, including architecture of JMM-CNNs and its formulation, regularized loss function, patch sampling strategy, network training procedure and face verification and identification pipeline.

3.1 JMM-CNNs

As illustrated in Fig. 1, the proposed JMM-CNNs consist of three parts: feature learning from the holistic image and the local patches sampled from the aligned image using a set of CNNs, feature fusion by a fully-connected layer, and multi-task learning of face recognition and pose estimation by two additional fully connected layers for each task respectively. To further enhance pose insensitiveness of the extracted identity feature, a regularization loss is added in softmax loss of the multi-task learning to constrain both tasks in feature space, which is detailed in Sect. 3.2. The proposed JMM-CNNs can also be considered as a two-stage facial feature extraction architecture. At the first stage, a set of CNNs are used to extract global feature and local features and these features are fused into a shared facial feature by the feature fusion subnetwork. And at the second stage, two fully connected networks are followed by the feature fusion subnetwork to refine the feature to learn task-specific features for face recognition and pose estimation, respectively.

We select two typical networks as the CNN for global and local feature learning: a modified AlexNet [33] and ResNet-18 [34]. The architecture of the modified AlexNet is shown in Fig. 2. The main modification of AlexNet is that Batch Normalization [35] is applied after each convolutional layer to accelerate the training process and Parametric Rectified Linear Units (PReLUs) [36] nonlinearity is used as the activation function for hidden neurons in all convolutional layers.

We assume a training dataset D with N face images and their labels, denoted as $D = \{(I_i, y_i^d, y_i^p)\}_{i=1}^N$, where I_i is the ith face image, y_i^d and y_i^p are the identity label and the pose label of the ith face image respectively. Let I, y^d and y^p represent set for I_i, y_i^d, y_i^p respectively, i.e. $I = \{I_i\}_{i=1}^N$,

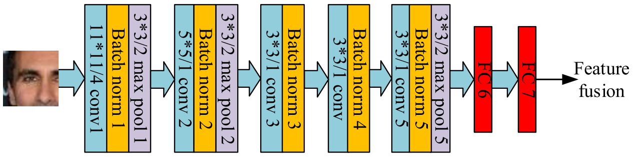


Fig. 2 The modified AlexNet architecture for global and local feature learning. Conv denotes convolutional layer, Batch norm denotes Batch Normalization layer, and FC denotes fully connected layer. Then through the final fully connected layer the output embedding of network is fed to feature fusion subnetwork.

$y^d = \{y_i^d\}_{i=1}^N$, and $y^p = \{y_i^p\}_{i=1}^N$. For each input image I_i, we denote the feature vector extracted by each CNN as $x_j = conv(I_i, \theta_c)$, $j \in (1, 2, \ldots M)$, where M is the number of CNNs (namely the number of image patches including the holistic image), $conv(\cdot)$ represents the feature extraction function defined in CNN, and θ_c denotes the parameter to be learned of each CNN. Then the extracted features are fused by concatenating and a fully connected layer, which can be formulated as

$$f_s(x) = W_f^T x + b_f, \tag{1}$$

where $x = [x_1, x_2, \ldots, x_M]$ denotes the concatenated feature, and W_f and b_f represent the weight matrices and bias of the fully connected layer for feature fusion respectively. The $f_s(x)$ is a shared feature for face recognition and pose estimation task.

Assume x_d and x_p be the output embedding of network for face recognition task and pose estimation task respectively, then they can be formulated as

$$\begin{aligned} x_d &= F_d(f_s(x), \theta_d) \\ x_p &= F_p(f_s(x), \theta_p), \end{aligned} \tag{2}$$

where $F_d(\cdot, \cdot)$ denotes the non-linear mapping function of the shared feature $f_s(x)$ and parameter θ_d, which maps from the shared feature $f_s(x)$ to the extracted identity sensitive feature x_d. Similarly, $F_p(\cdot, \cdot)$ denotes the mapping function that maps from $f_s(x)$ to the pose sensitive feature x_p.

We use softmax layer as classifier for both identity classification and pose estimation task. And then the loss function for face recognition task can be defined use the cross-entropy loss:

$$L_d(I, y^d) = -\frac{1}{N} \sum_{i=1}^{N} log \frac{exp((W_{y_i^d}^d)^T x_{di} + b_{y_i^d}^d)}{\sum_{c=1}^{C} exp((W_c^d)^T x_{di} + b_c^d)}, \tag{3}$$

where y_i^d indicate the corresponding identity label, and W^d and b^d represent the weight matrices and bias of the last layer for face recognition task respectively. C is the number of the subjects, and N is the number of training samples.

The loss function for pose estimation task can be formulated similarly as:

$$L_p(I, y^p) = -\frac{1}{N} \sum_{i=1}^{N} log \frac{exp((W_{y_i^p}^p)^T x_{pi} + b_{y_i^p}^p)}{\sum_{g=1}^{G} exp((W_g^p)^T x_{pi} + b_g^p)}, \tag{4}$$

where G is the number of pose degrees.

3.2 Regularized Loss Function

We aim to obtain facial feature that are sensitive to identity but insensitive to pose variance. In our method, the face recognition and pose estimation task share the same feature $f_s(x)$ obtained by the feature extraction subnetwork and the feature fusion subnetwork, while face recognition task pursues identity sensitive features and pose estimation task pursues pose sensitive features. For feature competition, the features sensitive to face recognition task should be insensitive to pose estimation task, and the features sensitive to pose estimation task should be insensitive to face recognition task. That is to say, there is conflict between the two features, which suggests the relationship of them should be negative correlation. Therefore, we introduce a cosine similarity regularizer over the two features as the regularization loss to constrain the correlation between them.

$$L_R(x_d, x_p) = \frac{(x_d)^T x_p}{||x_d||_2 ||x_p||_2} \tag{5}$$

By combining Eqs. (3), (4) and (5), we can define the regularized loss function as follows:

$$L(I, y^d, y^p) = L_d(I, y^d) + L_p(I, y^p) + \lambda L_R(x_d, x_p), \tag{6}$$

where λ is the regularization parameter to control the importance of the regularization term. Then given the training set D, our JMM-CNNs will aim to minimize the regularized loss function $L(I, y^d, y^p)$.

3.3 Patch Sampling Strategy

For multi-patch CNNs based face recognition, the image patch sampling strategy will affect the performance of face recognition. Since [17] has verified that sampling a small number of image patches uniformly in the semantic meaning can even outperform DeepID2 which randomly cropped a large number of patches (25 patches), we also uniformly sample few patches with the help of facial landmarks for effectiveness of the model. We crop 145 × 120 holistic image centering at the nose tip and five 100 × 100 image patches centering around the five facial landmarks, i.e. the two eye centers, the nose tip, and the two mouth corners, from the aligned face image, which is aligned to 230 × 230 using the five landmarks. Figure 3 shows the cropped holistic image and patches in our method. Note that non-frontal face is asymmetric, as shown in Fig. 3, so we leverage all the five patches corresponding to the five facial landmarks rather than only adopt patches of the left or the right half face. This sampling strategy can help us achieve an end-to-end

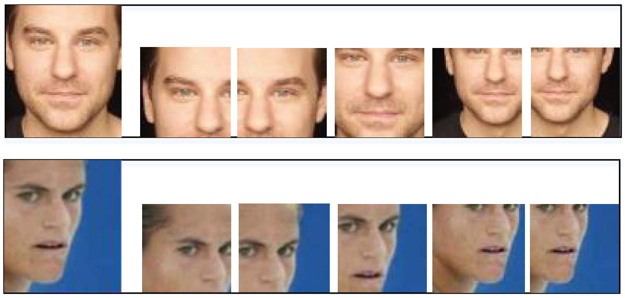


Fig. 3 Illustration of the cropped holistic image and patches. From left to right: holistic image, the patches centering at left eye center, right eye center, nose tip, left mouth corner and right mouth corner. Top row: the cropped holistic image and patches of frontal face. Bottom row: the cropped holistic image and patches of non-frontal face.

Table 1 Illustration of training algorithm for training the proposed JMM-CNNs

```
input: training set $D = \{(I_i, y_i^d, y_i^p)\}_{i=1}^N$, initialized parameter $\theta_c$,
$\theta_d, \theta_p, W_f, W_d, W_p$, learning rate $\eta(iter)$, $iter \leftarrow 0$
while not converge do
  $iter \leftarrow iter + 1$ sample training samples $(I_i, y_i^d, y_i^p)$ from $D$
  $L(I_i, y_i^d, y_i^p) = L_d(I_i, y_i^d) + L_p(I_i, y_i^p) + \lambda L_R(x_{di}, x_{pi})$
  $\nabla\theta_c = \nabla_{\theta_c} L(I_i, y_i^d, y_i^p)$
  $\nabla W_f = \nabla_{W_f} L(I_i, y_i^d, y_i^p)$
  $\nabla\theta_d = \frac{\partial L(I_i, y_i^d, y_i^p)}{\partial x_{di}} \times \frac{\partial x_{di}}{\partial \theta_d}$
  $\nabla\theta_p = \frac{\partial L(I_i, y_i^d, y_i^p)}{\partial x_{pi}} \times \frac{\partial x_{pi}}{\partial \theta_p}$
  $\nabla W_d = \frac{\partial L(I_i, y_i^d, y_i^p)}{\partial W_d}$,  $\nabla W_p = \frac{\partial L(I_i, y_i^d, y_i^p)}{\partial W_p}$
  update $W_d = W_d - \eta(iter) \cdot \nabla W_d$,  $W_p = W_p - \eta(iter) \cdot \nabla W_p$,
  $W_f = W_f - \eta(iter) \cdot \nabla W_f$,  $\theta_d = \theta_d - \eta(iter) \cdot \nabla\theta_d$,  $\theta_p = \theta_p -$
  $\eta(iter) \cdot \nabla\theta_p$,  $\theta_c = \theta_c - \eta(iter) \cdot \nabla\theta_c$
end while
output $\theta_c, \theta_d, \theta_p, W_f$
```

architecture and we do not need to train each CNN for one patch separately, once the training sample is preprocessed by facial landmark detection and face alignment which are normally essential steps in current face recognition pipeline.

3.4 Network Training

In the training set D, each face image is labeled with identity label and pose label y_i^d, y_i^p. The parameter of each CNN θ_c, the feature fusion network parameter W_f, and the multi-task learning network parameter θ_d and θ_p are trained by minimizing the loss function $L(I, y^d, y^p)$ using stochastic gradient descent (SGD) and standard back propagation algorithm. Training algorithm is illustrated in Table 1.

For the parameter θ_c of some CNN network and the feature fusion parameter W_f, the back propagation gradients of θ_c and W_f are

$$\begin{aligned}\nabla_{\theta_c} L(I, y^d, y^p) = (&\frac{\partial L(I, y^d, y^p)}{\partial x_d} \times \frac{\partial x_d}{\partial f_s(x)} + \\ &\frac{\partial L(I, y^d, y^p)}{\partial x_p} \times \frac{\partial x_p}{\partial f_s(x)}) \times \frac{\partial f_s(x)}{\partial x} \times \frac{\partial x}{\partial \theta_c} \quad (7)\end{aligned}$$

and

$$\begin{aligned}\nabla_{W_f} L(I, y^d, y^p) = (&\frac{\partial L(I, y^d, y^p)}{\partial x_d} \times \frac{\partial x_d}{\partial f_s(x)} + \\ &\frac{\partial L(I, y^d, y^p)}{\partial x_p} \times \frac{\partial x_p}{\partial f_s(x)}) \times \frac{\partial f_s(x)}{\partial W_f} \quad (8)\end{aligned}$$

respectively.

For the backward propagation, we need to calculate the derivative of the loss with respect to x_d and x_p. Let δ_d and δ_p denote backpropagation errors of final loss layer for each task respectively [37]. Different from other layers, we calculate the gradient of $L(I, y^d, y^p)$ respect to x_d and x_p using the chain rule as follows:

$$\begin{aligned}\frac{\partial L(I, y^d, y^p)}{\partial x_d} &= \frac{\partial L_d(I, y^d)}{\partial x_d} + \lambda \frac{\partial L_R(x_d, x_p)}{\partial x_d} \\ &= (W^d)^T \delta_d + \lambda(\frac{x^p}{||x_d||_2 ||x_p||_2} - \frac{(x_d^T x_p) x_d}{||x_p||_2 ||x_d||_2^3}) \quad (9) \\ \frac{\partial L(I, y^d, y^p)}{\partial x_p} &= \frac{\partial L_p(I, y^p)}{\partial x_p} + \lambda \frac{\partial L_R(x_d, x_p)}{\partial x_p} \\ &= (W^p)^T \delta_p + \lambda(\frac{x^d}{||x_d||_2 ||x_p||_2} - \frac{(x_d^T x_p) x_p}{||x_d||_2 ||x_p||_2^3}). \quad (10)\end{aligned}$$

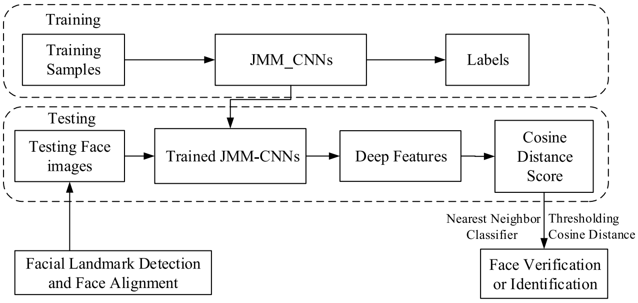

Training
Training Samples
JMM_CNNs
Labels
Testing
Testing Face images
Trained JMM-CNNs
Deep Features
Cosine Distance Score
Nearest Neighbor Classifier
Thresholding Cosine Distance
Facial Landmark Detection and Face Alignment
Face Verification or Identification

Fig. 4 The framework of training JMM-CNNs that are embed into the face verification or identification pipeline.

3.5 Face Verification and Identification

To evaluate the proposed JMM-CNNs model, we embed the feature learned by JMM-CNNs into the traditional face verification or identification pipeline of facial landmark detection, face alignment, feature extraction and face verification or identification. The general framework to train JMM-CNNs and the pipeline of face verification or identification are shown in Fig. 4. We first use TCDCN [38] to detect the five facial landmarks. Then the face images are aligned by an affine transformation matrix calculated using the detected facial landmarks and these corresponding landmarks on average face image. During the alignment procedure, the facial landmarks are also projected to the aligned image with the same affine transformation. After the face alignment procedure, the aligned face images with five facial landmarks are used as inputs of the proposed feature extraction model JMM-CNNs. And then the deep features can be taken from the output of the fully connected layer for face recognition task of the JMM-CNNs. Finally, we compute the score by Cosine Distance of two features, and use nearest neighbor classifier for face identification and threshold comparison for face verification, respectively.

4. Experimental Results

In this paper, we evaluate the proposed method on both constrained and unconstrained datasets. For constrained evaluation, we train and test the proposed model both on Multi-PIE dataset. For unconstrained evaluation, we train our model on CASIA-WebFace [39] dataset and test on three current popular and important unconstrained face recognition datasets: LFW, YTF and MegaFace Challenge. We firstly evaluate the effectiveness of the proposed model on Multi-PIE dataset. The experiments present the role of the multiple patches and

Table 2 Identification performance (%) on setting V of MultiPIE of each single CNN and the combinations of different number of CNNs. H, le, re, n, lm, and rm indicate holistic image, patches centering at left eye center, right eye center, nose tip, left corner and right corner of the mouth, respectively. Avg. rank-1 denotes the average rank-1 identification rate among the pose interval [−90°, 90°].

Single CNN	Patch Number	Avg. rank-1	Combination with best performance	Patch Number	Avg. rank-1
CNN-H	–	90.36	H+n	1	90.41–90.49
CNN-le	1	89.75	H+re+n	2	90.53–90.67
CNN-re	1	89.68	H+le+n+rm	3	90.75–90.84
CNN-n	1	89.59	except rm	4	90.89–90.96
CNN-lm	1	88.57	JMM-CNNs	5	91.85
CNN-rm	1	88.51	Randomly cropped	5	90.71

the multi-task learning. Then we evaluate face verification accuracies of the proposed model on LFW and YTF dataset, and our model achieves better results than traditional multi-patch based CNNs and MTL-based CNNs. We also conducted experiments on MegaFace Challenge, the largest face recognition benchmark, and our model outperforms most of the state-of-the-art models under the small training data protocol on MegaFace Challenge.

4.1 Datasets and Preprocessing

The proposed model is firstly trained on Multi-PIE dataset for effectiveness evaluation of multi-patch and multi-task learning, due to Multi-PIE has straightforward pose labels. The Multi-PIE dataset is composed of 754,200 images of 337 identities. Each identity was imaged under 15 poses including 13 raw angles from −90° to 90°, 20 different illuminations, and 6 different expressions. We mainly focus on yaw angle estimation, so the Multi-PIE setting V [40], which includes totally 301,600 face images of 337 identities under 13 raw angles, 20 illuminations, 1 expression, and 4 sessions, is selected to evaluate the proposed method. The first 200 identities (199,940 images) are used for training, and the remaining 137 identities are used for testing, where one image with frontal pose, neutral illumination and expression for each identity is selected as the gallery set (137 images) and the remaining as the probe set (101,523 images). We use TCDCN to detect the five facial landmarks and align each face to size 230 × 230 according to the detected five facial landmarks.

For unconstrained evaluation, we train our model on the publicly available CASIA-WebFace dataset. CASIA-WebFace dataset contains 49,414 face images from 10,575 different identities, the faces in which are all centered on the image. Since images in CASIA-WebFace dataset does not have pose labels, we use PIFA [41] to estimate the yaw angle as the pose label as in [15]. According to the yaw angle distribution on CASIA-Webface in [15], CASIA-Webface has 12% faces belong to large pose group, i.e. the pose group [−90°, −45°] and [45°, 90°]. We use the same facial landmark detection and face alignment methods as described in Sect. 3.5 to preprocess the set of training samples. After face alignment, we obtain the normalized faces which are resized to be 230 × 230.

For all models based on the modified AlexNet, the initial learning rate is set to 0.01 and decays by a factor of 0.5 for every 20,000 iterations. For the models based on ResNet-18, the learning rate starts at 0.05 and reduces with a factor of 0.5 for every 20,000 iterations. The pose degrees number G is set to 13.

We use LFW, YTF and MegaFace Challenge dataset as the unconstrained face recognition testing datasets. LFW dataset is an image dataset containing 13,233 face images with large variations in illumination, pose, expression and occlusion of 5,749 individuals. YTF dataset is a video database which includes 3,424 videos from 1,595 different identities. All videos in YTF are collected from YouTube and Face images in YTF dataset have larger variations in illumination, pose and expression, and lower resolution than those in LFW dataset. LFW and YTF are excellent benchmarks for face verification and identification in image and video. MegaFace consists of a gallery set and a probe set. The gallery set includes more than 1 million face images of about 690,000 identities. The probe set contains two subsets, Facescrub and FGNet. Facescrube consists of 100k images from 530 different individuals, while FGNet includes about 100 images of 82 different individuals. We use the Facescrube dataset as the probe set. Compared with LFW dataset, MegaFace is a more challenging database which includes images of richer scene and larger pose.

4.2 Effectiveness of Multiple Patches

To evaluate the effective of using multiple patches, we train models with different number of patches as input of the model on Multi-PIE dataset. In this experiment, the multi-task learning is included in all trained models, the modified AlexNet is used as the CNN network and we set the regularization parameter $\lambda = 0.1$.

Table 2 shows the performance comparison of each single CNN and combinations of different number of CNNs. For the performance of the combination of different number of patches, we provide the range of the results of all combinations and the combination with the best performance. For example, combinations of 2 image patches with ten kinds of combination achieve identification performance from 90.53% to 90.67% and the best performance 90.67% is the Avg. rank-1 identification rate of ‘H+re+n’, i.e. the combination of the holistic image and the patches centered at the right eye center and the nose tip. It can be seen from Table 2 that the combination of the CNN features of multiple patches with that from holistic image can obtain better

Table 3 Performance comparison on setting V of Multi-PIE. FR represents the face recognition task that is singly used without MTL, while Multi-task represents the proposed model with MTL of the two tasks.

Model	Parameter λ	Avg. rank-1 (%)
Single task: FR	–	86.43
Multi-task	$\lambda = 0$	88.01
Multi-task	$\lambda = 0.001$	88.12
Multi-task	$\lambda = 0.01$	89.06
Multi-task	$\lambda = 0.1$	91.85
Multi-task	$\lambda = 1$	89.97

performance than only extracting global features. Moreover in our experiment, the identification performance increases as the patches increase. We also train a model with randomly cropped 5 patches, and as shown in Table 2, the proposed method that uniformly samples patches with semantic meaning outperforms the randomly cropping manner.

4.3 Effectiveness of Multi-Task Learning and Parameter λ

In this experiment, we train models with single task and multi-task models with different regularization parameters to analyze the effect of multi-task learning and parameter λ on JMM-CNNs model. Here, we just change the latter part of the architecture of JMM-CNNs (i.e. the multi-task learning part) with the feature extraction part of 6 CNNs and feature fusion part unchanged. The modified AlexNet is also used as the CNN network here. Table 3 shows the performance comparison of single-task learning and multi-task learning with different regularization parameters. It can be seen that adding the pose estimation task is helpful to improve the performance of face recognition task, even in the circumstance that regularization term is zero. It also can be seen from Table 3 that the regularization parameter $\lambda = 0.1$ can achieve the best performance among the value interval $\{0, 0.001, 0.01, 0.1, 1\}$. The result is reasonable as the regularization term cannot work when λ is too small and the effect of the separate learning task will be weaken when λ is large.

Combining Table 2 and Table 3, it can be concluded that the combination of multi-patch and multi-task CNNs (5 patches and $\lambda = 0.1$) can increase the identification performance to 91.85% compared with the performance of 90.36% that singly uses multi-task CNN ($\lambda = 0.1$) or the performance of 86.43% that singly uses multi-patch CNNs (5 patches).

4.4 Performance on LFW and YTF Datasets

We retrain our model on CASIA-WebFace dataset and in this section conduct unconstrained face verification on two well-known datasets, LFW and YTF dataset. The proposed model is evaluated by comparing with the state-of-the-art CNN-based methods including VGGFace [42], DeepFace, DeepID, DeepID2, FaceNet, CenterFace [43], and MultiBatch [44], etc. And to further evaluate the effectiveness of the multi-patches feature fusion and multi-task learning of the proposed JMM-CNNs to unconstrained face recognition, we also conduct experiments of removing the multi-patches feature fusion or the pose estimation task of JMM-CNNs_Alex (JMM-CNNs based on the modified AlexNet) on LFW and YTF datasets. Following the unrestricted with labeled outside data protocol, we evaluate our model on 6000 face pairs from LFW dataset and 5000 video pairs from YTF dataset.

Table 4 Verification rates of different methods on LFW and YTF datasets. JMM-CNNs_Res and JMM-CNNs_Alex represent the proposed JMM-CNNs based on ResNet-18 and the modified AlexNet respectively. 'without MP' denotes removing the CNNs corresponding to the image patches in JMM-CNNs_Alex, i.e. only including the CNN-H with multi-task learning in JMM-CNNs_Alex, and 'without MTL' denotes removing the pose estimation task in JMM-CNNs_Alex.

Method	Data	#Net	Metric	LFW (%)	YTF (%)
DeepFace [13]	4M	3	Cosine	97.35	91.40
VGGFace [42]	2.6M	1	Euclidean	98.95	97.30
FaceNet [12]	200M	1	L2	99.65	95.10
MultiBatch [44]	2.6M	1	Euclidean	98.2	/
Xi Yin et al. [15]	0.49M	1	Cosine	98.27	/
DeepID2 [9]	300K	25	Joint-Bayes	98.97	93.20
CenterFace [43]	0.7M	1	Cosine	99.28	94.90
DeepID [8]	5.8M	100	Joint-Bayes	97.45	/
MMDFR [17]	0.47M	8	Joint-Bayes	99.02	/
without MP	0.49M	1	Cosine	98.59	93.09
without MTL	0.49M	6	Cosine	97.91	91.39
JMM-CNNs_Alex	0.49M	6	Cosine	99.14	94.93
JMM-CNNs_Res	0.49M	6	Cosine	99.76	95.74
JMM-CNNs_Res	2.6M	6	Cosine	99.83	97.18

From the results in Table 4, we conclude the following three observations. First, JMM-CNNs_Alex with multi-patch feature fusing achieves better performance than using the same model architecture but without multi-patch feature fusing. The verification rates are improved from (98.59% on LFW and 93.09% on YTF) to (99.14% on LFW and 94.93% on YTF) by local features fusing. Second, JMM-CNNs_Alex with multi-task learning beats the model with same architecture but without multi-task learning, through improving the performance by 1.23% on LFW and 3.54% on YTF. These two observations show the advantage of the combination of multi-patch and multi-task CNNs in the designed JMM-CNNs. Finally, JMM-CNNs_Res achieves better performance than JMM-CNNs_Alex and outperforms most of the state-of-the-art methods including multi-patch-based methods such as DeepID2 and multi-task-learning-based methods like Xi Yin et al., except VGGFace, although some of these methods apply larger training dataset or more networks. Since VGGFace employs relatively larger training dataset, we also train the JMM-CNNs_Res on a 2.6M dataset for relatively fair comparison. The dataset used by VGGFace is not publicly available, so we form a 2.6M training dataset by selecting about 2.11M images randomly from MultiPIE and supplementing the CASIA-WebFace dataset with these images. The results in Table 4 indicate that the JMM-CNNs_Res trained on the 2.6M dataset achieves a comparable performance with VGGFace.

4.5 Performance on MegaFace Dataset

To further evaluate the effectiveness of the proposed model, we also execute experiments on the standard settings of the MegaFace dataset, which is currently the largest face recognition benchmark. MegaFace has several testing scenarios (including identification and verification) under two protocols, large and small training set protocol. Since our training set is less than 0.5M images and 20K identities, it belongs to the protocol of small training set. There are different distractors (from 10 to 1M) in MegaFace gallery, which increases testing challenge. Here, we demonstrate the results on 1M distracters.

We compare our method with the typical participating methods in MegaFace challenge 1, such as Google-FaceNet v8, NTechLAB-facenx large, Barebones FR-cnn, and SIAT_MMLAB, etc., and three publicly released methods in recent years, CenterFace [43], Lightened CNN [46], and Attr-constr CNN [14]. The results are given in Table 5 and Fig. 5. Table 5 shows face identification and verification performance of methods including those under large training set protocol on MegaFace challenge. From these results, we have the observation that the proposed model outperforms most of the methods although it is trained under the small training set protocol. These results also show the advantage of the combination of multi-patch and multi-task CNNs in the designed JMM-CNNs. In Fig. 5, we demonstrate the results using Cumulative Match Characteristics (CMC) curves for face verification and Receiver Operating Characteristic (ROC) curves for face identification, respectively. It can be seen from Fig. 5, our method achieves a competitive result compared with other results under the small training set protocol of MegaFace dataset.

Table 5 Identification Accuracy (%) of different methods and Verification TAR (%) of different methods at 10^{-6} FAR on MegaFace with 1M distractors. Iden. Acc. indicates rank-1 identification accuracy, while Ver. Acc. indicates verification TAR for 10^{-6} FAR.

Method	Protocol	Iden. Acc.	Ver. Acc.
`NTechLAB-facenx large`	Large	73.30	85.08
`Google-FaceNet v8` [12]	Large	70.50	86.47
`Beijing FaceAll_Norm_1600`	Large	64.80	67.12
`Beijing FaceAll_1600`	Large	63.98	63.96
`Barebones FR-cnn`	Small	59.36	59.04
`NTechLAB-facenx_small`	Small	58.22	66.37
`3DiVi Company-tdvm6`	Small	33.71	36.93
`SIAT_MMLAB` [45]	Small	65.23	76.72
`Attr-constr` [14]	Small	77.74	79.24
`CenterFace` [43]	Small	65.23	76.52
`Lightened CNN` [46]	Small	67.11	77.64
`without MP`	Small	74.71	77.94
`without MTL`	Small	72.93	76.49
`JMM-CNNs_Alex`	Small	77.53	80.61
`JMM-CNNs_Res`	Small	78.22	81.53

5. Conclusions

In this paper, we propose a joint multi-patch and multi-task CNNs model which combines multi-patch CNNs and multi-task learning CNN in one framework. The proposed model can extract more descriptive and robust facial feature for face recognition by fusing global and local features and multi-task learning of face recognition and pose estimation. To make the proposed model have an end-to-end architecture, we use a simple patch sampling strategy that crop the image patches only depending on the facial landmarks. Meanwhile, we propose a regularized loss function to further enhance the pose insensitiveness of the extracted facial feature. Experiments on Multi-PIE, LFW, YTF and MegaFace dataset prove the advantage and effectiveness of the proposed model which outperforms most of the existing state-of-the-art face recognition models.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Grant No. 61502444), the youth project of science and technology research program of Chongqing Education Commission of China (Grant No. KJQN201801119), and the Scientific Research Foundation of Chongqing University of Technology (Grant No. 2017ZD58).

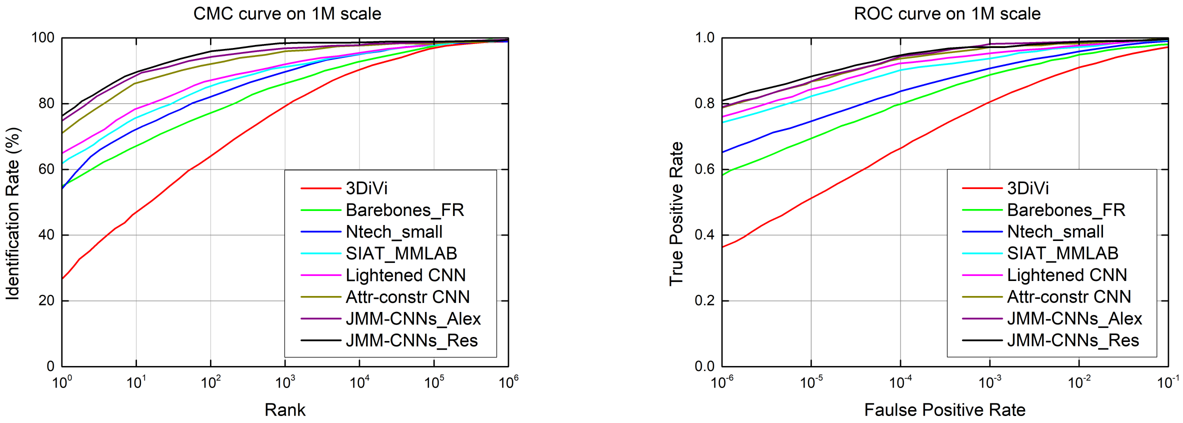


Fig. 5 CMC and ROC curves of different methods on MegaFace Challenge under the small training set protocol.

References

[1] H. Tan, B. Yang, and Z. Ma, "Face recognition based on the fusion of global and local HOG features of face images," Iet Computer Vision, vol.8, no.3, pp.224–234, 2013.

[2] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of Dimensionality: High-Dimensional Feature and Its Efficient Compression for Face Verification," Computer Vision and Pattern Recognition, pp.3025–3032, 2013.

[3] A. Timo, H. Abdenour, and P.I. Matti, "Face description with local binary patterns: application to face recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.28, no.12, pp.2037–2041, 2006.

[4] T. Ahonen, A. Hadid, and a. M. Pietikäinen, "Face recognition with local binary patterns," European Conference on Computer Vision, 2004.

[5] C.S. Inn, n. N.A., S.K. Phooi, and A.L. Minn, "Block-based deep belief networks for face recognition," International Journal of Biometrics, vol.4, no.2, pp.130–143, 2012.

[6] M. Kan, S. Shan, H. Chang, and X. Chen, "Stacked Progressive Auto-Encoders (SPAE) for Face Recognition Across Poses," 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp.1883–1890, 2014.

[7] P.J.S. Vega, R.Q. Feitosa, V.H.A. Quirita, and P.N. Happ, "Single Sample Face Recognition from Video via Stacked Supervised Auto-Encoder," 2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp.96–103, 2016.

[8] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.1891–1898, 2013.

[9] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," arXiv preprint arXiv:1406.4773, 2014.

[10] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," CVPR, pp.2892–2900, 2015.

[11] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face recognition with very deep neural networks," Computer Science, 2015.

[12] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," arXiv preprint arXiv:1503.03832v1, pp.815–823, 2015.

[13] Y. Taigman, M. Yang, M.A. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.1701–1708, 2013.

[14] Z. Wang, K. He, Y. Fu, R. Feng, Y.-G. Jiang, and X. Xue, "Multi-task Deep Neural Network for Joint Face Recognition and Facial Attribute Prediction," presented at the 2017 ACM, 2017.

[15] X. Yin and X. Liu, "Multi-Task Convolutional Neural Network for Pose-Invariant Face Recognition," IEEE Trans. Image Process., vol.27, no.2, pp.964–975, 2018.

[16] L. Tran, X. Yin, and X. Liu, "Disentangled Representation Learning GAN for Pose-Invariant Face Recognition," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1283–1292, 2017.

[17] C. Ding and D. Tao, "Robust Face Recognition via Multimodal Deep Face Representation," IEEE Trans. Multimedia, vol.17, no.11, pp.2049–2058, 2015.

[18] Y. Zhang, K. Shang, J. Wang, N. Li, and M.M.Y. Zhang, "Patch strategy for deep face recognition," IET Image Processing, vol.12, no.5, pp.819–825, 2018.

[19] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," 8th IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp.1–8, 2008.

[20] G.B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, 2008.

[21] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," Computer Vision and Pattern Recognition, pp.529–534, 2011.

[22] D. Miller, E. Brossard, S. Seitz, and I. Kemelmachershlizerman, "MegaFace: A million faces for recognition at scale," Computer Science, 2015.

[23] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli, "On the Use of SIFT Features for Face Authentication," 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06), p.35, 2006.

[24] W. Zhang, S. Shan, G. Wen, X. Chen, and H. Zhang, "Local Gabor Binary Pattern Histogram Sequence (LGBPHS): A novel non-statistical model for face representation and recognition," Tenth IEEE International Conference on Computer Vision, 2005.

[25] N. Kumar, A.C. Berg, P.N. Belhumeur, and S.K. Nayar, "Attribute and simile classifiers for face verification," IEEE International Conference on Computer Vision, pp.365–372, 2010.

[26] T. Berg and P.N. Belhumeur, "Tom-vs-Pete Classifiers and Identity-Preserving Alignment for Face Verification," Bmvc, 2012.

[27] C. Szegedy et al., "Going deeper with convolutions," arXiv preprint arXiv:1409.4842, pp.1–9, 2015.

[28] A. Liu, N. Xu, W. Nie, Y. Su, and Y. Zhang, "Multi-Domain and Multi-Task Learning for Human Action Recognition," IEEE Trans. Image Process., vol.28, no.2, pp.853–867, 2019.

[29] Y. He, J. Zhang, H. Shan, and L. Wang, "Multi-Task GANs for View-Specific Feature Learning in Gait Recognition," IEEE Trans. Inf. Forensics Security, vol.14, no.1, pp.102–113, 2019.

[30] J. Deng, J. Guo, and S. Zafeiriou, "Single-Stage Joint Face Detection and Alignment," 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp.1836–1839, 2019.

[31] R. Ranjan, V.M. Patel, and R. Chellappa, "HyperFace: A Deep Multi-task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.41, no.1, pp.121–135, 2019.

[32] R. Ranjan, S. Sankaranarayanan, C.D. Castillo, and R. Chellappa, "An All-In-One Convolutional Neural Network for Face Analysis," FG 2017 proceedings, vol.1, pp.17–24, 2017.

[33] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems, vol.25, no.2, pp.1097–1105, 2012.

[34] A. Khan and N. Wahab, "Deep residual learning for image recognition," CVPR, 2016.

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167, 2015.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," 2015 IEEE International Conference on Computer Vision (ICCV), pp.1026–1034, 2015.

[37] D. Rumelhart, G. Hinton, and R. Williams, "Learning Representations by Back-Propagating Errors," Nature, vol.323, pp.533–536, 10/09 1986.

[38] Z. Zhang, P. Luo, C.C. Loy, and X. Tang, "Facial Landmark Detection by Deep Multi-task Learning," presented at the European Conference on Computer Vision, 2014.

[39] D. Yi, Z. Lei, S. Liao, and S.Z. Li, "Learning face representation from scratch," Computer Science, 2014.

[40] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards Large-Pose Face Frontalization in the Wild," ICCV, pp.3990–3999, 2017.

[41] A. Jourabloo and X. Liu, "Pose-invariant face alignment via CNN-based dense 3D model fitting," International Journal of Computer Vision, pp.1–17, 04/19 2017.

[42] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," British Machine Vision Conference, 2015.

[43] B. Leibe, J. Matas, N. Sebe, and M. Welling, "A discriminative feature learning approach for deep face recognition," European Con-

ference on Computer Vision, 2016.

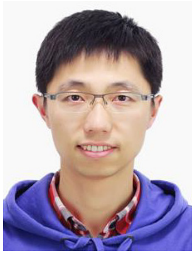
[44] O. Tadmor, Y. Wexler, T. Rosenwein, S. Shalev-Shwartz, and A. Shashua, "Learning a metric embedding for face recognition using the multibatch method," Proc. NIPS, 05/23 2016.

[45] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A Discriminative Feature Learning Approach for Deep Face Recognition," presented at the European Conference on Computer Vision (ECCV), 2016.

[46] X. Wu, R. He, and a. Z. Sun, "A lightened CNN for deep face representation," arXiv preprint arXiv:1511.02683, 2015.

Yanfei Liu received the B.S. and M.S. degrees in Automation from Chongqing University of Post and Telecommunication, China, in 2003 and 2006, respectively, and the Ph.D. degree in Circuits and System from Chongqing University, China, in 2012. During 2012-2017, she stayed in Chongqing institute of Green and Intelligent Technology, Chinese Academy of Science to study pattern recognition. She is currently an Associate Professor with Chongqing University of Technology, Chongqing, China.

Junhua Chen received the B.S. and M.S. degrees in Control Science and Engineering from Chongqing University of Posts and Telecommunications in 2003 and 2006, respectively. From 2006 to 2014, he was a research engineer of 3G/4G baseband chip product with Chongqing Chongyou Information Technology Group, Inc. He is currently a senior engineer with Key Laboratory of Industrial Internet of Things & Networked Control, Chongqing University of Posts and Telecommunications, Chongqing, China.



Yu Qiu received the B.S degree in Electronic Information Engineering, M.S degree in Communication and Information System, and Ph.D. degree in Circuit and System from Chongqing University in 2003, 2007 and 2011 respectively. He is currently an Associate Professor of Electrical Engineering in Chongqing Industry Polytechnic College. His research interests includes image processing and Industrial Automation.