1-1-2020

Low resolution face recognition using a two-branch deep convolutional neural network architecture

E. Zangeneh
Amirkabir University of Technology

Mohammad Rahmati
Amirkabir University of Technology

Y. Mohsenzadeh
Western University, ymohsenz@uwo.ca

Low Resolution Face Recognition Using a Two-Branch Deep Convolutional Neural Network Architecture

Erfan Zangeneh, Mohammad Rahmati, and Yalda Mohsenzadeh

***Abstract*—We propose a novel coupled mappings method for low resolution face recognition using deep convolutional neural networks (DCNNs). The proposed architecture consists of two branches of DCNNs to map the high and low resolution face images into a common space with nonlinear transformations. The branch corresponding to transformation of high resolution images consists of 14 layers and the other branch which maps the low resolution face images to the common space includes a 5-layer super-resolution network connected to a 14-layer network. The distance between the features of corresponding high and low resolution images are backpropagated to train the networks. Our proposed method is evaluated on FERET data set and compared with state-of-the-art competing methods. Our extensive experimental evaluations show that the proposed method significantly improves the recognition performance especially for very low resolution probe face images (11.4% improvement in recognition accuracy). Furthermore, it can reconstruct a high resolution image from its corresponding low resolution probe image which is comparable with the state-of-the-art super-resolution methods in terms of visual quality.**

***Index Terms*—low resolution face recognition, super-resolution methods, coupled mappings methods, deep convolutional neural networks**

I. INTRODUCTION

In the past few decades, face recognition has shown promising performance in numerous applications and under challenging conditions such as occlusion [1], variation in in pose, illumination, and expression [2]. While many face recognition systems have been developed for recognizing high quality face images in controlled conditions [3], there are a few studies focused on face recognition in real world applications such as surveillance systems with low resolution faces [4]. One important challenge in these applications is that high resolution (HR) probe images may not be available due to the large distance of the camera from the subject. Thus the performance of traditional face recognition systems which are developed for high quality images degrades considerably for these images with low resolution (LR) face regions [5]–[7]. These LR face images typically have a size smaller than 32×24 pixels with an eye-to-eye distance about 10 pixels [8].

Here, we focus on addressing the problem of recognizing low resolution probe face images when a gallery of high quality images is available. There are three standard approaches to address this problem. 1) down sampling the gallery images to the resolution of the probe images and then performing the recognition. However, this approach is suboptimal because the additional discriminating information available in the high resolution gallery images is lost. 2) The second approach is to obtain higher resolution probe images from the low resolution images, which are then used for recognition. Most of these super-resolution techniques aim to reconstruct a good high resolution image in terms of visual quality and are not optimized for recognition performance [9]. Some of the well known methods of this category are [10]–[13] 3) Finally, the third approach simultaneously transforms both the LR probe and the HR gallery images into a common space where the corresponding LR and HR images are the closest in distance; [14]–[17] are the well known methods of this approach. Fig. 1 summarizes the three general ways for low resolution face recognition (LR FR) problems.

In this paper, we use the third approach and propose a method that employs deep convolutional neural networks (DCNNs) to find a common space between low resolution and high resolution pairs of face images. Despite previous works that used linear equation as objective function to find two projection matrices, our work finds a nonlinear transformation from LR and HR to common space. In our proposed method, the distance of transformed low and high resolution images in the common space is used as an objective function to train our deep convolutional neural networks. Our proposed method also reconstructs good HR face images which are optimum for the recognition task. We evaluated the effectiveness of the proposed approach on the FERET database [18]. Our results show the proposed approach improves the matching performance significantly as compared to other state-of-the-art methods in the low resolution face recognition and the improvement becomes more significant for very low resolution probe images. The main contributions of this study can be summarized as:

- We proposed a novel nonlinear coupled mapping architecture using two deep convolutional neural networks to project the low and high resolution face images into a common space.
- Our proposed method offers high recognition accuracy compared to other state-of-the-art competing methods especially when the probe image is extremely low resolution.
- Our proposed coupled mappings method also offers high resolution version of the low resolution input image because of an embedded super-resolution CNN in its architecture.
- Our proposed method needs much less space compared

to the typical face recognition methods that use deep convolutional neural networks such as VGGnet [19]. This feature makes it applicable on regular systems with lower Memory.

In the following, we first review previous works presented in the domain of low resolution face recognition both super-resolution methods and coupled mapping methods in Section 2. In Section 3, we present our proposed method, network architecture and training procedure. Eventually, we present the experimental evaluation results in Section 4 and discussion and conclusion in Section 5.

II. PREVIOUS WORKS

In this section, we briefly review the related works in the literature of low resolution face recognition and also introduce deep convolutional neural networks.

Low resolution face recognition: To resolve the mismatch between probe and gallery images, most of studies concentrated on super-resolution approaches. The aim of these approaches is to obtain a HR image from the LR input and then use the obtained HR image for recognition. Some super-resolution studies suggested using face priors for image reconstruction. The learning method proposed by Chakrabarti et al. [20] for super-resolution uses kernel principal component analysis to derive prior knowledge about the face class. To achieve good reconstruction results, Liu et al. [11] presented a two-step statistical modeling approach for hallucinating a HR face image from a LR input image. Baker [21] also proposed a face hallucination method based on face priors. Freeman et al. [22] trained a patch-wise Markov network as a super-resolution prediction model. Yang et al. [13] used compressed sensing to reconstruct a super-resolution image from a low resolution input image. Zou et al. [10] proposed a super-resolution method which clusters low resolution face images and then a projection matrix corresponding to the assigned cluster maps LR image into HR space. They proposed two separate projection matrices for optimal visualization and recognition purposes. In [13], the authors suggested a sparse coding method to find a representation of the LR input patch in terms of its neighboring image patches; then the same representation coefficients were used to reconstruct the target HR patch based on the corresponding neighboring HR patches.

The other category of works on LR FR is known as coupled mappings method. These methods learn the transformations using a training set consisting of HR images and LR images of the same subjects. Given training data, the goal is to find transformations which minimizes the distances between the transformed LR and HR feature vectors, x_i^l and x_i^h, respectively. Most of coupled mappings methods use linear objective function as following [23]:

$$J(W_L, W_H) = \sum_{i=1}^{n}\sum_{j=1}^{n} ||W_L^T x_i^l - W_H^T x_j^h||^2 P_{ij} \qquad (1)$$

where n is the number of training images and $\{x_i^h\}_n^{i=1}$ and $\{x_i^l\}_n^{i=1}$ are corresponding extracted features of the HR and LR images, respectively. W_L and W_H denote the linear

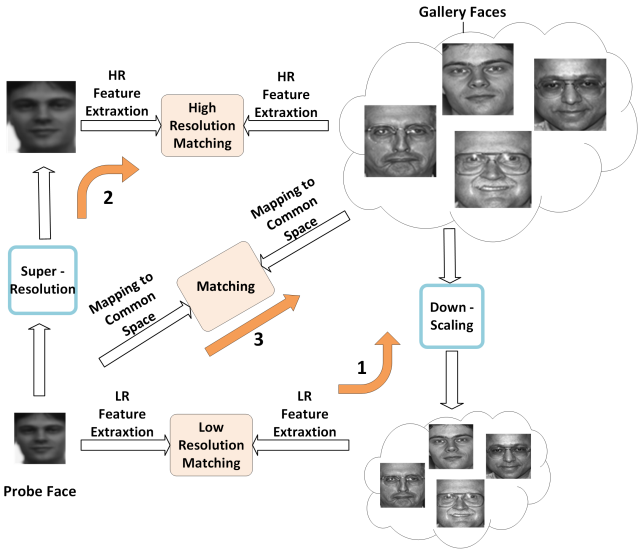

Fig. 1. Three general approaches for low resolution face recognition.

mappings of low resolution and high resolution feature vectors to the common space, respectively. P is a $n \times n$ penalty weighting matrix that preserves the local relationship between data points in the original feature spaces and it is defined on the neighborhoods of the data points as follows:

$$P_{ij} = \begin{cases} exp(-\frac{||x_i^h - x_j^h||^2}{\sigma^2}) & j \in C(i) \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

Here, $C(i)$ contains the indices of k nearest neighbors of x_i^h in high resolution space and σ is Gaussian function width which is defined as

$$\sigma = \frac{\alpha \sum_{i,j} ||x_i^h - x_j^h||^2}{n^2} \qquad (3)$$

where α is a scale parameter.Since it is assumed that HR feature space has more discriminative information, the goal of the above objective function is to find a common feature space similar to HR feature space. Finally after optimizing the above objective function, W_L and W_H will be found, and low and high resolution images can be transformed into common space with these mappings, respectively. P. Hennings et al. [14] proposed a joint objective function which aims to optimize both super-resolution and face recognition. While this method improves the recognition accuracy compared to two-step methods, its optimization procedure is slow. Mainly because its optimization procedure has to be executed for each test image with respect to each enrollment. In [15], the author used singular value decomposition (SVD) of face images in multi resolution to map low resolution images to high resolution space. Furthermore, the method improved both the hallucination and the recognition accuracy. Huang et al. [24] proposed a method which finds a common space for low resolution probe and high resolution gallery images and an objective function that guarantees the discriminability in the new common space. Biswas et al. [16] used multidimensional scaling transformation learning to find both low resolution and

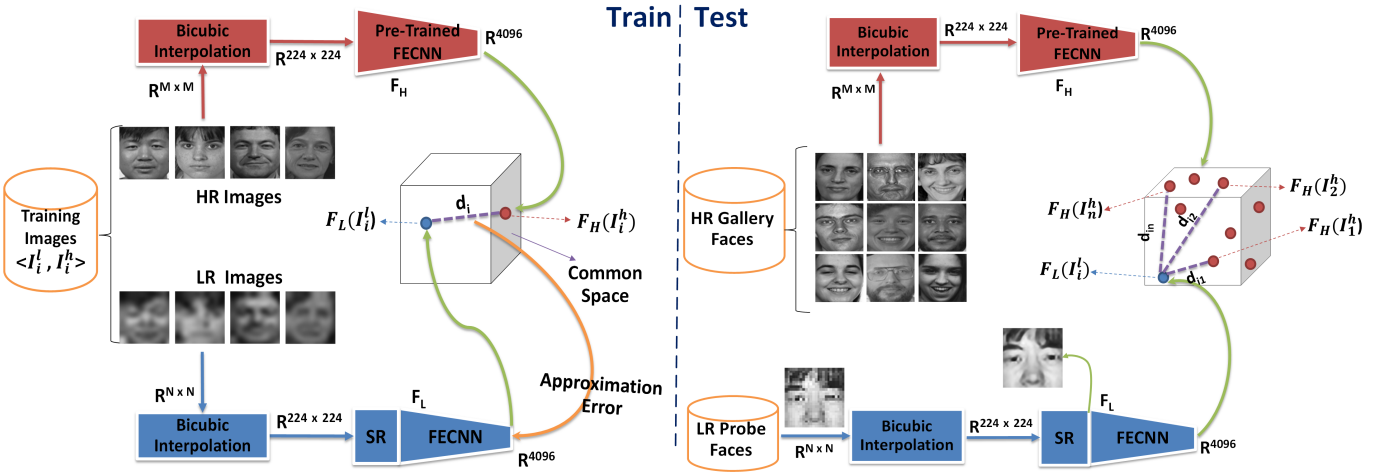

Fig. 2. Overview of our method. M and N denote dimensions of HR and LR images, respectively, and $M > N$.

high resolution projection matrices. The objective function of optimization problem enforces the same distance between low resolution and high resolution image pairs of a class in the common space as the distance of high resolution image pairs of that class. Huang et al. [24] used canonical correlation analysis (CCA) to project low resolution and high resolution images into a common space where a low resolution image and its correspond high resolution image are as close as possible. Later, Ren et al. [25] employed coupled nonlinear kernels to map the LR and HR face image pairs into an infinite common subspace. Also a coupled linear mappings method was presented in Zhou et al. [17] using the classical discriminant analysis. Shi et al. [26] proposed an optimization objective function including three terms, associated with the LR/HR consistency, intraclass compactness and interclass separability. Zhang et al. [27] introduced coupled marginal discriminant mappings (CMDM) method. The method uses gaussian similarity between pairs of class-mean samples from HR images to construct intraclass similarity matrix. The interclass similarity matrix is defined by the gaussian similarity between sample pairs from HR images in the same class. Also, Zhang et al. [27] solved the coupled mappings problem as an eigen decomposition problem that helps to achieve good recognition performance when faces are occluded. Mudunuri et al. [28] proposed a coupled mappings method that at first aligned faces by detecting eyes and then computed the SIFT descriptor of probe faces to transform them to a common space. Stereo matching cost function is then used to preserve distance in the transformed space across different illumination, pose and resolution.

In summary, coupled mappings methods achieve better recognition performance than super-resolution methods, but these methods do not aim at reconstructing a high resolution image from the low resolution input image. On the other hand, the main objective of super-resolution methods is to reconstruct a high quality image for visualization purposes which may not necessarily offer better recognition accuracy.

Deep convolutional neural networks: Although convolutional neural networks (CNN) were first presented three decades ago [29], since introduction of AlexNet [30] in 2012, deep CNNs have become explosively popular especially due to their success in computer vision. The main elements that help to this popularity are [31]:

- Participation of the best labs of top universities in computer vision challenges such as ILSVRC [32] and PASCAL VOC [33]
- Easy access to data with larger size such as ImageNet [34].
- Introduction of more efficient activation functions like the rectified linear unit (ReLU) [35], and exponential linear unit (ELU) [36] which help DCNNs in faster convergence.
- Existence of modern GPU like NVIDIA TITAN black X, and also efficient deep learning frameworks such as Caffe [37] and Tensorflow [38].

In the next section, we propose a coupled mappings method using deep convolutional neural networks for nonlinear mapping to a common space. Our method similar to other successful methods that use deep convolutional neural networks, benefits from the above mentioned advantages. In addition to offering high recognition performance, the proposed method also produces high resolution images from low resolution input images.

III. PROPOSED METHOD

Due to the difficulty of solving a nonlinear optimization problem, objective functions in previous coupled mappings methods (as discussed in Section 2) were modeled with a linear transformation. However, a nonlinear transformation of low resolution and high resolution to a common space can possibly result in a better performance. Here, we propose a nonlinear coupled mappings approach which uses two deep convolutional neural networks (DCNNs) to extract features from low resolution probe images and high resolution gallery images and project them into a common space. We use gradient based optimization to minimize the distance between the mapped HR and LR image pairs in the common space

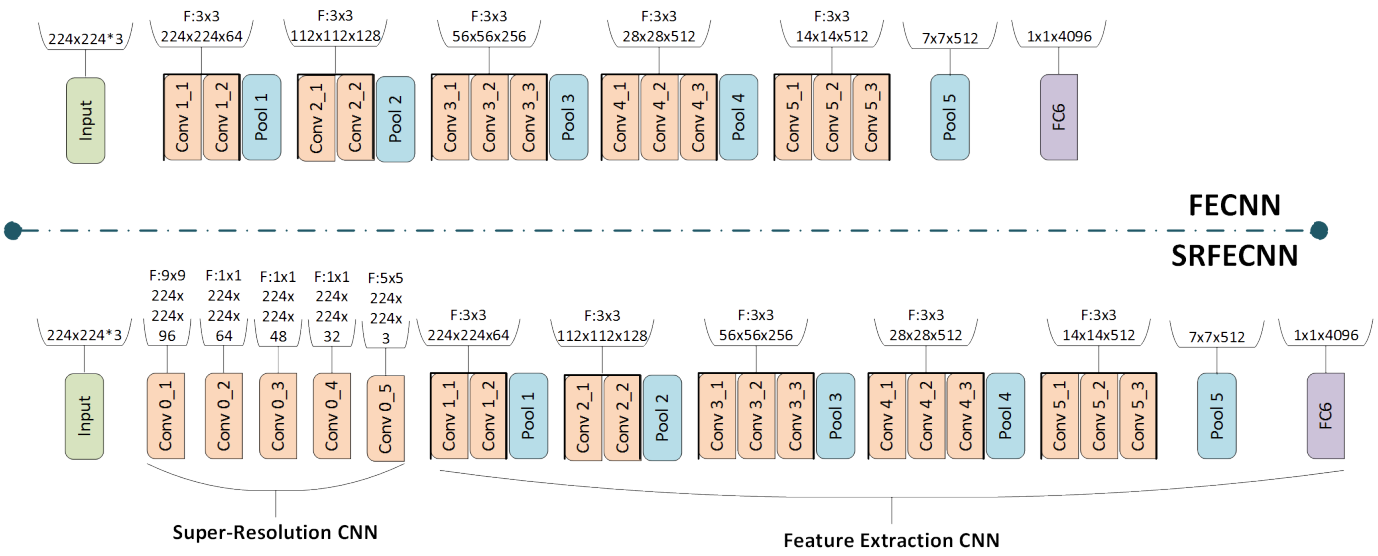

Fig. 3. Architecture of two deep convolutional neural networks in two branches of our proposed method.

with updating the weights of DCNN by backpropagation of the error.

Fig. 2 shows the overview of our proposed architecture. In training phase, we use a training image set that contain pairs of low resolution and high resolution images of the same person which can vary in different images under different conditions of illumination, pose and expression (not necessarily the same image only with different resolutions). In the next section, we present the architecture of the proposed method in detail.

A. Networks Architecture

Our method has a two branch architecture that one of them projects high-resolution images to the common space and the other one maps low-resolution images into this common space. In our method we use a DCNN known as VGGnet [9]. The most well-known configuration of this network has sixteen layers with thirteen convolutional layers and three fully connected layers. The last fully connected layer of VGGnet used for a specific classification task. In the top branch of our method (Fig. 2), we dropped out two last fully connected layers of this VGGnet and called it feature extraction convolutional neural network (FECNN). The input image of the top branch of our method is the high resolution image (I_i^h) that has to be in 224×224 dimensions (whenever input image size is different from 224×224, we use traditional bicubic interpolation method to obtain the required size). The output from the last layer is a feature vector with 4096 elements.

In the bottom branch of our method, we use a DCNN previously used for super-resolving low resolution images following by a second network which has a similar architecture as the network in the top branch. The first subnet has a similar architecture as DCNN that proposed by Dong et al. [31], but we extended this architecture from three layers to five layers, although the authors show there is no difference between a three layer architecture and a five layer one in terms of visual quality of reconstructed images, we found increasing of layers from three to five improves the recognition performance of our method. We call the first subnet of our bottom branch super-resolution net (SRnet). The output of the first subnet is fed into the second subnet (FECNN). Therefore, the top branch net of our method consists of fourteen layers and the bottom branch includes nineteen layers as shown in Fig. 3. The input of bottom branch net is the low resolution image(I_i^l) that has to be interpolated with the traditional interpolation method to the size of 224×224. Also, the output of SR subnet is an image with the size of 224×224.

As mentioned above, the FECNN net has the same architecture as VGGnet excluding the last two fully connected layers. Although the super resolution and feature extraction convolutional neural network (SRFECNN) has eighteen convolutional layers and one fully connected, the entire number of weights used in SRFECNN is much less than VGGnet. Table I shows all used weights for SRFECNN. Even though our proposed SRFECNN includes eighteen convolutional layers, because of less number of fully connected layers compared to VGGnet, it has less number of weights than VGGnet ($141M$ weights). Thus in testing phase when we need to load SRFECNN weights on memory, our proposed method needs much less space than VGGnet. This is an important feature which makes our proposed method applicable on systems with lower memory.

B. Common Subspace Learning

We trained our network in three stages as summarized below:

- First, we used trained VGGnet on face dataset [19] and then dropped the last two fully connected layers, because they are specific to the classification task the network is trained on. We called this network pre-trained FECNN and used it in both top and bottom branches of our architecture.

Table I
NUMBER OF USED WEIGHTS IN LAYERS OF SRFECNN.

Layer set	Parameters	Number of weights
Conv0_1	$F = 9 \times 9$ $Depth = 96$	$3 \times 9 \times 9 \times 96 = 23328$
Conv0_2	$F = 1 \times 1$ $Depth = 64$	$96 \times 1 \times 1 \times 64 = 6144$
Conv0_3	$F = 1 \times 1$ $Depth = 48$	$64 \times 1 \times 1 \times 48 = 2928$
Conv0_4	$F = 1 \times 1$ $Depth = 32$	$48 \times 1 \times 1 \times 32 = 1536$
Conv0_5	$F = 5 \times 5$ $Depth = 3$	$32 \times 5 \times 5 \times 3 = 2400$
Conv1 (2 Convs)	$F = 3 \times 3$ $Depth = 64$	$2(3 \times 3 \times 3 \times 64) = 3456$
Conv2 (2 Convs)	$F = 3 \times 3$ $Depth = 128$	$2(64 \times 3 \times 3 \times 128) = 147456$
Conv3 (3 Convs)	$F = 3 \times 3$ $Depth = 256$	$3(128 \times 3 \times 3 \times 256) = 884736$
Conv4 (3 Convs)	$F = 3 \times 3$ $Depth = 512$	$3(256 \times 3 \times 3 \times 512) = 3538944$
Conv5 (3 Convs)	$F = 3 \times 3$ $Depth = 512$	$3(512 \times 3 \times 3 \times 512) = 7077888$
FC6	$Depth = 4096$	$7 \times 7 \times 512 \times 4096 = 102760448$
All Layers		$114449264 \approx 114M$

- In the second step, we trained the SRnet of the bottom branch with a database of high and low resolution face image pairs. The details of used datasets are presented in the experimental evaluation section.
- The third step is the main training phase. We merged the two subnets namely SRnet and FECNN and a training database that contains pairs of low resolution and high resolution of same persons was fed into the bottom and top branches, respectively.

We considered the top branch FECNN net and the bottom branch SRFECNN as two nonlinear functions that project a high resolution image and low resolution image to a 4096 dimensional common space:

$$\phi_i^h = F_H(I_i^h) \tag{4}$$

$$\phi_i^l = F_L(I_i^l) \tag{5}$$

where $I_i^h \in R^{M \times M}$ and $I_i^l \in R^{N \times N}$ that $N < M$. During this phase of training $F_H(I_i^h)$ was considered fixed and did not change, but $F_L(I_i^l)$ was trained to minimize the distance between low and high resolution images of same subjects in the common space. With this aim, the distance was backpropagated into the bottom branch net (both FECNN and SRnet) as an error.

The main training procedure was repeated many times for all pairs of training images. We reduced learning rate of all layers to fine-tune the weights obtained in the first two training phases. However, the learning rate of first layers of FECNN is less than last layers of it, because in a specific problem, last layers of a DCNN have more discriminant information about the problem and the first layers of it have more general features that can change sparsely [39].

C. Reconstruct Input Image

Additionally, our method can reconstruct a high resolution image from the low resolution probe image. First subnet of the bottom branch used for super-resolution to produce a high resolution face image from the low resolution probe face to feed into FECNN. In the test phase, after feeding low resolution probe image into the bottom net we can extract corresponding high resolution face image from the last layer of SRnet.

D. Test Phase

At first in the testing phase, all high resolution gallery images are fed to the top branch net and mapped into the common space and the probe image is fed into the bottom branch net. The label of probe image is determined by following formulae

$$Label(I_i^l) = Label(I_k^h) \tag{6}$$

where k determined by

$$k = \arg\min_j \{d_{i,j}\}_{j=1}^{N_G} \tag{7}$$

where I_i^l is the low resolution probe image, I_k^h is the k^th high resolution gallery image and N_G denotes number of high resolution face gallery images.

IV. EXPERIMENTAL EVALUATION

The two primary tasks of face recognition are face identification and verification. In face identification, a query face is compared to the gallery face database to determine its identity. In face verification, the claimed identity of a query face is verified. In this section, all of the experiment results belong to face identification task. The experiments are designed to answer the following questions:

- How well does the proposed approach perform compared to the state-of-the-art super-resolution methods and coupled mappings approaches?
- How robust does the proposed approach perform across different resolutions of probe images?
- How robust is the proposed approach to variations in expression, illumination, and age?
- How does the proposed approach perform when the super-resolution subnet is excluded from the architecture?
- How well does the proposed method reconstruct a high resolution face image from the low resolution probe one?
- How is the convergence of the proposed approach in training phase effected when SR subnet is excluded from its architecture?

In order to demonstrate the effectiveness of our proposed method, we compare the face recognition performance of our method with state-of-the-art competing methods including one super-resolution (discriminative super-resolution (DSR) method [10]) and three coupled mappings approaches (coupled locality preserving mappings (CLPM) [23], nonlinear mappings on coherent features (NMCF) [24], and multidimensional scaling (MDS) [16] methods).

A. Data Description

In this section, we describe the datasets we used to train and evaluate our proposed method.

Training dataset: The details of datasets we used for training are presented in Table II In total we used 45315 face images with variations in pose, expression, illumination and age for training. From FERET dataset [40], we used 10585 images in training and the rest (3541 images) in the evaluation phase.

Evaluation dataset: We carried out our evaluations on part of FERET [18] face database. The FERET face data set contains 14126 face images from 1199 individuals. A subset of this data set including 3541 images is assigned for evaluation. This dataset includes four probe categories, each one assigned with a gallery set. All gallery face images are frontal. The four probe categories characteristics are explained below:

- The first probe category is called FB and it includes 1195 frontal face images. Its gallery set contains 1196 frontal face images with different expressions.
- The second probe category which is called $duplicateI$ contains all duplicate frontal images in the FERET database (722 images). The gallery is the same gallery as FB containing 1196 images.
- The third category is called fc which includes 194 images taken on the same day, but with a different camera and illumination condition. The gallery is the same gallery as FB containing 1196 images.
- The fourth category called $duplicateII$ consists of duplicate probe images which are taken at least with one year difference with acquisition of corresponding gallery image (different age condition). The gallery for $duplicateII$ probes is a subset of the gallery for other categories containing 864 images.

Table II
LIST OF DATASETS USED FOR TRAINING AND THEIR DESCRIPTION IN TERMS OF NUMBER OF IMAGES AND THEIR VARIABILITY IN CONDITIONS SUCH AS E:EXPRESSION, I:ILLUMINATION, AND P:POSE. * PLEASE NOTE THAT FERET DATASET CONTAINS 14126 IMAGES AND WE USED 10585 IMAGE FOR TRAINING, AND THE REST, 3541 IMAGES, FOR EVALUATION.

Databases	Number of images	Highlights
300-W [41]	600	in the wild, large variations in E&I&P
HELEN [42]	2330	in the wild, large variations in E&I&P ,and has occlusion
IBUG [41]	135	in the wild, large variations in E&I&P
AFW [43]	250	in the wild, large variations in E&I&P
Georgia Tech Face Database [44]	750	large variations in E&I&P
LFW [45]	13233	in the wild, large variations in E&I&P ,and scale
UMIST [46]	564	gray scale, and variations in pose and race
YALE B [47]	5760	gray scale, and variation in P&I
AT&T [48]	400	variation of time, eye glasses, and E&I
FERET [40]	14126*	changes in appearance through time, and P&I&E
CK+ [49]	10708	variation P&E

B. Training Phase

We used the pre-trained VGGnet weights [19] and dropped the last two fully connected layers to construct our FECNN. Also, before training of our two branches architecture, we trained SRnet on the training datasets described in Table II. For SRnet training, we first down-sampled faces from all of the training images to make the LR faces for the corresponding HR images in the dataset. The SRnet includes five convolutional layers and we trained the network with 45315 pairs of LR and HR face images. After training of FECNN and SRnet separately, we connected the pre-trained SR and FECNN subnets. Then we trained our proposed architecture using 45315 faces. In this main part of the training phase, we reduced learning rate of each layer in bottom branch to fine-tune the bottom net on training for coupled mappings purpose.

C. Robustness Against Expression, Illumination and Age Variations

In this experiment, we evaluated our proposed method on the four categories of FERET evaluation datasets described in Section IV-A. Since the FB images have different expression conditions, the fc set includes probe images with different illumination conditions and $duplicateII$ set contains probe images with different age conditions compared to the corresponding gallery images, we can evaluate the robustness of our proposed method against these variations as well. In this experiment, the HR face images with the size of 72×72 pixels are aligned with the positions of the two eyes. The LR images with size of 12×12 pixels are generated by the operation of down-sampling and smoothing on aligned HR face images.

Fig. 4 shows the cumulative match curve (CMC) for our method and four competing methods, DSR [10], MDS [16], NMCF [24], and CLPM [23]. The cumulative match score for rank k is a face identification measure which is defined as the recognition accuracy of the probe images when at least one of the k nearest neighbors of the HR gallery images belongs to the same individual as the LR probe image. The results presented in Fig. 4 shows that the recognition performance of our method is significantly better than other state-of-the-art methods. Fig. 4.a depicts the cumulative match curves on the FB dataset. As we explained in SectionIV-A, this dataset includes probe images different from gallery images only in terms of expression. The recognition accuracy of our proposed method in the rank 1 is 91.8%, while the best performance of the competing methods belongs to CLPM [23] with 90.1% recognition accuracy. Our proposed method outperforms the competing methods with 1.7% difference. Fig. 4.b depicts the CMC results on fc dataset. The probe images in this dataset vary in illumination compared to gallery images. Our proposed method outperform competing methods across all ranks sig-

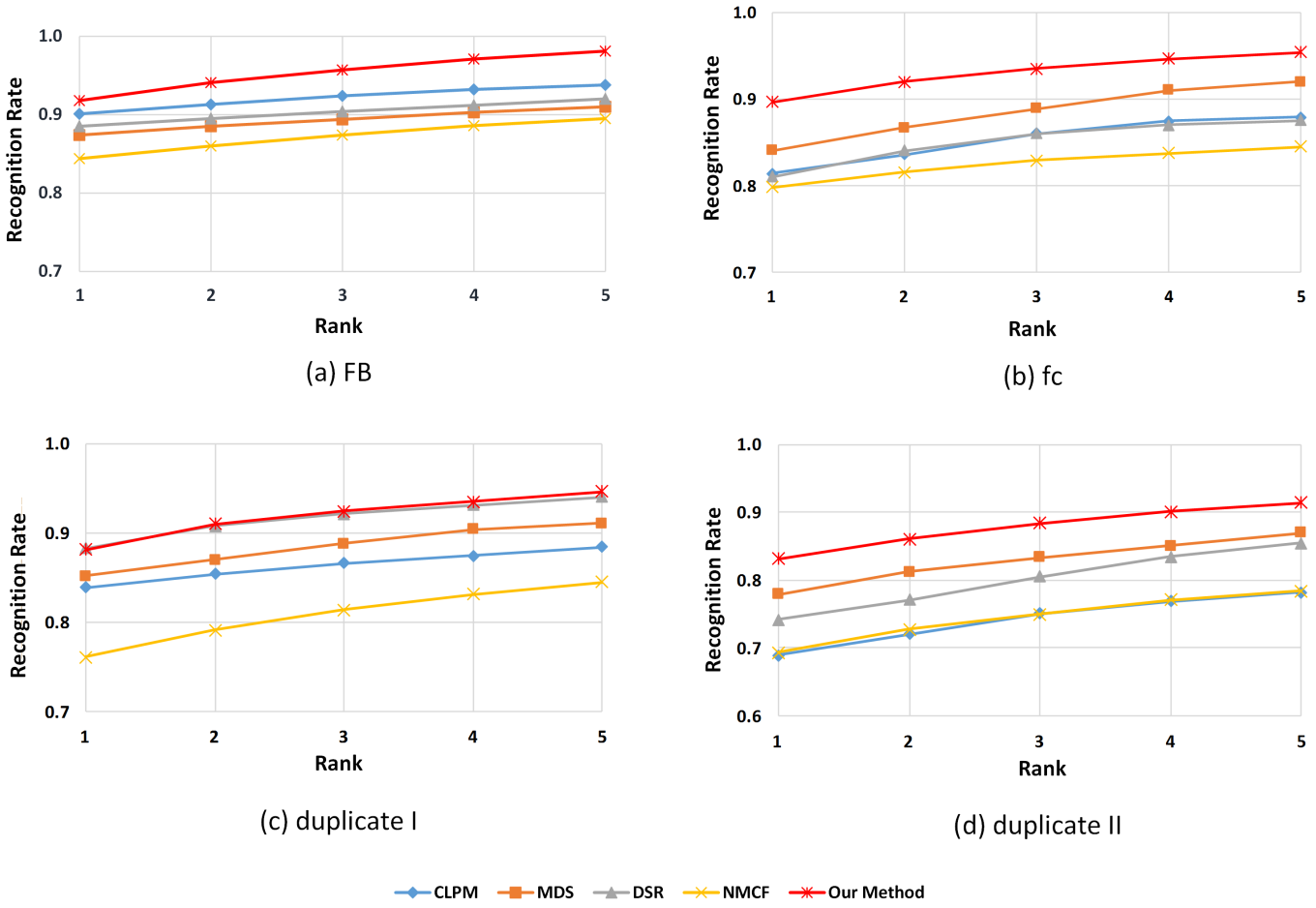

Fig. 4. Comparison of our proposed method with CLPM [23], MDS [16], DSR [10] and NMCF [24] in terms of recognition rates. Cumulative match curves on (a) FB, (b) fc, (c) duplicate I, and (d) duplicate II datasets.

nificantly. In rank 1, our method demonstrates an increase of 5.6% compared to the best competing method on fc dataset. This basically shows the efficiency of deep convolutional neural networks in feature extraction and generalization even in different illumination conditions. While the performance of our method is robust against the changes in illumination, the other competing methods performance drops significantly on fc dataset compared to FB. $DuplicateI$ includes images in similar condition as the gallery, but with slightly expression variation. On this dataset, the performance of our method is close to the best competing method (DSR [10]) (Fig. 4.c). The $duplicateII$ contains probe images with different age condition compared to gallery images. Our proposed method outperforms the best competing method (here MDS [16]) on rank 1 with 5.2% recognition accuracy (Fig. 4.d). Again, this shows the robustness of our proposed method against variations in age.

Taken together, our proposed method shows the best performance on FB, fc, and $duplicateII$ probe images and close to the best on $duplicateI$ dataset. Also our method shows robustness against variations in expression, illumination and age as shown in Fig. 4.b and d.

Table III
COMPARISON OF RANK 1 RECOGNITION ACCURACY ACROSS DIFFERENT PROBE IMAGE RESOLUTIONS.

	6×6	12×12	24×24	36×36
CLPMs [23]	64.4%	90.1%	93.4%	95.2%
MDS [16]	57.3%	87.4%	90.2%	92.2%
NMCF [24]	60.3%	84.4%	88.4%	91.1%
DSR [10]	69.4%	88.5%	90%	93%
Our Method	80.8%	91.8%	96.7%	98.8%

D. Evaluation on Different Probe Resolutions

Here, we evaluated the effectiveness of our proposed method on probe images with very low resolutions. In this experiment, we compared the performance of our method with state-of-the-art methods on FB probe dataset which all probe faces of this dataset are similar to gallery faces, but with slighly variation in expression. Thus appropriate to study the effect of variations in resolution. We considered four different resolutions, 6×6, 12×12, 24×24, and 36×36. Each time, we trained the SRnet

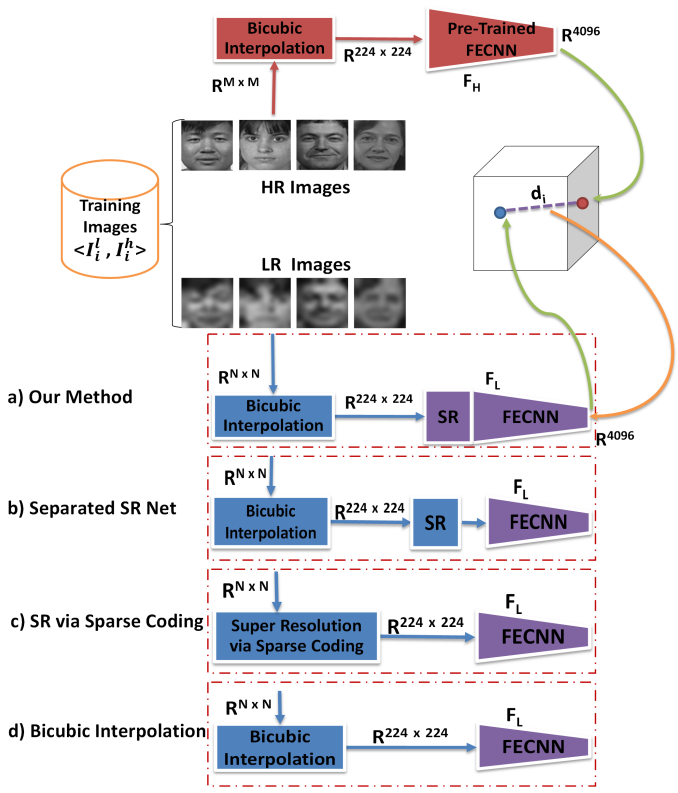

Fig. 5. Configurations with different super-resolution modules. Modules with violet color are involved in training phase. a) Configuration of our method. b) SR subnet is separated from SRFECNN. c) Using sparse coding for SR [13] d) Using only bicubic interpolation.

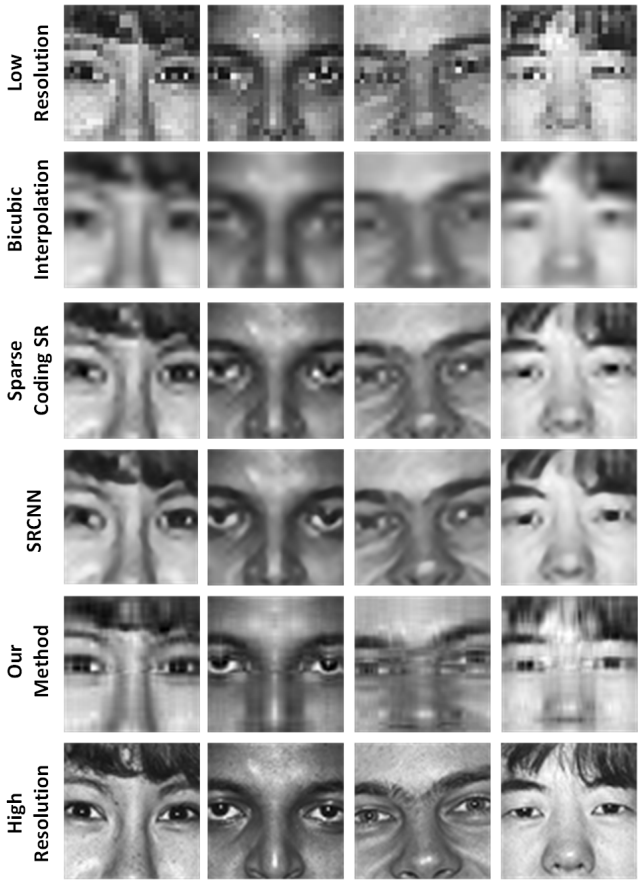

Fig. 6. Reconstructed Faces by different configurations in Fig. 5.

separately on training data with reduced images resolutions and then connected the SRnet to FECNN and retrained the bottom branch of our proposed method on each resolution condition separately. Table III shows the rank 1 recognition accuracy of our method compared to the competing methods on different resolution conditions evaluated on FB set. As can be seen, our proposed method outperforms all the competing methods on all four resolution conditions. The most significant improvement (11.4%) is on the very low resolution of 6×6 where our proposed method beats DSR [10], a method specifically proposed for the recognition of very low face images.

E. The Role of SR Subnet

As explained, the bottom branch net is consist of two nets, SR and FECNN. In training phase, both SR and FECNN nets are involved in the main training phase. In this experiment, we aim to study the impact of using SRnet and also its fine-tuning on the recognition performance of our method. Fig. 5 shows three different configurations that we compared our proposed method with them. Our proposed method configuration is depicted in Fig. 5.a where both SR and FECNN subnet are trained during the main training phase. In the configuration shown in Fig. 5.b, SRnet is separated from FECNN in bottom branch, and in the main training phase weights of SRnet are kept fixed. The configuration shown in Fig. 5.c employs sparse coding [13] method instead of the SRnet. Again only the FECNN is trained during the main training phase. The configuration illustrated in Fig. 5.d uses only a bicubic interpolation to map the low resolution input image to an image of size 224×224 and thus no super-resolution net is used. Therefore, in the training phase, only FECNN weights are updated. Table IV shows, the rank 1 recognition accuracy of the four different configurations (see Fig. 5). These results illustrate that using the SRnet in the configuration improves the performance (see the second row of Table IV). Furthermore, involving the SRnet in the main training phase improves the recognition performance considerably (our proposed method in Table IV). Especially, when the resolution of probe set is very low, the recognition

Table IV
COMPARISON OF RANK 1 RECOGNITION ACCURACY FOR DIFFERENT SR MODULE CONFIGURATIONS ACROSS DIFFERENT PROBE IMAGE RESOLUTIONS.

	6×6	12×12	24×24	36×36
Only Bicubic	66.7%	82.1%	88.6%	93.9%
Separated SR Subnet	75.4%	89.7%	95.3%	97.9%
SR via Sparse Coding	73.9%	88.4%	94.1%	97%
Our Method	80.8%	91.8%	96.7%	98.8%

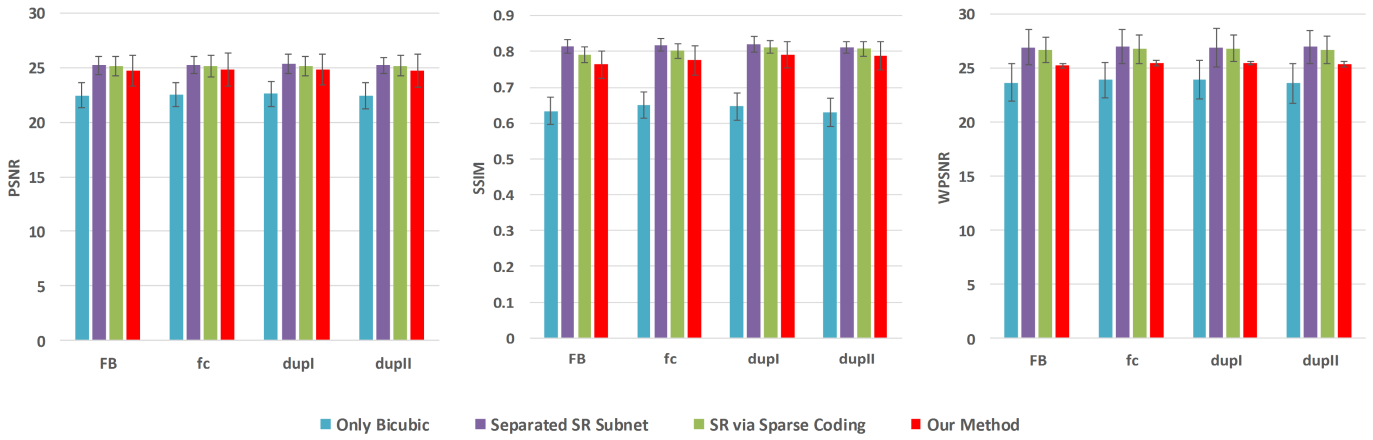

Fig. 7. Visual quality comparison of reconstructed HR faces in terms of PSNR, SSIM and WPSNR, while scale factor is 3.

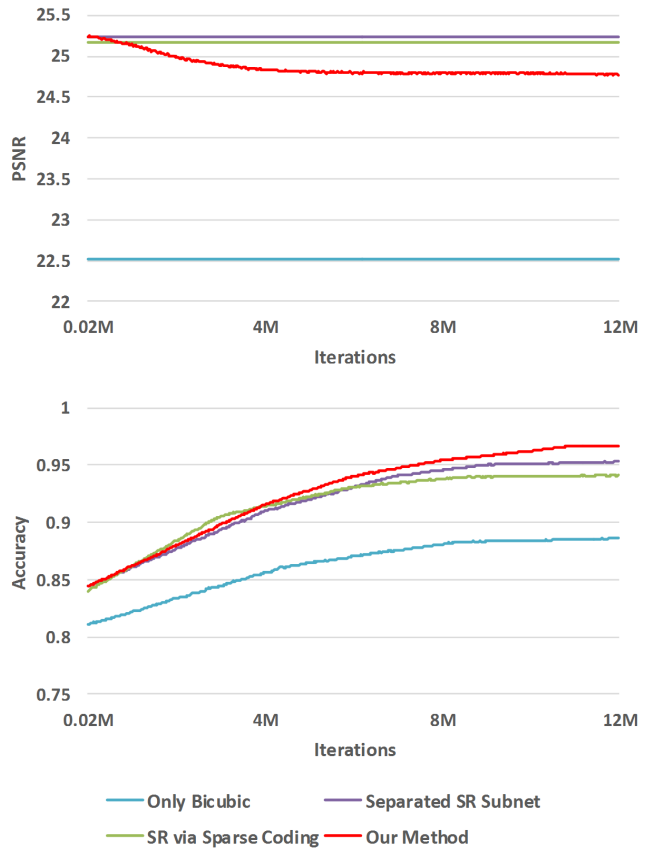

Fig. 8. Changes in visual quality of reconstructed HR faces and recognition accuracy during training.

performance of our method is considerably higher than other configurations. Together, we can conclude the employment and training of SRnet improves the recognition performance of our proposed method architecture especially for probe images with very low resolutions.

F. Evaluation on Reconstructed HR Face

Despite other coupled mappings methods, our proposed method can also reconstruct a high resolution face from the low resolution one. In this experiment, we aim to evaluate our method in terms of high resolution face reconstruction. Here, we again compare the performance of our method with the three configurations introduced in Fig. 5 in terms of visual quality of reconstructed face images. The size of low resolution images used in this section is 24×24 pixels. Fig. 6 shows some examples of reconstructed face images by each method. To compare visual enhancement of the four methods, peak signal to noise ratio (PSNR), structural similarity index (SSIM) and weighted peak signal to noise ratio (WPSNR [50]) metrics are used. As shown in Fig. 7, when SRnet is separated from FECNN net, the reconstructed face images have the best visual quality and sparse coding is the second. Our method places in the third position in these results, however the differences between reconstructed face images by our method in comparison with the top two methods is small. As discussed in Section IV-E, the recognition accuracy of our proposed method is much better compared to other configurations. This shows that the visual quality of super-resolved face images is compromised for better recognition performance in our proposed method. One interesting point is that the variance of PSNR and SSIM is higher for our method compared to other three competing methods. This shows that in some cases like the first two examples (on the left) in Fig. 6, the visual quality has improved while in others like the other two examples, the quality has degraded. In other words, the changes in SRnet has been in a direction to help the recognition performance eventually which is not necessarily in the direction of visual enhancement. Fig. 8 compares the changes of visual quality of reconstructed face images and recognition accuracy, during training for the four methods. As can be seen at the end of the training phase, our method achieved the best recognition performance but the third place in the visual quality of reconstructed face images.

V. Conclusion

In this paper, we presented a novel coupled mappings approach for the recognition of low resolution face images using deep convolutional neural networks. The main idea of our method is to use two DCNNs to transform low resolution

probe and high resolution gallery face images into a common space where the distances between all faces belong to the same individual are closer than distances between faces belong to different persons. Our proposed method demonstrates significant improvement in recognition accuracy compared to the state-of-the-art coupled mapping methods (CLPM [23], NMCF [24], MDS [16]) and super resolution method (DSR [10]). Our proposed method shows significant improvement and robustness against variations in expression, illumination and age. Our method also outperforms competing methods across various resolutions of probe images and it shows even more considerable performance improvement (11.4%) when applied on very low resolution images of 6×6 pixels. Our proposed method also offers HR image reconstruction which its visual quality is comparable with state-of-the-art super-resolution methods. The required space for our trained model is much less than the traditional deep convolutional neural networks trained for face recognition like VGGnet and thus our proposed low resolution face recognition method is applicable on systems with regular memory.

REFERENCES

[1] H. Jia and A. M. Martinez, "Support vector machines in face recognition with occlusions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 136–141.

[2] A. M. Martínez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Transactions on Pattern analysis and machine intelligence*, vol. 24, no. 6, pp. 748–763, 2002.

[3] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM computing surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[4] A. Pnevmatikakis and L. Polymenakos, *Far-field, multi-camera, video-to-video face recognition*. INTECH Open Access Publisher, 2007.

[5] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1991, pp. 586–591.

[6] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[7] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[8] Z. Wang, Z. Miao, Q. J. Wu, Y. Wan, and Z. Tang, "Low-resolution face recognition: a review," *The Visual Computer*, vol. 30, no. 4, pp. 359–386, 2014.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[10] W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 327–340, 2012.

[11] C. Liu, H.-Y. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *International Journal of Computer Vision*, vol. 75, no. 1, pp. 115–134, 2007.

[12] W. Liu, D. Lin, and X. Tang, "Hallucinating faces: Tensorpatch super-resolution and coupled residue compensation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 478–484.

[13] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[14] P. H. Hennings-Yeomans, S. Baker, and B. V. Kumar, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in *Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[15] M. Jian and K.-M. Lam, "Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 11, pp. 1761–1772, 2015.

[16] S. Biswas, K. W. Bowyer, and P. J. Flynn, "Multidimensional scaling for matching low-resolution face images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 10, pp. 2019–2030, 2012.

[17] C. Zhou, Z. Zhang, D. Yi, Z. Lei, and S. Z. Li, "Low-resolution face recognition via simultaneous discriminant analysis," in *Biometrics (IJCB)*, 2011, pp. 1–6.

[18] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[19] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, vol. 1, no. 3, 2015, p. 6.

[20] A. Chakrabarti, A. Rajagopalan, and R. Chellappa, "Super-resolution of face images using kernel pca-based prior," *IEEE Transactions on Multimedia*, vol. 9, no. 4, pp. 888–892, 2007.

[21] S. Baker and T. Kanade, "Hallucinating faces," in *Automatic Face and Gesture Recognition*, 2000, pp. 83–88.

[22] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *International journal of computer vision*, vol. 40, no. 1, pp. 25–47, 2000.

[23] B. Li, H. Chang, S. Shan, and X. Chen, "Low-resolution face recognition via coupled locality preserving mappings," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 20–23, 2010.

[24] H. Huang and H. He, "Super-resolution method for face recognition using nonlinear mappings on coherent features," *IEEE Transactions on Neural Networks*, vol. 22, no. 1, pp. 121–130, 2011.

[25] C.-X. Ren, D.-Q. Dai, and H. Yan, "Coupled kernel embedding for low-resolution face image recognition," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3770–3783, 2012.

[26] J. Shi and C. Qi, "From local geometry to global structure: Learning latent subspace for low-resolution face image recognition," *IEEE Signal Processing Letters*, vol. 22, no. 5, pp. 554–558, 2015.

[27] P. Zhang, X. Ben, W. Jiang, R. Yan, and Y. Zhang, "Coupled marginal discriminant mappings for low-resolution face recognition," *Optik-International Journal for Light and Electron Optics*, vol. 126, no. 23, pp. 4352–4357, 2015.

[28] S. P. Mudunuri and S. Biswas, "Low resolution face recognition across variations in pose and illumination," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 5, pp. 1034–1040, 2016.

[29] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[31] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*, 2014, pp. 184–199.

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[33] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[35] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[36] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *ICLR*, vol. abs/1511.07289, 2016.

[37] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[38] M. Abadi, P. Barham, J. Chen, Z. Chen, and v. . a. y. . . u. . h. t. . W. b. . h. b. . d. others title = TensorFlow: A system for large-scale machine learning, journal = CoRR.

[39] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, 2014, pp. 818–833.

[40] National institute of standards and technology. the color feret database. [Online]. Available: http://www.itl.nist.gov/iad/humanid/colorferet/home.html

[41] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 397–403.

[42] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. Huang, "Interactive facial feature localization," *Computer Vision–ECCV*, pp. 679–692, 2012.

[43] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 2879–2886.

[44] Georgia tech face database. [Online]. Available: ftp://ftp.ee.gatech.edu/pub/users/hayes/facedb/

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

[46] D. B. Graham and N. M. Allinson, "Characterising virtual eigensignatures for general purpose face recognition," in *Face Recognition*, 1998, pp. 446–456.

[47] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[48] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Applications of Computer Vision*, 1994, pp. 138–142.

[49] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 94–101.

[50] S. Voloshynovskiy, A. Herrigel, N. Baumgaertner, and T. Pun, "A stochastic approach to content adaptive digital image watermarking," in *International Workshop on Information Hiding*, 1999, pp. 211–236.