Expert Systems With Applications 139 (2020) 112854

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

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

Erfan Zangeneh[a], Mohammad Rahmati[a], Yalda Mohsenzadeh[b,*]

[a] Department of Computer Engineering and Information Technology at Amirkabir University of Technology, Tehran, Iran
[b] Department of Computer Science, Brain and Mind Institute, Western University, London, ON, Canada

ARTICLE INFO

*Article history:*
Received 4 February 2019
Revised 26 July 2019
Accepted 28 July 2019
Available online 30 July 2019

*Keywords:*
Low resolution face recognition
Super-resolution methods
Coupled mappings methods
Deep convolutional neural networks

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, LFW, and MBGC datasets 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 (5% 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.



## 1. Introduction

In the past few decades, face recognition has shown promising performance in numerous applications and under challenging conditions such as occlusion (Jia & Martinez, 2009), variation in pose, illumination, and expression (Martínez, 2002). While many face recognition systems have been developed for recognizing high quality face images in controlled conditions (Zhao, Chellappa, Phillips, & Rosenfeld, 2003), there are a few studies focused on face recognition in real world applications such as surveillance systems with low resolution faces (Pnevmatikakis & Polymenakos, 2007). 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. 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 (Simonyan & Zisserman, 2014). Some of the well known methods of this category are Liu, Shum, and Freeman (2007); Liu, Lin, and Tang (2005); Zou and Yuen (2012) and Yang, Wright, Huang, and Ma (2010). (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; (Biswas, Bowyer, & Flynn, 2012; Hennings-Yeomans, Baker, & Kumar, 2008; Jian & Lam, 2015; Zhou, Zhang, Yi, Lei, & Li, 2011) 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 the proposed method, the distance of transformed low and high resolution images in the common space is used as an objective function to train the deep convolutional neural networks. The proposed method also reconstructs good HR face images which are

* Corresponding author.
*E-mail addresses:* zangeneh.erfan@aut.ac.ir (E. Zangeneh), rahmati@aut.ac.ir (M. Rahmati), ymohsenz@uwo.ca (Y. Mohsenzadeh).

https://doi.org/10.1016/j.eswa.2019.112854
0957-4174/

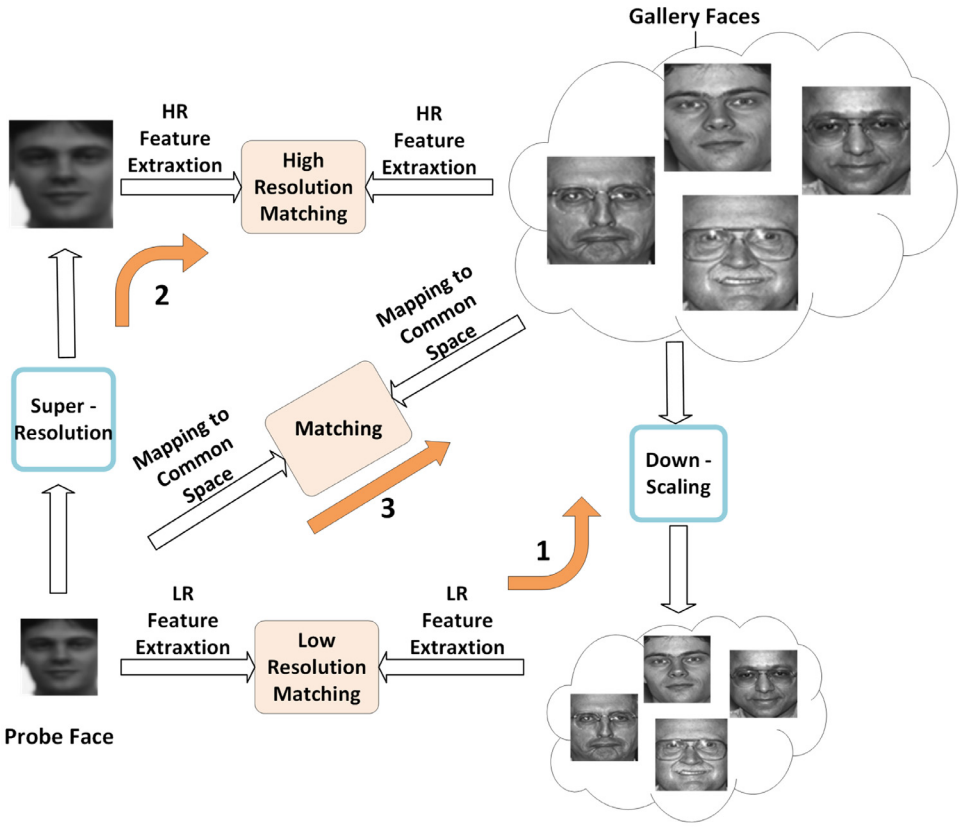


**Fig. 1.** Three general approaches for low resolution face recognition.

optimum for the recognition task. We evaluated the effectiveness of the proposed approach on the FERET (Phillips, Moon, Rizvi, & Rauss, 2000), LFW (Huang, Ramesh, Berg, & Learned-Miller, 2007), and MBGC (Phillips et al., 2009) datasets. Our results show the proposed approach improves the matching performance significantly compared to other state-of-the-art methods in the low resolution face recognition and other state of the art face recognition methods. 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.
- The proposed method offers high recognition accuracy compared to other state-of-the-art competing methods especially when the probe image is extremely low resolution. Please see Table 7 which summarizes the comparisons.
- The 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.

## 2. 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. 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. To achieve good reconstruction results, Liu et al. (2007) presented a two-step statistical modeling approach for hallucinating a HR face image from a LR input image. In Yang et al. (2010), 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. In Li, Prieto, Mery, and Flynn (2019), the authors implemented a framework to improve super resolution methods using generated LR images from a Generative Adversarial Network (GAN). While the super resolved images generated by these methods were improved in terms of visual quality, they were not optimized for face recognition performance. Yu, Fernando, Hartley, and Porikli (2018) proposed a super-resolution method for mapping of a LR image to a HR one using an exemplar dataset. They generated multiple HR face images from each LR image, then their method combined these candidate HR face images to generate one HR image. This method improved the performance of recognition slightly, when the LR face image was not very low resolution.

In Heinsohn, Villalobos, Prieto, and Mery (2019), the authors introduced a new dataset (called AR-LQ) for low resolution face recognition, and proposed a new method based on sparse representations to reconstruct a super resolution face image from LR one using a dictionary that is trained on different levels of image blurriness. He, Cao, Song, Sun, and Tan (2019) proposed a method that combines the output of a texture inpainting component and a pose correction component. Their inpainting component inpaints a super-resolution face image from near infrared visible image textures. Their pose correction component maps any pose in NIR image to frontal pose.

The other category of works on LR FR is known as coupled mappings methods. 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 a

transformation 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 (Li, Chang, Shan, & Chen, 2010):

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

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 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:

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

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

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

where $\alpha$ 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 the common space with these mappings, respectively.

Huang and He (2011) 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. (2012) used multidimensional scaling transformation learning to find both low resolution and 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. Yang, Yang, Gao, and Liao (2018) proposed a method based on MDS method (Biswas et al., 2012), but with changing of optimization problem formulation and embedding intraclass and between class metrics they achieve a better performance than MDS method. Huang and He (2011) 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. Mudunuri and Biswas (2016) 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. The authors of Lu, Jiang, and Kot (2018) suggested a framework for face recognition that implements a Residual Network for matching two low and high resolution face images into common space, but they did not use any super resolution method for LR image super resolving purpose. Abdollahi Aghdam, Bozorgtabar, Kemal Ekenel, and Thiran (2019) proposed a deep convolutional neural network (DCNN) for low resolution face recognition. They used eight DCNNs with different training datasets and employed the features in the last layer of these networks for matching between LR probe face and gallery HR faces. In this method, they show using training face images with various resolutions will improve the performance of any DCNN-based face recognition method.

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. In the next section, we propose a coupled mappings method using deep convolutional neural networks for nonlinear mapping to a common space. The proposed 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.

## 3. 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 with updating the weights of DCNN by backpropagation of the error. Fig. 2 shows the overview of the 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.

### 3.1. Networks architecture

The proposed 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 (Simonyan & Zisserman, 2014). 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 $224 \times 224$ dimensions (whenever input image size is different from $224 \times 224$, 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, Loy, He, and Tang (2014), 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

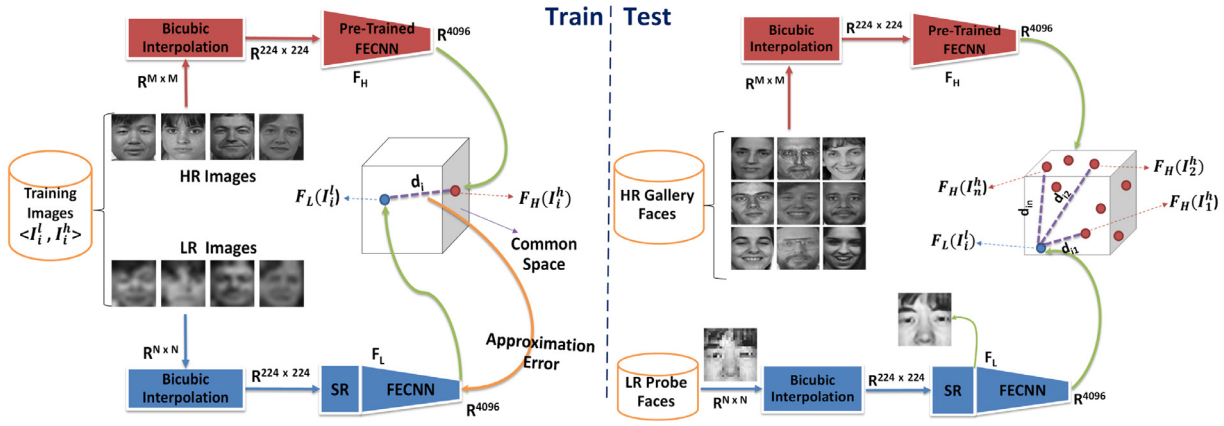


**Fig. 2.** Overview of the proposed method. $M$ and $N$ denote dimensions of HR and LR images, respectively, and $M > N$.

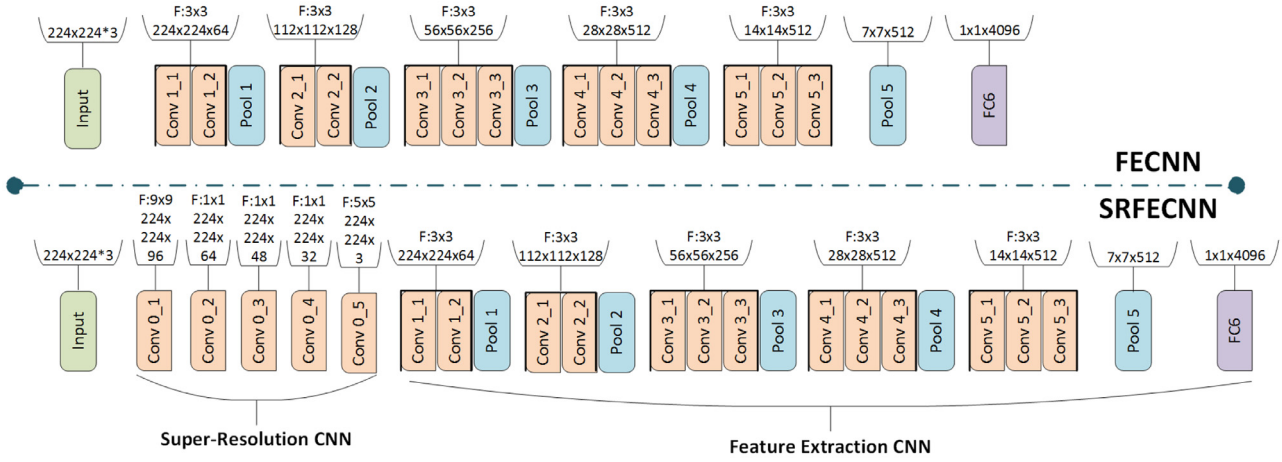


**Fig. 3.** Architecture of two deep convolutional neural networks in two branches of our proposed method.

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 $224 \times 224$. Also, the output of SR subnet is an image with the size of $224 \times 224$. 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 1 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.

**Table 1**
Number of used weights in layers of SRFECNN.

| Layer set | Parameters | Number of weights |
|---|---|---|
| Conv0_1 | $F = 9 \times 9$<br>Depth $= 696$ | $3 \times 9 \times 9 \times 96 = 23,328$ |
| Conv0_2 | $F = 1 \times 1$<br>Depth $= 64$ | $96 \times 1 \times 1 \times 64 = 6144$ |
| Conv0_3 | $F = 1 \times 1$<br>Depth $= 48$ | $64 \times 1 \times 1 \times 48 = 2928$ |
| Conv0_4 | $F = 1 \times 1$<br>Depth $= 32$ | $48 \times 1 \times 1 \times 32 = 1536$ |
| Conv0_5 | $F = 5 \times 5$<br>Depth $= 3$ | $32 \times 5 \times 5 \times 3 = 2400$ |
| Conv1 (2 Convs) | $F = 3 \times 3$<br>Depth $= 64$ | $2(3 \times 3 \times 3 \times 64) = 3456$ |
| Conv2 (2 Convs) | $F = 3 \times 3$<br>Depth $= 128$ | $2(64 \times 3 \times 3 \times 128) = 147,456$ |
| Conv3 (3 Convs) | $F = 3 \times 3$<br>Depth $= 256$ | $3(128 \times 3 \times 3 \times 256) = 884,736$ |
| Conv4 (3 Convs) | $F = 3 \times 3$<br>Depth $= 512$ | $3(256 \times 3 \times 3 \times 512) = 3,538,944$ |
| Conv5 (3 Convs) | $F = 3 \times 3$<br>Depth $= 512$ | $3(512 \times 3 \times 3 \times 512) = 7,077,888$ |
| FC6 | Depth $= 4096$ | $7 \times 7 \times 512 \times 4096 = 102,760,448$ |
| All layers | | $114,449,264 \approx 114M$ |

### 3.2. Common subspace learning

We trained our network in three stages as summarized below:

- First, we used trained VGGnet on face dataset (Parkhi, Vedaldi, & Zisserman, 2015) 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.

- In the second step, we trained the SRnet of the bottom branch with a dataset 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 dataset 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 net as two nonlinear functions that project a high resolution image and low resolution image to a 4, 096 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 (Zeiler & Fergus, 2014).

### 3.3. 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.

### 3.4. 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.

### 3.5. Implementation

All the experimental evaluation run on Intel Core i7-5930K, NVIDIA Titan X, and Pytorch with cudnn 5.1.10 is selected as the deep framework for implementing the proposed method. In All Experiment Stochastic Gradient Descent(SGD) is the optimizer, also used distance function is L2norm and learning rate for each layer is changed based on Table 2.

**Table 2**
Learning rate changes in all layers in all 3 steps.

| Network | SRNet | FECNN | | | |
|---|---|---|---|---|---|
| Layer Set | Conv0 | Conv1,2 | Conv3,4 | Conv5 | FC |
| Initial value | $10^{-5}$ | $10^{-6}$ | $5\times10^{-6}$ | $10^{-5}$ | $5\times10^{-4}$ |
| Final value | $10^{-6}$ | $5\times10^{-8}$ | $10^{-7}$ | $5\times10^{-7}$ | $10^{-s6}$ |

## 4. Experimental evaluation

The experiments are designed to compare performance of the proposed method against the state-of-the-art super-resolution and coupled mappings approaches when the resolution of probe face image is very low, and further how robust the proposed approach performs against variations in expression, illumination, and age. Next, in two experiments, we show the role of super-resolution subnet in recognition accuracy and reconstructing the high resolution face image. Finally, the proposed method is compared with the state-of-the-art high resolution face recognition using in the wild datasets. For a fair comparison, in all experiments presented in this section, the proposed method and competing methods are trained and evaluated following the same procedure. Please note that in our experiments we used the trained weights of the competing methods for initialization, and then we trained their models on the same training dataset as our proposed method and then evaluated them with exactly the same methodology as our proposed method.

### 4.1. Data description

**Training dataset:** The details of datasets we used for training are presented in Table 3. In total we used 90,897 face images with variations in pose, expression, illumination and age for training. From FERET dataset (Freeman, Pasztor, & Carmichael, 2000), we used 10,585 images in training and the rest (3541 images) in the evaluation phase.

**Evaluation datasets:** We carried out our evaluations on LFW (Huang et al., 2007), MBGC (Phillips et al., 2009), and FERET (Phillips et al., 2000) face dataset. The LFW (Huang et al., 2007) dataset contains 13,233 face images of 5749 subjects. 1680 subjects of this dataset have two or more face images. The MBGC (Huang et al., 2007) dataset includes images and videos. One image of each 147 subjects used as gallery, and the captured face images from videos used as probe set. The FERET face dataset contains 14,126 face images from 1199 individuals. A subset of this dataset including 3541 images is assigned for evaluation. This dataset includes four probe categories. The *FB* probe set includes 1195 frontal face images with different expressions. The second probe category which is called *duplicateI* contains all duplicate frontal images in the FERET dataset (722 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 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).

### 4.2. Training phase

We used the pre-trained VGGnet weights (Parkhi et al., 2015) 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 3. 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

**Table 3**
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, P:pose, and R:race. * Please note that FERET dataset contains 14,126 images and we used 10,585 image for training, and the rest, 3541 images, for evaluation.

| Datasets | # of images | Highlights |
|---|---|---|
| 300-W (Sagonas, Tzimiropoulos, Zafeiriou, & Pantic, 2013) | 600 | in the wild, variations in E&I&P |
| HELEN (Le, Brandt, Lin, Bourdev, & Huang, 2012) | 2330 | in the wild, variations in E&I&P |
| IBUG (Sagonas et al., 2013) | 135 | in the wild, variations in E&I&P |
| AFW (Zhu & Ramanan, 2012) | 250 | in the wild, variations in E&I&P |
| Georgia Tech DB (Freeman et al., 2000) | 750 | variations in E&I&P |
| PubFig (Kumar, Berg, Belhumeur, & Nayar, 2009) | 58, 797 | in the wild, variations in E&I&P |
| UMIST (Graham & Allinson, 1998) | 564 | gray scale, variations in P&R |
| YALE B (Georghiades, Belhumeur, & Kriegman, 2001) | 5760 | gray scale, variation in P&I |
| AT&T (Samaria & Harter, 1994) | 400 | wearing eyeglasses, variation in E&I |
| FERET (Freeman et al., 2000) | 14, 126* | variation in P&I&E |
| CK+ (Lucey et al., 2010) | 10, 708 | variation P&E |

trained the network with 90,897 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 90,897 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.

### 4.3. 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 4.1. 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 $72 \times 72$ pixels are aligned with the positions of the two eyes. The LR images with size of $12 \times 12$ 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 (Zou & Yuen, 2012), MDS (Biswas et al., 2012), NMCF (Huang & He, 2011), and CLPM (Li et al., 2010). 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 Section 4.1, this dataset includes probe images different from gallery images only in terms of expression. The recognition accuracy of our

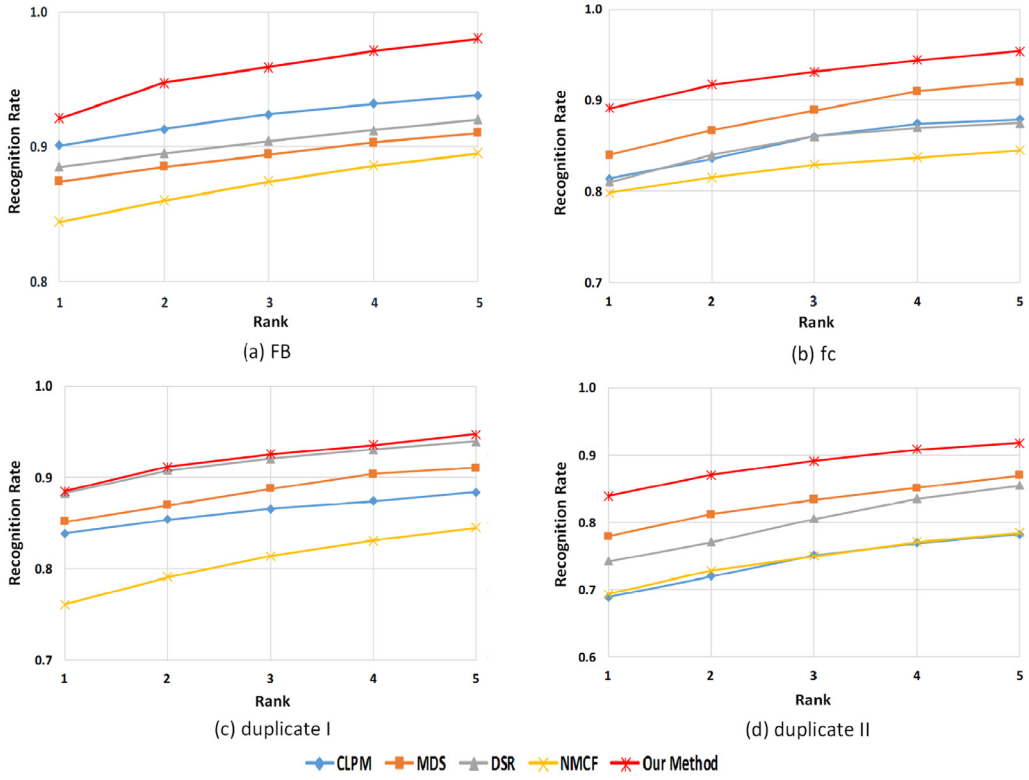


**Fig. 4.** Comparison of our proposed method with CLPM (Li et al., 2010), MDS (Biswas et al., 2012), DSR (Zou & Yuen, 2012) and NMCF(Huang & He, 2011) in terms of recognition rates. Cumulative match curves on (a) FB, (b) fc, (c) duplicate I, and (d) duplicate II datasets.

proposed method in the Rank-1 is 92.1%, while the best performance of the competing methods belongs to CLPM (Li et al., 2010) with 90.1% recognition accuracy. Our proposed method outperforms the competing methods with 2% 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 significantly. In Rank-1, our method demonstrates an increase of 5.1% 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 slightly better than competing method (DSR Zou & Yuen, 2012) (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 Biswas et al., 2012) on Rank-1 with 5.9% 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 all probe sets *FB, fc, duplicateI*, and *duplicateII*. Also our method shows robustness against variations in expression, illumination and age as shown in Fig. 4(b) and (d).

**Table 4**
Comparison of Rank-1 recognition accuracy across different probe image resolutions on FERET dataset.

| | $6\times6$ | $12\times12$ | $24\times24$ | $36\times36$ |
|---|---|---|---|---|
| CLPMs (Li et al., 2010) | 64.4% | 90.1% | 93.4% | 95.2% |
| MDS (Biswas et al., 2012) | 57.3% | 87.4% | 90.2% | 92.2% |
| NMCF (Huang & He, 2011) | 60.3% | 84.4% | 88.4% | 91.1% |
| DSR (Zou & Yuen, 2012) | 69.4% | 88.5% | 90% | 93% |
| DMS (Yang et al., 2018) | 76.4% | 91.5% | 97% | 99.4% |
| bASR (Heinsohn et al., 2019) | 74.1% | 88.3% | 93.6% | 96.9% |
| CSF (He et al., 2019) | 76.7% | 91.4% | 95.9% | 98.1% |
| Our method | 81.4% | 92.1% | 96.7% | 99.2% |

### 4.4. 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 set which all probe faces of this set are similar to gallery faces, but with slightly variation in expression. Thus appropriate to study the effect of variations in resolution. We considered four different resolutions, $6\times6$, $12\times12$, $24\times24$, and $36\times36$. Each time, we trained the SRnet 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 4 shows the

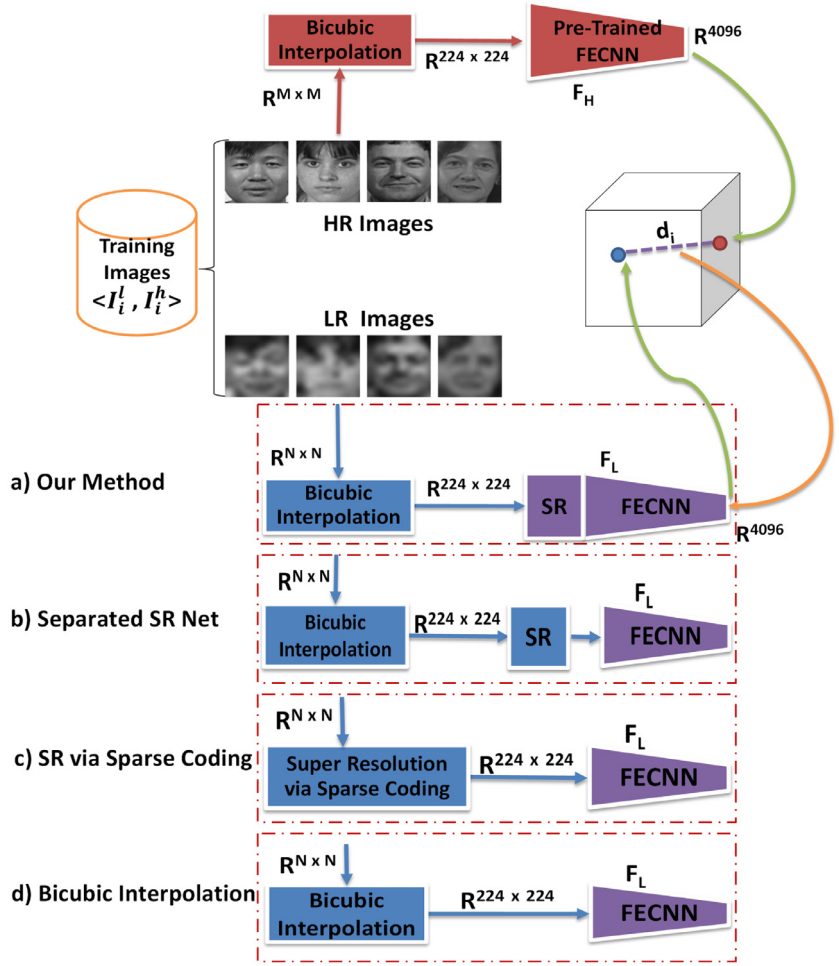


**Fig. 5.** Configurations with different super-resolution modules. Modules with violet color are involved in training phase. (a) Configuration of our method. (b) SR subnet is separated from SRFECNN. (c) Using sparse coding for SR (Yang et al., 2010) (d) Using only bicubic interpolation.

**Table 5**
Comparison of Rank-1 recognition accuracy for different SR module configurations across different probe image resolutions.

| | $6\times6$ | $12\times12$ | $24\times24$ | $36\times36$ |
|---|---|---|---|---|
| Only Bicubic | 66.8% | 81.9% | 88.9% | 93.6% |
| Separated SR Subnet | 75.8% | 89.3% | 95.4% | 97.6% |
| SR via Sparse Coding | 74.1% | 88.5% | 94.4% | 96.8% |
| Our Method | 81.4% | 92.1% | 96.7% | 99.2% |

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 (12%) is on the very low resolution of $6\times6$ where our proposed method beats DSR (Zou & Yuen, 2012), a method specifically proposed for the recognition of very low face images.

### 4.5. 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 (Yang et al., 2010) 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 $224\times224$ and thus no super-resolution net is used. Therefore, in the training phase, only FECNN weights are updated. Table 5 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 5). Furthermore, involving the SRnet in the main training phase improves the recognition performance considerably (our proposed method in Table 5). Especially, when the resolution of probe set is very low, the recognition 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.

### 4.6. 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 $24\times24$ 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 Voloshynovskiy, Herrigel, Baumgaertner, & Pun, 1999) metrics are used. As shown in Fig. 7, when SRnet is separated

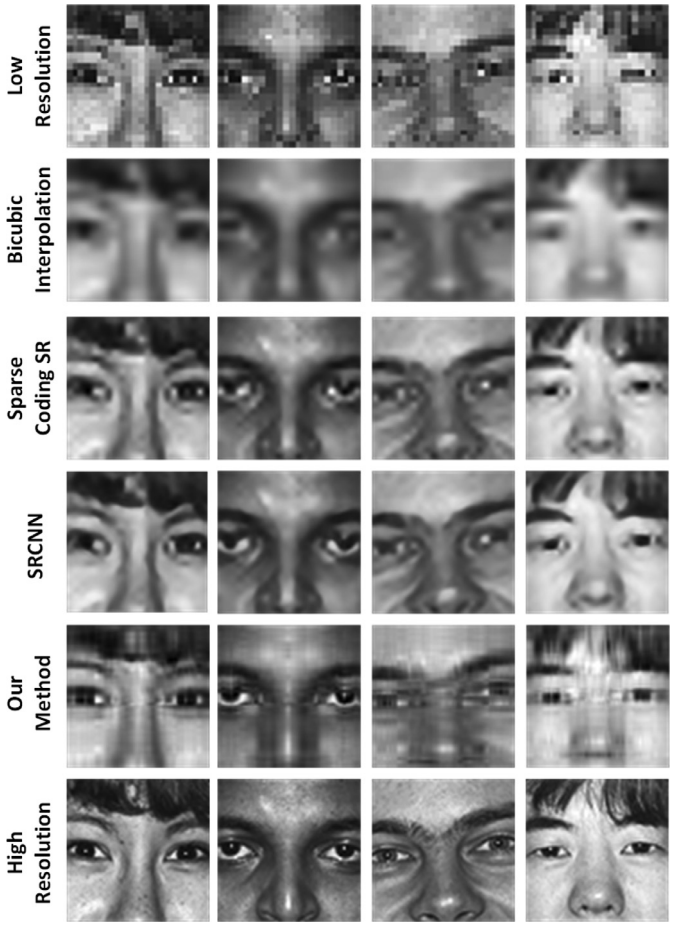


**Fig. 6.** Reconstructed Faces by different configurations in Fig. 5.

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 4.5, 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.

### 4.7. Compare on LFW

In this experiment, we compare the proposed method performance with state-of-the-art deep convolutional neural networks on face recognition. Also we evaluate performance of the proposed method when faces are in the wild. In this experiment we compare the proposed method with state-of-the-art methods in Low Resolution Face Recognition(LRFR) such as Reference LRFR (Mudunuri & Biswas, 2016), NMCF (Huang & He, 2011), DSR (Zou & Yuen, 2012) and in High Resolution Face Recognition(HRFR) such as DeepFace-Ensemble (Taigman, Yang, Ranzato, & Wolf, 2014), DeepID (Sun, Wang, & Tang, 2014), MMDFR

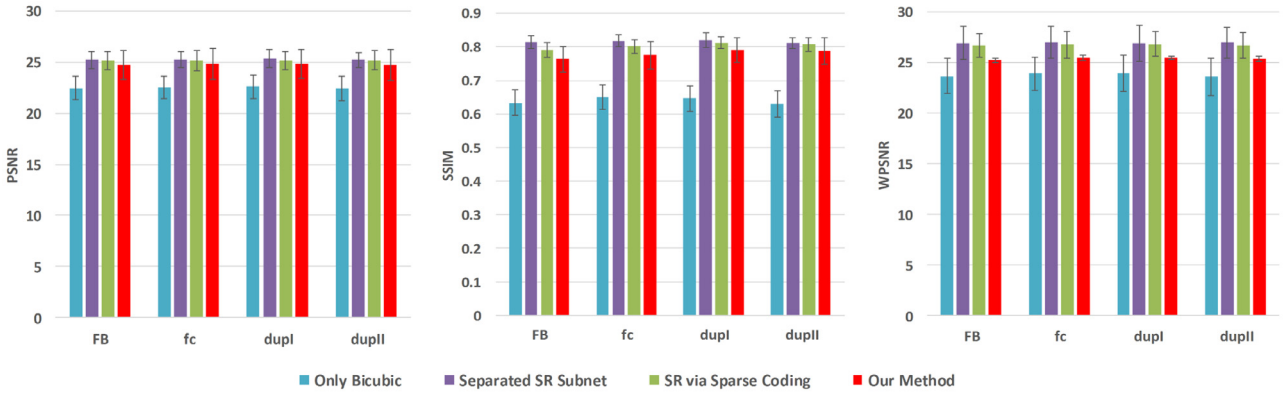


**Fig. 7.** Visual quality comparison of reconstructed HR faces in terms of PSNR, SSIM and WPSNR, while scale factor is 3.

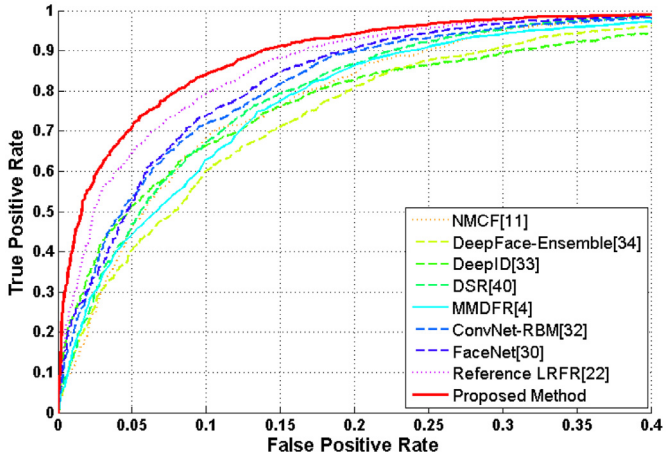


**Fig. 8.** ROC curve for LRFR on LFW dataset.

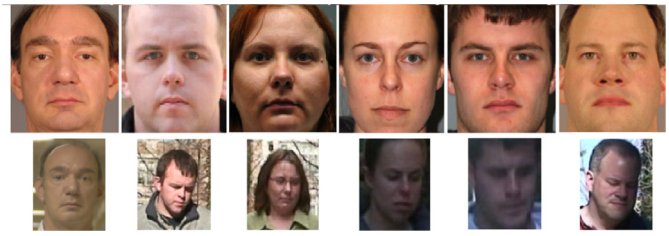

**Fig. 9.** Sample images of MBGC dataset. First row shows some gallery images. Second row shows some probe images taken from videos of corresponding individual.

(Ding & Tao, 2015), ConvNet-RBM (Sun, Wang, & Tang, 2013), FaceNet (Schroff, Kalenichenko, & Philbin, 2015).

We used unconstrained configuration of LFW protocol Huang and Learned-Miller (2014) for testing that allows the proposed method to be trained with another training dataset. For test procedure, we split all LFW images into 10 sets randomly and each evaluation set consists of 300 matched and 300 mismatched pairs. In this experiment, first image of each pair down-sampled to 8 while the size of the second image of this pair is 64. Finally average of recognition accuracies for 10 evaluation sets shown in Fig. 8. Results show the proposed method outperforms state-of-the-art methods on high resolution and low resolution face recognition tasks. Although the HRFR methods which compared on this experiment, resulted in good accuracy on LFW dataset for high resolution face recognition, their performance drops (Mudunuri & Biswas, 2016) significantly when input faces are low resolution.

### 4.8. Compare on MBGC

In this experiment we evaluate the proposed method on Multiple Biometric Grand Challenge (Phillips et al., 2009) to demonstrate its performance when images taken by surveillance cameras. This dataset consists images and videos from 147 subjects that all images are frontal and only one image per subject used as gallery images and 5 captured faces per subject from videos used as probe images. Thus probe set consists frontal and non-frontal face images, with poor resolution and illumination. Fig. 9 shows some sample images of gallery and probe set.

We did not use MBGC (Phillips et al., 2009) dataset in training phase. The experiment repeated 10 times and in each time 70 randomly subjects were used for testing. Table 6 shows average Rank-1 accuracy of the proposed method and other state-of-the-art methods for low resolution face recognition.

In this experiment, we compared the proposed method with other state-of-the-art low resolution face recognition methods. The

**Table 6**
Comparison of average Rank-1 accuracy on MBGC(Phillips et al., 2009) dataset.

| Method | Average Rank-1 accuracy |
|---|---|
| CLPM (Li et al., 2010) | 41.19% |
| MDS (Huang & He, 2011) | 39.48% |
| KISSME (Koestinger, Hirzer, Wohlhart, Roth, & Bischof, 2012) | 49.15% |
| Reference LRFR (Mudunuri & Biswas, 2016) | 50.57% |
| SA (Yu et al., 2018) | 62.41% |
| Heterogeneous LRFR (Mudunuri & Biswas, 2017) | 60.52% |
| DCR (Lu et al., 2018) | 67.13% |
| bASR (Heinsohn et al., 2019) | 64.71% |
| CSF (He et al., 2019) | 66.93% |
| Our method | 68.64% |

**Table 7**
Summary Table. Comparison of all state-of-the-art methods with the proposed method.

| | Super resolution based | | Coupled mapping based | | | | HR based trained on LR images | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | DSR[a] | SA[b] | CLPM[c] | MDS[d] | DMS[e] | DCR[f] | Deep-Face[g] | Face-Net[h] | MMDFR[i] | Our method |
| FERET $6 \times 6$ | 69.4% | 74.8% | 64.4% | 57.3% | 76.4% | 76.3% | 63.0% | 61.7% | 62.2% | **81.4%** |
| FERET $12 \times 12$ | 88.5% | 90.1% | 90.1% | 87.4% | 91.5% | 91.7% | 89.7% | 88.0% | 87.5% | **92.1%** |
| FERET $24 \times 24$ | 90.0% | 96.4% | 93.4% | 90.2% | 97.0% | 95.8% | 94.3% | **97.2%** | 95.6% | 96.7% |
| LFW $8 \times 8$ | 66.6% | 71.8% | 62.8% | 65.4% | 70.1% | 73.4% | 59.8% | 67.2% | 61.4% | **76.3%** |
| MBGC $12 \times 12$ | 52.9% | 62.4% | 41.2% | 36.5% | 63.6% | 67.1% | 62.7% | 61.9% | 60.5% | **68.64%** |

[a] Zou and Yuen (2012).
[b] Yu et al. (2018).
[c] Li et al. (2010).
[d] Biswas et al. (2012).
[e] Yang et al. (2018).
[f] Lu et al. (2018).
[g] Parkhi et al. (2015).
[h] Schroff et al. (2015).
[i] Ding and Tao (2015).

proposed method outperforms on Deep Coupled Residual Network (Lu et al., 2018) that reported the best recognition accuracy on MBGC (Phillips et al., 2009) with more than 1.5%. This experiment illustrates the efficient performance of the proposed method when probe images are in the wild or taken by surveillance cameras.

## 5. Discussion and 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. We evaluated our proposed method in 8 experiments on FERET (Phillips et al., 2000), LFW (Huang et al., 2007), and MBGC (Phillips et al., 2009). Comparisons of the proposed method with 9 state-of-the-art methods which are grouped in three approaches are summarized in Table 7. These three approaches include (1) super resolution based methods which first generate a HR face image from LR probe face, then use this super resolved face images for matching with HR Gallery images; (2) common space mapping methods which map LR probe face and HR gallery faces into a common space where they are comparable and can be tested for matching; (3) methods that are designed for HR face recognition task, but we trained them on LR face images to be comparable with our method.Our proposed method demonstrates significant improvement in recognition accuracy compared to the state-of-the-art coupled mapping methods (CLPM Li et al., 2010, NMCF Huang & He, 2011, MDS Biswas et al., 2012, DSM Yang et al., 2018, Reference LRFR Mudunuri & Biswas, 2016, Heterogeneous LRFR Mudunuri & Biswas, 2017), discriminative super resolution (DSR Zou & Yuen, 2012), Mapping using Supplementary Attributes (Yu et al., 2018), and deep residual network (Lu et al., 2018) method. As shown in Table 7, the proposed method outperforms the other state-of-the-art methods especially when probe face image has very low resolution. 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 (5%) when applied on very low resolution images of $6 \times 6$ pixels.The proposed method shows the best performance when faces are in the wild and also taken by surveillance cameras. Although the state-of-the-art high resolution face recognition methods achieved well performance on challenging datasets, this category of methods cannot compete with the state-of-the-art low resolution face recognition methods, like our proposed method on low resolution face images. Our proposed method also offers HR image reconstruction which its visual quality is comparable with state-of-the-art super-resolution methods.

## Declaration of Competing Interest

The authors declare that no competing interests exist.

## Credit authorship contribution statement

**Erfan Zangeneh:** Conceptualization, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Mohammad Rahmati:** Conceptualization, Investigation, Methodology, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Yalda Mohsenzadeh:** Conceptualization, Investigation, Methodology, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing.

## References


Abdollahi Aghdam, O., Bozorgtabar, B., Kemal Ekenel, H., & Thiran, J.-P. (2019). Exploring factors for improving low resolution face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 0–0.

Biswas, S., Bowyer, K. W., & Flynn, P. J. (2012). Multidimensional scaling for matching low-resolution face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(10), 2019–2030.

Ding, C., & Tao, D. (2015). Robust face recognition via multimodal deep face representation. *IEEE Transactions on Multimedia, 17*(11), 2049–2058.

Dong, C., Loy, C. C., He, K., & Tang, X. (2014). *Learning a deep convolutional network for image super-resolution* (pp. 184–199).

Freeman, W. T., Pasztor, E. C., & Carmichael, O. T. (2000). Learning low-level vision. *International Journal of Computer Vision, 40*(1), 25–47.

Georghiades, A. S., Belhumeur, P. N., & Kriegman, D. J. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(6), 643–660.

Graham, D. B., & Allinson, N. M. (1998). Characterising virtual eigensignatures for general purpose face recognition. In *Face recognition* (pp. 446–456).

He, R., Cao, J., Song, L., Sun, Z., & Tan, T. (2019). Cross-spectral face completion for NIR-VIS heterogeneous face recognition. arXiv:1902.03565.

Heinsohn, D., Villalobos, E., Prieto, L., & Mery, D. (2019). Face recognition in low-quality images using adaptive sparse representations. *Image and Vision Computing, 85*, 46–58.

Hennings-Yeomans, P. H., Baker, S., & Kumar, B. V. (2008). *Simultaneous super-resolution and feature extraction for recognition of low-resolution faces* (pp. 1–8).

Huang, G. B., & Learned-Miller, E. (2014). Labeled faces in the wild: Updates and new reporting procedures. *Tech. Rep*. Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA.

Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *Technical Report, Technical Report 07-49*. University of Massachusetts, Amherst.

Huang, H., & He, H. (2011). Super-resolution method for face recognition using nonlinear mappings on coherent features. *IEEE Transactions on Neural Networks, 22*(1), 121–130.

Jia, H., & Martinez, A. M. (2009). *Support vector machines in face recognition with occlusions* (pp. 136–141).
Jian, M., & Lam, K.-M. (2015). Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition. *IEEE Transactions on Circuits and Systems for Video Technology, 25*(11), 1761–1772.
Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2012). *Large scale metric learning from equivalence constraints* (pp. 2288–2295).
Kumar, N., Berg, A. C., Belhumeur, P. N., & Nayar, S. K. (2009). *Attribute and simile classifiers for face verification*.
Le, V., Brandt, J., Lin, Z., Bourdev, L., & Huang, T. (2012). Interactive facial feature localization. In *Computer vision–ECCV* (pp. 679–692).
Li, B., Chang, H., Shan, S., & Chen, X. (2010). Low-resolution face recognition via coupled locality preserving mappings. *IEEE Signal Processing Letters, 17*(1), 20–23.
Li, P., Prieto, L., Mery, D., & Flynn, P. J. (2019). On low-resolution face recognition in the wild: Comparisons and new techniques. *IEEE Transactions on Information Forensics and Security*.
Liu, C., Shum, H.-Y., & Freeman, W. T. (2007). Face hallucination: Theory and practice. *International Journal of Computer Vision, 75*(1), 115–134.
Liu, W., Lin, D., & Tang, X. (2005). Hallucinating faces: Tensorpatch super-resolution and coupled residue compensation, *2*, 478–484.
Lu, Z., Jiang, X., & Kot, A. (2018). Deep coupled resnet for low-resolution face recognition. *IEEE Signal Processing Letters, 25*(4), 526–530.
Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). *The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression* (pp. 94–101).
Martínez, A. M. (2002). Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(6), 748–763.
Mudunuri, S. P., & Biswas, S. (2016). Low resolution face recognition across variations in pose and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(5), 1034–1040.
Mudunuri, S. P., & Biswas, S. (2017). *Dictionary alignment for low-resolution and heterogeneous face recognition* (pp. 1115–1123).
Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. *British Machine Vision Conference, 1*(3), 6.
Phillips, P. J., Flynn, P. J., Beveridge, J. R., Scruggs, W. T., O'toole, A. J., Bolme, D., et al. (2009). *Overview of the multiple biometrics grand challenge* (pp. 705–714).
Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(10), 1090–1104.
Pnevmatikakis, A., & Polymenakos, L. (2007). *Far-field, multi-camera, video-to-video face recognition*. INTECH Open Access Publisher.
Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). *300 faces in-the-wild challenge: The first facial landmark localization challenge* (pp. 397–403).
Samaria, F. S., & Harter, A. C. (1994). *Parameterisation of a stochastic model for human face identification* (pp. 138–142).
Schroff, F., Kalenichenko, D., & Philbin, J. (2015). *Facenet: A unified embedding for face recognition and clustering* (pp. 815–823).
Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556.
Sun, Y., Wang, X., & Tang, X. (2013). *Hybrid deep learning for face verification* (pp. 1489–1496).
Sun, Y., Wang, X., & Tang, X. (2014). *Deep learning face representation from predicting 10,000 classes* (pp. 1891–1898).
Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). *Deepface: Closing the gap to human-level performance in face verification* (pp. 1701–1708).
Voloshynovskiy, S., Herrigel, A., Baumgaertner, N., & Pun, T. (1999). *A stochastic approach to content adaptive digital image watermarking* (pp. 211–236).
Yang, F., Yang, W., Gao, R., & Liao, Q. (2018). Discriminative multidimensional scaling for low-resolution face recognition. *IEEE Signal Processing Letters, 25*(3), 388–392.
Yang, J., Wright, J., Huang, T. S., & Ma, Y. (2010). Image super-resolution via sparse representation. *IEEE Transactions on Image Processing, 19*(11), 2861–2873.
Yu, X., Fernando, B., Hartley, R., & Porikli, F. (2018). Super-resolving very low-resolution face images with supplementary attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 908–917).
Zeiler, M. D., & Fergus, R. (2014). *Visualizing and understanding convolutional networks* (pp. 818–833).
Zhao, W., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Computing Surveys (CSUR), 35*(4), 399–458.
Zhou, C., Zhang, Z., Yi, D., Lei, Z., & Li, S. Z. (2011). *Low-resolution face recognition via simultaneous discriminant analysis* (pp. 1–6).
Zhu, X., & Ramanan, D. (2012). *Face detection, pose estimation, and landmark localization in the wild* (pp. 2879–2886).
Zou, W. W., & Yuen, P. C. (2012). Very low resolution face recognition problem. *IEEE Transactions on Image Processing, 21*(1), 327–340.