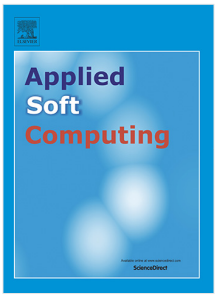
Journal Pre-proof



Multi-scale patch based representation feature learning for low-resolution face recognition

Guangwei Gao, Yi Yu, Meng Yang, Pu Huang, Qi Ge, Dong Yue

PII: S1568-4946(20)30123-X
DOI: https://doi.org/10.1016/j.asoc.2020.106183
Reference: ASOC 106183

To appear in: *Applied Soft Computing Journal*

Received date : 19 September 2019
Revised date : 8 January 2020
Accepted date : 10 February 2020

Please cite this article as: G. Gao, Y. Yu, M. Yang et al., Multi-scale patch based representation feature learning for low-resolution face recognition, *Applied Soft Computing Journal* (2020), doi: https://doi.org/10.1016/j.asoc.2020.106183.

Multi-scale Patch based Representation Feature Learning for Low-Resolution Face Recognition

Guangwei Gao[a,b,c,*], Yi Yu[b], Meng Yang[d], Pu Huang[a], Qi Ge[e], Dong Yue[a]

[a]*Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing, China*
[b]*Digital Content and Media Sciences Research Division, National Institute of Informatics, Tokyo, Japan*
[c]*Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou, China*
[d]*School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China*
[e]*College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China*

Abstract

In practical video surveillance, the quality of facial regions of interest is usually affected by the large distances between the objects and surveillance cameras, which undoubtedly degrade the recognition performance. Existing methods usually consider the holistic representations, while neglecting the complementary information from different patch scales. To tackle this problem, this paper proposes a multi-scale patch based representation feature learning (MSPRFL) scheme for low-resolution face recognition problem. Specifically, the proposed MSPRFL approach first exploits multi-level information to learn more accurate resolution-robust representation features of each patch with the help of a training dataset. Then, we exploit these learned resolution-robust representation features to reduce the resolution discrepancy by integrating the recognition results from all patches. Finally, by considering the complementary discriminative ability from different patch scales, we try to fuse the multi-scale outputs by learning scale weights via an ensemble optimization model. We further verify the efficiency of the proposed MSPRFL on low-resolution face recognition by the comparison experiments on several commonly used face datasets.

Keywords: Face recognition, low-resolution, feature learning, multi-scale patch

1. Introduction

Face image recognition, as one of the most commonly used biometrics technologies, has become the research hotspot of the pattern recognition community in past decades [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. Generally, most of the current methods perform well on the cases that the acquired region of interest (ROI) has high image resolution and contains enough

*Corresponding author
Email address: csggao@gmail.com (Guangwei Gao)

discriminative information for recognition tasks. However, in real-world robotics and video surveillance applications, the observed faces generally have low-resolution (LR) together with pose and illumination variations, while the referenced faces are always enrolled with high resolution (HR). The main challenge is to match an LR probe face with limited details against HR gallery faces. We name this kind of problem as low-resolution face recognition (LRFR). An alternative solution is down-sampling the HR galleries and then matching in the same resolution space. In this way, the resolution discrepancy is reduced at the expense of losing the discriminative facial details in the gallery.

Generally, there are two typical categories to address the LRFR problem. One is super-resolution approaches, which first synthesize the target HR faces from the observed LR image, and then utilize traditional face recognition approaches in the common resolution domain. The other is resolution-robust feature extraction methods, which directly extract discriminative features from respective domains, thus obtain better performance than super-resolution methods.

Super-resolution approaches devote to synthesize the target HR faces from the input LR image to alleviate the resolution gap. Recently, most methods mainly focus on the reasonable design of regularization terms. Jiang et al. [11] incorporated smooth prior to their objective to obtain stable reconstruction weights. Recently, they further proposed an efficient context-patch based face hallucination method via thresholding locality-constrained representation and reproducing learning strategy [12]. Liu et al. [13] proposed to super-resolve the target face images and suppress noise simultaneously. Rajput et al. [14] proposed an iterative sparsity and locality-constrained representation model for robust face image super-resolution. To hallucinate HR face from a real low-quality face, a definition-scalable inference method was proposed in [15]. These previous super-resolution approaches fail to acquire the high-quality face from the recognition respective, lacking discriminative facial details for the subsequent face recognition tasks.

Resolution-invariant approaches just consider the face recognition problem directly. Jian et al. [16] proposed a singular value decomposition-based framework to perform both hallucination and recognition simultaneously. Yang et al. [17] presented a discriminative multidimensional scaling method to take both intraclass distances and interclass distances into account. Wang et al. [18] studied the LR image recognition problem with the assistance of HR samples and proposed partially coupled networks to combat the domain mismatch problem. Banerjee et al. [19] designed a deep network-based method to synthesize realistic HR images for recognition. Mudunuri et al. [20] propose a convolutional neural network architecture to recognize LR images as well as generalize well to images of unseen categories. Later, Mudunuri et al. [21] further proposed an orthogonal dictionary alignment method with a re-ranking scheme to enhance the recognition accuracy. Recently, Li et al. [22] analyzed different metric learning methods for LRFR and provided a comprehensive analysis of the experimental results in real surveillance applications. Zangeneh et al. [23] presented a two-branch deep convolutional neural network for LRFR task. However, the aforementioned methods usually take the holistic images into account when extracting resolution-robust features, while the complementary information from different patch scales is overlooked.

In this paper, to tackle the above problems, we formulate a multi-scale patch based

representation feature learning (MSPRFL) scheme to further promote the performance of low-resolution face recognition. Due to the various degradations from HR to LR space, it is impractical to extract sufficient distinct features directly from the LR faces for ideal matching of LR probes against HR galleries. On the contrary, we propose to learn resolution-robust representation features on patches with the help of a pre-collected training dataset. Then, the new learned resolution-robust features are exploited to recognize the LR probe by integrating the recognition results from all patches. Because it is usually unreasonable to predefine an approving patch scale for all experimental cases and patches from different scales may provide the complementary discriminative ability for recognition, we further devote to fuse the multi-scale outputs by learning scale weights via an ensemble optimization model. Experimental results verified the superiority of our MSPRFL in recognizing the low-resolution faces compared to several state-of-the-art approaches.

We organize the rest of our paper as follows. Section 2 details our method, including local feature extraction, patch-based representation feature learning, recognition strategy, and multi-scale fusion. Section 3 evaluates the effectiveness of the proposed method and further provides some discussions. The conclusions are given in Section 4.

2. The proposed MSPRFL

The local patch-based trick has been largely used in many communities [7, 12, 24, 25, 26, 27, 28, 29, 30, 31]. The reasons are two-fold: (i) Compared with the global features, local features are more robust to local variations such as pose, expression, and illumination. (ii) Since the local dictionary based linear system seems to be under-determined, the local patch-based linear system may obtain more accurate representation than the global image-based one. Based on these two observations, in this section, we will detail our proposed method. We first introduce the local feature extraction strategy in subsection 3.1 and then give our patch-based representation feature learning formulation in subsection 3.2. We next introduce the classification procedure based on these representation features in subsection 3.3. Finally, we combine the multi-scale outputs in subsection 3.4 to achieve the best recognition performance. The whole framework of the proposed MSPRFL method is shown in Fig. 1.

2.1. Local feature extraction

The most commonly used feature descriptors to depict the image patches can be the raw luminance values of pixels. Generally, from the viewpoint of perception, the human visual system is more focused on the high-frequency component of the object. Also, the raw patch features can not reveal the potential relationship between LR patches and their HR counterparts. Inspired by the work in [32], we also exploit the gradients maps of the local patches as the feature descriptor due to its effectiveness and simplicity. The 1-D filters used here are composed of:

$$\begin{aligned} f_1 &= [-1, 0, 1], & f_2 &= f_1^T, \\ f_3 &= [1, 0, -2, 0, -1], & f_4 &= f_3^T, \end{aligned} \tag{1}$$

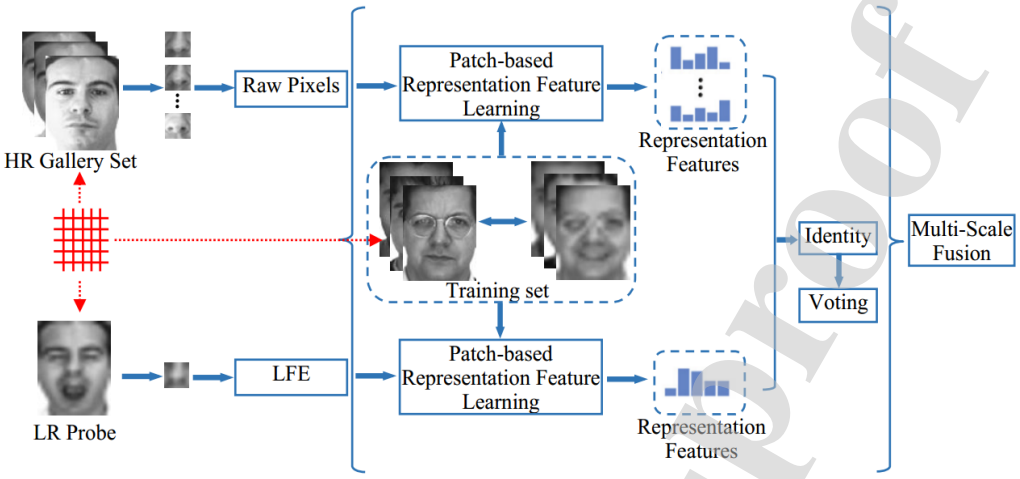


Figure 1: Overview of the proposed MSPRFL method (the red grid denotes the patch dividing strategy).

in which the superscript T denotes the transpose operation. For each LR patch, we can obtain four feature vectors using these four filters, which are then concatenated into one vector. In practice, we do not directly conduct these four filters on those extracted image patches; instead, we filter the global image with these four filters in the horizontal and vertical directions by

$$G_{ij} = A_i * f_j, j = 1, 2, 3, 4, \tag{2}$$

where $A_i \in \Re^{p \times q}$ is the i^{th} LR image, $*$ operator stands for convolution. Thus, for each LR image, four gradient maps are obtained. Then, at each location, we concatenate the four patches extracted from these gradient maps to form the feature descriptor by

$$g_i = [g_{i1}; g_{i2}; g_{i3}; g_{i4}], \tag{3}$$

where patch $g_{ij} \in \Re^{d \times 1}$ comes from the gradient maps $G_{ij}(j = 1, 2, 3, 4)$ with the same location. Specifically, the features for each LR image patch region also involve their neighboring relationship, which has been verified to be helpful for patch-based recognition strategy [24]. Finally, $g_i \in \Re^{4d \times 1}$ is considered as the local patch feature representation from the LR image A_i.

2.2. Patch-based representation feature learning

Let $A_l = \{A_l^1, A_l^2, \ldots, A_l^M\}$ and $A_h = \{A_h^1, A_h^2, \ldots, A_h^M\}$ be the LR and HR training faces, Y_l and Y_h be the test probe and gallery face images. For A_h and Y_h, their overlapped patches at location (i, j) are represented as $\{A_h^m(i, j) \in \Re^{d \times 1} | 1 \le i \le R, 1 \le j \le C\}$ and $\{Y_h(i, j) \in \Re^{d \times 1} | 1 \le i \le R, 1 \le j \le C\}$. Using the local feature extraction (LFE) strategy described in subsection 2.1, the feature descriptors of overlapped position patches are also extracted from A_l and Y_l, and represented as $\{A_l^m(i, j) \in \Re^{4d \times 1} | 1 \le i \le R, 1 \le j \le C\}$ and

$\{Y_l(i,j) \in \Re^{4d\times 1} | 1 \le i \le R, 1 \le j \le C\}$, where R and C denote the numbers of patch in every row and column.

For simplicity, we omit the index (i,j) in the next section. For each patch Y_l and Y_h, they can be collaboratively with

$$Y_l = \sum_{m=1}^{M} A_l^m \alpha_l^m + e_l,\ Y_h = \sum_{m=1}^{M} A_h^m \alpha_h^m + e_h, \tag{4}$$

where e_l and e_h denote the representation error. We can get the representation weights of the patch Y_l and Y_h by considering the subsequent constrained least square problem

$$\begin{aligned} \alpha_l^* &= \arg\min_{\alpha_l} \left\| Y_l - \sum_{m=1}^{M} A_l^m \alpha_l^m \right\|_2^2,\ s.t. \sum_{m=1}^{M} \alpha_l^m = 1, \\ \alpha_h^* &= \arg\min_{\alpha_h} \left\| Y_h - \sum_{m=1}^{M} A_h^m \alpha_h^m \right\|_2^2,\ s.t. \sum_{m=1}^{M} \alpha_h^m = 1. \end{aligned} \tag{5}$$

In our experiments, we find that the problem (5) may have an unstable solution. Researchers [26, 33] have made many efforts on regularization constraints and suggested that locality constraints are more powerful than sparsity constraints in exposing the inherent geometry of the nonlinear manifold. Thus, we also introduce the locality constraint and rewrite our objective as

$$\begin{aligned} \alpha_l^* &= \arg\min_{\alpha_l} \left\{ \left\| Y_l - \sum_{m=1}^{M} A_l^m \alpha_l^m \right\|_2^2 + \lambda_l \sum_{m=1}^{M} [d_l^m \alpha_l^m]^2 \right\}, \\ \alpha_h^* &= \arg\min_{\alpha_h} \left\{ \left\| Y_h - \sum_{m=1}^{M} A_h^m \alpha_h^m \right\|_2^2 + \lambda_h \sum_{m=1}^{M} [d_h^m \alpha_h^m]^2 \right\}, \end{aligned} \tag{6}$$

where λ_l and λ_h are regularization parameters, and $d_l^m = \|Y_l - A_l^m\|_2^2$ is the metric between LR probe and each LR training atom (d_h^m has the similar definition).

Actually, from the viewpoint of feature extraction, the relational vector α_l^* returns a representation feature containing the representation similarity of a probe LR patch to each LR training patch. Also, for a gallery HR patch, α_h^* returns a representation feature vector of representation similarity to each HR training patch. Using this representation similarity learning strategy, for each LR probe and HR gallery, we can study resolution-robust discriminative representation features for the following recognition task.

We will detail how to obtain the above relational feature vector α_l^* (α_h^* can be obtained in the same way). The problem (6) can be reformulated as the following form:

$$\alpha_l^* = \arg\min_{\alpha_l} \{ \| Y_l - A_l \alpha_l \|_2^2 + \lambda_l \| D_l \alpha_l \|_2^2 \}, \tag{7}$$

where D_l is a diagonal matrix with the size of $M \times M$, and with entries

$$D_l^{mm} = d_l^m, 1 \leq m \leq M. \tag{8}$$

Following [34], we can derive the analytical solution of objective (7) as

$$\alpha_l = (G_l + \lambda_l D_l^2) \backslash ones(M, 1), \tag{9}$$

where variable $ones(M, 1)$ denotes an M-dimensional vector with entries of ones, and the symbol "\" represents the left matrix division operation. while $G_l = C^T C$ is a covariance matrix with

$$C = Y_l \cdot ones(M, 1)^T - A_l. \tag{10}$$

The final representation feature vector is given by rescaling to satisfy $\sum_{m=1}^{M} \alpha_l^m = 1$.

2.3. Recognition procedure

With the help of a referenced LR-HR training patch pair A_l and A_h, we can obtain the representation features from the LR probe patch y and the HR gallery patch set X_h for low-resolution face recognition purposes. Concerning all elements in X_h, their representative and discriminative features over the training patch set A_h can be denoted as $F = [F_1, F_2, \ldots, F_c]$, where F_k is the subset of the k^{th} class, with each column of F_k is a representation feature vector from class k. Concerning probe patch y, we use a column vector x to denote its representation feature over A_l. Then, the representation of x over F is

$$\rho^* = \arg\min_{\rho} \{\|x - F\rho\|_2^2 + \eta\|\rho\|_2^2\}. \tag{11}$$

η is the balance parameter. The solution of problem (11) is $\rho^* = \left(F^T F + \eta \cdot I\right)^{-1} F^T x$. Thus, for each class, we can calculate its regularized reconstruction by

$$r_i(x) = \|x - F_i \cdot \rho_i^*\|_2^2 / \|\rho_i^*\|_2^2, \tag{12}$$

where vector ρ_i^* denotes the weights corresponding to the i^{th} class. Finally, the recognition output of the probe patch y is Identity $(y) = \text{argmin}_i \{r_i\}$.

2.4. Multi-scale fusion

In our experiments, we find that it is usually unreasonable to find a suitable patch scale for various experimental configurations (e.g., different databases). The recognition results of patch based representation feature learning (PRFL) versus various patch scales and testing gallery scales on two widely used face databases are shown in Fig. 2, from which, we can make the following findings. First, for different datasets, the suitable patch scale always varies a lot. Second, for different testing gallery sample scale per person, the suitable patch scale also varies a lot. To handle those above troubles, we propose to adaptively fuse the multi-scale complementary recognition results from different patch scales for further performance improvement.

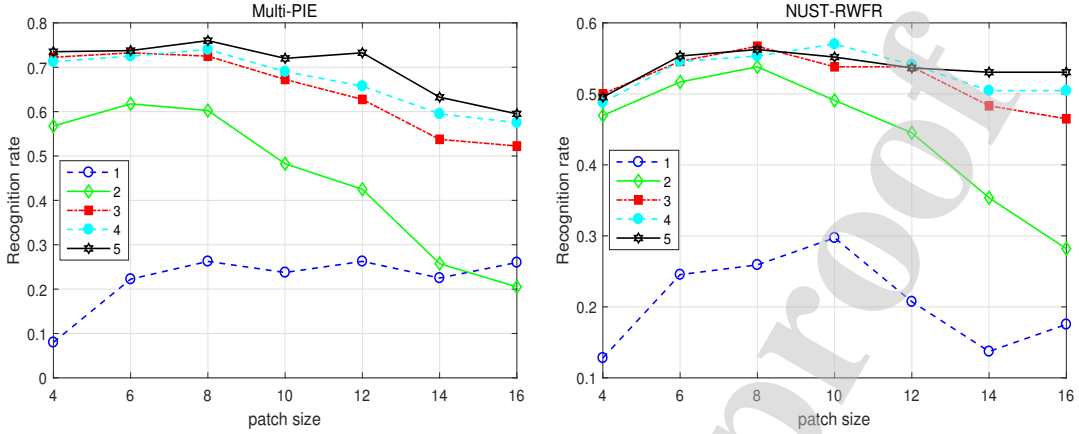


Figure 2: The recognition results of PRFL based on different patch size and testing gallery sample size.

With the help of an example set $T = \{(x_i, z_i)\}\,(i = 1, 2, \ldots, n)$ and s scales, we can define a decision matrix D [35] by

$$d_{ij} = f(h_{ij}, z_i) = \begin{cases} +1, & \text{if } h_{ij} = z_i \\ -1, & \text{if } h_{ij} \neq z_i \end{cases}, \tag{13}$$

where z_i denotes the true label of example x_i and $h_{ij}(i = 1, 2, \ldots, n, j = 1, 2, \ldots, s)$ denote the recognition outputs of samples x_i on s scales. Then, we can define the ensemble loss of x_i as

$$l_{x_i} = l(\varepsilon(x_i)) = l\left(\sum_{j=1}^{s} \beta_j d_{ij}\right), \tag{14}$$

where $\beta = [\beta_1, \beta_2, \ldots, \beta_s]^T$ is the weight parameters of s scales and $\sum_{j=1}^{s} \beta_j = 1$. Considering the whole set T, its ensemble loss is denoted as

$$l(S) = \sum_{i=1}^{n} l_{x_i} = \sum_{i=1}^{n} [1 - \sum_{j=1}^{s} \beta_j d_{ij}]^2 = \| e_1 - D\beta \|_2^2, \tag{15}$$

where $e_1 = [1, 1, \ldots, 1]^T$, whose length is s.

The goal of ensemble optimization is to minimize equation (15). However, directly minimizing equation (15) will lead to unstable solution. Inspired by the robustness of sparse coding [36], we impose l_1-regularized constraint on the objective to obtain the adaptive fusion weights:

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta} \{ \| e_1 - D\beta \|_2^2 + \gamma \| \beta \|_1 \} \\ s.t. &\sum_{j=1}^{s} \beta_j = 1, \beta_j > 0, \end{aligned} \tag{16}$$

Figure 3: Example images from the LFW face dataset. Top: HR gallery faces; Bottom: LR probe face.

Figure 4: Example images from the Multi-PIE face dataset. Top: HR gallery faces; Bottom: LR probe faces.

where γ is the regularization parameter. By some simple algebraic derivations, equation (16) can be reformulated as

$$\hat{\beta} = \arg\min_{\beta} \left\{ \left\| \hat{e} - \hat{D}\beta \right\|_2^2 + \gamma \|\beta\|_1 \right\} \; s.t. \; \beta_j > 0, j = 1, 2, \cdots, s, \tag{17}$$

where $\hat{e} = [e_1; 1]$, $\hat{D} = [D; e_2]$ and $e_2 = [1, 1, \ldots, 1]$ have a length of s. Problem (17) can be easily solved by the widely used l_1_ls solver [37]. Once we gain the optimal values of the scale weights, the recognition result of the probe sample x_i is Identity $(x_i) = \arg\max_k \{\sum \beta_j | h_{ij} = k\}$.

3. Experiments and discussions

In this section, we perform experiments on three public face sets (LFW, Multi-PIE, and real-world NUST-RWFR) to validate the superiority of the proposed method for recognizing the low-resolution faces. Without loss of generality, we treat the original high-quality face images as HR galleries, while the downsampled and then upscaled faces are taken as LR probes.

3.1. Datasets descriptions

The Labeled Faces in the Wild (LFW) [38] is a database of face photographs designed for studying the problem of unconstrained face recognition. The database contains more than

Figure 5: Example images from the NUST-RWFR face dataset. Top: HR gallery faces; Bottom: LR probe faces.

13,000 face images of 5,749 persons collected from the web. LFW-a contains the same images available in the original LFW database after alignment using a commercial face alignment software. We gathered the objects that have more than ten samples to form a dataset with 158 objects from LFW-a. All the face parts are manually cropped and resized to 165×120. Fig. 3 exhibits several example faces from this set.

The Multi-PIE face dataset [39] collects face images from 337 subjects in four separate sessions together with expression, pose and illumination variations. In our experiment, we choose a subset that contains 164 individuals from session 3. For each person, 10 samples with neutral expressions and another 10 samples with smile expressions are used. We resize all the images to 100×80. Fig. 4 depicts some examples of face images from this dataset.

The real-world database, NUST-RWFR face database [40], collects 2400 color faces from 100 subjects with different lighting conditions, facial expressions, and blurring. All the images are acquired in a real-world situation in two separate periods, and each period includes 12 samples. The image qualities in the first period are relatively good, while that in the second period are poor. We manually crop the face region of each image and resize them to 80×80. Some samples are listed in Fig. 5.

3.2. Ablation study

In this part, we investigate the effect of local feature extraction and multi-scale ensemble learning. In our method, we use the first- and second-order gradient maps of the local patches as the feature descriptors due to its effectiveness and simplicity. Then, these extracted features are used for the discriminative representation feature leaning on a given training dataset. Also, we propose to adaptively fuse the multi-scale complementary recognition results from different patch scales for further performance improvement using multi-scale ensemble learning.

To illustrate the effect of local feature extraction, we first give an example on the NUST-RWFR face dataset. Fig. 6(a) illustrates the combination coefficients ρ of the learned representation features without (the top row) or with (the bottom row) feature extraction for a query patch from subject 1. Fig. 6(b) shows the corresponding residuals for these 100 subjects. From Fig. 6, we can see that the coefficients obtained by our method with local feature extraction are much sparse than those without local feature extraction, and the

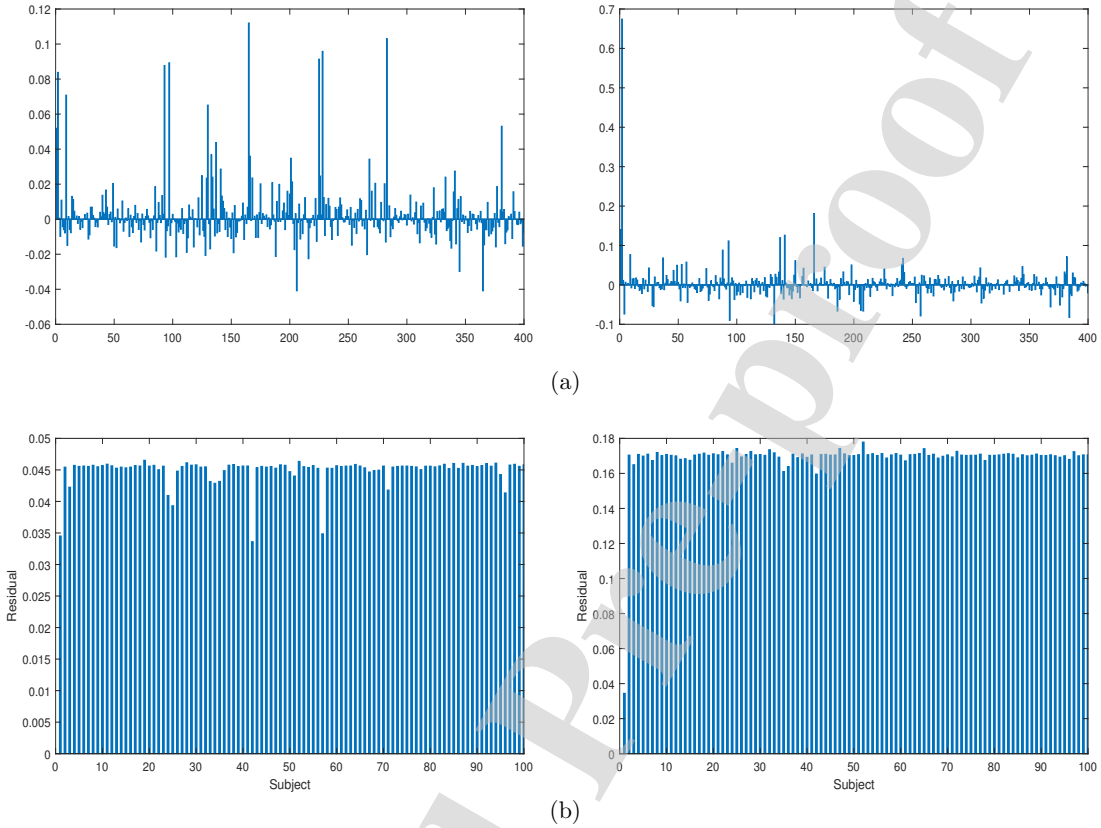
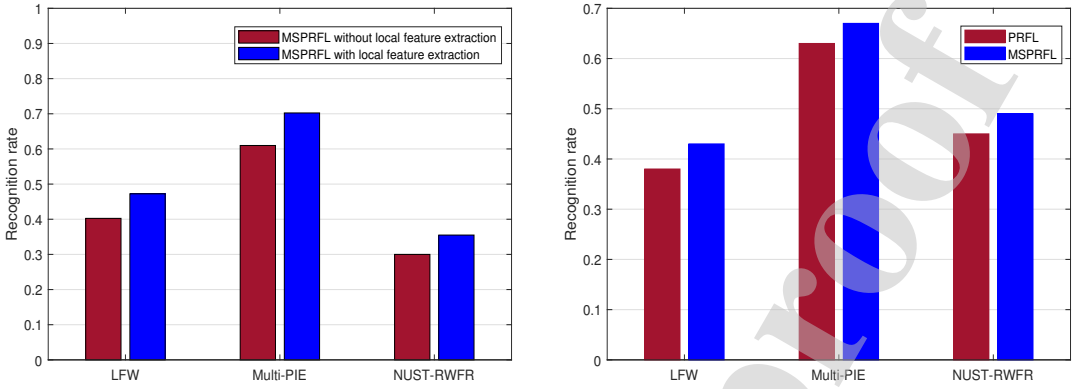


Figure 6: An explanatory example from the NUST-RWFR dataset. (a) Combination coefficients of a query patch from subject 1. (b) The residuals of the query patch from subject 1. The left column indicates results without feature extraction. While the right column indicates results with feature extraction. Large coefficients correspond to the correct subject and the smallest residual is related to subject 1.

dominant coefficients are related to subject 1. Thus, the smallest residual in Fig. 6(b) (the bottom row) corresponds to the correct label (subject 1). This example verifies that with the local feature extraction, the learned representation features are more discriminative. We also give the recognition results of our method with or without local feature extraction on three datasets in Fig. 7(a). The recognition results further demonstrate that our method with local feature extraction can obtain better performance. To investigate the effect of multi-scale ensemble learning, we give the recognition results of PRFL and its multi-scale version (i.e., MSPRFL) on three datasets in Fig. 7(b). From Fig. 7(b), we can observe that compared with PRFL, by fusing the multi-scale recognition results from different patch scales, MSPRFL indeed promotes the performance improvement.



(a) (b)

Figure 7: The effect of (a) local feature extraction and (b) multi-scale ensemble learning in our method.

3.3. Experimental settings

In our method, we use seven scales, which have the sizes of 4×4, 6×6, 8×8, 10×10, 12×12, 14×14 and 16×16. Parameter γ (in Eq. (17)) is fixed as 0.1 for all databases. For simplicity, we set $\lambda_l = \lambda_h$ in our experiments. As in [24], we also separate the whole gallery set into the new probe subset (one sample per subject) and the new gallery subset (the remainder of the gallery set) for scale weight learning purpose.

For the LFW face dataset, the HR face samples have a size of 48×48. We first downsample the HR samples by a scaling factor of K (K is 2, 4, 8) and then upsampled them back to the original resolution to serve as the corresponding LR versions. We divide the dataset into three parts. We treat the first part (3 images per person) as the training subsets A_h and A_l. The second part (5 images per person) is used as the HR gallery set X_h. The remainders (2 images per person) are used as the LR probe set Y_l. For each scaling factor K, the tests are performed 10 runs for each method.

For the Multi-PIE face dataset, we set the size of the HR images as 32×24, 44×32, 64×48 and 100×80 in this experiment. All the HR face samples are first downsampled by a scaling factor of 4 and then upscaled back to the primal size to serve as the LR probes. As for each person, we randomly choose 3 samples with neutral expressions as the training subsets A_h and A_l, another 4 samples with neutral expressions as the HR gallery subset X_h, and the remaining 4 samples with smile expressions as the LR probe subset Y_l. For each image size, the tests are performed 10 runs for each method.

For the NUST-RWFR face dataset, the size of the HR face images is 48×48. We first downsample the HR images to the size 12×12 and then upsampled them back to the original resolution by bicubic interpolation to serve as the LR faces. For each class, we randomly pick 4 samples as the training subsets A_h and A_l, another 4 samples as the LR probe subset Y_l, and the rest K (K is 6, 8, 10, 12, 14) samples as the HR gallery set X_h. For each K, we perform the tests 10 runs for each method.

Table 1: The recognition results (%) on the LFW database.

Methods	×2	×4	×8
LcBR	47.48±3.49	32.17±3.49	17.25±2.79
ISLcR	47.47±3.56	32.76±3.60	17.29±2.51
TLcR	47.52±3.63	32.87±3.47	17.55±2.62
RPCN	47.71±3.45	33.36±3.63	20.37±2.34
DAlign	47.82±3.48	33.69±3.76	20.48±2.75
Centerloss	48.52±3.87	33.95±3.49	20.62±2.60
TDCNN	48.58±3.57	33.65±3.36	20.54±2.19
MSPRFL	**50.54±2.52**	**36.90±2.77**	**24.25±1.17**

Table 2: The recognition results (%) on the Multi-PIE database.

Methods	32 × 24	44 × 32	64 × 48	100 × 80
LcBR	47.57±7.29	61.33±7.38	74.31±7.15	84.26±4.61
ISLcR	48.19±7.48	61.16±7.90	75.16±7.82	84.60±5.82
TLcR	48.47±7.45	61.79±7.91	75.52±7.35	84.65±4.25
RPCN	49.19±7.53	62.83±7.16	76.75±6.83	85.34±4.58
DAlign	50.27±7.66	62.96±7.34	76.87±6.58	85.53±4.91
Centerloss	50.59±7.08	63.15±6.47	77.32±6.36	86.85±4.89
TDCNN	50.93±7.25	63.14±6.35	77.43±6.55	86.93±3.75
MSPRFL	**54.88±6.67**	**67.80±7.02**	**80.80±6.05**	**89.98±3.65**

3.4. Comparison results

The effectiveness of our method is evaluated by comparing it with two types of state-of-the-arts: the first is super-resolution methods, including LcBR [13], ISLcR [14] and TLcR [12] approach. These vision-based methods cascade the super-resolved HR faces with one well-known baseline of deep-learning based recognition methods, i.e., DFLA [41], for recognition tests. The second one is resolution-invariant feature extraction methods that just use LR images as the probe, including RPCN [18], DAlign [21], Centerloss [22] and TDCNN [23] approach. It is worthy to note that in all experiments, for resolution-invariant methods, we exploit the same gallery and probe sets. Concerning super-resolution methods, we exploit the same training sets.

Tables 1-3 show the average recognition rates and std of the respective approaches in all cases. For a better demonstration, we also list some super-resolved results in Fig. 8. As seen in Fig. 8, the hallucinated faces usually have some ghosting artifacts around mouth, eye and face contours. Recognition results in Tables 1-3 also demonstrate that directly feeding the hallucinated HR faces into classifier engine seems to contribute less to recognition. On the contrary, the resolution-robust feature extraction methods (i.e., RPCN, DAlign, Centerloss, and TDCNN) consider the discriminative abilities of features, getting higher recognition rates than super-resolution methods. The quantitative compared results also show that our MSPRFL outperforms super-resolution methods and resolution-robust feature extrac-

Table 3: The recognition results (%) on the NUST-RWFR database.

Methods	6	8	10	12	14
LcBR	38.73±3.45	40.49±3.52	43.36±3.76	46.12±3.32	47.09±3.17
ISLcR	39.58±4.76	40.43±3.98	43.38±3.88	46.40±3.56	47.24±3.77
TLcR	39.90±3.34	40.95±3.48	43.90±3.45	46.78±3.34	47.65±3.49
RPCN	40.76±2.61	41.49±3.54	44.48±3.21	47.80±3.60	48.18±3.33
DAlign	40.96±2.93	42.25±3.50	44.68±3.20	48.14±3.18	48.27±3.39
Centerloss	41.65±2.85	42.81±3.14	45.68±3.22	48.45±3.00	48.68±3.22
TDCNN	41.46±3.97	42.87±3.30	45.26±3.32	48.15±2.99	48.21±3.24
MSPRFL	**44.80±2.23**	**45.90±2.83**	**49.25±2.45**	**50.43±2.50**	**51.65±2.59**

Figure 8: Face hallucination results on the NUST-RWFR dataset. From first to the fifth columns: the observed LR inputs, the super-resolved results of Bicubic interpolation, LcBR [13], ISLcR [14], and TLcR [12]. The last two columns denote the HR probe and one HR gallery.

tion methods, respectively. These achievements confirm that by integrating the recognition results of all patches and further taking full advantage of the complementary discriminative ability from various patch sizes, our proposed MSPRFL can dramatically enhance the recognition performance.

3.5. Parameter analysis

In this section, we mainly investigate the effect of parameters used in our approach. We also perform experiments on above mentioned three datasets (i.e., LFW, Multi-PIE, and NUST-RWFR) and the experimental configurations are the same as that in the above experiments. In this experiment, we just vary one parameter while fixing the other one. We divide the whole training set into three parts. The first part (one image per person) is used as the probe set, the second part (one image per person) is treated as the gallery set, and the remainders are used as the dictionary.

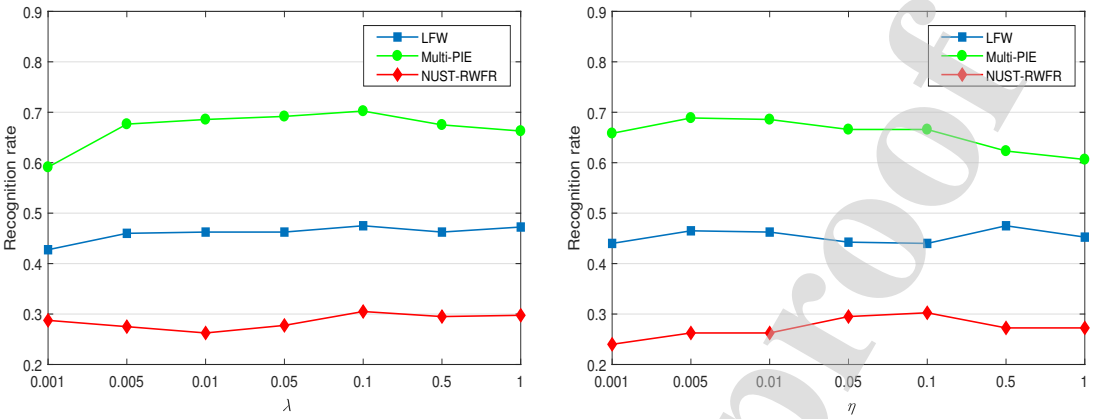


Figure 9: The recognition results of MSPRFL with different values of parameters in different face datasets.

Fig. 9 plots the recognition rates of MSPRFL with different values of parameters λ and η in different face datasets. As seen in Fig. 9, MSPRFL always achieves higher recognition results when the parameter λ is set around 0.1. For LFW and Multi-PIE database, MSPRFL always gives higher recognition results when the parameter η is set around 0.005. For the NUST-RWFR database, MSPRFL can obtain higher recognition results when the parameter η is set around 0.1.

3.6. Running time comparison

In this part, we compare the computational cost of different methods. For simplicity, we only give the test results of one face image from the NUST-RWFR face database. The tests are also performed 10 runs for each method. The average running time of each method is listed in Fig. 10. We can see that the position-patch based super-resolution methods (i.e., LcBR, ISLcR, and TLcR) require similar running time since they both requires a few matrix multiplications steps. The deep based methods (i.e., RPCN, DAlign, Centerloss, and TDCNN) runs faster since the network can be trained offline. Due to the patch based representation feature learning, our method requires much more time than deep based methods. However, our method runs faster than super-resolution based methods since the scale weight can be learned offline.

4. Conclusions

In this work, we present a new model named multi-scale patch based representation feature learning (MSPRFL) for low-resolution face recognition purposes. In the proposed method, the multi-level information of patches and the multi-scale outputs are thoroughly utilized. More specially, the proposed MSPRFL takes full advantage of multi-level information to learn more accurate resolution-robust representation features of each patch. Moreover, the recognition results of all patches are then combined by voting strategy. Finally, we

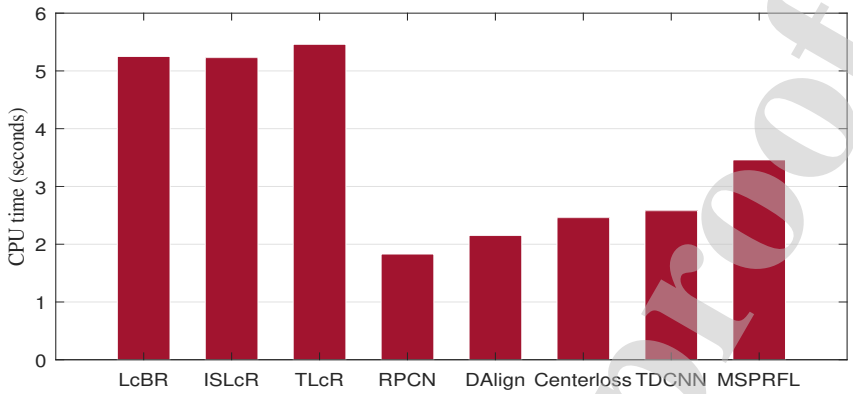


Figure 10: The average running time of each method on NUST-RWFR face database.

further fuse the multi-scale outputs by taking full advantage of the complementary discriminative information from different patch scales. Experiments on several public face datasets have illustrated the effectiveness of our MSPRFL.

Acknowledgement

The author would like to thank the editor and the anonymous reviewers for their detailed comments and constructive suggestions which greatly contribute to this paper. This work was supported in part by the National Key Research and Development Program of China under Project no. 2018AAA0100102; the National Natural Science Foundation of China under Grant nos. 61972212, 61772568, 61972213 and 61702197; the Six Talent Peaks Project in Jiangsu Province under Grant no. RJFW-011; the Natural Science Foundation of Jiangsu Province under Grant no. BK20190089; the open fund project of Science and Technology on Space Intelligent Control Laboratory under Grant no. 6142208180302; and the Open Fund Project of Provincial Key Laboratory for Computer Information Processing Technology (Soochow University) (no. KJS1840).

References

[1] G. Gao, J. Yang, X.-Y. Jing, F. Shen, W. Yang, D. Yue, Learning robust and discriminative low-rank representations for face recognition with occlusion, Pattern Recognition 66 (2017) 129–143.

[2] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, Y. Xu, Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (1) (2017) 156–171.

[3] G. Gao, J. Yang, S. Wu, X. Jing, D. Yue, Bayesian sample steered discriminative regression for biometric image classification, Applied Soft Computing 37 (2015) 48–59.

[4] S. P. Mudunuri, S. Biswas, Low resolution face recognition across variations in pose and illumination, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (5) (2016) 1034–1040.

[5] X. Yin, X. Liu, Multi-task convolutional neural network for pose-invariant face recognition, IEEE Transactions on Image Processing 27 (2) (2018) 964–975.

[6] J. Zhao, J. Han, L. Shao, Unconstrained face recognition using a set-to-set distance measure on deep learned features, IEEE Transactions on Circuits and Systems for Video Technology 28 (10) (2018) 2679–2689.
[7] C. Peng, N. Wang, J. Li, X. Gao, Dlface: Deep local descriptor for cross-modality face recognition, Pattern Recognition 90 (2019) 161–171.
[8] J. Yang, L. Zhang, Y. Xu, J.-y. Yang, Beyond sparsity: The role of l1-optimizer in pattern classification, Pattern Recognition 45 (3) (2012) 1104–1118.
[9] Z. Liu, J. Wang, G. Liu, L. Zhang, Discriminative low-rank preserving projection for dimensionality reduction, Applied Soft Computing 85 (2019) 105768.
[10] J. Yang, D. Chu, L. Zhang, Y. Xu, J. Yang, Sparse representation classifier steered discriminative projection with applications to face recognition, IEEE Transactions on Neural Networks and Learning Systems 24 (7) (2013) 1023–1035.
[11] J. Jiang, J. Ma, C. Chen, X. Jiang, Z. Wang, Noise robust face image super-resolution through smooth sparse representation, IEEE Transactions on Cybernetics 47 (11) (2017) 3991–4002.
[12] J. Jiang, Y. Yu, S. Tang, J. Ma, A. Aizawa, K. Aizawa, Context-patch based face hallucination via thresholding locality-constrained representation and reproducing learning, IEEE Transactions on Cybernetics 50 (1) (2020) 324–337.
[13] L. Liu, C. P. Chen, S. Li, Y. Y. Tang, L. Chen, Robust face hallucination via locality-constrained bi-layer representation, IEEE Transactions on Cybernetics 48 (4) (2018) 1189–1201.
[14] S. S. Rajput, K. Arya, V. Singh, Robust face super-resolution via iterative sparsity and locality-constrained representation, Information Sciences 463 (2018) 227–244.
[15] X. Hu, P. Ma, Z. Mai, S. Peng, Z. Yang, L. Wang, Face hallucination from low quality images using definition-scalable inference, Pattern Recognition 94 (2019) 110–121.
[16] M. Jian, 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 25 (11) (2015) 1761–1772.
[17] F. Yang, W. Yang, R. Gao, Q. Liao, Discriminative multidimensional scaling for low-resolution face recognition, IEEE Signal Processing Letters 25 (3) (2018) 388–392.
[18] Z. Wang, S. Chang, Y. Yang, D. Liu, T. S. Huang, Studying very low resolution recognition using deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4792–4800.
[19] S. Banerjee, S. Das, Lr-gan for degraded face recognition, Pattern Recognition Letters 116 (2018) 246–253.
[20] S. Prasad Mudunuri, S. Sanyal, S. Biswas, Genlr-net: Deep framework for very low resolution face and object recognition with generalization to unseen categories, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 489–498.
[21] S. P. Mudunuri, S. Venkataramanan, S. Biswas, Dictionary alignment with re-ranking for low-resolution nir-vis face recognition, IEEE Transactions on Information Forensics and Security 14 (4) (2019) 886–896.
[22] P. Li, L. Prieto, D. Mery, P. J. Flynn, On low-resolution face recognition in the wild: Comparisons and new techniques, IEEE Transactions on Information Forensics and Security 14 (8) (2019) 2000–2012.
[23] E. Zangeneh, M. Rahmati, Y. Mohsenzadeh, Low resolution face recognition using a two-branch deep convolutional neural network architecture, Expert Systems with Applications 139 (2020) 112854.
[24] P. Zhu, L. Zhang, Q. Hu, S. C. Shiu, Multi-scale patch based collaborative representation for face recognition with margin distribution optimization, in: Proceedings of European Conference on Computer Vision, Springer, 2012, pp. 822–835.
[25] X. Ma, J. Zhang, C. Qi, Hallucinating face by position-patch, Pattern Recognition 43 (6) (2010) 2224–2236.
[26] J. Jiang, R. Hu, Z. Wang, Z. Han, Noise robust face hallucination via locality-constrained representation, IEEE Transactions on Multimedia 16 (5) (2014) 1268–1281.
[27] C. Zhao, X. Wang, W. K. Wong, W. Zheng, J. Yang, D. Miao, Multiple metric learning based on

bar-shape descriptor for person re-identification, Pattern Recognition 71 (2017) 218–234.
[28] S. Soltanpour, B. Boufama, Q. J. Wu, A survey of local feature methods for 3d face recognition, Pattern Recognition 72 (2017) 391–406.
[29] Y. Chen, J. Wang, X. Chen, M. Zhu, K. Yang, Z. Wang, R. Xia, Single-image super-resolution algorithm based on structural self-similarity and deformation block features, IEEE Access 7 (2019) 58791–58801.
[30] H. Lu, Y. Li, T. Uemura, H. Kim, S. Serikawa, Low illumination underwater light field images reconstruction using deep convolutional neural networks, Future Generation Computer Systems 82 (2018) 142–148.
[31] X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, X. Li, Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval, IEEE Transactions on Cybernetics (2019) 1–14.
[32] J. Yang, J. Wright, T. S. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE Transactions on Image Processing 19 (11) (2010) 2861–2873.
[33] K. Yu, T. Zhang, Y. Gong, Nonlinear learning using local coordinate coding, in: Proceeding of Advances in Neural Information Processing Systems, 2009, pp. 2223–2231.
[34] Z. Fan, D. Zhang, X. Wang, Q. Zhu, Y. Wang, Virtual dictionary based kernel sparse representation for face recognition, Pattern Recognition 76 (2018) 1–13.
[35] J. C. Gower, G. B. Dijksterhuis, et al., Procrustes problems, Vol. 30, Oxford University Press on Demand, 2004.
[36] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2) (2009) 210–227.
[37] K. Koh, S.-J. Kim, S. Boyd, An interior-point method for large-scale l1-regularized logistic regression, Journal of Machine Learning Research 8 (2007) 1519–1555.
[38] G. B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled faces in the wild: A database for studying face recognition in unconstrained environments, Tech. Rep. 07-49, University of Massachusetts, Amherst (October 2007).
[39] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, Image and Vision Computing 28 (5) (2010) 807–813.
[40] J. Qian, J. Yang, Y. Xu, Local structure-based image decomposition for feature extraction with applications to face recognition, IEEE Transactions on Image Processing 22 (9) (2013) 3591–3603.
[41] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: Proceeding of European conference on computer vision, Springer, 2016, pp. 499–515.

Highlights

1. The paper focused on face recognition scenarios where the testing images have low- resolutions.
2. The paper proposed a multi-scale patch based representation feature learning scheme to exploit multi-level information to learn more accurate resolution-robust representation features of each patch for low-resolution face recognition problem.
3. Experimental results demonstrate the effectiveness of our method.

Declaration of interest statement:

The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.