Pattern Recognition 90 (2019) 161–171

Contents lists available at ScienceDirect

Pattern Recognition

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

DLFace: Deep local descriptor for cross-modality face recognition

Chunlei Peng[a], Nannan Wang[b], Jie Li[c], Xinbo Gao[d,*]

[a] State Key Laboratory of Integrated Services Networks, School of Cyber Engineering, Xidian University, Xi'an, Shaanxi 710071, PR China
[b] State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an, Shaanxi 710071, PR China
[c] Video and Image Processing System Laboratory, School of Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, PR China
[d] State Key Laboratory of Integrated Services Networks, School of Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, PR China

ARTICLE INFO

Article history:
Received 27 September 2018
Revised 7 December 2018
Accepted 25 January 2019
Available online 28 January 2019

Keywords:
Cross modality
Local descriptor
Composite sketch
ID photo
NIR-VIS matching

ABSTRACT

Cross-modality face recognition aims to identify faces across different modalities, such as matching sketches with photos, low resolution face images with high resolution images, and near infrared images with visual lighting images, which is challenging because of the modality gap caused by texture, resolution, and illumination variations. Existing approaches either utilized hand-crafted approaches which ignore inherent data distribution characteristic, or applied deep learning-based algorithms on holistic face images with facial local information ignored. In this paper, we propose a deep local descriptor learning framework for cross-modality face recognition, which aims to learn discriminant and compact local information directly from raw facial patches. A novel cross-modality enumeration loss is proposed to eliminate the modality gap on local patch level, which is then integrated into a convolutional neural networks for deep local descriptor extraction. The proposed deep local descriptor can be easily applied to any traditional face recognition systems, and we use Fisherface as an example in the paper. Extensive experiments on six widely used cross-modality face recognition datasets demonstrate the superiority of proposed method over state-of-the-art methods.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

Face recognition is an important problem for public security and law enforcement. Despite the great progress in traditional face recognition, the problem of cross-modality face recognition is also crucial in real-world applications. Here cross-modality faces represent face images obtained from different sources, *e.g.*, different lighting condition, resolution, or different sensors. For example, there are many circumstances that the photo of the suspect is not available, and the description of suspect is the only clue. Based on the description about the suspect, law enforcement agencies can invite forensic artist to draw hand-drawn sketches or make composite sketches with software. FBI recently released a hand-drawn forensic sketch of the suspect in their poster[1] for information leading to the recovery of the victim. The law enforcement offices, for example Pasco County Sheriff's Office[2], often release computer generated composite sketches created based on victim's description, which is used to help the public identify the suspect. Traditional face recognition algorithms always fail on this issue because of the shape and texture differences between sketches and photos. There are also situations where low-quality surveillance video about the suspect is available. Matching low resolution faces with high resolution faces is an important yet challenging problem, too. In order to cope with varying lighting conditions and nighttime, near infrared images (NIR) are used because of their robust to lighting environment [1]. Because the NIR image and visible light images (VIS) are captured by different sensors, NIR-VIS face recognition is a popular and challenging task in cross-modality face recognition.

The major challenging problem in cross-modality face recognition is the modality gap between cross-modality face pairs, as illustrated in Fig. 1. A number of researches have been devoted to cope with the aforementioned modality gap, including synthesis based methods, subspace projection based methods, and local feature descriptor based methods. However, most hand-crafted descriptors are manually designed based on human experience, ignoring the inherent data distribution from raw images. The hand-crafted features or metrics may not be optimal to deal with large intra-class variation and small inter-class variation. With the significant progress of deep learning in face recognition research, a number of deep learning based cross-modality face recognition

* Corresponding author.
E-mail addresses: clpeng@xidian.edu.cn (C. Peng), nnwang@xidian.edu.cn (N. Wang), leejie@mail.xidian.edu.cn (J. Li), xbgao@mail.xidian.edu.cn (X. Gao).

[1] https://www.fbi.gov/wanted/kidnap/ruochen-liao.

[2] https://www.wfla.com/news/pasco-county/pasco-sheriff-s-office-releases-composite-sketch-of-possible-kidnapping-attempt-suspect/1623092222.

https://doi.org/10.1016/j.patcog.2019.01.041
0031-3203/© 2019 Elsevier Ltd. All rights reserved.

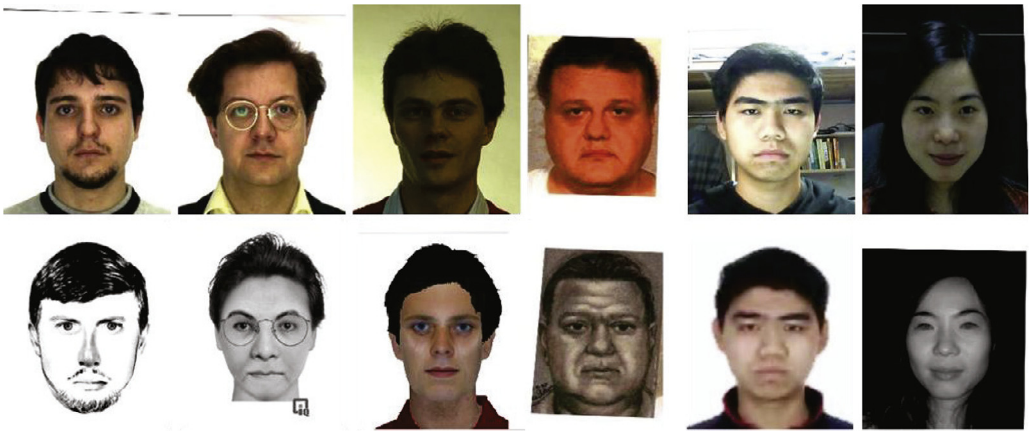


Fig. 1. Illustration of challenges and complexities in cross-modality face recognition. The first three columns show photos with corresponding composite sketches generated by difference software [2–4], where the composite sketches contain shape deformation and are lack of detail textures. The fourth column shows a pair of mug shot photo and forensic sketch [5], where the forensic sketch possesses shape and texture variations. The fifth column demonstrates a pair of high resolution camera image and low resolution ID card image [6], where the modality gap lies in the huge resolution discrepancy. The last column shows a pairs of NIR-VIS images [7], where the modality gap originates from different image acquisition sensors.

techniques are proposed, which can learn deep features directly from raw data and achieve promising results on multiple cross-modality scenarios. However, existing deep learning based approaches suffer the same drawback that the holistic face images are taken as the input. Therefore, the local discriminant information in face images is ignored.

Motivated by the great success of local feature descriptor based methods among conventional approaches, it is natural to ask: can we design a deep local descriptor which not only possesses the ability of learning directly from raw data with deep neural networks, but also is robust and compact to describe discriminant local information in cross-modality faces?

In this paper, we propose a Deep Local descriptor for cross-modality face recognition, refer to as DLFace. The proposed DLFace can directly learn deep local descriptor from raw local facial patches, which is not only modality invariant but also compact and robust to different cross-modality scenarios. Firstly, multiple local image patches are extracted around the facial landmarks detected by face alignment algorithm. Secondly, these local facial patches are fed into convolutional neural networks for deep local descriptor learning. In order to better minimize the distance of intra-class cross-modality local patches, in the meanwhile maximize the distance of inter-class both cross-modality and same-modality local patches, we propose an *enumeration loss* function and an *EnumerateNet* network in this paper. The input of EnumerateNet is a set of cross-modality local patches in a batch, and the enumeration loss is to ensure intra-class cross-modality variation not only smaller than inter-class cross-modality variation but also inter-class same-modality variation. The inspiration of proposed enumeration loss is that the modality gap on local patch level should be eliminated by our convolutional neural networks. We further add a compactness constraint into loss function to avoid overfitting and reduce redundancy of the output deep local descriptor from EnumerateNet. Finally, the obtained deep local descriptor can be applied to any traditional face recognition systems (famous Fisherface used in this paper). The proposed DLFace is evaluated on six popular benchmarks. Experimental results show that it significantly outperforms current state-of-the-art in cross-modality face recognition.

The contributions of this paper are:

1. We develop a deep local descriptor learning framework, namely DLFace, for cross-modality face recognition, which can learn discriminant local information from raw image patches. To the best of our knowledge, our work is the first to consider deep local descriptor for cross-modality face recognition.
2. We propose an enumeration loss function to eliminate modality gap on local patch level. The proposed loss function together with compactness constraint are integrated into a convolutional neural network named EnumerateNet.
3. Extensive cross-modality face recognition experiments show that DLFace outperforms state-of-the-art methods, which demonstrate the effectiveness of our method.

The rest of this paper is organized as follows. In Section 2, we briefly review related works on cross-modality face recognition. Then in Section 3, we describe our DLFace framework in details and evaluate it with substantial experiments in Section 4. Section 5 concludes the paper.

2. Related work

In this section, we briefly review important works related to cross-modality face recognition, including traditional hand-crafted approaches and deep learning based approaches.

2.1. Hand-crafted approaches

Conventional cross-modal face recognition approaches consist of three categories: synthesis based methods, subspace projection based methods, and local feature descriptor based methods.

Synthesis based methods are usually devoted to transforming cross-modal faces into the same modality, where common face recognition algorithms can be applied directly. Tang et al. [8] firstly proposed to use eigen-transformation for whole face synthesis between photos and hand-drawn sketches. Liu et al. [9] and Chen et al. [10] later extended cross-modal face synthesis to local patch level by using locally linear embedding theory. Wang et al. [11] proposed to transform heterogeneous images via sparse feature selection and support vector regression, and recently designed a fast cross-modal face synthesis algorithm [12] based on random sampling strategy. In order to take the relationship among neighboring local patches into consideration, a series of Markov networks based approaches [13–18] were proposed then. A more systematic interpretation about synthesis based cross-modal face recognition can be found in [19].

Subspace projection based methods attempt to find a common subspace where the modality gap is minimized. It began with a common discriminant feature extraction method proposed by Lin et al. [20]. A number of regression based methods [21–23] were later designed to learn mapping function for cross-modal face pairs. The inter-class and intra-class correlations of cross-modal faces were then exploited by Kan et al. [24,25], which can contribute to the process of discriminant metric learning. Huo et al. [6,26,27] proposed cross-modal metric learning techniques to minimize intrapersonal cross-modality distances while forcing a margin for interpersonal cross-modality distances.

Local feature descriptor based methods focus on encoding cross-modal face images with local descriptors, which are then used for recognition. Liao et al. [28] firstly utilized different of Gaussian filter and multi-block local binary patterns for NIR and VIS matching. Later other local descriptors, such as scale invariant feature transform (SIFT) feature, multi-scale local binary patterns (MLBP) feature, local radon binary pattern (LRBP), multi-scale circular Weber's local descriptor (MCWLD) were exploited on cross-modal face recognition [29–32]. Roy et al. [33] presented a local maximum quotient pattern to capture modality invariant facial features. Except for the hand-crafted local descriptors, there are also several learning-based local features [5,34] proposed recently, where the feature learning frameworks were carefully hand-crafted for cross-modal scenarios. Peng et al. [35] extended face recognition from single sketch to multiple stylistic sketches, where more than two modalities are considered. Yu et al. [36,37] explored multiview representations to cope with cross-modality problems.

The hand-crafted approaches mentioned above still suffer from several drawbacks. The performance of synthesis based methods depend heavily on the quality of synthesized images. The mapping procedure in subspace projection based methods may lose discriminative information for recognition. Although the local feature descriptor based methods can achieve relatively higher accuracies than other kinds of conventional methods, their performance still heavily relies on human experience when designing these hand-crafted features, ignoring the inherent data distribution characteristic. Therefore, the hand-crafted features or metrics can be not optimal to deal with large intra-class variation and small inter-class variation. With the help of deep learning, we can learn the inherent distribution character and image features directly from raw data.

2.2. Deep learning based approaches

Deep learning was introduced into cross-modal face recognition scenario [4,38,39] since 2015. Zhang et al. [39] proposed an end-to-end network for cross-modal face synthesis. Jiao et al. [40] later proposed a modified convolutional neural network for cross-modal face synthesis problem. Yi et al. [38] firstly utilized Restricted Boltzmann Machines (RBMs) to learn a shared representation on common subspace. Considering the lack of training data in composite sketch recognition, Mittal et al. [4] designed a transfer learning approach. Liu *et al.* [41] later combined transfer learning with triplet loss to generate deep representation for NIR-VIS face recognition. Sarfraz et al. [42] conducted deep perceptual mapping on densely computed hand-crafted features from cross-modal faces. With the growing attentions on generative adversarial networks (GAN), researchers also applied GAN on cross-modal face synthesis and recognition [43,44]. Wang et al. [45] designed an effective postprocessing method for GAN based cross-modal face synthesis. Neural style transfer and heterogeneous knowledge transfer are also conceivable paths to cope with cross-modality problem. Gatys et al. [46] proposed to use image representations from convolutional neural networks to produce new images combining the content from source photo and the style in the artworks. Isola et al. [47] proposed to perform image-to-image translation via a conditional GAN. Zhu et al. [48] attempted to perform unpaired image-to-image translation by introducing cycle consistency loss into GAN. Heterogeneous knowledge transfer [49,50] usually aims to transform the features from different modalities into a common subspace, so that the difference between heterogeneous modalities is reduced. These neural style transfer and heterogeneous transfer techniques are also possible solutions for cross-modality face recognition.Recently He et al. [51–53] proposed a series of deep learning methods for NIR-VIS matching problem. Galea et al. [2] proposed a very deep convolutional neural network for software-generated composite sketch recognition, together with a newly extended large scale composite sketch database published online. Lin et al. [54] incorporated similarity measure matrix into deep architecture which enabling an end-to-end way of model optimization. Besides the aforementioned cross-modality face recognition scenarios, there are other approaches dealing with cross-modality face recognition situations such as matching clear still face image with blurred video [55], and profile faces with frontal faces [56].

Despite the great progress of deep learning based cross-modal face recognition, existing approaches suffer a common drawback: the deep representation is always learned from the holistic face, with facial local information ignored. *Therefore, our method is inspired, on one hand, by the former success of hand-crafted local descriptors on representing the details of cross-modal face images. On the other hand, motivated by the recent progress of convolutional neural networks on holistic face representation, we aim to design a deep local descriptor learning network, which can learn robust and compact deep local descriptors for cross-modal face recognition.*

3. Methodology

3.1. Problem formulation

Without losing generality and for ease of representation, we take face sketch-photo recognition as an example to introduce our method. We assume that there is a collection of M sketch-photo pairs, denoted as $\{(\mathbf{s}^1, \mathbf{p}^1), \ldots, (\mathbf{s}^M, \mathbf{p}^M)\}$, where $\mathbf{s}^m$ is a face sketch image and $\mathbf{p}^m$ is a face photo image, $m = 1, 2, \ldots, M$. As the face sketch $\mathbf{s}^m$ and face photo $\mathbf{p}^m$ are generated from different sources, they have different texture statistical properties and follow different distributions. Thus, they cannot be directly compared for cross-modality face recognition. We aim to find an embedding $f(x)$, from an input image x to a feature space where sketches and photos can be directly comparable, independent of image modality. Considering the complexity of face structure, it is easier to finish the above goal on local patch level. Specific to our DLFace method proposed in this paper, we strive to learn the embedding function with deep convolutional networks. The most important part lies in the loss function, which can directly reflect the objectives of our model. We will explain in the following sections about our proposed loss function and the architecture of our DLFace.

3.2. Cross-modality enumeration loss

In this section, we present a new cross-modality enumeration loss, which is motivated by the commonly used triplet loss in face recognition [57]. In triplet loss based networks, the inputs are usually a series of tuplets $\{x_i^a, x_i^p, x_i^n\}$, where x_i^a represents an anchor image, and x_i^p and x_i^n represent positive and negative sample images respectively. The triplet loss is to ensure that x_i^a is closer to x_i^p than x_i^n. Suppose there are N tuples selected for training, the triplet loss function can be defined as follows:

$$L = \sum_{i}^{N} [\, \| f(x_i^a) - f(x_i^p) \|_2^2 - \| f(x_i^a) - f(x_i^n) \|_2^2 + \alpha]_+, \qquad (1)$$

where $[z]_+ = \max(z, 0)$, and α is a manually defined minimum margin between positive samples and negative samples. There are two crucial issues in triplet loss: the triplet selection and margin setting. Existing methods [2,57] usually use an online hard positive/negative picking strategy for triplet selection and the margin α is defined manually.

However, directly applying triplet loss to cross-modality face recognition scenario brings the following dilemma: how to choose reasonable triplets in cross-modality scenario. After dividing face images into patches, we suppose that there are N pairs of face sketch-photo patches $\{(\mathbf{y}_1, \mathbf{x}_1), (\mathbf{y}_2, \mathbf{x}_2), \dots, (\mathbf{y}_N, \mathbf{x}_N)\}$ in a batch for network training, where $\mathbf{y}_i$ and $\mathbf{x}_i$ represent the sketch-photo patches from the same identity, $i = 1, 2, \dots, N$. For an anchor sketch patch $\mathbf{y}_i$, the corresponding positive sample should be the photo patch $\mathbf{x}_i$. But for the negative sample, there are photo patches $\mathbf{x}_j$ and sketch patches $\mathbf{y}_j$ in different modalities, where $j \neq i$. A straightforward way is to select the negative sample from photo patches $\mathbf{x}_j$, which can ensure that anchor sketch patch $\mathbf{y}_i$ is closer to photo patch $\mathbf{x}_i$ than other photo patch $\mathbf{x}_j$ of different identity. In this way, the triplet loss in cross-modality situation is defined as follows:

$$L = \sum_{i}^{N} [\|f(\mathbf{y}_i) - f(\mathbf{x}_i)\|_2^2 - \|f(\mathbf{y}_i) - f(\mathbf{x}_j)\|_2^2 + \alpha]_+, \qquad (2)$$

where $j \neq i$.

Inspired by the fact that modality gap can be smaller on local patch level and traditional hand-crafted local descriptors used to focus on reducing the modality gap, we propose a novel cross-modality enumeration loss in this paper. The purpose of the proposed loss function is to totally eliminate the modality gap on local patch level with the help of deep convolutional network. Specifically, the enumeration loss aims to make the anchor sketch patch $\mathbf{y}_i$ is closer to photo patch $\mathbf{x}_i$ than not only other photo patch $\mathbf{x}_j$, but also other sketch patch $\mathbf{y}_j$. In this way, the embedding distance of $\mathbf{y}_i$ with $\mathbf{x}_i$ should be smaller than $\mathbf{y}_i$ with any other photo or sketch patches in the batch, *i.e. enumerate all the combinations in a batch and the anchor with the positive sample should be closest*.

The cross-modality enumeration loss function is defined as below. We firstly consider the inter-class cross-modality constraint L_c, which is similar with a straightforward triplet loss. The rationale is that intra-class cross-modality variation should be smaller than inter-class cross-modality variation:

$$L_c = \sum_{i}^{N} \sum_{j \neq i}^{N} [\|f(\mathbf{y}_i) - f(\mathbf{x}_i)\|_2^2 - \|f(\mathbf{y}_i) - f(\mathbf{x}_j)\|_2^2]_+ \qquad (3)$$

Then we introduce the inter-class same-modality constraint L_s, which is designed to help eliminate the modality gap on image patch level. The rationale is that intra-class cross-modality variation should also be smaller than inter-class same-modality variation:

$$L_s = \sum_{i}^{N} \sum_{j \neq i}^{N} [\|f(\mathbf{y}_i) - f(\mathbf{x}_i)\|_2^2 - \|f(\mathbf{y}_i) - f(\mathbf{y}_j)\|_2^2]_+ \qquad (4)$$

The above two constraints could help train our deep local descriptor while ignoring the modality gap during the training phase. However, in our experiments it is difficult to converge based on the combination of these two loss functions. We further introduce a compactness term C in our proposed Enumeration loss, which ensures that each dimension in the generated deep local descriptor evenly distributed as much as possible, so that the obtained descriptor is more compact and informative. This compactness term is defined by calculating the differences between each element in $f(\mathbf{y}_i)$ and its mean value:

$$C = \sum_{i}^{N} \sum_{r}^{R} \| |f_r(\mathbf{y}_i)| - \bar{f}(\mathbf{y}_i)\|_2^2 \qquad (5)$$

Here $f_r(\mathbf{y}_i)$ is the rth element in $f(\mathbf{y}_i)$, and $\bar{f}(\mathbf{y}_i)$ represents the mean value of all elements in $f(\mathbf{y}_i)$. R is the dimension of the output deep local descriptor $f(\mathbf{y}_i)$. The compactness constraint is commonly used in binary descriptor learning [58]. The purpose of introducing the compactness term here is to help avoid overfitting during our network training. In our experiments, the network is difficult to converge without the compactness term in our enumeration loss. Meanwhile, the compactness term can also help reduce redundancy, thus make our deep local descriptor discriminative and informative.

Table 1
The architecture of EnumerateNet. The output size is based on height × width × depth and #Params refer to the number of parameters in the network. Conv: convolution layer, BN: batch normalization, ReLU: rectified linear unit activation.

Layer name	Filter size, stride, pad	Output size	#Params
Upsampling	–	32 × 32 × 1	–
Conv	3 × 3,1,1	32 × 32 × 32	0.2 K
BN+ReLU	–	32 × 32 × 32	–
Conv	3 × 3,1,1	32 × 32 × 32	9 K
BN+ReLU	–	32 × 32 × 32	–
Conv	3 × 3,2,1	16 × 16 × 64	18 K
BN+ReLU	–	16 × 16 × 64	–
Conv	3 × 3,1,1	16 × 16 × 64	36 K
BN+ReLU	–	16 × 16 × 64	–
Conv	3 × 3,2,1	8 × 8 × 128	73 K
BN+ReLU	–	8 × 8 × 128	–
Conv	3 × 3,1,1	8 × 8 × 128	147 K
BN+ReLU	–	8 × 8 × 128	–
Conv	8 × 8,1,0	1 × 1 × 128	1,048 K
BN	–	1 × 1 × 128	–
Total	–	–	1,331 K

Finally, our Enumeration loss function is defined by combining them together:

$$L = L_c + L_s + \lambda C \qquad (6)$$

Here λ is used to balance the effect of the last compactness constraint in our loss.

A visualization comparison between triplet loss and proposed enumeration loss is shown in Fig. 2. Instead of taking triplets as the input, our enumeration loss considers all the combination with the anchor in a batch. Therefore, the enumeration loss do not need to worry about carefully selecting hard positive or negative samples, and all the positive and negative samples in a batch can be taken into consideration. Furthermore, our enumeration loss do not need specific margin setting, which can avoid the drawbacks of using bad values of margin in triplet loss. By the way, because our enumeration loss is designed for local patch feature learning, the size of input patches are small which makes the requirements of computation and memory relatively acceptable.

3.3. Framework of DLFace

The architecture of deep convolutional network used in this paper, namely EnumerateNet, in shown in Table 1. The EnumerateNet contains one upsampling layer, seven convolution layers, seven batch normalization layers and six rectified linear unit (ReLU) activation layers. Because the size of input face patches are small, we add an upsampling layer at the beginning of EnumerateNet for ease of network training. There are about 1,331 K parameters in the network and the dimension of output descriptor is 128. In our experiments we observed that our networks performs better without

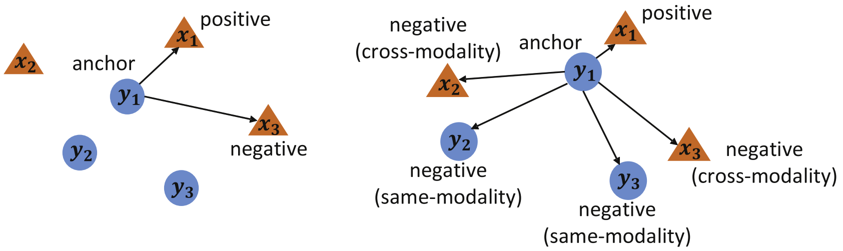


Fig. 2. Triplet loss (left figure) only takes triplets into consideration each time, and proposed enumeration loss (right figure) considers all the combinations in a batch that related to the anchor.

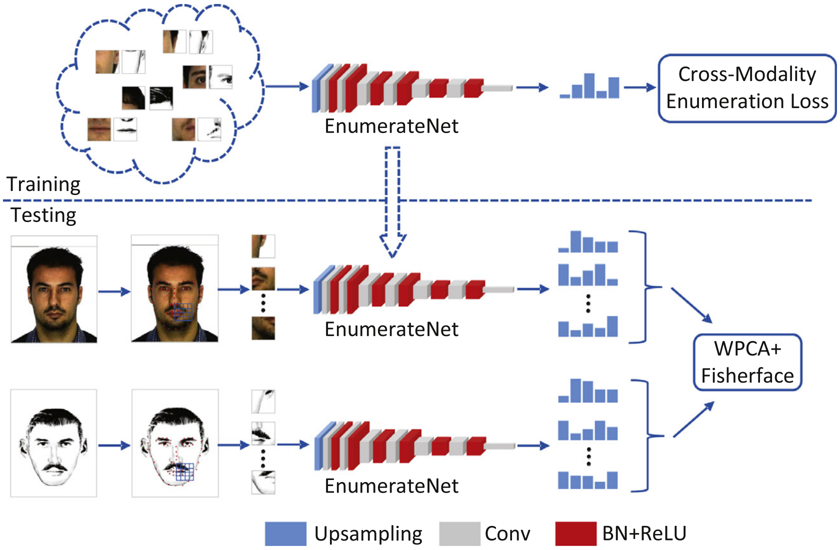


Fig. 3. Framework of proposed DLFace during training and testing procedure.

full connected layers, therefore we only utilize a stack of convolution layer together with batch normalization and ReLU activation in proposed EnumerateNet.

The framework of proposed DLFace for EnumerateNet training and testing procedure is shown in Fig. 3. Training data contains a number of face sketch patches with corresponding face photo patches, which are randomly extracted from face sketch-photo dataset. The network is firstly trained on a large-scale patch descriptor learning dataset, Brown dataset [59], with standard triplet loss. Based on the pre-trained model, we implement the enumeration loss on our face sketch-photo patches data.

In testing procedure, the EnumerateNet can be regarded as a deep local descriptor extractor, and can be applied to any face recognition frameworks. In this paper, we present a simple framework of extracting features around facial landmarks and utilize famous Fisherface [60] for recognition. Benefited from recent progress in face alignment algorithms that can accurately locate dense facial landmarks in real-time, we adopt [61] to automatically detect 66 landmarks around facial components. Features extracted around landmarks mainly focus on the description of local facial components, which is robust to pose, expression, and occlusion variations. Besides these facial landmarks, we further use a set of dense holistic keypoints to obtain a holistic-level description, which is complementary to component-level description. Thus, fusion of the two can help improve the performance. For each landmark or keypoint, we extract a grid of local patches around it, and each local patch is encoded to deep local descriptor by EnumerateNet. Finally, all the obtained deep local descriptors are concatenated to represent the face sketch or photo image. We apply whitened principal component analysis (WPCA) to the concatenated descriptors first, and standard Fisherface is utilized for recognition then. In our experiments, we find the output of the last convolution layer performs better than after batch normalization. The implementation details of our DLFace will be introduced in the experimental section.

4. Experiments

We conduct experiments on six widely used cross-modality face recognition datasets: PRIP-VSGC dataset [3], e-PRIP dataset [4], UoM-SGFS dataset [2], forensic sketch dataset [5], NJU-ID dataset [6], and CASIA NIR-VIS 2.0 dataset [7]. The first three datasets include composite sketch-photo pairs, where the composite sketches are created by software. The forensic sketch dataset contains mug shot photos and forensic sketches collected from real-world criminal cases. The NJU-ID dataset is used for evaluating face recognition across different resolutions, while the CASIA NIR-VIS 2.0 dataset is evaluated for matching near infrared images (NIR) with visible light images (VIS). Examples of the datasets used in this paper are shown in Fig. 1.

4.1. Datasets

PRIP-VSGC dataset: The PRIP-VSGC dataset includes 123 subjects. There are one composite sketch created using IdentiKit with one corresponding photo from the AR dataset [62]. To evaluate on this dataset, we randomly divide it into two splits, with 48 subjects for training and 75 subjects for testing. In order to better mimic real-world cross-modality face recognition scenario, we further enlarge the gallery with 10,000 photos following the same protocol in [5], where the enlarged 10,000 photos are collected from the FERET

Table 2
Performance when using different numbers of local patches around landmarks on PRIP-VSGC dataset (rank−10, %).

Grid	2 × 2	3 × 3	4 × 4	5 × 5	6 × 6
Accuracy	68.91	74.87	76.40	76.73	76.68

dataset (2,722), XM2VTS dataset (1,180), CAS-PEAL dataset (3,098), and LFW dataset (3,000).

e-PRIP dataset: The e-PRIP dataset also includes 123 subjects, with the same photos from the AR dataset as PRIP-VSGC. The composite sketches in e-PRIP are created using another FACES software. We follow the same protocol on PRIP-VSGC by dividing the dataset to 48 subjects for training and 75 subjects for testing. Experiment on enlarged gallery is also conducted.

UoM-SGFS dataset: The newly extended UoM-SGFS dataset contains 600 subjects, with photos obtained from the FERET dataset [63]. There are two sets of composite sketches, where each set contains one sketch for one subject. SetA contains the composite sketches created by EFIT-V software, and SetB contains the composite sketches edited by Corel PaintShop Pro-X7 based on SetA. We follow the standard protocol provided by the dataset [2], with 450 subjects for training and 150 subjects for testing. We also populate the gallery with the enlarged gallery in [5], where the 2722 subjects from FERET are removed here.

Forensic sketch dataset: The forensic sketch dataset contains 168 real-world forensic sketches with corresponding mug shot photos. The forensic sketches are drawn by forensic artist based on the description of eyewitness. Here we follow the same protocol in [5] by taking 112 subjects for training and 56 subjects for testing. The 10,000 enlarged gallery is also applied on this dataset.

NJU-ID dataset: The NJU-ID dataset consists of 256 subjects. For each subject, there is one low-resolution card image with corresponding one high-resolution image from digital camera. We follow the same protocol as [6] by randomly dividing the dataset into 10 folds and 10-fold cross-validation is used.

CASIA NIR-VIS 2.0 dataset: The CASIA NIR-VIS 2.0 dataset is a large-scale NIR-VIS face recognition dataset, with 725 subjects. Each subject has 1–22 VIS images and 5–50 NIR images. There are many-to-many correlations between NIR and VIS. We follow the standard protocol of 10-fold experiments provided by the dataset [7].

4.2. Parameter settings and implementation details

We first introduce the parameters used in this paper. The experiments here are conducted on PRIP-VSGC dataset. All the face images used in this paper are firstly aligned based upon five facial points (centers of two eyes, nose tip, and two corners of mouth). Each face image is then cropped to 200 × 250 based on the five facial points. The grid around each landmark is 4 × 4, and the local patch size is 20 × 20. We conduct experiments by using different numbers of local patches around each landmark for face recognition on the PRIP-VSGC dataset, as shown in Table 2. When we take 2 × 2 or 3 × 3 patches around landmarks, the recognition performance is poor because the local facial character around each landmark is not fully explored. When more local patches are considered around the landmarks, the computational cost will increase. Therefore, we choose 4 × 4 which could achieve rank-10 accuracy of 76.40%. For each input patch, we subtract the pixel values by the mean and divided by the standard deviation before input into the network. Since our method mainly focuses on extracting deep features from local image patch, the size of the local patch is crucial to the performance. We evaluate our DLFace with different local patch size, as shown in left top figure of Fig. 4. With the increase of the local patch size, the discriminative power decreases. We therefore set the size of local patch to 20 × 20. In the recognition phase, the whitened principal component analysis is applied on the obtained features, where 99% of variance is retained. Considering that there are M sketch-photo pairs utilized for training the classifier, $M-1$ eigenvectors will be kept. Then standard Fisherface will be taken as the recognition algorithm.

Table 3
The ablation study for DLFace (%).

Descriptor	PRIP-VSGC Rank-10	NJU-ID Rank-1	CASIA NIR-VIS 2.0 Rank-1
MLBP	60.93	30.46	86.12
SIFT	61.60	30.72	87.02
Triplet loss	64.67	33.87	93.55
Enumeration loss	76.40	43.46	98.68

Robustness of DLFace with inaccurately detected landmarks: Based on our visual observation, most of the facial landmarks can be detected accurately using [61]. Sometimes there may be a little deviation. In order to evaluate the effect of inaccurate detection on facial landmark, we conduct an experiment by randomly move the detected facial landmarks on each image to any direction within 5 pixels. Here the matched patches from two modalities are moved with random direction and pixels. Experimental result of our DLFace with random landmark movement is shown in right top figure of Fig. 4. It can be seen that the performance only decrease a little with the inaccurate landmark detection. The reason is that the generated deep local descriptor in DLFace is informative, which makes our DLFace robust to the landmark detection.

Effectiveness of enumeration loss: We compare proposed loss with the straightforward triplet loss in Eq. (2). Experimental result is shown in left bottom figure of Fig. 4. From the figure we can observe that our proposed Enumeration loss performs better than traditional triplet loss. This improvement originates from two aspects. Firstly, our Enumeration loss takes all the combinations in a batch into consideration, which intends to eliminate the modality gap on image patch level. On the contrary, triplet loss needs to carefully select proper triplets to train the network, which heavily affects its performance. Secondly, we add a compactness term in our proposed loss, which can help avoid overfitting during the training procedure, while reduce redundancy and improve generalization of the obtained deep local descriptor. Therefore, our proposed Enumeration loss performs better than triplet loss, which evaluates the effectiveness of the proposed loss function.

Effectiveness of deep local descriptor: We compare our deep local descriptor with two classical hand-crafted descriptors MLBP and SIFT. Experimental results are shown in right bottom figure of Fig. 4. Hand-crafted features are manually designed based on human experience, which ignore the inherent data distribution characteristic and cannot deal with large intra-class variation and small inter-class variation. Our proposed deep local descriptor directly learn the data distribution characteristic from raw local patches, which is more discriminant and robust to cross-modality face recognition.

Ablation study for DLFace: In order to better illustrate the contribution of proposed loss in our deep local descriptor, we conduct the ablation study on three representative cross-modality face recognition datasets (PRIP-VSGC, NJU-ID, and CASIA NIR-VIS 2.0) as shown in Table 3. We replace the deep local descriptor using cross-modality enumeration loss proposed in our framework with hand-crafted local descriptors as well as deep local descriptor using traditional triplet loss. From the table we can observe that hand-crafted multiscale LBP (MLBP) and SIFT features perform poor on three cross-modality datasets, which results from the hand-crafted procedure ignoring inherent data distribution. We also replace our proposed enumeration loss with standard triplet

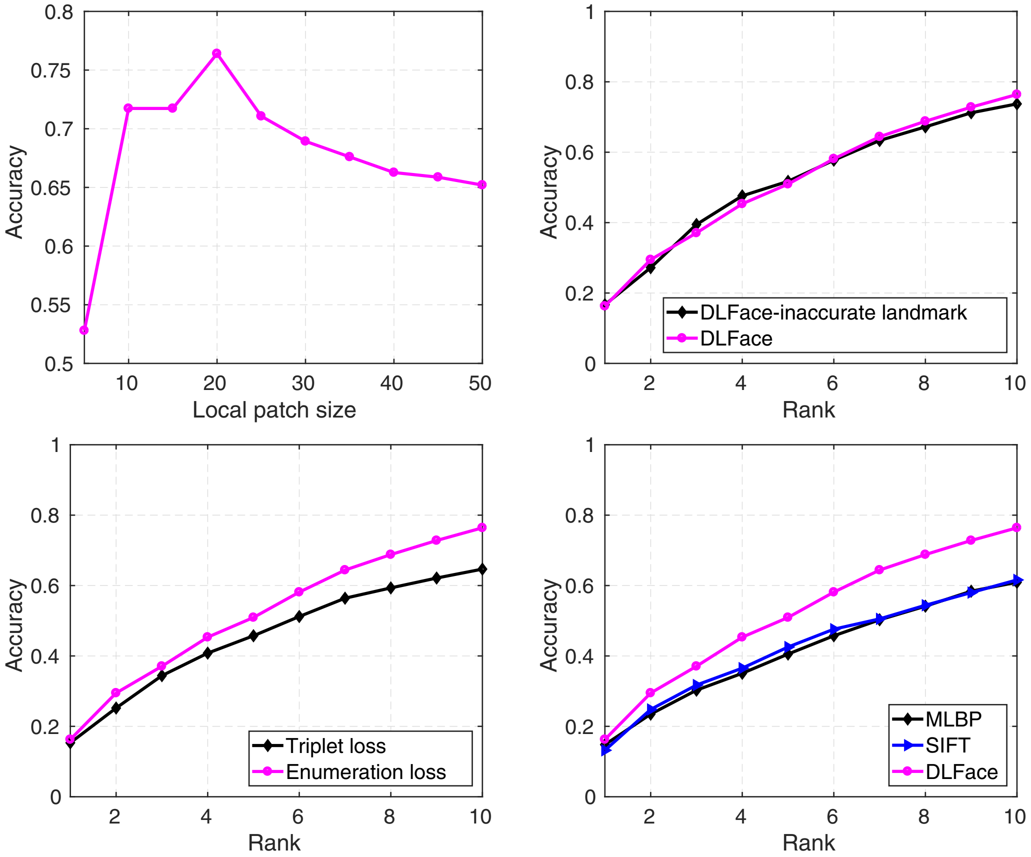


Fig. 4. Left top figure evaluates the effect of local patch size; right top figure evaluates the robustness of DLFace for inaccurate landmark detection; left bottom figure illustrates the effectiveness of proposed enumeration loss compared with triplet loss; right bottom figure illustrates the effectiveness of proposed deep local descriptor compared with classical hand-crafted descriptors.

loss. However, proper triplets are needed to be carefully chosen which can heavily affect the performance. By taking all the combinations in a batch into consideration, our proposed cross-modality enumeration loss achieves best performance on all three datasets. This ablation study clearly shows that our proposed deep local descriptor with enumeration loss is more suitable for cross-modality face recognition.

We train the network using stochastic gradient descent (SGD) algorithm and the training data are augmented with rotation and flipping. Our network starts training with learning rate of 0.01 and the batch size is 1000. The compactness term in the enumeration loss is crucial for our network training. In the experiments we set the weight of this term λ as 0.0001. From our observation, with the increase of weight λ, the discriminative power of deep local descriptor decreases as well as the recognition performance. However, the network is easier to be overfitting if the weight λ is too small. Actually, the network cannot converge if we ignore the compactness term by setting λ to 0.

The most time-consuming part in our method lies in the process of extracting the deep local descriptor. During the training process of the neural network, it is easy to converge with around 50 epochs, which takes about 5 hours. Since the structure of our neural network is simple, it is also very efficient to extract deep local descriptor in testing phase. We compare the extraction time cost for one descriptor with hand-crafted SIFT and MLBP descriptors per patch by averaging 10K patches, and the time comparison is shown in Table 4. The experiment is conducted on an NVIDIA GTX TITAN X GPU in Ubuntu 16.04. It takes 1.0445ms to extract descriptor per patch for our method.

Table 4
Comparison with hand-crafted descriptors for extraction time per patch. (ms).

Method	MLBP	SIFT	Ours
Time	0.7845	0.1029	1.0445

Following we will compare our DLFace with state-of-the-art methods on six cross-modality face recognition datasets. Without specific notification, the experimental results are reported based on 10 random splits.

4.3. Comparison with existing methods

Results on PRIP-VSGC Dataset: We first compare the proposed DLFace with existing methods on PRIP-VSGC dataset, including both traditional hand-crafted methods and deep learning based methods. Hand-crafted MCWLD [29] and SSD [64] perform poor with the best rank-10 accuracy of 45%. The deep learning based approaches TLDNN [4] and CNN [65] can achieve accuracies 52% and 51.5% respectively with the help of transfer learning. The recently proposed DEEPS [2] utilized a very deep CNN model integrated with morphed faces and transfer learning and can achieve the state-of-the-art accuracy of 54.90%. However, all existing deep learning based approaches are performed on holistic faces, ignoring local facial details. Our DLFace achieves the highest rank-10 accuracy of 76.40%, outperforming DEEPS [2] 21.5% as shown in Table 5. Because these existing deep learning based methods [2,4,65] take the whole face as the input to neural networks, local information is ignored. Our DLFace can learn deep local descrip-

Table 5
Comparison with state-of-the-art methods on PRIP-VSGC dataset (rank-10, %).

Method	Accuracy	Method	Accuracy
MCWLD [29]	15.40	SSD [64]	45.30
TLDNN [4]	52.00	CNN [65]	51.50
DEEPS [2]	54.90	**DLFace**	**76.40**

Table 6
Comparison with state-of-the-art methods on e-PRIP dataset (rank-10, %).

Method	Accuracy	Method	Accuracy
MCWLD [29]	24.00	SSD [64]	53.3
TLDNN [4]	60.20	CNN [65]	65.60
DEEPS [2]	80.80	**DLFace**	**82.80**

Table 7
Comparison with state-of-the-art methods on UoM-SGFS dataset SetA (%).

Method	Rank-1	Rank-10	Rank-50
PCA [60]	2.80	8.40	17.73
HAOG [66]	13.60	37.33	52.67
CBR [67]	5.73	18.80	43.33
LGMS [68]	21.87	51.20	72.40
VGG-Face [69]	9.33	31.07	
D-RS [31]	22.13	49.33	69.87
D-RS+CBR [3]	25.87	56.00	76.27
DEEPS [2]	31.60	66.13	86.00
DLFace	**64.80**	**92.13**	**97.60**

Table 8
Comparison with state-of-the-art methods on UoM-SGFS dataset SetB (%).

Method	Rank-1	Rank-10	Rank-50
PCA [60]	5.33	9.87	18.67
HAOG [66]	21.60	42.27	57.07
CBR [67]	7.60	25.47	48.27
LGMS [68]	43.47	73.60	86.93
VGG-Face [69]	16.13	48.00	72.80
D-RS [31]	40.80	70.80	86.40
D-RS+CBR [3]	42.93	75.87	90.13
DEEPS [2]	52.17	82.67	94.00
DLFace	**72.53**	**94.80**	**98.80**

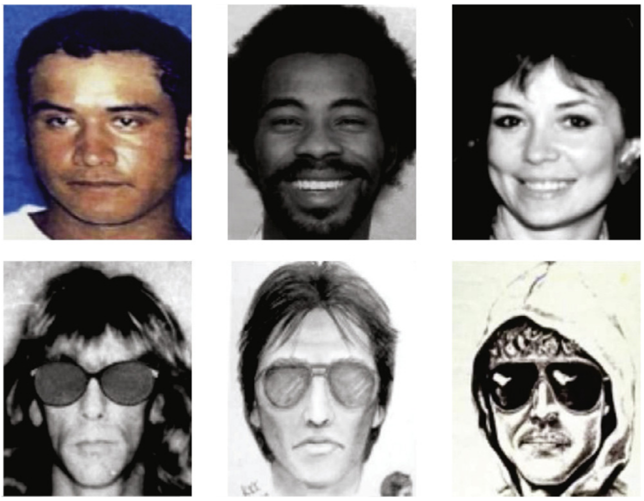


Fig. 5. Example images with different pose, expression, and occlusions evaluated in forensic sketch dataset.

tor on facial patches, thus achieves the highest performance. We further conduct the experiment with enlarged gallery. Peng et al. [5] proposed a graphical representation and achieved rank-50 of 51.22% on this dataset. Our DLFace can achieve rank-50 of 58.93% with the enlarged gallery.

Results on e-PRIP Dataset: We also compare our DLFace with existing methods on another composite sketch dataset, *i.e.* e-PRIP dataset. As shown in Table 6, the hand-crafted MCWLD [29] and SSD [64] achieve rank-10 accuracies of 24% and 53.3% respectively. The transfer learning based TLDNN can achieve a rank-10 accuracy of 60.2% while the CNN based method achieves rank-10 accuracy of 65.6%. The state-of-the-art DEEPS technique achieves rank-10 accuracy of 80.80% on this challenging dataset. With the help of the proposed deep local description, our DLFace can also outperform existing methods including deep learning based approaches by achieving rank-10 of 82.80%. Furthermore, we conduct the experiment with the enlarged gallery and can achieve rank-50 of 70.00% on e-PRIP dataset.

Results on UoM-SGFS Dataset: Compared with previous PRIP-VSGC and e-PRIP which contains 123 subjects, the UoM-SGFS is a large scale composite sketch dataset including 600 subjects and 1200 composite sketches. There are two sets of sketches in this dataset. SetA contains sketches created by software, and SetB contains sketches from SetA that are lightly altered to make them more realistic. Therefore, accuracies on SetB are usually higher. In this paper we follow similar protocol in [2] and randomly choose 1521 photos to populate the gallery. We demonstrate rank-1, rank-10, and rank-50 accuracies of existing methods as well as our proposed DLFace in Table 7 for Set A and Table 8 for Set B respectively. From the table we can observe that the traditional PCA [60], HAOG [66], and LGMS [68] perform poor on this large scale composite sketch dataset. The CBR [67] method considers the representation of facial components, but hand-crafted local descriptors are utilized in CBR [67] and only achieves rank-50 of 43.33% and 48.27% on SetA and SetB respectively. The baseline deep learning based VGG-Face [69] cannot fully extract valuable identity information from composite sketches. By combining facial component extraction with hand-crafted D-RS [31] method, rank-50 accuracies of 76.26% and 90.13% are achieved by [3]. The state-of-the-art DEEPS method is designed specifically for matching composite sketches to face photographs and rank-50 accuracies of 86% and 94% are achieved. Our DLFace extracts facial local information with the proposed enumeration loss based deep local descriptor, and can achieve rank-50 accuracyes of 97.6% and 98.8% respectively, which are higher than state-of-the-art methods.

Results on forensic sketch dataset: In order to mimic real-world law enforcement scenarios, we also evaluate our algorithm on forensic sketch dataset [5] where the forensic sketches and mug shot photos are collected from real-world cases. There is varying face perception bias during the creation of forensic sketches, and the mug shot photos are collected from different sources with varying resolutions. There are also severe lighting, pose, expression, background and occlusion variations in this dataset. A few example images from the forensic sketch dataset are shown in Fig. 5. Experiments on this challenging dataset can help illustrate the superiority of our DLFace technique. Experimental comparison of our DLFace with existing methods on rank-10 and rank-50 accuracies are shown in Table 9. Two baseline PCA [60] and Fisherface [60] methods achieve rank-50 accuracies of 12.68% and 17.14% respectively. Two state-of-the-art methods, D-RS method [31] and graphical representation based method [5], report their results on these forensic sketches with rank-50 accuracies of 20.80% and 31.96%. Note that the images in this forensic datasets are collected from real-world law enforcement scenarios, and only several papers reported their performance on this dataset. The proposed DL-

Table 9
Comparison with state-of-the-art methods on forensic sketch dataset (%).

Method	Rank-10	Rank-50
PCA [60]	6.07	12.68
Fisherface [60]	10.71	17.14
D-RS [31]	11.21	20.80
G-HFR [5]	18.21	31.96
DLFace	**40.73**	**57.64**

Table 10
Comparison with state-of-the-art methods on NJU-ID dataset (rank-1, %).

Method	Accuracy	Method	Accuracy
PCA [60]	23.00	CSR [22]	29.30
CCA [70]	20.30	CDFE [20]	24.60
MvDA [24]	16.50	CMML [25]	20.30
HMLCR [71]	27.70	ESPAC [6]	20.80
CML [27]	30.90	**DLFace**	**43.46**

Table 11
Comparison with state-of-the-art methods on CASIA NIR-VIS 2.0 dataset (rank-1, %).

Method	Accuracy		
CSR [22]	33.80	D-RS [31]	28.20
KDSR [72]	37.50	PCA+HCA [7]	23.70
LCFS [73]	35.40	H2(LBP3) [74]	43.80
C-DFD [75]	65.80	CDFL [75]	71.50
Gabor+RBM [38]	86.20	Recon.+UDP [76]	78.50
CEFD [77]	85.60	TRIVET [41]	95.70
VGG-Face [69]	62.09	CNN [65]	85.90
IDNet [78]	87.10	SeetaFace [79]	68.03
CenterLoss [80]	87.69	Lightened CNN [81]	91.88
IDR-128 [51]	97.33	**DLFace**	**98.68**

Face can achieve a rank-50 accuracy of 57.64%, which is almost twice higher than current state-of-the-art method.

Results on NJU-ID dataset: Besides face sketch-photo recognition scenarios, we also evaluate our DLFace on cross-resolution face recognition problem. The NJU-ID dataset contains low resolution face images from ID cards and high resolution faces from digital camera. There is server blurring affect on low resolution face images, and the lighting conditions are varying. Existing methods perform poor on this challenging dataset. We compare our method with several state-of-the-arts on ID photo recognition as shown in Table 10. The baseline PCA [60], CSR [22], and CCA [70] methods perform poor on ID photo recognition problem. The subspace projection based methods CDFE [20] and MvDA [24] can achieve rank-1 accuracies of 24.6% and 16.5%. The accuracies of metric learning based methods CMML [25], HMLCR [71], and ESPAC [6] are lower than 30%. The state-of-the-art CML method [27] can achieve a rank-1 accuracy of 30.90% on this dataset. Because our method utilizes deep local descriptor to capture facial details, the proposed DLFace is robust to varying cross-modality matching scenarios. Our DLFace achieves a rank-1 accuracy of 43.46% which is the highest accuracy on this dataset.

Results on CASIA NIR-VIS 2.0 Dataset: In order to evaluate the robustness and effectiveness of the proposed DLFace, we further conduct experiments of matching NIR images with VIS faces on CASIA NIR-VIS 2.0 dataset. This is a challenging large scale NIR-VIS face recognition dataset, with totally 725 subjects. Table 11 shows the rank-1 accuracy of different methods on this large dataset. Traditional methods including CSR [22], D-RS [31], KDSR [72], PCA+HCA [7], LCFS [73] and H2(LBP3) [74] achieve rank-1 accuracies lower than 50% on this dataset. The baseline deep learning based approach CNN [65] can achieve 85.9% here. The state-of-the-art IDR-128 [51] learns invariant deep representations for NIR-VIS matching and achieves a rank-1 accuracy of 97.33%. However, the invariant deep representation in [51] is still learnt from holistic face images. In this paper we propose to learn deep representations from local facial patches which can better capture detail information for recognition. Our DLFace achieves a rank-1 accuracy of 98.68%, which is comparable to state-of-the-art methods on NIR-VIS matching. All of the experimental results shown in this section suggest that our DLFace is effective for various cross-modality face recognition problems, thus demonstrates the superiority of the proposed method in this paper.

5. Conclusion

A deep local descriptor is proposed in this paper for cross-modality face recognition. We propose an EnumerateNet network integrating an enumeration loss function to generate deep local descriptor from local image patches. Instead of selecting negative samples for network training, we enumerate all the combinations in a batch and ensure the anchor with the positive sample to be closet. We further incorporate a compactness constraint term in our enumeration loss to avoid overfitting as well as reduce redundancy of the generated deep local descriptor. Our proposed deep local descriptor is easy to use in traditional face recognition systems, and we provide the performance of our descriptor applied in Fisherface algorithm. Experiments are conducted on multiple cross-modality face recognition scenarios, including composite sketch-photo recognition, forensic sketch-photo recognition, low-high resolution face recognition, and NIR-VIS face recognition, and our method could outperform state-of-the-art methods which validates the effectiveness of this paper.

The usage of our proposed deep local descriptor is easy by simply replacing the role of traditional hand-crafted descriptor such as SIFT with our deep local descriptor in the face recognition systems. It is also possible to apply our deep local descriptor to other tasks such as image retrieval and person re-identification. However, there are still weaknesses in our method which could be improved in the future. For example, we only utilize a simple convolutional neural network in our method. It is promising to further improve the ability of the deep local descriptor if the network can be improved with more advanced architectures including recurrent neural network and residual neural network. On the other hand, the training process in our method ignores the inherent facial component information, which is valuable for face recognition. In the future we will explore training our EnumerateNet on facial components respectively to further improve the performance. Furthermore, extending our deep local descriptor based cross-modality face recognition framework by metric learning algorithms is also worth investigating in the future.

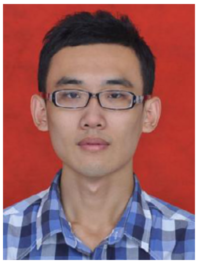
Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61806152, 61876142, 61432014, U1605252, 61772402, 61501339, 61671339, and 61702394, in part by the National Key Research and Development Program of China under Grant 2016QY01W0200, in part by Key Industrial Innovation Chain in Industrial Domain under Grant 2016KTZDGY04-02, in part by National High-Level Talents Special Support Program of China under Grant CS31117200001, in part by Young Elite Scientists Sponsorship Program by CAST (under Grant 2016QNRC001), in part by Natural Science Basic Research Plan in Shaanxi Province of China (under Grant 2017JM6085 and 2017JQ6007), in part by Young Talent fund of University Association for Science and Technology in Shaanxi, China, in part by the Fundamental Research Funds for the Central Universities under Grant XJS17086, in part by CCF-Tencent Open Fund, in part by China 111 Project (Grant No. B16037), in part by the China PostDoctoral Science Foundation under Grant 2018M631124, and in part by the Xidian University-Intellifusion Joint Innovation Laboratory of Artificial Intelligence.

References

[1] R. Ghiass, O. Arandjelovic, A. Bendada, X. Maldague, Infrared face recognition: a comprehensive review of methodologies and databases, Pattern Recognit. 47 (9) (2014) 2807–2824.
[2] C. Galea, R. Farrugia, Matching software-generated sketches to face photographs with a very deep CNN, morphed faces, and transfer learning, IEEE Trans. Inf. Forens. Secur. 13 (6) (2018) 1421–1431.
[3] S. Klum, H. Han, B. Klare, A.K. Jain, The facesketchID system: matching facial composites to mugshots, IEEE Trans. Neural Netw. Learn. Syst. 9 (12) (2014) 2248–2263.
[4] M. Paritosh, M. Vasta, R. Singh, Composite sketch recognition via deep network-a transfer learning approach, in: Proceedings of the IAPR International Conference on Biometrics, 2015, pp. 251–256.
[5] C. Peng, X. Gao, N. Wang, J. Li, Graphical representation for heterogeneous face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2) (2017) 301–312.
[6] J. Huo, Y. Gao, Y. Shi, W. Yang, H. Yin, Ensemble of sparse cross-modal metrics for heterogeneous face recognition, in: Proceedings of the ACM International Conference on Multimedia, 2016, pp. 1405–1414.
[7] S. Li, D. Yi, Z. Lei, S. Liao, The CASIA NIR-VIS 2.0 face database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2013, pp. 348–353.
[8] X. Tang, X. Wang, Face sketch synthesis and recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2003, pp. 687–694.
[9] Q. Liu, X. Tang, H. Jin, H. Lu, S. Ma, A nonlinear approach for face sketch synthesis and recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 1005–1010.
[10] J. Chen, D. Yi, J. Yang, G. Zhao, S. Li, M. Pietikainen, Learning mappings for face synthesis from near infrared to visual light images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 156–163.
[11] N. Wang, J. Li, D. Tao, X. Li, X. Gao, Heterogeneous image transformation, Pattern Recognit. Lett. 34 (1) (2013) 77–84.
[12] N. Wang, X. Gao, J. Li, Random sampling for fast face sketch synthesis, Pattern Recognit 76 (2018) 215–227.
[13] C. Peng, X. Gao, N. Wang, D. Tao, X. Li, J. Li, Multiple representations-based face sketch-photo synthesis, IEEE Trans. Neural Netw. Learn. Syst. 27 (11) (2016) 2201–2215.
[14] N. Wang, D. Tao, X. Gao, X. Li, J. Li, Transductive face sketch-photo synthesis, IEEE Trans. Neural Netw. Learn. Syst. 24 (9) (2013) 1364–1376.
[15] X. Wang, X. Tang, Face photo-sketch synthesis and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (11) (2009) 1955–1967.
[16] H. Zhou, Z. Kuang, K. Wong, Markov weight fields for face sketch synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1091–1097.
[17] N. Wang, M. Zhu, J. Li, B. Song, Z. Li, Data-driven vs. model-driven: fast face sketch synthesis, Neurocomputing 257 (2017) 214–221.
[18] N. Wang, X. Gao, L. Sun, J. Li, Anchored neighborhood index for face sketch synthesis, IEEE Trans. Circuits Syst. Video Technol. 28 (9) (2018) 2154–2163.
[19] N. Wang, X. Gao, L. Sun, J. Li, Bayesian face sketch synthesis, IEEE Trans. Image Process. 26 (3) (2017) 1264–1274.
[20] D. Lin, X. Tang, Inter-modality face recognition, in: Proceedings of the European Conference on Computer Vision, 2006, pp. 13–26.
[21] D. Gong, Z. Li, J. Liu, Y. Qiao, Multi-feature canonical correlation analysis for face photo-sketch image retrieval, in: Proceedings of the ACM International Conference on Multimedia, 2013, pp. 617–620.
[22] Z. Lei, S. Li, Coupled spectral regression for matching heterogeneous faces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1123–1128.
[23] Z. Lei, C. Zhou, D. Yi, A. Jain, S. Li, An improved coupled spectral regression for heterogeneous face recognition, in: Proceedings of the IAPR International Conference on Biometrics, 2012, pp. 7–12.
[24] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 808–821.
[25] A. Mignon, F. Jurie, CMML: a new metric learning approach for cross modal matching, in: Proceedings of the Asian Conference on Computer Vision, 2012, pp. 1–14.
[26] J. Huo, Y. Gao, . Shi, W. Yang, H. Yin, Heterogeneous face recognition by margin-based cross-modality metric learning, IEEE Trans. Cybern. 48 (6) (2017) 1814–1826.
[27] J. Huo, Y. Gao, Y. Shi, H. Yin, Cross-modal metric learning for AUC optimization, IEEE Trans. Neural Netw. Learn. Syst. 29 (10) (2018) 4844–4856.
[28] S. Liao, D. Yi, Z. Lei, R. Qin, S. Li, Heterogeneous face recognition from local structures of normalized appearance, in: Proceedings of the IAPR International Conference on Biometrics, 2009, pp. 209–218.
[29] H. Bhatt, S. Bharadwaj, R. Singh, M. Vatsa, Memetically optimized MCWLD for matching sketches with digital face images, IEEE Trans. Inf. Forens. Secur. 7 (5) (2012) 1522–1535.
[30] H. Galoogahi, T. Sim, Face sketch recognition by local radon binary pattern, in: Proceedings of the IEEE International Conference on Image Processing, 2012, pp. 1837–1840.
[31] B. Klare, A. Jain, Heterogeneous face recognition using kernel prototype similarities, IEEE Trans. Pattern Anal. Mach. Intell. 35 (6) (2013) 1410–1422.
[32] B. Klare, Z. Li, A. Jain, Matching forensic sketches to mug shot photos, IEEE Trans Pattern Anal Mach Intell 33 (3) (2011) 639–646.
[33] H. Roy, D. Bhattacharjee, A novel quaternary pattern of local maximum quotient for heterogeneous face recognition, Pattern Recognit. Lett. 113 (2018) 19–28.
[34] W. Zhang, X. Wang, X. Tang, Coupled information-theoretic encoding for face photo-sketch recognition, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2011, pp. 513–520.
[35] C. Peng, X. Gao, N. Wang, J. Li, Face recognition from multiple stylistic sketches: scenarios, datasets, and evaluation, Pattern Recognit. 84 (2018) 262–272.
[36] J. Yu, Y. Rui, D. Tao, Click prediction for web image reranking using multimodal sparse coding, IEEE Trans. Image Process. 23 (5) (2014) 2019–2032.
[37] J. Yu, Y. Rui, B. Chen, Exploiting click constraints and multiview features for image reranking, IEEE Trans Multimed. 16 (1) (2014) 159–168.
[38] Y. Di, Z. Lei, S. Li, Shared representation learning for heterogeneous face recognition, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2015, pp. 1–7.
[39] L. Zhang, L. Lin, X. Wu, S. Ding, L. Zhang, End-to-End photo-sketch generation via fully convolutional representation learning, in: Proceedings of the ACM International Conference on Multimedia Retrieval, 2015, pp. 627–634.
[40] L. Jiao, S. Zhang, L. Li, F. Liu, W. Ma, A modified convolutional neural network for face sketch synthesis, Pattern Recognit. 76 (2018) 125–136.
[41] X. Liu, L. Song, X. Wu, T. Tan, Transferring deep representation for NIR-VIS heterogeneous face recognition, in: Proceedings of the IAPR International Conference on Biometrics, 2016, pp. 1–8.
[42] M. Sarfraz, R. Stiefelhagen, Deep perceptual mapping for cross-modal face recognition, Int. J. Comput. Vis. 122 (3) (2017) 426–438.
[43] L. Song, M. Zhang, X. Wu, R. He, Adversarial discriminative heterogeneous face recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018.
[44] Z. Yi, H. Zhang, P. Tan, . Gong, Dualgan: Unsupervised dual learning for image–to-image translation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017.
[45] N. Wang, W. Zha, J. Li, X. Gao, Back projection: an effective postprocessing method for GAN-based face sketch synthesis, Pattern Recognit. Lett. 107 (2018) 59–65.
[46] L. Gatys, A. Ecker, M. Bethge, Image style transfer using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2414–2423.
[47] P. Isola, J. Zhu, T. Zhou, A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[48] J. Zhu, T. Park, P. Isola, A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of IEEE International Conference on Computer Vision, 2017.
[49] Y. Luo, Y. Wen, T. Liu, D. Tao, General heterogeneous transfer distance metric learning via knowledge fragments transfer, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2017.
[50] Y. Luo, Y. Wen, T. Liu, D. Tao, Transferring knowledge fragments for learning distance metric from a heterogeneous domain, IEEE Trans. Pattern Anal. Mach. Intell. (2018).
[51] R. He, X. Wu, Z. Sun, Learning invariant deep representation for NIR-VIS face recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2017, pp. 2000–2006.
[52] X. Wu, L. Song, R. He, T. Tan, Coupled deep learning for heterogeneous face recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018.
[53] R. He, X. Wu, Z. Sun, T. Tan, Wasserstein cnn: learning invariant features for Nir-vis face recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2018). 1–1.
[54] L. Lin, G. Wang, W. Zuo, X. Feng, L. Zhang, Cross-domain visual matching via generalized similarity measure and feature learning, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1089–1102.
[55] C. Ding, D. Tao, Trunk-branch ensemble convolutional neural networks for video-based face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2018) 1002–1014.
[56] C. Ding, C. Xu, D. Tao, Multi-task pose-invariant face recognition, IEEE Trans. Image Process. 24 (3) (2015) 980–993.
[57] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: a unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.
[58] J. Lu, V. Liong, X. Zhou, J. Zhou, Learning compact binary face descriptor for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (10) (2015) 2041–2056.
[59] M. Brown, G. Hua, S. Winder, Discriminative learning of local image descriptors, IEEE Trans. Pattern Anal. Mach. Intell. 33 (1) (2011) 43–57.
[60] P. Belhumeur, J. Hespanda, D. Kiregeman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[61] X. Cao, Y. Wei, F. Wen, J. Sun, Face alignment by explicit shape regression, Int. J. Comput. Vis. 107 (2) (2014) 177–190.

[62] A. Martinez, R. Benavente, The AR face database, Technical Report, CVC Technical Report #24, 1998.
[63] P. Phollips, H. Moon, P. Rauss, S. Rizvi, The FERET evaluation methodology for face recognition algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 22 (10) (Oct. 2000) 1090–1104.
[64] P. Mittal, A. Jain, G. Goswami, R. Singh, M. Vatsa, Recognizing composite sketches with digital face images via SSD dictionary, in: Proceedings of the IAPR International Conference on Biometrics, 2014, pp. 1–6.
[65] S. Saxena, J. Verbeek, Heterogeneous face recognition with CNNs, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 483–491.
[66] H. Galoogahi, T. Sim, Inter-modality face sketch recognition, in: Proceedings of the IEEE International Conference on Multimedia and Expo, 2012, pp. 224–229.
[67] H. Han, B.F. Klare, K. Bonnen, A.K. Jain, Matching composite sketches to face photos: a component-based approach, IEEE Trans. Inf. Forensics Secur. 8 (1) (2013) 191–204.
[68] C. Galea, R.A. Farrugia, Face photo-sketch recognition using local and global texture descriptors, in: Proceedings of the European Signal Processing Conference, 2016.
[69] O. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: Proceedings of the British Machine Vision Conference, 2015, p. 6.
[70] H. Hotelling, Relations between two sets of variate, Biometrika 28 (3–4) (1936) 321–377.
[71] L. Wu, L. Du, B. Liu, G. Xu, Y. Ge, Y. Fu, J. Li, Y. Zhou, H. Xiong, Heterogeneous metric learning with content-based regularization for software artifact retrieval, in: Proceedings of the IEEE International Conference on Data Mining, 2014, pp. 610–619.
[72] X. Huang, Z. Lei, M. Fan, X. Wang, S.Z. Li, Regularized discriminative spectral regression method for heterogeneous face matching, IEEE Trans. Image Process. 22 (1) (2013) 353–362.
[73] K. Wang, R. He, W. Wang, L. Wang, T. Tan, Learning coupled feature spaces for cross-modal matching, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2088–2095.
[74] M. Shao, Y. Fu, Cross-modality feature learning through generic hierarchical hyperlingual-words, IEEE Trans. Neural Netw. Learn. Syst. 28 (2) (2016) 451–463.
[75] Y. Jin, J. Lu, Q. Ruan, Coupled discriminative feature learning for heterogeneous face recognition, IEEE Trans. Inf. Forensi. Secur. 10 (3) (2015) 640–652.
[76] F. Juefei-Xu, D.K. Pal, M. Savvides, NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2015.
[77] D. Gong, Z. Li, W. Huang, X. Li, D. Tao, Heterogeneous face recognition: a common encoding feature discriminant approach, IEEE Trans. Image Process. 26 (5) (2017) 2079–2089.
[78] C. Reale, N.M. Nasrabadi, H. Kwon, R. Chellappa, Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 54–62.
[79] X. Liu, M. Kan, W. Wu, S. Shan, X. Chen, VIPLFAcenet: an open source deep face recognition SDK, Front. Comput. Sci. 11 (2) (2017) 208–218.
[80] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 499–515.
[81] X. Wu, R. He, Z. Sun, T. Tan, A light CNN for deep face representation with noisy labels, 2015, http://arxiv.org/abs/1511.02683).



Chunlei Peng received the B. Sc degree in electronic and information engineering from Xidian University, Xi'an, China, in 2012. He received his Ph.D. degree in information and telecommunications engineering in 2017. Now, he works with the School of Cyber Engineering at Xidian University. From September 2016 to September 2017, he has been a visiting Ph.D. student with the Duke University, NC, USA. His current research interests include computer vision, pattern recognition, and machine learning.

Nannan Wang received the B. Sc degree in information and computation science from Xi'an University of Posts and Telecommunications in 2009. He received his Ph.D. degree in information and telecommunications engineering in 2015. Now, he works with the state key laboratory of integrated services networks at Xidian University. From September 2011 to September 2013, he has been a visiting Ph.D. student with the University of Technology, Sydney, NSW, Australia. His current research interests include computer vision, pattern recognition, and machine learning. He has published more than 10 papers in refereed journals and proceedings including International Journal of Computer Vision, IEEE T-NNLS, T-IP, T-CSVT etc.

Jie Li received the B.Sc. degree in electronic engineering, the M.Sc. degree in signal and information processing, and the Ph.D. degree in circuit and systems, from Xidian University, Xi'an, China, in 1995, 1998, and 2004, respectively. She is currently a Professor in the School of Electronic Engineering, Xidian University, China. Her research interests include image processing and machine learning. In these areas, she has published around 50 technical articles in refereed journals and proceedings including IEEE T-NNLS, T-IP, T-CSVT, Information Sciences etc.

Xinbo Gao (M'02-SM'07) received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a Research Fellow at the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Post-Doctoral Research Fellow at the Department of Information Engineering, the Chinese University of Hong Kong, Hong Kong. Since 2001, he has been at the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor of Ministry of Education, a Professor of Pattern Recognition and Intelligent System, and the Director of the State Key Laboratory of Integrated Services Networks, Xi'an, China. His current research interests include multimedia analysis, computer vision, pattern recognition, machine learning, and wireless communications. He has published five books and around 200 technical articles in refereed journals and proceedings. Prof. Gao is on the Editorial Boards of several journals, including Signal Processing (Elsevier), and Neurocomputing (Elsevier). He served as the General Chair/Co-Chair, Program Committee Chair/Co-Chair, or PC Member for around 30 major international conferences. He is currently a fellow of the Institution of Engineering and Technology.