Dynamic Feature Learning for Partial Face Recognition

Lingxiao He[1,2], Haiqing Li[1,2], Qi Zhang[1,2], and Zhenan Sun[1,2,3]
[1] CRIPAC & NLPR, CASIA [2] University of Chinese Academy of Sciences, Beijing, P.R. China
[3] Center for Excellence in Brain Science and Intelligence Technology, CAS
{lingxiao.he, hqli, qi.zhang, znsun}@nlpr.ia.ac.cn

Abstract

Partial face recognition (PFR) in unconstrained environment is a very important task, especially in video surveillance, mobile devices, etc. However, a few studies have tackled how to recognize an arbitrary patch of a face image. This study combines Fully Convolutional Network (FCN) with Sparse Representation Classification (SRC) to propose a novel partial face recognition approach, called Dynamic Feature Matching (DFM), to address partial face images regardless of size. Based on DFM, we propose a sliding loss to optimize FCN by reducing the intra-variation between a face patch and face images of a subject, which further improves the performance of DFM. The proposed DFM is evaluated on several partial face databases, including LFW, YTF and CASIA-NIR-Distance databases. Experimental results demonstrate the effectiveness and advantages of DFM in comparison with state-of-the-art PFR methods.

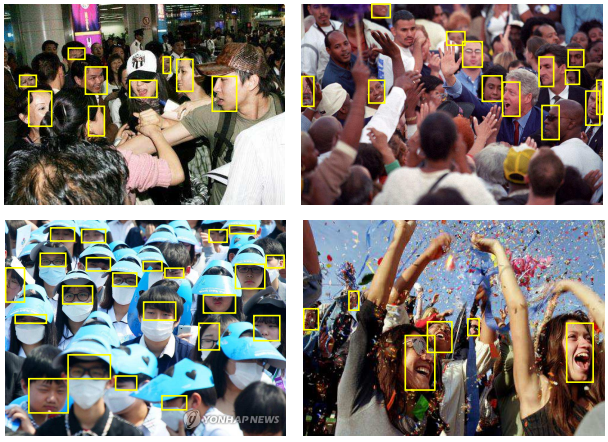

Figure 1. Partial face images are produced in unconstrained environments. A face may be 1) occluded by sunglasses, a hat and a scarf; 2) captured in various poses; 3) positioned partially out of cameras filed of view.

1. Introduction

Face recognition performance has been improved due to the rapid development of deep Convolution Neural Networks (CNNs), but it is assumed to have a full face observation as the input of current methods [16, 26, 32, 34, 35, 36]. However, the assumption of face recognition on full and frontal images is easily violated in real-world applications. Partial face recognition (PFR) has become an emerging problem with increasing requirements of identification from CCTV cameras and embedded vision systems in mobile devices, robots and smart home facilities. However, PFR is a challenging problem without a solution from traditional face recognition approaches. Examples of partial face images are illustrated in Fig. 1. Partial face images occur when a face is 1) occluded by objects such as faces of other individuals, sunglasses, hats, masks or scarves; 2) captured in the various pose without user awareness; 3) positioned partially outside camera' view. Therefore, only arbitrary-sizes face patches are presented in these captured images. In some situations, surveillance videos are vital clues for case investigation. For example, when a crime takes place, the surveillance footage often discovers only a partial face of a criminal suspect because the suspect may want to hide his/her appearance deliberately. From this perspective, it is important to develop a face recognition system that can work for holistic faces as well as partial faces. Excitingly, many face detection algorithms [4, 19, 22] are available to detect visible face patches. Therefore, we can effectively take advantage of these partial information to accomplish identity authentication.

Existing face recognition algorithms based on deep networks require fixed-size face images as inputs (e.g. 224×224 in VGGFace) because deep networks with fully-connected layers require fixed-size inputs by their definition. Therefore, most of them cannot directly deal with partial face images of arbitrary sizes (e.g. 120×160). To handle the problem, traditional methods usually re-scale arbitrary-size input images to fixed-size face images. However, the performance would suffer from deterioration extremely due to the undesired geometric deformation. The Multi-Scale Region-based Convolutional Neural Network

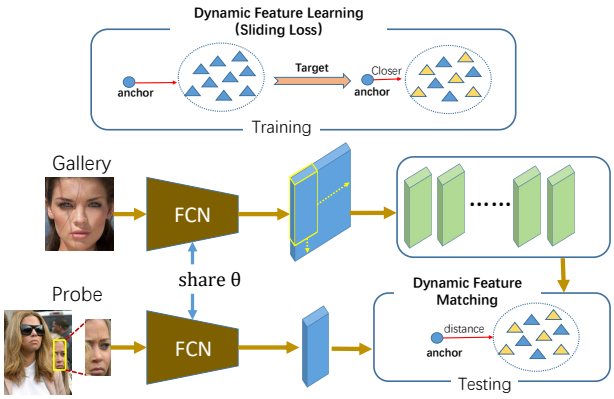

Figure 2. Proposed partial face recognition framework: Dynamic Feature Matching (DFM). Note that the sliding window on feature maps of gallery image shares the same size of the probe feature map. Therefore, the quantity and dimension of these sub-feature maps vary across different input images. And the triangles in yellow color represent the selected sub-feature maps.

(MR-CNN) [10] offers a solution of partial face recognition. It decomposes a face/partial face into several sub-region proposals and then extracts features of each sub-region proposal via deep convolutional networks. Finally, it achieves partial face recognition using region-to-region matching. Although MR-CNN achieves remarkable performance on several partial face databases, the computational cost is extensive because it repeatedly runs a deep convolutional network for each sub-region proposal. To improve computational efficiency, Sliding Window Matching (SWM) [43] introduces another solution for partial face recognition by setting up a sliding window of the same size as the probe image that is used to search for the most similar region within each gallery image. However, the computational efficiency of SWM is still extensive because computing the features of each region within the sliding window repeatedly is inevitable.

Deep Convolutional Neural Networks, as feature extractors in face recognition problems, require fixed-size images as inputs. Therefore, they could not directly accept partial face images with inconsistent sizes as input. Even though MR-CNN and SWM are able to address the problem, computation efficiency becomes the most difficult bottleneck in practical applications. In this paper, we introduce a PFR method: Dynamic Feature Matching (DFM) that can handle an arbitrary patch of a face image, and achieve impressive accuracy with high-efficiency. Fig. 2 illustrates the architecture of DFM. Firstly, Fully Convolutional Network (FCN) [33] is applied to extract spatial feature maps of given gallery and probe faces because FCN is applicable to arbitrary input image size while generating spatial feature maps with corresponding sizes of the input image. To extract more discriminative features, we transfer recent successful face model VGGFace [26] to an FCN by discarding the non-convolutional portion of networks. The last pooling layer is used as a feature extractor regardless of the size/scale of the input face. Secondly, we set up a sliding window of the same size as the probe feature maps to decompose the gallery feature maps into several gallery sub-feature maps (the dimension of probe feature maps and each gallery sub-feature maps are equal). Finally, Sparse Representation Classification (SRC) imposed by similarity-guided constraint provides a feasibility scheme where the probe feature maps are linearly represented by these gallery sub-feature maps for achieving alignment-free dynamic feature matching. DFM has a great advantage in computational speed since the convolutional layers are forwarded once on the entire/partial face, which improves the speed over 20× compared to MR-CNN. Given a new probe, we only require decomposing the entire gallery feature maps corresponding to the probe feature maps size. In addition, we propose a loss function called sliding loss based on DFM that encourages feature maps extracted from faces of the same identity to be similar.

The major contributions of our work are three-fold:

1. The proposed partial face recognition approach: Dynamic Feature Matching (DFM) combines FCN with SRC, achieving state-of-the-art performance in computational efficiency and recognition accuracy.
2. The proposed DFM can not only work for holistic faces but also can deal with partial faces of arbitrary-size without requiring face alignment.
3. We propose a sliding loss based on DFM that can learn more discrimination deep features using arbitrary-size input images. Comprehensive analysis shows the proposed sliding loss is more effective.

The remainder of this paper is organized as follows. In Sec. 2, we review the related work about FCN, SRC and existing PFR algorithms. Sec. 3 introduces technical details of Dynamic Feature Matching and Sliding Loss. Sec. 4 shows experimental results and analyzes the performance in computational efficiency and accuracy. Finally, we conclude our work in Sec. 5.

2. Related Work

Deep Neural Networks. Convolutional Neural Networks (CNNs) have been widely applied into some vision tasks including image classification [9, 13], object detection [29, 28] and semantic segmentation [17]. DeepFace [35] first bring CNNs into face recognition, improving the performance of face recognition greatly. With success of CNNs in face recognition, Some face networks such as VGGFace [26], Light CNN [39], FaceNet [32] and SphereFace [16] are proposed to further improve the performance of face recognition. FCN only contains convolutional layers and pooling layers, which has been applied into spatially dense

tasks including semantic segmentation [6] and object detection [5, 27]. Besides, He *et al.* [8] introduce a spatial pyramid pooling (SPP) layer imposed on FCN to produce fixed-length representation from arbitrary-size inputs.

Sparse Representation Classification. Wright *et al.* [38] introduce a well-known SRC for face recognition, achieving a robust performance under occlusions and illumination variations. Similar studies [42, 40] based on SRC about face recognition have also been conducted. SRC is also applied to signal classification [12], visual tracking [20] and visual classification [41].

Partial Face Recognition. Most keypoint-based approaches [11, 15, 37] have been proposed for PFR. Hu *et al.* [11] propose an approach based on SIFT descriptor [18] representation that does not require alignment, and the similarities between a probe patch and each face image in a gallery are computed by the instance-to-class (I2C) distance with the sparse constraint. Liao *et al.* [15] propose an alignment-free approach called multiple keypoints descriptor SRC (MKD-SRC), where multiple affine invariant keypoints were extracted for facial features representation and sparse representation based on classification (SRC) [38] is used for classification. Weng *et al.* [37] propose a Robust Point Set Matching (RPSM) method based on SIFT descriptor, SURF descriptor [2] and LBP [1] histogram for partial face matching. Their approach first aligns the partial faces and then computes the similarity of the partial face image and a gallery face image. The performance of keypoint-based methods is far from satisfaction with local descriptors. Besides, region-based models [3, 7, 21, 23, 24, 30, 31] also offer a solution for PFR, they only require face subregions as inputs, such as eye [30], nose [30], half (left or right portion) of the face [7], or the periocular region [25]. He *et al.* [10] propose a Multi-Scale Region-based CNNs (MR-CNN) model and achieve the highest performance (85.97%) for PFR on CASIA-NIR-Distance database [10]. However, these methods require the presence of certain facial components and pre-alignment. To this end, we propose an alignment-free PFR algorithm DFM that can achieve high accuracy with high-efficiency.

3. Our Approach

3.1. Fully Convolutional Network

Typical face recognition networks mainly consist of convolutional layers and fully-connected layers. The convolutional layers operate in a sliding-window manner and generate spatial outputs. The fully-connected layers produce fixed-dimension feature representation and throw away spatial coordinates. Therefore, face recognition networks with fully-connected layers cannot learn feature representation from arbitrary-size inputs and the fixed-length resulting feature vector is equivalent to the pre-trained dimension. For example, VGGFace [26] requires a 224×224 face image as the input to generate 4096-dimension feature vector. In fact, we find that the requirement of fixed-sizes comes from the fully-connected layers that demand fixed-length vectors as inputs. To process arbitrary-size face images, fully-connected layers in deep networks are discarded to evolve into a Fully Convolutional Network (FCN) which can accept arbitrary-size inputs. The designed FCN is implemented on the basis of a successful face recognition model VGGFace that is able to generate robust deep features, as shown in Table 1. The FCN contains convolution and pooling layers. The last pooling layer in the FCN generate spatial outputs (these outputs are known as *feature maps*.). For example, 7×7 and 5×6 spatial feature maps can be extracted by FCN from 224×224 and 160×200 face images, respectively. Therefore, FCN can infer correspondingly-size spatial feature maps without the limitation of input sizes.

Table 1. Fully convolutional network configuration. For each convolutional layers, the filter size, number of filters, stride and padding are indicated.

type	input	conv	conv	mpool	conv
name	-	conv1-1	conv1-2	pool1	conv2-1
support	-	3	3	2	3
filt dim	-	3	64	-	64
num filts	-	64	64	-	128
stride	-	1	1	1	1
pad	-	1	1	0	1
type	conv	mpool	conv	conv	conv
name	conv2-2	pool2	conv3-1	conv3-2	conv3-3
support	3	2	3	3	3
filt dim	128	-	128	256	256
num filts	128	-	256	256	256
stride	1	2	1	1	1
pad	1	0	1	1	1
type	mpool	conv	conv	conv	mpool
name	pool3	conv4-1	conv4-2	conv4-3	pool4
support	2	3	3	3	2
filt dim	-	256	512	512	-
num filts	-	512	512	512	-
stride	2	1	1	1	2
pad	0	1	1	1	0
type	conv	conv	conv	mpool	
name	conv5-1	conv5-2	conv5-3	pool5	
support	3	3	3	2	
filt dim	512	512	512	-	
num filts	512	512	512	-	
stride	1	1	1	2	
pad	1	1	1	0	

3.2. Dynamic Feature Matching

This section will introduce the detail of Dynamic Feature Matching (DFM). We first introduce probe feature maps extraction and dynamic gallery dictionary construction.

Probe: Given an arbitrary-size probe face image (anchor), spatial feature maps $\mathbf{p}$ of size $w \times h \times d$ are generated by

FCN, where w, h, d denote the width, the height and the channel of $\mathbf{p}$, respectively.

Gallery: Spatial feature map $\mathbf{g}_c$ are generated by FCN for subject c in the gallery.

Distinctly, it fails to compute the similarity of $\mathbf{p}$ and $\mathbf{g}_c$ on account of feature dimension-inconsistent. For this reason, we set up a sliding window of the same size as $\mathbf{p}$ at a stride of s to decompose $\mathbf{g}_c$ into several sub-feature maps. As shown in Fig. 2, the corresponding k sub-feature maps are denoted by $\mathbf{G}_c = [\mathbf{g}_{c_1}, \mathbf{g}_{c_2}, \ldots, \mathbf{g}_{c_k}]$. We stretch $\mathbf{p}$ and each sub-feature maps in $\mathbf{G}_c$ into a M-dimension vector, where $M = w \times h \times d$. Since the size of the probe is arbitrary, thus the dimension and number of $\mathbf{G}_c \in \mathbb{R}^{M \times k}$ are various dynamically corresponding to the probe size.

To achieve feature matching without alignment, we transfer the problem into error reconstruction where $\mathbf{p}$ can be linearly represented by $\mathbf{G}_c$. Therefore, we wish to compute coefficients of $\mathbf{p}$ with respect to $\mathbf{G}_c$, and we denote the coefficients as $\mathbf{w}_c$. The reconstruction error represents the matching score. Finally, we solve $\mathbf{w}_c$ by minimizing the reconstruction error. We define the reconstruction error as follows

$$\mathcal{L}(\mathbf{w}_c) = ||\mathbf{p} - \mathbf{G}_c\mathbf{w}_c||_2^2 \tag{1}$$

where $\mathbf{w}_c \in \mathbb{R}^{k \times 1}$ and $\mathbf{p} \in \mathbb{R}^{M \times 1}$. In order to solve $\mathbf{w}_c$, we impose two constraints, sparsity and similarity, on the reconstruction process.

Sparse constraint controls the sparsity of coding vector $\mathbf{w}_c$, since few sub-feature vector should be used for reconstructing the feature vector $\mathbf{p}$. Therefore, we constraint $\mathbf{w}_c$ using ℓ_1-norm: $||\mathbf{w}_c||_1$.

Similarity-guided constraint. There is a shortcoming in above reconstruction process where it is free to use any sub-feature vectors to reconstruct the feature $\mathbf{p}$. In other words, all feature vector in $\mathbf{G}_c$ are selected for reconstructing $\mathbf{p}$, it does not consider the similarities between $\mathbf{p}$ and each feature vector in $\mathbf{G}_c$ because the aim of reconstruction process is to minimize the reconstruction error by linearly combination of $\mathbf{G}_c$. Thus, some dissimilar feature vectors will be selected because a linear combination could produce the minimum reconstruction error. To this end, we impose the similarity-guided constraint on reconstruction term for selecting similar feature vector and excluding dissimilar feature vectors automatically. For computing the similarity between $\mathbf{p}$ and $\mathbf{G}_c$, we first normalize the $\mathbf{p}$ and $\mathbf{g}_{c_i}$ to 1 using ℓ_2-norm. For the k sub-feature vectors, similarity score vector is denoted as $\mathbf{p}^T\mathbf{G}_c \in \mathbb{R}^{1 \times k}$ by cosine similarity measure. Distinctly, the more similar between $\mathbf{p}$ between $\mathbf{g}_{c_i}$, the more likely $\mathbf{g}_{c_i}$ is selected, otherwise, $\mathbf{g}_{c_i}$ is excluded. Therefore, the reconstruction coefficients $\mathbf{w}_c$ are positively correlated to $\mathbf{p}^T\mathbf{G}_c$. Then, the similarity-guided constraint is defined as $\mathbf{p}^T\mathbf{G}_c\mathbf{w}_c$.

Thus, the sparse representation formulation finally is defined as:

$$\mathcal{L}(\mathbf{w}_c) = ||\mathbf{p} - \mathbf{G}_c\mathbf{w}_c||_2^2 - \alpha\mathbf{p}^T\mathbf{G}_c\mathbf{w}_c + \beta||\mathbf{w}_c||_1 \tag{2}$$

where α and β are constants that control the strength of similarity-guided constraint and sparsity constraint, respectively.

For the optimization of $\mathbf{w}_c$, we transform Eq. (2) to the following formulation

$$\mathcal{L}(\mathbf{w}_c) = \frac{1}{2}\mathbf{w}_c{}^T\mathbf{G}_c^T\mathbf{G}_c\mathbf{w}_c - (1 + \frac{\alpha}{2})\mathbf{p}^T\mathbf{G}_c\mathbf{w}_c + \frac{\beta}{2}||\mathbf{w}_c||_1. \tag{3}$$

We utilize the feature-sign search algorithm [14] to solve an optimal $\mathbf{w}_c$. Non-zero values of $\mathbf{w}_c$ represents the selected sub-feature vectors that are used for reconstruction. After solving the optimal sparse coefficients. Then, we adopt the following dynamic matching method to determine the identity of the probe image:

$$\min_c r_c(\mathbf{p}) = ||\mathbf{p} - \mathbf{G}_c\mathbf{w}_c||_2 - \alpha\mathbf{p}^T\mathbf{G}_c\mathbf{w}_c. \tag{4}$$

Eq. (4) applies a sum fusion among reconstruction error and weighted matching scores, which determines the identity of the probe image and returns the result with least score. When a new probe partial face needs to be authenticated, we re-decompose the gallery feature maps according to different-sized probe images without repeatedly computing gallery features. Therefore, the gallery dictionary varies dynamically. Some convolutional layers in FCN greatly reduce the size of output feature maps. Therefore, the computation of gallery feature maps decomposition is high-efficiency. For example, if the size of $\mathbf{p}$ is 5×5, the size of $\mathbf{g}_c$ is 7×7, and the stride is $s = 1$ pixel, we perform decomposition operation $k = 9$ times per a gallery image. DFM is a high-efficiency and high-accuracy PFR method. By sharing computation, it is able to achieve fast feature extraction. Besides, DFM is an alignment-free method, it can deal with an arbitrary patch of a face image and does not require to know priori location information.

3.3. Sliding Loss

In Eq. (2), FCN parameters in convolution layers are fixed. We want to update convolution parameters θ to improve the discrimination of deep features generated by FCN. So, we propose a new loss function called **Sliding Loss** based on dynamic feature matching formulation in Eq. (2). The aim of Sliding Loss is to learn the coefficients $\mathbf{w}_c$ and convolution parameters θ to improve the discriminative of deep features by minimizing Eq. (2). It effectively ensures that the two feature maps from the same identity are close to each other while those extracted from different identities stay away. Thus, the proposed Sliding Loss is

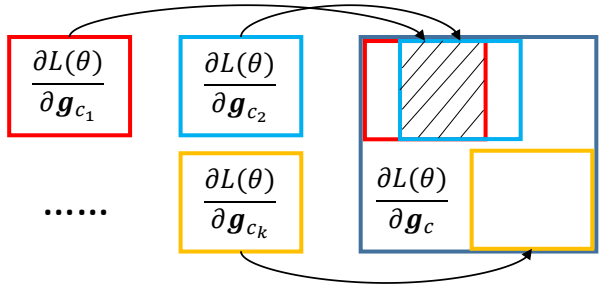

Figure 3. The gradients of $\mathcal{L}(\theta)$ with respect to $\mathbf{g}_c$. It is computed by concatenating $\frac{\partial \mathcal{L}(\theta)}{\partial \mathbf{G}_c}$.

finally defined as

$$\mathcal{L}(\mathbf{w}_c, \theta) = y_c(||\mathbf{p} - \mathbf{G}_c\mathbf{w}_c||_2^2 - \alpha \mathbf{p}^T \mathbf{G}_c \mathbf{w}_c) + \beta||\mathbf{w}_c||_1. \tag{5}$$

$y_c = 1$ means that $\mathbf{p}$ and $\mathbf{G}_c$ are from the same identity. In this case, it minimizes $||\mathbf{p} - \mathbf{G}\mathbf{w}||_2^2 - \alpha \mathbf{p}^T \mathbf{G}_c \mathbf{w}_c$. $y_c = -1$ means different identities, thus it minimizes $-(||\mathbf{p} - \mathbf{G}\mathbf{w}||_2^2 - \alpha \mathbf{p}^T \mathbf{G}_c \mathbf{w}_c)$.

3.4. Optimization.

We employ alternating optimization algorithm to solve the Sliding Loss. We first optimize $\mathbf{w}_c$ and then optimize θ.

Step 1: fix θ, optimize $\mathbf{w}_c$. The aim of this step is to learn the reconstruction coefficients $\mathbf{w}_c$. We rewrite Eq. (5) the following:

$$\mathcal{L}(\mathbf{w}_c) = \frac{1}{2} y_c {\mathbf{w}_c}^T \mathbf{G}_c^T \mathbf{G}_c \mathbf{w}_c - y_c(1 + \frac{\alpha}{2}) \mathbf{p}^T \mathbf{G}_c \mathbf{w}_c + \frac{\beta}{2}||\mathbf{w}_c||_1, \tag{6}$$

and utilize the feature-sign search algorithm [14] to solve an optimal $\mathbf{w}_c$.

Step 2: fix $\mathbf{w}_c$, optimize θ to update FCN parameters to obtain discriminative features. $||\mathbf{p} - \mathbf{G}_c\mathbf{w}_c||_2^2$ ensures that $\mathbf{p}$ is closer to combination of all selected sub-feature vectors. $\alpha \mathbf{p}^T \mathbf{G}_c \mathbf{w}_c$ aims to decrease the distance between $\mathbf{p}$ and each selected sub-feature vector. The two terms contribute to improving the discrimination of deep features. The gradients of $\mathcal{L}(\theta)$ with respect to $\mathbf{p}$ and $\mathbf{G}_c$ are computed as

$$\begin{cases} \frac{\partial \mathcal{L}(\theta)}{\partial \mathbf{p}} = 2(\mathbf{p} - \mathbf{G}_c\mathbf{w}_c) - \alpha \mathbf{G}_c \mathbf{w}_c \\ \frac{\partial \mathcal{L}(\theta)}{\partial \mathbf{G}_c} = -2(\mathbf{p} - \mathbf{G}_c\mathbf{w}_c)\mathbf{w}_c^T - \alpha \mathbf{p} \mathbf{w}_c^T, \end{cases} \tag{7}$$

where $\frac{\mathcal{L}(\theta)}{\partial \mathbf{G}_c} = [\frac{\mathcal{L}(\theta)}{\partial \mathbf{g}_{c_1}}, \frac{\mathcal{L}(\theta)}{\partial \mathbf{g}_{c_2}}, \ldots, \frac{\mathcal{L}(\theta)}{\partial \mathbf{g}_{c_k}}]$, $\mathbf{p}$ and $\mathbf{G}_c$ share the same FCN parameter θ. Then, we concatenate the $\frac{\mathcal{L}(\theta)}{\partial \mathbf{G}_c}$ at stride of s pixels to obtain the gradients with respect to $\mathbf{g}_c$, as shown in Fig. 3. Clearly, FCN supervised by sliding loss is trainable and can be optimized by standard Stochastic Gradient Descent (SGD). In Algorithm 1, we summarize detail of feature learning with sliding loss in FCN.

Algorithm 1: Feature learning with Sliding Loss

Input:
Training data $\{\mathbf{p}, \mathbf{G}_c\}$. Initialized parameter θ in convolution layers. The parameters of similarity-guided constraint λ and sparsity strength β. Learning rate α. The number of iteration $t \leftarrow 0$

Output: The parameter θ.

1: **while** not converge **do**
2: $t + 1 \leftarrow t$
3: Compute the sliding loss by $\mathcal{L}(\mathbf{w}, \theta)$.
4: Update the sparse coefficients $\mathbf{w}$.
5: Update the gradients of $\mathcal{L}(\mathbf{w}, \theta)$ with respect to $\mathbf{p}$ and $\mathbf{g}$.
6: Update the parameters θ by $\theta^{t+1} = \theta^t - \alpha(\frac{\partial \mathcal{L}}{\partial \mathbf{p}} \frac{\partial \mathbf{p}}{\partial \theta^t} + \frac{\partial \mathcal{L}}{\partial \mathbf{g}_c} \frac{\partial \mathbf{g}_c}{\partial \theta^t})$
7: **end while**

We put feature matching and learning into a unified framework. We train an end-to-end FCN with sliding loss, which can effectively improve the performance of DFM. Compared to softmax loss, sliding loss can learn distance information between an arbitrary patch of a face image and all faces of a subject.

4. Experiments

4.1. Experiment Settings

Network Architecture. Fully Convolutional Network (FCN) is implemented on basis of successful face model VGGFace that is able to generate robust features. We remove all non-convolutional layers to obtain an FCN that contains 13 convolutional layers followed by ReLU layers, as shown in Table 2. The last pooling layer in the FCN (*pool5*) is used as feature extractor.

Training and testing. Our implementation is based on the publicly available code of *MatConvnet* [40]. All experiments in the paper are trained and tested on PC with 16GB RAM, i7-4770 CPU @ 3.40GHz. The framework of training and testing is illustrated in Fig. 4. In training term, we use publicly available web-collected training database CASIA-WebFace to train FCN with sliding loss. 2,000 group face images are selected to fine-tune on the pre-trained FCN model. Each group has an arbitrary-size face image (anchor) and 5 holistic face images of the same subject as the anchor face image. The model is trained with the batch of 20 and 10^{-4} learning rate. Besides, we set $\alpha = 0.6$ and $\beta = 0.4$ in training term, respectively. In testing term, the test set consists of a probe set and a gallery set. The probe set is composed of all arbitrary-size face images and

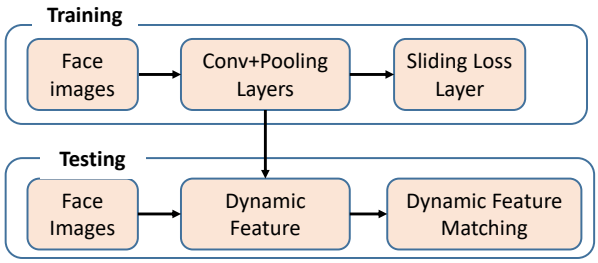

Figure 4. Training with sliding loss and testing with dynamic feature matching.

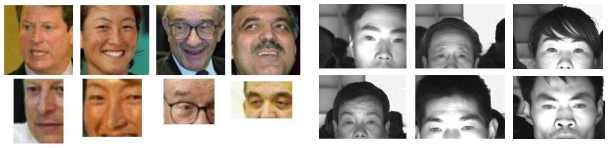

Figure 5. (a) Examples of face images and partial face images in LFW. (b) Example images in CASIA-NIR-Distance.

the holistic face images are used as the gallery set. We extract spatial feature maps from the *pool5* layer. And then we use Eq. (3) and Eq. (7) to achieve alignment-free partial face recognition.

Evaluation Protocol. In order to show the performance of the proposed method, we provide the average Cumulative Match Characteristic (CMC) curves for close-set experiment and Receiver Operating Characteristic (ROC) curves for verification experiment.

Databases and Settings. 1) A simulated partial face database named Partial-LFW that is based on LFW databases is used for evaluation. LFW database contains 13,233 images from 7,749 individuals. Face images in LFW have large variations in pose, illumination, and expression, and may be partially occluded by other faces of individuals, sunglasses, etc. There is no partial face database under visible illumination publicly available. Therefore, we synthetically generate the partial face database. The partial-LFW gallery set contains 1,000 holistic face images from 1,000 individuals. The probe set share same subjects with the gallery set, but for each individual they contain different images. Gallery face images are re-scaled to 224×224. To generate partial face images as the probe set, an arbitrary-size region at random position of a random size is cropped from a holistic face image. Fig. 5(a) shows some partial face images and holistic faces images in Partial-LFW. 2) Partial-YTF is used for evaluation, which is based on YTF. YTF database includes 3.424 videos form 1,595 different individuals. The gallery of Partial-YTF contains 200 holistic face images from 200 individuals. The probe set consists of individuals in the gallery with different images (all probe images are arbitrary-size face patches that are generated by randomly cropping). 3) CASIA-NIR-Distance database [10] is also used for evaluation. CASIA-NIR-Distance database is a newly proposed partial face database, which includes 276 subjects. Each subject has a sequence of face images, only half of which contain the entire facial region of the subject. Partial face images are captured by the infrared camera with the subject presenting an arbitrary region of the face. Besides the variation in presented partial faces, face images in the CASIA-NIR-Distance database are taken with different distances, view angles, scales, and lighting conditions. The examples of partial face images in CASIA-NIR-Distance are shown in Fig. 5(b).

Table 2. Performance comparison with 1,000 classes.

Method	Rank-1	TPR@FPR=0.1%
MR-CNN [10]	24.7%	17.6%
MKDSRC-GTP [43]	1.1%	0.7%
I2C [11]	6.8%	0.3%
VGGFace [26]	20.9%	18.1%
DFM	**27.3%**	**29.8%**

4.2. Experiment on LFW

In the section, we focus on five aspects below, 1). comparison to the state-of-the-art methods; 2). face verification using DFM; 3). evaluation on mixed partial and holistic face images; 4). influence of the cropped region size and the area; 5) parameter analysis of DFM. Comprehensive experiments in this section are conducted on Partial-LFW database.

4.2.1 Comparison to the State-of-the-Art

To verify the performance of DFM, we conduct experiments on LFW database with 1,000 classes. Images in the probe set are all partial face images that are produced by cropping from holistic face images. For comparison, the Multi-Scale Region-based CNN (MR-CNN) [10], and two key-point-based algorithms: MKDSRC-GTP [43] and I2C [11] are considered in this experiment. Besides, we compare VGGFace by re-scaling the probe images to 224×224. We set $\alpha = 2.1$ and $\beta = 0.1$ for DFM.

Table 2 shows the experimental results on LFW database. DFM achieves 27.3% rank-1 accuracy which shows clearly that DFM performs much better than these conventional PFR methods. The keypoint-based algorithms do not perform well because the feature representation based on local keypoint descriptor is not robust. MR-CNN requires aligned partial face images and run CNNs many times so that it does not achieve good performance.

To handle arbitrary-size face images, traditional methods usually re-scale the arbitrary-size faces to fixed-size face images. We analysis the influence of face deformation. VGGFace model [26] is used for comparison, and the

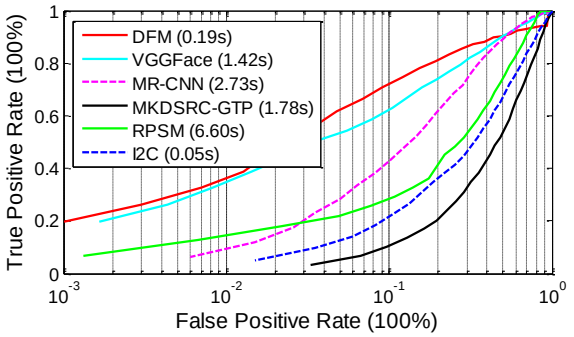

Figure 6. The performance of partial face verification.

Table 3. Experimental results using holistic and mixed database.

Probe set	Rank-1	TPR@FPR=0.1%	TPR@FPR=1%
Probe-1	56.8%	61.0%	78.4%
Probe-2	41.9%	45.2%	65.9%
Probe-3	27.3%	29.8%	52.6%

probe images are re-scaled to 224×224 to meet the input size of VGGFace. Table 2 shows the performance of DFM and VGGFace model. The VGGFace model performs worse than DFM because face image stretching would produce unwanted geometric deformation. DFM retains the spatial information which it effectively avoids the effect of deformation.

4.2.2 Face verification on LFW

Face verification aims to verify whether a pair of face images come from the same individual or not. In this experiment, we follow the Labeled Faces in the Wild (LFW) benchmark test protocol[1], where the database is divided into 10 subsets for cross-validation, with each subset containing 300 pairs of genuine matches and 300 pairs of impostor matches for verification. We synthetically generate a series of image pairs. In a pair image, one is an original full image without processing and the other one is a partial face produced by cropping an arbitrary-size region at random position of a random size from the other raw image.

Fig. 6 shows the ROC curves of various PFR algorithms and runtime of verifying a pair of images is illustrated in Fig. 6. DFM performs the best compared to other partial face algorithms. Besides, DFM shows competitive performance in computational efficiency compared to other PFR methods, it cost 0.19 seconds to verify a pair of face images.

[1] http://vis-www.cs.umass.edu/lfw/pairs.txt

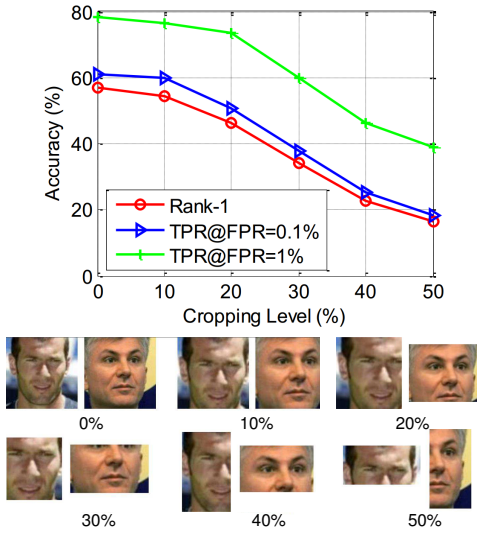

Figure 7. Experimental results using different cropping level faces.

4.2.3 Evaluation on holistic face images and mixed partial and holistic face images

Additional experiment on LFW with 1,000 classes is conducted. Three different probe sets are constructed in our experiment. **Probe-1:** 1,000 holistic faces; **Probe-2:** 500 random cropped partial face images and 500 holistic face images; **Probe-3:** 1,000 random cropped partial faces. Table 3 shows the experimental results, which suggests that face incompletion would influence the performance of face recognition.

4.2.4 Influence of the cropped region size and the area

We add additional experiment for evaluating the influence of the cropped region size and the size on LFW with 1,000 classes (each class contains one gallery image and one probe image). Fig. 7 shows the examples of different cropped sizes of face images and experimental results, respectively. Fig. 8 shows the examples of different areas of face images and Table 4 shows the experimental result using different areas.

It is not surprising to observe accuracy degradation when the size of cropped patches goes down, as there is less information retained in the image patches. Besides, according to the results, eye regions (upper) tend to contain more information that is helpful for identification. The drastic perfor-

Figure 8. Examples of different face areas.

Table 4. Experimental result using different areas.

Area	Rank-1	TPR@FPR=0.1%	TPR@FPR=1%
Upper	39.2%	44.8%	66.5%
Down	7.8%	8.6%	19.8%
Right	24.2%	27.0%	48.5%
Left	27.6%	31.5%	56.3%

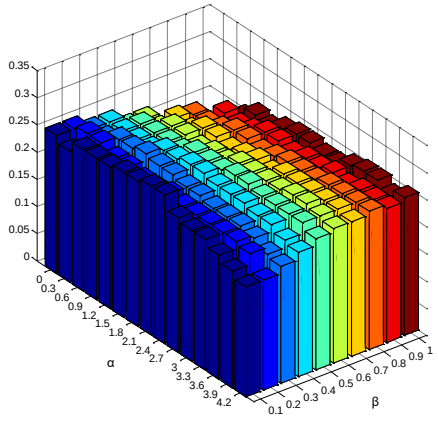

Figure 9. Evaluate of the weighting parameters α and β.

mance degradation when using lower region of the face with the cropping level over 20% as the probe image justifies our observations.

4.2.5 Parameter Analysis

To evaluate the influence of the strength of similarity-guided constraint α and sparsity constraint β, we conduct comprehensive experiment on LFW with different α and β. We set value of α by from 0 to 4.2 at stride of 0.3 and value of β by from 0.1 to 1 at stride of 0.1, respectively. Fig. 9 shows that DFM performs best with $\alpha = 2.1$ and $\beta = 0.1$, and it achieves 27.3% and rank-1 accuracy. When $\beta = 0.1$, the performance of DFM stays the higher rank-1 accuracy. And the performance of DFM first rises up and then moves down as α increases.

4.3. Partial Face Recognition on CASIA-NIR-Distance and Partial-YTF

CASIA-NIR-Distance and Partial-YTF are used to evaluate the DFM.. We compare MKDSRC-GTP, RPSM, I2C, VGGFace, and FaceNet. one image per subject is selected to construct the gallery set and one different image per subject is selected to construct the probe set. Specifically, since none of the holistic faces of some subjects is captured by the iris recognition system at a distance, thus partial face images may exist in the gallery set, which increases the challenge of partial face recognition.

Table 5 shows the performance of the proposed DFM on CASIA-NIR-Distance. DFM achieves 94.96% rank-1 accuracy, which shows clearly that DFM performs much

Table 5. Performance comparison on CASIA-NIR-Distance and Partial-YTF.

Method	CASIA-NIR-Distance, p=276		
	$r = 1$	$r = 3$	$r = 5$
MR-CNN [10]	85.97	88.13	89.93
VGGFace [26]	35.25	38.85	40.65
MKDSRC-GTP [43]	83.81	85.25	86.69
RPSM [15]	77.70	80.22	83.45
I2C [11]	71.94	75.18	79.50
DFM	**94.96**	**96.40**	**97.84**
Method	Partial-YTF, p=200,		
	$r = 1$	$r = 3$	$r = 5$
MR-CNN [10]	57.00	65.50	71.00
VGGFace [26]	36.00	52.00	59.50
MKDSRC-GTP [43]	43.50	53.50	55.50
RPSM [15]	50.50	54.50	55.50
I2C [11]	51.50	55.50	57.00
DFM	**61.00**	**69.50**	**74.00**

better than other PFR approaches. The reasons are analyzed as follows: 1) DFM could represent a partial face more robustly in comparison with key-point-based algorithms MKDSRC-GTP (83.81%), RPSM (77.70%) and I2C (71.94%); 2) VGGFace achieves 28.13% rank-1 accuracy, they perform worse than DFM. The reason is that requirement of fixed-size input would generate unwanted face deformation. Similar to the results on CASIA-NIR-Distance, DFM perform the best on Partial-YTF.

5. Conclusion

We have proposed a novel approach called Dynamic Feature Matching (DFM) to address partial face recognition. Fully Convolutional Network (FCN) is used in generating spatial features with sharing computation regardless of the arbitrary size inputs. Dynamic feature dictionary that corresponds to the size of the probe is generated. In terms of feature matching, the similarity-guided constraint imposed on SRC provides an alignment-free matching, which effectively improves the performance of partial face recognition. Besides, we provided a sliding loss based on DFM that can improve the discrimination of deep features generated by FCN. The proposed DFM method has exhibited promising results on two simulated partial face databases, and CASIA-NIR-Distance database acquired from iris recognition systems. Furthermore, DFM can be easily extended to other visual recognition tasks such as partial person re-identification.

Acknowledgments This work is supported by the Beijing Municipal Science and Technology Commission (Grant No. Z161100000216144) and National Natural Science Foundation of China (Grant No. 61427811, 61573360). Special thanks to Yunfan Liu who supports our experiments.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(12):2037–2041, 2006. 3

[2] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *European Conference on Computer Vision (ECCV)*, pages 404–417, 2006. 3

[3] I. Cheheb, N. Al-Maadeed, S. Al-Madeed, A. Bouridane, and R. Jiang. Random sampling for patch-based face recognition. In *Proceedings of the IEEE International Workshop on Biometrics and Forensics (IWBF)*, pages 1–5, 2017. 3

[4] Y. Chen, L. Song, and R. He. Masquer hunter: Adversarial occlusion-aware face detection. *arXiv preprint arXiv:1709.05188*, 2017. 1

[5] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 3

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014. 3

[7] S. Gutta, V. Philomin, and M. Trajkovic. An investigation into the use of partial-faces for face recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2002. 3

[8] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(9):1904–1916, 2015. 3

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016. 2

[10] L. He, H. Li, Q. Zhang, Z. Sun, and Z. He. Multiscale representation for partial face recognition under near infrared illumination. In *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2016. 2, 3, 6, 8

[11] J. Hu, J. Lu, and Y.-P. Tan. Robust partial face recognition using instance-to-class distance. In *Visual Communications and Image Processing (VCIP)*, 2013. 3, 6, 8

[12] K. Huang and S. Aviyente. Sparse representation for signal classification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 609–616, 2007. 3

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012. 2

[14] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, pages 801–808, 2007. 4, 5

[15] S. Liao, A. K. Jain, and S. Z. Li. Partial face recognition: Alignment-free approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(5):1193–1205, 2013. 3, 8

[16] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. *arXiv preprint arXiv:1704.08063*, 2017. 1, 2

[17] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3431–3440, 2015. 2

[18] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004. 3

[19] U. Mahbub, V. M. Patel, D. Chandra, B. Barbello, and R. Chellappa. Partial face detection for continuous authentication. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2991–2995. IEEE, 2016. 1

[20] X. Mei and H. Ling. Robust visual tracking and vehicle classification via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(11):2259–2272, 2011. 3

[21] H. Neo, C. Teo, and A. B. Teoh. Development of partial face recognition framework. In *International Conference on Computer Graphics, Imaging and Visualization (CGIV)*, pages 142–146, 2010. 3

[22] M. Opitz, G. Waltner, G. Poier, H. Possegger, and H. Bischof. Grid loss: Detecting occluded faces. In *European Conference on Computer Vision (ECCV)*, pages 386–402. Springer, 2016. 1

[23] W. Ou, X. Luan, J. Gou, Q. Zhou, W. Xiao, X. Xiong, and W. Zeng. Robust discriminative nonnegative dictionary learning for occluded face recognition. *Pattern Recognition Letters*, 2017. 3

[24] K. Pan, S. Liao, Z. Zhang, S. Z. Li, and P. Zhang. Part-based face recognition using near infrared images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 3

[25] U. Park, A. Ross, and A. K. Jain. Periocular biometrics in the visible spectrum: A feasibility study. In

IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS), pages 1–6, 2009. 3

[26] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *British Machine Vision Conference (BMVC)*, volume 1, page 6, 2015. 1, 2, 3, 6, 8

[27] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 91–99, 2015. 3

[28] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(6):1137–1149, 2017. 2

[29] S. Ren, K. He, R. Girshick, X. Zhang, and J. Sun. Object detection networks on convolutional feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(7):1476–1481, 2017. 2

[30] K. Sato, S. Shah, and J. Aggarwal. Partial face recognition using radial basis function networks. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 288–293, 1998. 3

[31] M. Savvides, R. Abiantun, J. Heo, S. Park, C. Xie, and B. Vijayakumar. Partial & holistic face recognition on frgc-ii data using support vector machine. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2006. 3

[32] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 1, 2

[33] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(4):640–651, 2017. 2

[34] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1891–1898, 2014. 1

[35] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1708, 2014. 1, 2

[36] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision (ECCV)*, pages 499–515. Springer, 2016. 1

[37] R. Weng, J. Lu, and Y.-P. Tan. Robust point set matching for partial face recognition. *IEEE Transactions on Image Processing (TIP)*, 25(3):1163–1176, 2016. 3

[38] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 31(2):210–227, 2009. 3

[39] X. Wu, R. He, and Z. Sun. A lightened cnn for deep face representation. arxiv preprint. *arXiv preprint arXiv:1511.02683*, 4, 2015. 2

[40] Y. Xu, D. Zhang, J. Yang, and J.-Y. Yang. A two-phase test sample sparse representation method for use with face recognition. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 21(9):1255–1262, 2011. 3

[41] X.-T. Yuan, X. Liu, and S. Yan. Visual classification with multitask joint sparse representation. *IEEE Transactions on Image Processing (TIP)*, 21(10):4349–4360, 2012. 3

[42] L. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation: Which helps face recognition? In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 471–478, 2011. 3

[43] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong. Partial person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 4678–4686, 2015. 2, 6, 8