A pairwise learning strategy for video-based face recognition

Meng Zhang[1] Rujie Liu[1]
Hajime Nada[2] Hidetsugu Uchida[2] Tomoaki Matsunami[2] Narishige Abe[2]
[1]Fujitsu R&D Center Co.,LTD., Beijing, China
[2]Fujitsu Laboratories LTD., Japan
{zhangmeng, rjliu}@cn.fujitsu.com,
{nada.hajime, u.hidetsugu, t.matsunamj, abe.narishige}@fujitsu.com

Abstract

In recent years, large-scale datasets together with the emergence of deep learning have led to the immense success of face recognition. However, face recognition in surveillance scenarios is still challenging due to severe blur, dramatic occlusion, richer pose, and illuminations. Meanwhile, owing to the source of data and cleaning strategies, existing large-scale datasets are inevitably affected by label noise. In this paper, a pairwise learning strategy is proposed to overcome the challenges of abundant variants in video-based face recognition (VFR). In addition, an online effective example mining (OEEM) method is designed to eliminate noisy samples to force the model focus more on effective examples during training. Experimental results on LFW, COX and one selfie dataset validate the effectiveness of the proposed approach.

1. Introduction

Although considerable progress has been witnessed in face recognition area due to the emergence of many deep learning-based approaches [1 - 8], most of them were designed for still image based recognition (SIFR). When extended from still image-based face recognition (SIFR) to video-based face recognition (VFR), many methods tended to ignore the peculiarities of videos compared to SIFR, however, VFR is significantly more challenging. Images in standard SIFR datasets are usually captured under good conditions or even framed by professional photographers, e.g., the Labeled Faces in the Wild (LFW) [9] database. On the contrary, the image quality of video frames tends to be much lower and the faces in the video exhibit much richer variations because video acquisition is less constrained. In particular, subjects in videos are usually moving, resulting in serious motion blur, out-of-focus blur, and a large range of pose variations. Hence, it is necessary to design a model to overcome challenges for effective and robust video face recognition.

Large scale datasets have been commonly recognized as one of the key factors promoting the advance of face recognition, and many datasets of still face photos have been released in recent years, such as MegaFace [10] and MS-Celeb-1M [11]. Unfortunately, there is no such large-scale datasets for VFR. Therefore, how to effectively use the still data in VFR has become a hot research topic. A number of recent VFR studies attempt to make use of redundant information of still image such as frame quality evaluation [12 - 13], and extracting high-quality face representations [14-16]. Frame quality evaluation is mainly utilized for key frame selection from video clips, such that only a subset of best quality frames is selected for efficient face recognition. Extracting high-quality face representations have been introduced to reduce the impact of severe blur in VFR, which has been widely used in VFR applications.

Another problem related to data is the noisy label. As can be seen in [17], considerable incorrect identity labels can be found in both MegaFace and MS-Celeb-1M, where some erroneous labels are easy to remove while many of them are hard to be cleaned. The noisy label is a pervasive problem, since well-annotated datasets in large-scale are prohibitively expensive, which motivates researchers to resort to cheap but imperfect alternatives.

How to reduce noisy samples to help an algorithm focus more on effective samples will be a reasonable direction toward efficient utilization of imperfect dataset. We propose a Pairwise CNN (P-CNN) to efficiently learn face representations for VFR, where two CNN networks are incorporated, i.e., Base CNN (B-CNN) and Reinforcement CNN (R-CNN). B-CNN is trained to learn face representations from still face images directly, while R-CNN is trained simultaneously to learn face representations from video-like images. The video like images are artificially generated by manipulating the still images to simulate video condition such as adding motion blur and out of focus blur. During training, a pairwise learning strategy is adopted to force the output of R-CNN to be equal with B-CNN. In this way, the model can be trained to reduce the influence of the video-based variance.

In order to reduce the influence of noisy labels, an OEEM method is used to eliminate noisy and easy samples and force the model focus more on effective samples during training.

Our main contribution can be summarized as three folds:

(1) P-CNN structure: Incorporates B-CNN and R-CNN networks to efficiently learn face representations for VFR.
(2) A pairwise learning strategy: to overcome the challenges of large variants in VFR.
(3) OEEM: reducing noisy samples and easy samples to help an algorithm focus more on effective samples learning.

The paper is organized as follows. Section 2 briefly reviews related work on VFR and noisy label problems. Section 3 describes the proposed P-CNN framework, the pairwise learning strategy and OEEM method. Finally, we present experimental results in Section 4 and conclude this paper in Section 5.

2. Related work

We review the literature in two parts: 1) video-based face recognition, and 2) noisy label problems.

2.1. Video-based Face Recognition

Existing studies on VFR can be categorized into two classes: methods that exploit redundant information contained between video frames, and methods that extract high-quality face representations from each frame.

For the first category, a number of VFR studies have been proposed. The sequence-based methods [18-19] aim to extract person-specific facial dynamics from continuous video frames, which means that they rely on robust face trackers. The dictionary-based methods [20-21] construct redundant dictionaries using video frames and employ sparse representation-based classifiers for classification. Frame quality evaluation is mainly utilized for key frame selection from video clips, such that only a subset of best quality frames is selected for efficient face recognition [12-13].

For the second category, extracting high-quality face representations has always been a core task in face recognition [22]. In contrast to still face images, video frames usually suffer from severe image blur because of the relative motion between the subjects and the cameras. Deblur-based methods [14] estimate a blur kernel from the blurred image and then deblur the face image prior to feature extraction. Blur-robust feature extraction-based methods [23] were proposed to employ the blur-insensitive Local Phase Quantization (LPQ) descriptor for facial feature extraction, which has been widely used in VFR applications.

To the best of our knowledge, the majority of existing works directly employ CNN models trained on large-scale still image databases for VFR. One important reason is that large amounts of real-world video training data are still lacking, and direct CNN training using small volume of real-world video data is prone to overfitting. However, few CNN-based method has been used to handle the occlusion and blur problem in VFR. Therefore, a pairwise learning strategy is proposed to overcome the challenges of abundant variants in VFR.

2.2. Noisy label Problems

Large-scale datasets, such as MegaFace and MS-Celeb-1M datasets, play a main role in driving the recent development of face recognition, because the DNN based face recognition has become a data driven approach. However, existing large-scale datasets inevitably contain label noises owing to the source of data and cleaning strategies as well as the cost of manual cleaning. It was reported that face image datasets that are more than a million scale typically have a noise ratio higher than 30% [17]. Label noise may also bring instability to the model. A-Softmax, which usually achieves a better result on a clean dataset, was reported to be inferior to Center loss [5] and Softmax in the high-noise region.

Some methods [17] have been proposed in the literature to deal with noisy label problems, which can generally be classified into three categories. In the first category, robust loss functions [24] are designed for the classification tasks, in order to learn classification models robust to the presence of label noise. In the second category [25], however, the quality of training data is improved by identifying mislabeled instances. The third category methods [26] directly models the distribution of noisy labels during learning, with the advantage of leveraging the information about noisy labels during learning.

Just as mentioned above, collecting and cleaning a large scale dataset requires tremendous efforts, as a result, leaving the data in this form will be more reasonable in that it will encourage researchers to devise new learning methods that can naturally deal with the inherent noises.

3. Approach

In this section, we first describe the P-CNN, which incorporates B-CNN and R-CNN network, to efficiently learn face representations robust to occlusion and blur. Then, a pair-wise training strategy is introduced to effectively optimize the P-CNN parameters. Finally, an OEEM method is used to force the model focus more on effective samples during training.

3.1. Pairwise CNN (P-CNN)

The P-CNN structure incorporates B-CNN and R-CNN network, where, the two networks are trained to learn face representations for still face images and video images respectively. As we know, motion blur and out-of-focus blur are two important characteristics of video images, due to

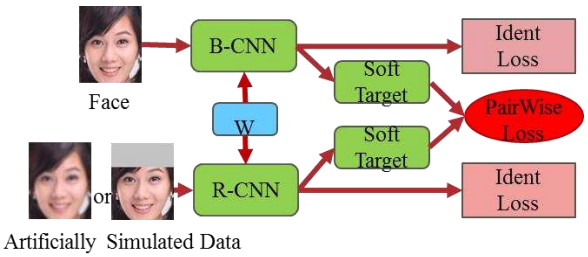

Figure 1, The Proposed Framework of P-CNN

face movement or mobile device camera shake during exposure, as well as the limited depth of field (DOF) of cameras and the large motion range of faces in videos. To solve such problems, a direct idea is to train a network with enough video data to learn such characteristic. Unfortunately, no such video corpus is available, so, we artificially generate video like images by manipulating the still images such as adding motion blur and out-of-focus blur. In learning, a Siamese network [27] is adopted as the basic framework of the P-CNN, to force the R-CNN and B-CNN networks have similar outputs, as shown in figure 1.

In this paper, P-CNN network is implemented based on ResNet, while the B-CNN and R-CNN share the same structure and weights. The main parameters are tabulated in table 1. In the table， Conv1.x, Conv2.x, Conv3.x and Conv4.x denote convolution units that may contain multiple convolution layers, and the residual units are shown in double-column brackets, e.g., $\left[3\times3,256\right]\times4$ denotes 4 cascaded convolution layers with 256 filters of size 3x3. S2 denotes that the stride is 2. FC is the abbreviation of fully connected layer.

3.2. A pairwise learning strategy

A pair-wise training strategy is proposed to effectively optimize the P-CNN parameters. Firstly, artificially simulated images are generated from still face images. The motion blur was simulated by one-dimensional local averaging of neighboring pixels [28], while the out-of-focus blur was simulated by a Gaussian kernel [16]. For the case of occlusion, images was generated by randomly removing some forehead region of the face.

The original still image and the artificially generated images are then input to the B-CNN and R-CNN networks respectively for feature extraction. The output softmax of B-CNN and R-CNN are denoted as follows:

$$P_B = soft\max(a_B) \tag{1}$$

$$P_R = soft\max(a_R) \tag{2}$$

where a_B and a_R are the pre-softmax activations vector of B-CNN and R-CNN.

P_R is forced to have similar value to P_B during training.

TABLE 1 P-CNN Network Parameters

Layer	20-layer CNN	64-layer CNN
Conv1.x	$\left[3\times3,64\right]\times1,S2$ $\left[\begin{matrix}3\times3,64\\3\times3,64\end{matrix}\right]\times1$	$\left[3\times3,64\right]\times1,S2$ $\left[\begin{matrix}3\times3,64\\3\times3,64\end{matrix}\right]\times3$
Conv2.x	$\left[3\times3,128\right]\times1,S2$ $\left[\begin{matrix}3\times3,128\\3\times3,128\end{matrix}\right]\times2$	$\left[3\times3,128\right]\times1,S2$ $\left[\begin{matrix}3\times3,128\\3\times3,128\end{matrix}\right]\times8$
Conv3.x	$\left[3\times3,256\right]\times1,S2$ $\left[\begin{matrix}3\times3,256\\3\times3,256\end{matrix}\right]\times4$	$\left[3\times3,256\right]\times1,S2$ $\left[\begin{matrix}3\times3,256\\3\times3,256\end{matrix}\right]\times16$
Conv4.x	$\left[3\times3,512\right]\times1,S2$ $\left[\begin{matrix}3\times3,512\\3\times3,512\end{matrix}\right]\times1$	$\left[3\times3,512\right]\times1,S2$ $\left[\begin{matrix}3\times3,512\\3\times3,512\end{matrix}\right]\times3$
FC1	512	512

Since P_B might be very close to the one hot code representation of the sample's true label, a relaxation parameter $\tau>1$ is introduced to soften the signal arising from the output of the B-CNN, and thus, more information can be provided during training. Similarly, the relaxation is also applied to the output of R-CNN. Thus, the soften softmax, P_B^τ and P_R^τ, can be obtained:

$$P_B^\tau = soft\max(\frac{a_B}{\tau}) \tag{3}$$

$$P_R^\tau = soft\max(\frac{a_R}{\tau}) \tag{4}$$

In order to make P_R^τ have similar value to P_B^τ, the pairwise loss was designed:

$$L_p = H(P_B^\tau, P_R^\tau) \tag{5}$$

where, $H(\bullet)$ represents cross entropy.

Besides the pairwise loss, traditional softmax loss functions are also applied to both B-CNN and R-CNN, denoted as LB and LR. The final loss function is thus the combination of the above three losses, as follows:

$$L = L_B + L_R + \lambda L_p \tag{6}$$

3.3. OEEM

By analyzing the loss during training, it was found that the easy samples usually have small loss values and thus contribute little to the gradient [29-30].In addition, we further found that the loss values of noise samples were very large, especially after several training epochs. Therefore, different from the traditional method of cleaning data first and then training model, we do OEEM in face classification

Figure 2, Images sample form O-LFW dataset

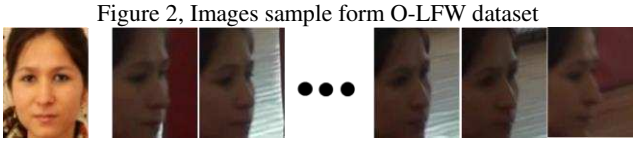

Figure 3, Images sample form COX dataset

task in the training process. Specifically, in each mini-batch, the loss values of all samples are computed and ranked in the forward propagation phase. Based on the loss values, the top ∂ percent of them is removed as noise samples and the bottom β percent of them is removed as easy samples. In the backward propagation phase, the gradient values are computed only for the remained effective samples. This method can not only ignore the easy samples that are less helpful to strengthen the recognition capability, but also eliminate noisy samples that may hamper the discrimination of the model. The OEEM strategy could help the algorithm automatically focus more on effective examples in learning. Moreover, the gradients calculation with a smaller mini-batch size becomes faster.

4. Experiment and results

4.1. Experiments Settings

The network model described in section 3.1 is used in our experiments. In prior to feature extraction, MTCNN [30] is adopted to detect the 5 landmark points in the face, after that, faces are aligned by similarity transformation according to the landmarks and cropped to 112x96 to remove the background information. Since B-CNN and R-CNN share the same network structure and weights, the deep features are extracted from the output of FC1 layer. The score is computed by the cosine distance of two features.

Training datasets: CASIA-WebFace and MS-Celeb-1M are two widely used datasets for training face recognition model. CASIA-WebFace contains 500K photos of 10K celebrities and it is semi-automatically cleaned via tag-constrained similarity clustering. MS-Celeb-1M contains 100K celebrities who are selected from the 1M celebrity list in terms of their popularities. Public search engines are then leveraged to provide approximately 100 images for each celebrity, resulting in about 10M web images. In its initial form, this data is deliberately left uncleaned for several reasons. Later, some researchers have made much effort to try to remove the incorrect labels and photos to avoid pitfalls in training. Unfortunately, some noise is still remained, e.g., the estimated noise percentage in MS-Celeb-1M is more than 45% [17].

Frontal Pose lighting WithoutGlasses WithGlasses

Figure 4, Image pair examples of FRDCMobile dataset.

Testing datasets: three datasets are used in the testing, two public datasets that are the well-known LFW [9] and COX Face DB [31], and one private dataset which is called FRDCMobile in this paper.

LFW is a widely used dataset for benchmarking face recognition approaches, which consists of 13,000 facial images of 1,680 celebrities.

The COX Face database incorporates 1,000 still images and 3,000 videos of 1,000 subjects. A high-quality camera in well-controlled conditions was used to capture still images to simulate ID photos, and the videos were taken while the subjects were walking in a large gym to simulate surveillance. Three cameras at different locations were installed to capture videos of the walking subject simultaneously. An example of video clip from COX is shown in Fig.3. Videos captured by the three cameras create three subsets, denoted as Cam1, Cam2, and Cam3. The standard matching protocols [31] proposed by the author, i.e., still-to-video (S2V), video-to-still (V2S), and video-to-video (V2V), were adopted in face identification evaluation.

The private dataset, FRDCMobile, contains photos captured by mobile phone from 1200 Chinese people. With the consideration of the variations in head pose, lighting, and accessory, different capturing conditions were designed such as with or without glasses, frontal or non-front, normal or dark lighting, etc. Some examples of FRDCMobile DB are shown in Figure 4. With the above setting, totally 88 photos were obtained for each subject, which leads to more than 100K image pairs in the identification test.

To verify the effectiveness of the proposed P-CNN in case of occlusion, we randomly add occlusion to the forehead region of the LFW face image to simulate the occlusion image. The new dataset is denoted as occlusion-LFW (O-LFW). The image pairs are kept same with the original LFW dataset, as shown in the fig.2.

4.2. Verification of OEEM strategy

To evaluate the contribution of the proposed OEEM strategy, we only use B-CNN to train the ResNet20 model with and without OEEM respectively. Besides this, both CASIA-WebFace and MS-Celeb-1M are used as the

TABLE 2 Accuracy (ACC) and TPR@FAR=0.1% on LFW, training with CASIA-WebFace dataset

Model	ACC	TPR@FAR=0.001
WO OEEM	99.22%+-0.40%	97.01%
W OEEM	99.35%+-0.34%	98.60%

TABLE 3 TPR@FAR=0.1% on FRDCMobile, training with CASIA-WebFace dataset

Model	TPR@FAR=0.001		
	Frontal	Pose	Light
WO OEEM	92.9%	75.9%	91.4%
W OEEM	94.5%	78.7%	91.9%

TABLE 4 Accuracy (ACC) and TPR@FAR=0.1% on LFW, training with Ms-CeleB-1M dataset

Model	ACC	TPR@FAR=0.001
WO OEEM	99.50%+0.34%	98.30%
W OEEM	99.55%+-0.29%	99.12%

TABLE 5 TPR@FAR=0.1% on FRDCMobile, training with Ms-CeleB-1M dataset

Model	TPR@FAR=0.001		
	Frontal	Pose	Light
WO OEEM	97.0%	84.2%	94.7%
W OEEM	98.9%	89.9%	97.8%

training corpus respectively, to reduce the risk of occasionality.

OEEM implementation follows the instructions described in section 3.3, and different parameter values are tested for ∂ and β in order to get the best accuracy. In our experiments, the value of ∂ is set to be $\{0.05, 0.1, 0.15, 0.2, 0.25\}$ while the value of β is set to be $\{0.1, 0.2, 0.3, 0.4\}$. It is worth noting that the best ∂ of CASIA-WebFace and MS-Celeb-1M are 0.1 and 0.15 respectively, while the best β are 0.3 and 0.2 respectively. The best ∂ of CASIA-WebFace is smaller than MS-Celeb-1M, which may be CASIA-WebFace is cleaner than MS-Celeb-1M. However, the best β of CASIA-WebFace is larger than MS-Celeb-1M, which may be MS-Celeb-1M have more rich pose, illuminations or another variants than CASIA-WebFace. Table 2-5 shows the identification accuracy on LFW and FRDCMobile data with CASIA-WebFace and Ms-CeleB-1M as training DB respectively.

On LFW, the models trained by CASIA-WebFace and MS-Celeb-1M with OEEM strategy outperform without OEEM by 0.13% and 0.05% respectively. On FRDCMobile, the models trained by CASIA-WebFace and MS-Celeb-1M with OEEM also outperform without OEEM, especially on FRDCMobile pose. The results suggested that the accuracy with OEEM is better than that without OEEM. However, the improvement of CASIA-WebFace is more obvious than MS-Celeb-1M on LFW. This is because the accuracy of model trained by MS-Celeb-1M has been very high on LFW, so the improvement is not obvious. Although the performance of model trained by MS-Celeb-1M is higher than model trained by CASIA-WebFace without OEEM strategy on FRDCMobile, the improvement of model trained by MS-Celeb-1M is still more obvious than model trained by CASIA-WebFace with OEEM strategy. This is because the noise ratio of MS-Celeb-1M is higher than CASIA-WebFace. It is suggested that OEEM can help an algorithm focus more on effective samples learning.

TABLE 6 Accuracy (ACC) and TPR@FAR=0.1% on O-LFW

Model	ACC	TPR@FAR=0.001
B-CNN	99.38+-0.31%	98.83%
P-CNN	99.58+-0.32%	99.13%

TABLE 7 TPR@FAR=0.1% on FRDCMobile

Model	TPR@FAR=0.001			
	WO-WO*	W-W*	WO-W*	W-WO*
B-CNN	99.9%	99.9%	99.6%	99.5%
P-CNN	99.9%	99.9%	99.9%	99.8%

* 'Wo' means without glasses, 'W' means with glass. WO-WO means gallery without glasses and probe without glasses; W-W means gallery with glass and probe with glass; WO-W means gallery without glasses and probe with glass; W-WO means gallery with glass and probe without glasses.

4.3. Effectiveness for Occlusion

This experiment is designed to evaluate the effectiveness of the proposed P-CNN for the case of occlusion. Because of the lack of publicly occluded face data sets, where, O-LFW and FRDCMobile datasets were used for test. Similar to previous test, ResNet20 is selected again as the basic network. The B-CNN model is firstly trained as a benchmark, based on which the P-CNN model is built using pairwise learning strategy. The results on the two datasets, O-LFW and FRDCMobile, are shown in Table 6 and Table 7.

On O-LFW, both the accuracy and true positive rate (TPR@FAR=0.001) of P-CNN is higher than B-CNN by around 0.2% and 0.3% respectively. In the case of accessory variations, e.g. with vs. without glasses, P-CNN outperforms B-CNN by around 0.3% and 0.3% on FRDCMobile set respectively. It is interesting to observe that although the artificially occluded images was generated by randomly adding occlusion, the performance of wearing glasses is improved. The result convincingly shows that the proposed pairwise learning strategy is essential to achieve occlusion-robustness in face recognition.

4.4. Performance on COX Face DB

In this part, we mainly focus on the evaluation of the proposed strategies in video face recognition, where, three experiments are designed: (1) B-CNN-WO, where B-CNN model is trained without blur data augmentation; (2) B-CNN-W, where data augmentation method is adopted by adding blur to the training data randomly; (3) P-CNN, which applies the proposed pairwise learning strategy. The

TABLE 9 Rank-1 Identification Rates (%) under the V2S Setting for Different Strategies on the COX Face Database.

Model	V2S_1	V2S_2	V2S_3
B-CNN-WO	93.01±0.49	85.40±0.77	98.01±0.14
B-CNN-W	93.13±0.57	88.59±0.97	97.77±0.17
P-CNN	95.01±0.20	90.50±0.70	98.91±0.14

TABLE 10 Rank-1 Identification Rates (%) under the V2S Setting for Different Methods on the COX Face Database

Model	V2S_1	V2S_2	V2S_3
PSCL[31]	38.60±1.39	33.20±1.77	53.26±0.80
LERM[32]	45.71±2.05	42.80±1.86	58.37±3.31
VGG Face[33]	88.36±1.02	80.46±0.76	90.93±1.02
TBE-CNN[16]	93.57± 0.65	93.69±0.51	98.96±0.17
P-CNN(ResNet20)	95.01±0.20	90.50±0.70	98.91±0.14
P-CNN(ResNet64)	97.69±0.22	93.83±0.46	99.46±0.12

COX dataset is used for evaluation, and the experimental results are shown in Table 9. In this table, the number *i* in V2S_*i* denotes the video captured by camera *i* is used as probe, while the still images are kept as the gallery.

It was shown from Table 9 that the performance of B-CNN-W is better than B-CNN-WO, but the improvement is not obvious, even slightly decreased in V2S_3. However, P-CNN achieves the best performance in comparison with training model directly.

Finally, we compare the face verification performance of P-CNN with state-of-the-art approaches on the COX Faces database. Generally speaking, larger backbone networks yield higher accuracy. At present, the mainstream face recognition uses large models, for example, TBE-CNN[16] implementation is based on GoogLeNet. Therefore, besides the ResNet20 P-CNN model built in the previous experiments, an even deeper model, ResNet64, is also built. The rank-1 identification rate is adopted as the comparison criterion, and the results are tabulated in Table10. In comparison with the strategy of training model directly by augmented data, Table 10 shows that our proposed method can achieve better performance, and it get superior result than state-of-the-art methods.

5. Discussion

In this paper, we propose a pairwise learning strategy to overcome challenges in VFR. In addition, we propose an OEEM method to reduce noisy samples to help model focus more on effective samples during training. The noisy label problem is pervasive since some noisy labels are easy to remove while many of them are hard to be cleaned. Therefore, well-annotated datasets in large-scale are prohibitively expensive and time-consuming to collect. That motivates researchers to resort to cheap but imperfect alternatives. While our method is designed for video face recognition, it can also be applied in other computer vision tasks, especially for other face applications such face detection, tracking, which is an interesting future work. Although some strategies have been studied for noisy label problem, massive noisy label is still an open issue for deep learning methods.

References

[1] Sun, Yi, et al. "Deep learning face representation by joint identification-verification." Advances in neural information processing systems. 2014.

[2] Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. "Deep face recognition." BMVC. Vol. 1. No. 3. 2015.

[3] Sun, Yi, Xiaogang Wang, and Xiaoou Tang. "Deep learning face representation from predicting 10,000 classes." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[4] Sun, Y., et al. "Face recognition with very deep neural networks." (2015).

[5] Wen, Yandong, et al. "A discriminative feature learning approach for deep face recognition." European Conference on Computer Vision. Springer, Cham, 2016.

[6] Liu, Weiyang, et al. "Sphereface: Deep hypersphere embedding for face recognition." The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 1. 2017.

[7] Wang, Hao, et al. "Cosface: Large margin cosine loss for deep face recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[8] Deng, Jiankang, et al. "Arcface: Additive angular margin loss for deep face recognition." arXiv preprint arXiv:1801.07698 (2018).

[9] Learned-Miller, Erik, et al. "Labeled faces in the wild: A survey." Advances in face detection and facial image analysis. Springer, Cham, 2016. 189-248.

[10] Kemelmacher-Shlizerman, Ira, et al. "The megaface benchmark: 1 million faces for recognition at scale." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[11]Guo, Yandong, et al. "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition." European Conference on Computer Vision. Springer, Cham, 2016.

[12] J. Phillips, J. R. Beveridge, D. S. Bolme, B. Draper, G. H. Givens, Y. M. Lui, S. Cheng, M. N. Teli, H. Zhang et al., "On the existence of face quality measures," in Proc. IEEE Int. Conf. Biometrics, Theory, Appl. Syst.,2013, pp. 1–8.

[13] S. Mau, S. Chen, C. Sanderson, and B. C. Lovell, "Video face matching using subset selection and clustering of probabilistic multi-region histograms," arXiv preprint arXiv:1303.6361, 2013.

[14] M. Nishiyama, A. Hadid, H. Takeshima, J. Shotton, T. Kozakaya, and O. Yamaguchi, "Facial deblur inference

using subspace analysis for recognition of blurred faces," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 4, pp. 838–845, 2011.
[15] R. Gopalan, S. Taheri, P. Turaga, and R. Chellappa, "A blur-robust descriptor with applications to face recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 6, pp. 1220–1226, 2012.
[16]Ding, Changxing, and Dacheng Tao. "Trunk-branch ensemble convolutional neural networks for video-based face recognition." IEEE transactions on pattern analysis and machine intelligence 40.4 (2018): 1002-1014.
[17] Wang, Fei, et al. "The devil of face recognition is in the noise." Proceedings of the European Conference on Computer Vision (ECCV). 2018.
[18] M. Bicego, E. Grosso, and M. Tistarelli, "Person authentication from video of faces: a behavioral and physiological approach using pseudo hierarchical hidden markov models," in Advances in Biometrics, 2006, pp. 113–120.
[19] A. Hadid and M. Pietik¨ainen, "Combining appearance and motion for face and gender recognition from videos," Pattern Recognit., vol. 42, no. 11, pp. 2818–2827, 2009.
[20] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa, "Dictionarybased face recognition from video," in Proc. Eur. Conf. Comput. Vis., 2012, pp. 766–779.
[21] L. Liu, L. Zhang, H. Liu, and S. Yan, "Toward large-population face identification in unconstrained videos," IEEE Trans. Circuits Syst. Video Technol., vol. 24, no. 11, pp. 1874–1884, 2014.
[22] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," ACM Trans. Intell. Syst. Technol., vol. 7, pp. 37:1–37:42,2016.
[23] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila, "Recognition of blurred faces using local phase quantization," in Int. Conf. Pattern Recognit., 2008, pp. 1–4.
[24] E. Beigman and B. B. Klebanov, "Learning with annotation noise," in Annual Meeting of the ACL, 2009.
[25] D. R. Wilson and T. R. Martinez, "Instance pruning techniques," in International Conference on Machine Learning, 1997.
[26] N. D. Lawrence and B. Sch¨olkopf, "Estimating a kernel fisher discriminant in the presence of label noise," in International Conference on Machine Learning, 2001.
[27] Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification[C]//Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005, 1: 539-546.
[28] R. Wang and D. Tao, "Recent progress in image deblurring," arXiv preprint arXiv:1409.6838, 2014.
[29] Shrivastava, Abhinav, Abhinav Gupta, and Ross Girshick. "Training region-based object detectors with online hard example mining." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
[30]Zhang, Kaipeng, et al. "Joint face detection and alignment using multitask cascaded convolutional networks." IEEE Signal Processing Letters 23.10 (2016): 1499-1503.
[31] Z. Huang, S. Shan, R.Wang, H. Zhang, S. Lao, A. Kuerban, and X. Chen, "A benchmark and comparative study of video-based face recognition on cox face database," IEEE Trans. Image Process., vol. 24, no. 12, pp. 5967–5981, 2015.
[32] Z. Huang, R. Wang, S. Shan, and X. Chen, "Learning euclidean-toriemannian metric for point-to-set classification," in Proc. IEEE Conf.
Comput. Vis. Pattern Recognit., 2014, pp. 1677–1684
[33] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in Proc. Brit. Mach. Vis. Conf., 2015, pp. 1–12.