2018 13th IEEE International Conference on Automatic Face & Gesture Recognition

What is the Challenge for Deep Learning in Unconstrained Face Recognition?

Guodong Guo[1,2] and Na Zhang[2]
[1] Beijing Advanced Innovation Center for Imaging Technology, Beijing 100048, China.
[2] Dept. of Comput. Sci. and Electr. Eng., West Virginia University, Morgantown, WV 26506, USA.

***Abstract*— Recently deep learning has become dominant in face recognition and many other artificial intelligence areas. We raise a question: Can deep learning truly solve the face recognition problem? If not, what is the challenge for deep learning methods in face recognition? We think that the face image quality issue might be one of the challenges for deep learning, especially in unconstrained face recognition. To investigate the problem, we partition face images into different qualities, and evaluate the recognition performance, using the state-of-the-art deep networks. Some interesting results are obtained, and our studies can show directions to promote the deep learning methods towards high-accuracy and practical use in solving the hard problem of unconstrained face recognition.**

I. INTRODUCTION

Deep learning (DL) [1], [2] has recently become dominant in a wide variety of biometrics problems and many other artificial intelligence (AI) areas. One of the greatest successes of DL has been in face recognition (FR) where the accuracies have been improved greatly over the traditional methods [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15]. We raise a question: Can deep learning really solve the face recognition problem? Or, can we say that, given the great success of DL, the face recognition problem has been solved or almost solved?

To answer this question, and get a better understanding of the performance of DL methods in FR, we perform an empirical study with designed experiments accordingly.

In face recognition, it is well-known that the hard problem is unconstrained face matching, in which we believe that the face image quality variations are probably the biggest issue that makes the problem hard. Based on this view, our conjecture is that the face image quality issue may still be a grand challenge even for the recently developed DL methods.

In FR with traditional features, it is well-known that the face image quality has a big influence on recognition accuracy; In DL features, however, a large dataset with face images of different quality for each subject, is used to train the deep models. Will the quality still be an issue?

In our empirical study, we design the face recognition experiments with matching across different face image qualities, which is seldom done in an explicit way in previous face recognition approaches. In practice, however, one can meet the cross-quality face matching problem frequently. For example, in the FBI's interstate photo system (IPS), millions of mugshot photos could be matched to face images collected from social media web sites where the wild photos may have a wide variety of qualities.

This was partly supported by a NSF-CITeR grant and a WV HEPC grant.

Since there is no existing face database, which is purposely assembled with annotated face image qualities, we partition some recent, public face databases into different face image qualities, using an automated face image quality assessment method. After the quality partition, the face images from the same subject are divided into different qualities, such as low, middle, and high. Then we can perform face recognition experiments across quality changes.

For the deep learning techniques, we select some representatives of the state-of-the-art. To avoid any bias in training and parameter tuning, we adopted the already-trained face models that have reported very high accuracies in the popular face database LFW (labeled faces in the wild) [16].

The contributions of our work include:

- An important problem is raised for deep learning, through investigating the impact of face image quality changes on deep learning techniques;
- "Annotations" of face image qualities are performed on two public face databases, which is for the *first time* to perform quality partition, to the best of our knowledge. This partition can be useful for examining the face image quality issue in unconstrained face recognition;
- The design of cross-quality face recognition protocols is useful to discover the real challenges in unconstrained face recognition, rather than simply saying "in the wild" where a number of high quality face images may be matched to each other with high accuracies;
- An evaluation of the performance of the state-of-the-art deep learning techniques in cross-quality face recognition, disclosing the capability of deep learning methods in cross-quality face matching, an important problem but not well-studied yet, in unconstrained FR.

The paper is organized as follows. In Section II, we introduce the automated annotations of face image qualities on two public databases, using a face image quality assessment method; The protocols of cross-quality face matching are designed as well. In Section III, we briefly describe the representative deep models that we used for the evaluation. In Section IV, the FR evaluations are executed. A discussion is given in Section V, and finally we draw some conclusions.

978-1-5386-2335-0/18/$31.00 ©2018 IEEE
DOI 10.1109/FG.2018.00070

II. Face Image Quality, Dataset, and Protocol

Probably the LFW [16] is the most popular face database used for face recognition in the wild. However, each subject has a very limited number of face images in LFW, making it difficult to investigate different variations of face image qualities for each subject. Further, "in the wild" does not mean low quality face images. There could be high quality face photos collected in the wild. In our view, the key issue in unconstrained face recognition is the face image quality changes. We should examine the impact of face image quality changes in order to have a better understanding of the difficulty, rather than simply saying "in the wild", or "unconstrained". If the majority of face images are with high quality in an unconstrained face database, the recognition accuracies might be high, concealing the true challenges. Furthermore, if multiple unconstrained face databases are available, how to compare their levels of recognition difficulty? Our quality assessment approach could give indications of how challenging each database could be.

In our study, we explicitly partition face images into different qualities, and then evaluate the performance of face recognition across quality variations. We believe that this is the way to find the real challenges in unconstrained FR.

A. Face Image Partition based on Quality

Face image quality assessment is an active research in face recognition, e.g., [17]. We selected to use a recent approach [18] to measure the face image quality for each face image. The key idea in the method [18] is that the relative qualities between pairs of face images are measured and used as the input to a ranking-based support vector machine (SVM) learning method. After learning, each test face image can be used as input, and the SVM function can output a quality score, in the range of 0 to 100. The higher the score value, the higher the face image quality.

Given the quality scores, we divide the face images into three quality levels: low, middle and high. When the quality scores are below 30, the face images are classified as low quality; When a quality score is greater than or equal to 30 but less than 60, the face image is classified into the middle quality; If the quality score is above 60, the face image is considered as high quality. The threshold values of 30 and 60 are selected based on a visual check of the face image qualities, and the three-category classification is to make the quality issue manageable in our empirical study.

B. Quality Partition on Two Databases

The quality partition of face images is performed on two public databases, the IJB-A and FaceScrub. The two databases were assembled recently, where each subject has many face images available with various quality changes. These databases are significantly different from the LFW, more appropriate to investigate some critical issues in unconstrained face recognition. The traditional CMU-PIE database [19] is not appropriate for studying unconstrained face recognition, since it was collected under a controlled environment.

1) IJB-A: The IARPA Janus Benchmark A (IJB-A) database [20], is a publicly available face in the wild dataset, containing 500 celebrities of 21,230 face images. The face regions were also manually localized. The IJB-A dataset contains a wider geographic variation of subjects, and their original protocol for face recognition has no consideration of face image quality issue. We performed an automated face image quality assessment for IJB-A, and the resulting partition is shown in Table I. There are more face images with the middle level quality, and a much smaller number of high quality face images.

TABLE I
Quality partition of face images in IJB-A.

Quality Set	# of Images	# of Subjects
High	1,543	500
Middle	13,491	483
Low	6,196	489

To illustrate the quality partition of face images, some example faces from IJB-A are shown in Fig. 1, where the qualities are changed from high to middle, and to low, shown from top to the bottom.

2) FaceScrub: The FaceScrub database [21] was collected from the Internet through searching for public figures. It consists of a total of 106,863 face images of 530 celebrities, about 200 images per subject. There are 55,306 face images of 265 males and 51,557 face images of 265 females.

After performing the face image quality partition on the whole database, we found that the FaceScrub database has a large percentage of good quality face images. Specifically, there are more than 70% of face images with high quality, and about 25% of the photos are with middle level quality. This is a case to show that the face images "in the wild" are not necessarily with low qualities.

Considering the cost of time and memory requirement in running the code, and matching the database size to IJB-A, we randomly selected 10,089 face images in high quality, and 10,444 face images in middle quality. For the low quality face images, we keep as many as possible, resulting in 362 low quality face images. In total, the selected face database from FaceScrub contains 20,895 face images of 530 subjects. See Table II for the numbers after quality partition in the partially selected FaceScrub dataset. It can be observed that the number of low quality face images is much smaller than the middle and high levels.

C. Recognition Protocol

We design recognition protocols with both identification and verification. In either case, the matching of faces is always across quality changes. In identification, we have gallery and probe sets where the face image quality is different between the two sets. In verification, we generate all positive and negative pairs, where face photos of different quality are put into each pair.

1) Face Identification: Face identification is to match between the gallery and probe face images. Three types of

Fig. 1. Illustration of the face images in IJB-A that are partitioned into different qualities: Top: high, Middle: middle, and Bottom: low quality.

TABLE II
QUALITY PARTITION OF THE SELECTED FACE IMAGES IN FACESCRUB.

Quality Set	# of Images	# of Subjects
High	10,089	530
Middle	10,444	530
Low	362	232

identification are developed: (1) matching between low and high quality faces, where the gallery contains high quality face images of all subjects, while the probe set contains low quality face images of all available subjects; (2) matching between middle and high, where the gallery contains high quality face images of all subjects, while the probe set contains middle quality face images; (3) matching between low and middle, where the gallery contains middle quality face images, while the probe set contains low quality face images of the subjects.

The identification protocol is used for both the IJB-A and FaceScrub databases. For similarity measure, we use the cosine similarity, computed between two faces A and B,

$$Similarity = \frac{A \cdot B}{\|A\|_2 \|B\|_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

where n is the total number of deep features extracted from each face image, using each of the deep models.

The identification performance is measured by the Cumulative Match Curve (CMC) [22].

2) Face Verification: Face verification is to have a set of pairwise comparisons between face images. In our design, each pair of faces are with different qualities.

All pairs are generated for verification. For the IJB-A database, in the verification of low to high quality faces, there are 18,978 positive pairs and 9,541,450 negative pairs; in middle to high quality verification, there are 41,642 positive pairs and 20,774,971 negative pairs. Since in the identification experiments, we found that the match from low to high quality has similar performance to the case of from low to middle, we do not include the case of low to middle in our verification study.

For the FaceScrub database, there are 6,676 positive pairs and 3,645,542 negative pairs in low to high quality face verification; There are 193,745 positive pairs and 105,175,771 negative pairs in middle to high quality verification. Table III shows the number of pairs in each verification.

TABLE III
THE NUMBER OF PAIRS IN VERIFICATION FOR THE TWO DATASETS.

DataSet	Pairs	Low vs. High	Middle vs. High
IJB-A	Positive Pairs	18,978	41,642
	Negative Pairs	9,541,450	20,774,971
FaceScrub	Positive Pairs	6,676	193,745
	Negative Pairs	3,645,542	105,175,771

The cosine function is used for similarity measure between two faces. The verification accuracies are computed with respect to FAR=0.01 and 0.001 (FAR: false accept rate). And the Receiver Operating Characteristic (ROC) curves are drawn to show the performance visually.

III. DEEP LEARNING METHODS

We choose four representative deep models, VGGFace [23], Light CNN [24], CenterLoss [25], and FaceNet [13],

for our evaluation and comparisons. To avoid any bias in training and parameter tuning, we adopted their already-trained face models for these deep networks, which have reported similarly high accuracies in LFW. These four deep face models can be considered as the representatives of the state-of-the-art in face recognition.

A. Light CNN

The Light CNN model [24] introduces a concept called Max-Feature-Map (MFM) operation, which is a special case of maxout. The MFM is defined to simulate neural inhibition for a compact representation and feature filter selection. It suppresses a neuron by a competitive relationship. Light CNN also integrates Network in Network (NIN) and small convolution kernel sizes in order to achieve better performance in terms of speed and storage space.

There are three types of Light CNN architectures (4-layer, 9-layer and 29-layer) with a 256-D representation. In our study, we use a 29-layer Light CNN with residual blocks of two 3*3 convolution layers and two MFM operations without batch normalization.

B. FaceNet

The FaceNet [13] directly learns an embedding mapped from the input to a Euclidean space in which the Euclidean distance indicates the face similarity. It uses triplets of tightly cropped face patches generated by a novel online triplet mining method to train the network, and its output is a compact 128-D embedding. The rectified linear units are used as the non-linear activation function. FaceNet is constructed with a batch input, a deep convolutional network, L2 normalization, and the triplet loss layers. Note the used FaceNet is from a public domain, since the original is private.

C. VGGFace

The VGGFace [23] is a deep network inspired by the work in [26]. It contains a long sequence of convolutional layers. This network is bootstrapped as classifiers. Each training face image is associated with a score vector generated by the final fully-connected layer containing N linear predictors, one per identity. The network computes the empirical softmax log-loss to compare the scores with the ground-truth class identity. VGGFace uses a triplet loss function in training to improve the overall performance, which is similar to the FaceNet [13]. The output is L2 normalized.

D. CenterLoss

The CenterLoss model [25] introduces a new loss function called center loss. It learns a center of deep features in each class and minimizes the distances between the deep features and their corresponding class centers. The CenterLoss model is trained with joint supervision of the softmax and center losses. A hyper parameter is used to balance the two supervision signals. The joint supervision enlarges the inter-class feature difference, reduces the intra-class variation, and enhances the discriminative power.

IV. Face Recognition Evaluation

To evaluate the face recognition performance of the representative deep models, we use the protocols introduced in Section II. Our emphasis is the cross-quality face matching, in order to understand the behaviors of various deep models in different cases, and discover the challenges for deep learning methods in unconstrained face recognition.

We perform both identification and verification experiments on the two databases, IJB-A and FaceScrub.

A. Identification

In face identification, the gallery and query faces are with different face image qualities. The CMC curves of the recognition results on IJB-A are shown in Fig. 2, where the results from each deep model are shown in one sub-figure. For each deep model, the identification is executed in three cases: low quality to high, low to middle, and middle to high. From Fig. 2, one can see clearly that the matching from middle to high is significantly higher than the other two cases, no matter which deep model is used for face image representation. This consistent difference indicates that the deep models can perform much better in matching middle quality to high, while significantly worse in matching low quality to high, or low to middle.

One can also notice that the four deep models perform differently in our identification experiments, although they can perform similarly well on the LFW [16]. Through the quality partition of face images, we can have some deep insights into the capability of different deep learning methods, and dig deeper the problem of unconstrained face recognition. For instance, different models perform quite differently in different cross-quality scenarios. The VGGFace, Light CNN, and CenterLoss models have similar recognition accuracies in the matching between middle and high quality faces, while the VGGFace can perform better in the other two cases: from low to high or from low to middle quality.

In the two cases of matching from low quality to high and low to middle, the recognition accuracies are close, while the matching between low and high is slightly less accurate than the matching between low and middle.

On the FaceScrub database, the identification results are shown in Fig. 3. The behaviors of the four deep models are similar to those on IJB-A. That is, the recognition accuracies in matching between middle and high quality are significantly better than the other two cases, no matter which deep model is used. Again, this indicates that the face image quality changes can be a big challenge for deep learning methods in unconstrained face recognition.

B. Verification

We also perform face verification on the two databases, based on the protocols presented in Section II. The ROC curves are used to measure and compare the verification performance. The verification results on IJB-A and FaceScrub are shown in Figs. 4 and 5, respectively. Similar observations can be obtained on the two databases: The VGGFace, Light CNN, and CenterLoss models can perform equally well in

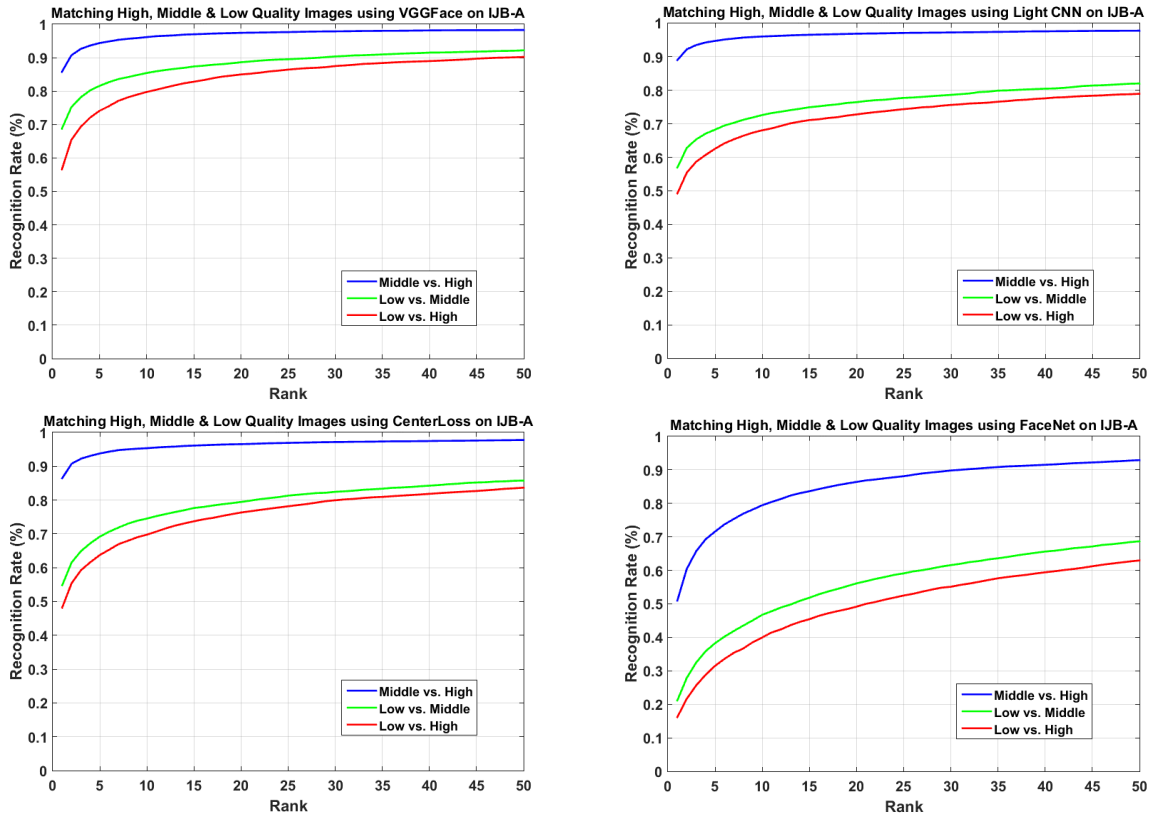

Fig. 2. Face identification across quality changes on the IJB-A database: Top Left: VGGFace, Top Right: Light CNN, Bottom Left: CenterLoss, and Bottom Right: FaceNet. The matching is performed in three cases for each model: Middle vs. High, Low vs. Middle, and Low vs. High.

matching between middle and high quality face pairs, while they perform much worse in the case of matching between low quality and high quality face pairs. The Gabor features are used as the baseline method for face matching, which is popular in traditional face recognition approaches, but it performs much worse than any of the deep models.

The verification accuracies over all positive and negative pairs at different FARs are also computed and shown in Table IV. The upper half is on the IJB-A database, and the bottom half is on the FaceScrub. All four deep models are evaluated at two different FARs. The accuracies of matching between low quality and high quality face pairs are much worse than matching between middle and high quality face pairs.

TABLE IV
VERIFICATION ACCURACIES AT FAR = 0.01 AND 0.001, RESPECTIVELY.

DataSet	Model	Low vs. High		Middle vs. High	
		FAR=0.01	0.001	0.01	0.001
IJB-A	VGGFace	0.605	0.367	0.858	0.675
	Light CNN	0.566	0.402	0.905	0.808
	CenterLoss	0.521	0.313	0.859	0.692
	FaceNet	0.257	0.100	0.586	0.330
FaceScrub	VGGFace	0.595	0.389	0.837	0.662
	Light CNN	0.503	0.330	0.896	0.811
	CenterLoss	0.493	0.341	0.914	0.814
	FaceNet	0.219	0.075	0.633	0.350

V. DISCUSSION

In training the deep networks for face representation, typically a variety of face images with different qualities are used in the training set. For example, the WebFace database [27] is often used for training the face models, which contains face images of different qualities for each subject. Theoretically, the deep networks have "seen" face images of various or mixed qualities in learning, they may already build some kinds of "connections" between faces of different qualities. However, in practice, the matching between different qualities is not trivial. For instance, in our evaluations of the representative deep models, it shows that the deep models still have difficulty in matching face images from low to high qualities, even though the matching from middle to high can get very high accuracies. Thus, we can say that the deep models can allow quality changes to some degrees, but not too large, for the test face images.

Based on our evaluation of cross-quality FR, we believe that one of the grand challenges for deep learning is the significant quality changes between face images in matching. Based on this observation, one promising direction for deep learning is to further improve its capability in building the relations between face images with large quality gaps.

Our quality partition of face images can also be useful for comparing multiple unconstrained face databases, even

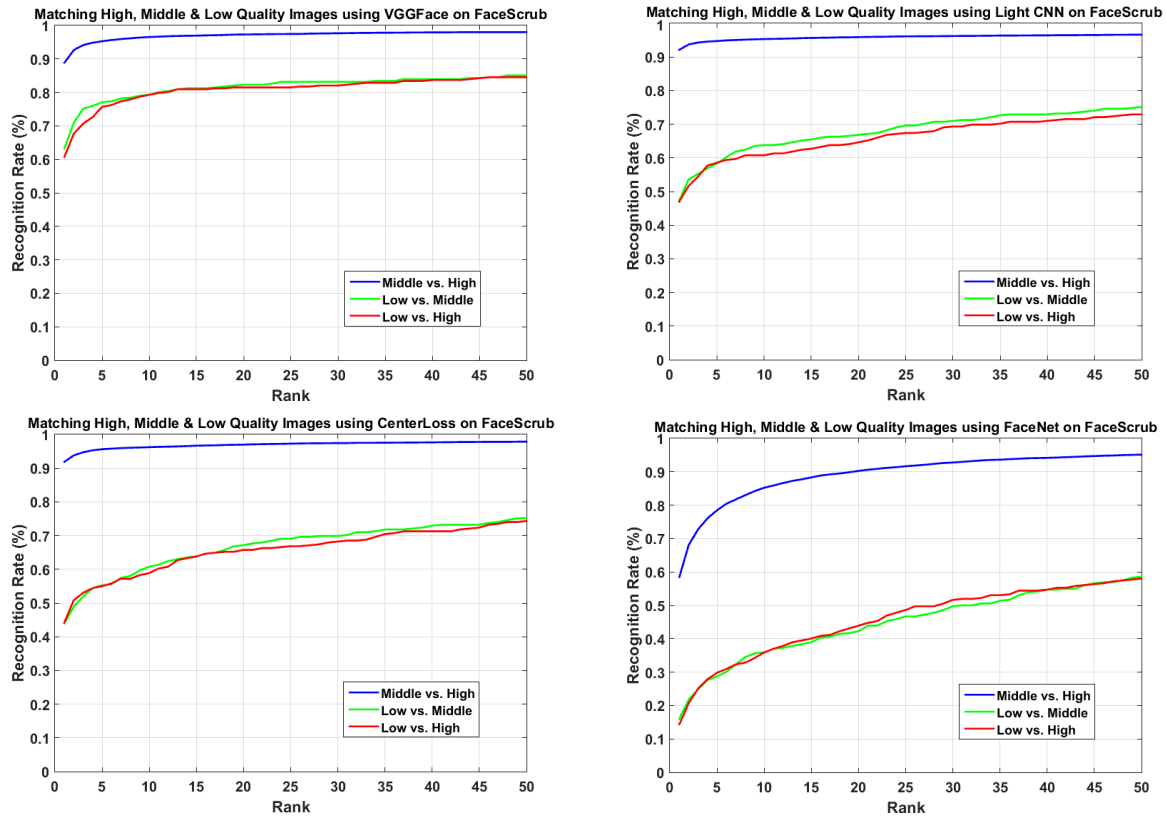

Fig. 3. Face identification across quality changes on the FaceScrub database: Top Left: VGGFace, Top Right: Light CNN, Bottom Left: CenterLoss, and Bottom Right: FaceNet. The matching is performed in three cases for each model: Middle vs. High, Low vs. Middle, and Low vs. High.

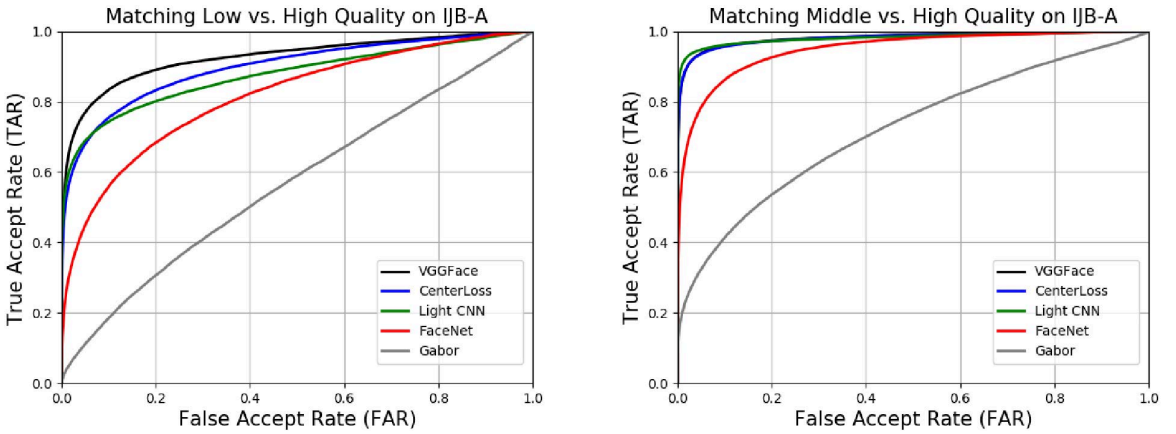

Fig. 4. Verification in face image pairs across quality changes: Low vs. High (left) and Middle vs. High (right), on IJB-A dataset.

without performing any FR experiments. For instance, if one unconstrained face dataset has much more high quality face images than others, it may be easier to perform face recognition on this dataset, and the recognition accuracies might be high without a big effort. Furthermore, in assembling an unconstrained face database, one can check the percentage of low quality face images, and thus to control the levels of challenges for the new database. For example, if the low quality face images are removed from the two databases, IJB-A and FaceScrub, both databases could report high recognition accuracies with the current methods, based on our FR evaluation.

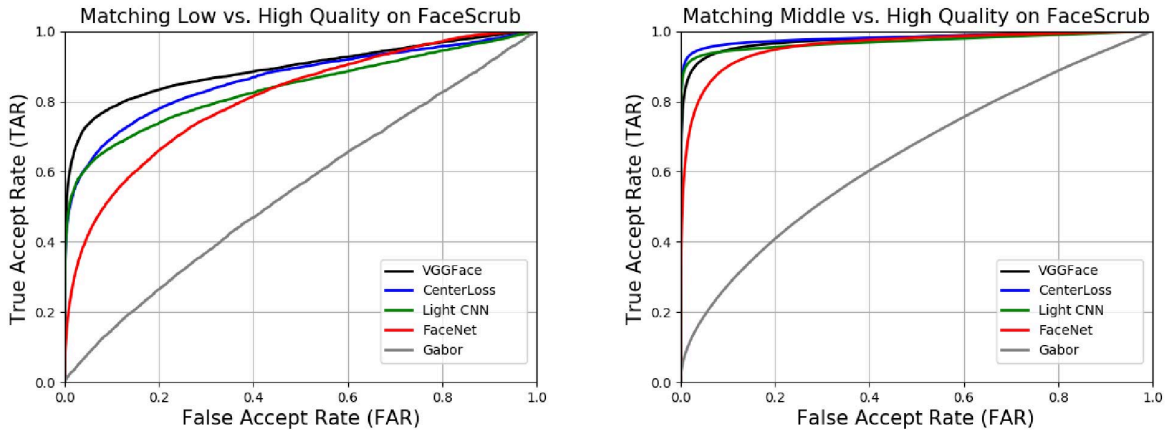

Fig. 5. Verification in face image pairs across quality changes: Low vs. High (left) and Middle vs. High (right), on FaceScrub.

VI. Conclusion

We have proposed to partition face images based on quality for investigating critical issues in unconstrained face recognition. Based on quality partition, we have developed FR protocols for cross-quality face identification and verification on two public databases. Some representative deep learning methods have been evaluated under our settings for unconstrained FR. We have shown that the face image quality variations are a grand challenge for deep learning in performing unconstrained FR, even though a variety of face images have been fed into the training of deep networks. Our study suggests the direction to promote deep learning techniques towards high-accuracy recognition in practice.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[3] G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *IEEE Conf. on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2518–2525.

[4] Y. Sun, X. Wang, and X. Tang, "Hybrid deep learning for face verification," in *Proc. of the IEEE Int'l Conf. on Computer Vision*, 2013, pp. 1489–1496.

[5] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Recover canonical-view faces in the wild with deep neural networks," *arXiv preprint arXiv:1404.3543*, 2014.

[6] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 1875–1882.

[7] H. Fan, Z. Cao, Y. Jiang, Q. Yin, and C. Doudou, "Learning deep face representation," *arXiv preprint arXiv:1403.2802*, 2014.

[8] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[9] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, 2014, pp. 1988–1996.

[10] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 2892–2900.

[11] E. Zhou, Z. Cao, and Q. Yin, "Naive-deep face recognition: Touching the limit of lfw benchmark or not?" *arXiv preprint arXiv:1501.04690*, 2015.

[12] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015.

[13] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2015, pp. 815–823.

[14] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang, "Targeting ultimate accuracy: Face recognition via deep embedding," *arXiv preprint arXiv:1506.07310*, 2015.

[15] C. Lu and X. Tang, "Surpassing human-level face verification performance on lfw with gaussianface." in *AAAI*, 2015, pp. 3811–3819.

[16] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.

[17] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *Computer Vision and Pattern Recognition Workshops*. IEEE, 2011, pp. 74–81.

[18] J. Chen, Y. Deng, G. Bai, and G. Su, "Face image quality assessment based on learning to rank," *IEEE signal processing letters*, vol. 22, no. 1, pp. 90–94, 2015.

[19] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression (pie) database," in *Proc. of Int'l Conf. on Automatic Face and Gesture Recognition*. IEEE, 2002, pp. 53–58.

[20] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 1931–1939.

[21] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *IEEE Int'l Conf. on Image Processing*. IEEE, 2014, pp. 343–347.

[22] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior, "The relation between the roc curve and the cmc," in *Fourth IEEE Workshop on Automatic Identification Advanced Technologies*. IEEE, 2005, pp. 15–20.

[23] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition." in *BMVC*, vol. 1, no. 3, 2015, p. 6.

[24] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *arXiv preprint arXiv:1511.02683*, 2015.

[25] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[27] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.