HIGH-RESOLUTION FACE RECOGNITION VIA DEEP PORE-FEATURE MATCHING

Shun-Cheung Lai[1], Minna Kong[1], Kin-Man Lam[1], and Dong Li[2]

[1]Department of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Hong Kong
[2]School of Automation, Guangdong University of Technology, Guangzhou, China

ABSTRACT

Because of the advancement of capturing devices, both image resolution and image quality have been significantly improved. Efficiently utilizing facial information is beneficial in enhancing the performance of face recognition methods. For high-resolution face images, pore-scale facial features can be observed. The positions and local patterns of pore features are biologically discriminative, so they can be explored for face identification. In this paper, we extend the previous work on pore-scale features, by proposing a new learning-based descriptor, namely PoreNet. Experiment results show that our proposed descriptor achieves an excellent performance on two high-resolution face datasets, namely Bosphorus and Multi-PIE. More importantly, our proposed method significantly outperforms the state-of-the-art Convolutional Neural Network (CNN)-based face recognition method, when query faces are highly occluded. The code of our proposed method is available at: `https://github.com/johnnysclai/PoreNet`.

Index Terms— Face recognition, high-resolution face recognition, feature extraction, pore-scale facial feature.

1. INTRODUCTION

Nowadays, it is easy to obtain high-quality face images with resolution in the size of over 10 megapixels. What would be interesting to know is how we can make use of the information gain to achieve highly accurate face recognition. Face recognition is a well-studied research topic, because the human face is the most outstanding biometrics. Recent deep Convolutional Neural Network (CNN)-based approaches [1, 2, 3, 4] achieve over 99% verification rate on the Labeled Faces in the Wild (LFW) [5] benchmark. Despite this remarkable improvement, face recognition is still challenging. For example, identifying the similar facial parts between two face images is essential for forensics and law enforcement, and it will solve the occluded face recognition problem as well. However, the existing CNN-based face recognition models cannot achieve this capability, because the global representation is used, *i.e.* a face image is represented by a single feature vector.

Our work is an extension of [6, 7]. In the previous works, local facial patterns are represented by using a variant of the Scale Invariant Feature Transformation (SIFT) [8] descriptor, namely Pore-SIFT (PSIFT). However, it is a handcrafted descriptor and is not optimized from local skin patterns. [6, 7] had attempted to learn the discriminative feature representations by using Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA), but the improvements are limited as the learning algorithms are not trained in an end-to-end manner.

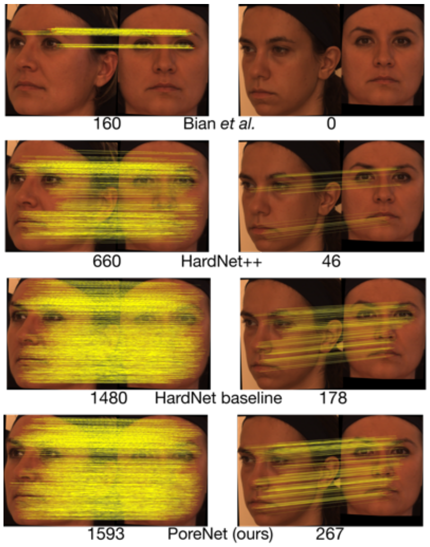

Fig. 1: Matching results, from top to bottom, based on 4 different methods. The gallery and query faces have a 30° pose difference. The two subjects in the left column are of the same subject, while the two subjects in the right column are of different subjects. The number of matched keypoints is listed at the bottom of the two subjects being matched.

Our work aims to learn a robust local descriptor for facial skin patches, so that pore patterns or facial skin regions can be matched. In this paper, we propose a descriptor by a specially designed CNN to learn the local feature representations, namely PoreNet. Compared to other existing CNN-based local descriptors [9, 10, 11], which extract features from generic local patches, PoreNet is optimized for pore pattern representations.

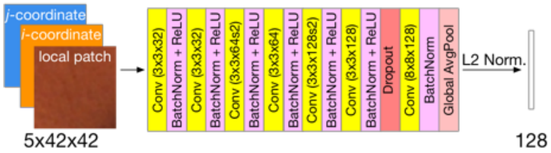

Fig. 2: The network architecture of PoreNet, adapted from HardNet [11]. Each skin patch is concatenated with the i-coordinate and the j-coordinate channels, corresponding to the (x, y) coordinates, as the input. Stride of 2 is denoted as s2.

In summary, our contributions are threefold: 1. A learning-based descriptor is proposed to extract the pore-feature representations from skin patches. 2. A simple yet effective feature matching framework is proposed for high-resolution face recognition. This matching method, with the high-performance descriptors proposed, can greatly improve the recognition rate, and is robust to different variations. 3. We have demonstrated that our proposed method is robust to the highly occluded face recognition problem, while the performance of those state-of-the-art CNN-based methods drops significantly.

2. REVIEW ON HIGH-RESOLUTION FACE RECOGNITION

In the early study of high-resolution (HR) face recognition, Lin and Tang [12] used Gabor filters to extract features from local skin regions. By considering the cheek and forehead regions, an accuracy of over 60% is achieved on a HR face dataset collected in their laboratory. In the study of soft biometrics traits, [13] suggested that facial marks, including moles, scars, freckles, wrinkles, and birthmarks, are distinctive and useful for face recognition. However, usually, there are only a few of these microscopic features on faces, and they cannot be observed when the resolution is not sufficiently high. Recently, Li and Lam [6] proposed a pore-scale feature extraction framework, adapted from SIFT detector and descriptor, to establish a dense set of correspondences from a pair of HR face images of the same identity. Furthermore, the correspondences of the pore-scale facial features have been explored for face recognition [7] and 3D face reconstruction [14].

3. PROPOSED METHOD

Our proposed method follows the standard paradigm of local feature matching, including: 1. pore detection (Section 3.1), 2. pore description (Section 3.2) and 3. matching and outlier rejection (Section 3.4). Our work is an extension of [6, 7]. We aim to improve the HR face recognition accuracy by using a robust, end-to-end learning-based local descriptor. To emphasize the importance of the local descriptor in our discussion, the pore detection and outlier rejection scheme are kept as simple as possible. In the proposed method, the similarity score of two faces is based on the number of inliers because we expect that matching two distinct subjects should produce a sparse set of false matches. Experiments have shown that, with a high-performance local descriptor, a simple similarity measurement is sufficient to achieve an excellent face recognition performance.

	HardNet baseline	Ours		
RGB input?	✗	✓	✓	✓
With coordinate channels?	✗	✗	✓	✓
Input size = 42×42?	✗	✗	✗	✓
FPR95	18.27	11.64	11.01	10.37

Table 1: Ablation study of each component. Performance is measured on the test set.

Motivated by the success of the learning-based local descriptors [9, 10, 11], we propose to learn a CNN-based local descriptor for extracting features from local facial skin patches, namely PoreNet. For this research, a local facial skin patch dataset was first constructed for training and evaluation. The details will be discussed in Section 3.3.

3.1. Pore detection

Pore detection means to locate the pore features in facial skins. Pore features are small, dark, and blob-shaped, including pores, moles, fine wrinkles, and hairs. Therefore, we employ a scale-normalized Laplacian of Gaussian (LoG) blob detector to locate potential pore keypoints. This is more accurate than the PSIFT detector [6], because the PSIFT detector employs Difference of Gaussians (DoG), which is an approximation of LoG. Our detector uses the first octave to construct 10 scale-space images, where the scale $\sigma = 1/\sqrt{2}, 1.5/\sqrt{2}, \ldots, 5.5/\sqrt{2}$, in order to detect the pore features of diameter d from 3 to 10 pixels (no keypoint is detected at the first and the last layers), given that $d = 2\sqrt{2}\sigma$. Each local maximum is compared to its 26 neighbors in a $3 \times 3 \times 3$ region, and its value should be larger than a threshold τ. A 2×2 Hessian matrix $\mathbf{H}$ is computed for each keypoint position. If the value of $(\mathrm{Tr}\,(\mathbf{H}))^2/\mathrm{Det}\,(\mathbf{H})$ is greater than a threshold r, the keypoint is considered unstable and will then be discarded. We set τ and r at 0.006 and 5, respectively, by experiments. These values are not optimal as the roughness of facial skin is different from person to person. However, empirically, we found that most of the pore features can be detected with these settings. Therefore, we fix these two threshold values in all of the experiments.

3.2. Pore descriptor

The network architecture of PoreNet is shown in Fig. 2. It is adapted from the HardNet [11] with minimal modifications (0.086% of parameters added), but the performance is substan-

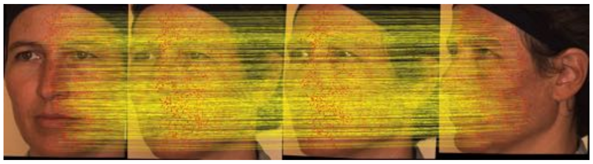

Fig. 3: Visualization of the matched correspondences of a subject across 4 adjacent poses. 729 inliers are detected for this subject.

tially improved. We keep other settings the same as HardNet, including weight initialization, the optimization method, and the loss function. We conducted an ablation study to demonstrate the performance of each component of our proposed method. We consider the trained-from-scratch HardNet model as our baseline, which produces the false positive rate at 95% true positive rate (FPR95) of 18.27%. In all the experiments, except for the indicated change(s), other settings remained unchanged. The experiment results are tabulated in Table 1.

First, our model extracts features from color patches instead of grayscale patches. This is because most face images are almost color images. Second, we provide the scale information, σ, to network, so as to better understand the local patterns. As long as the face images are of similar size, the scale of the matched keypoints should be more or less the same. Inspired by [15], we explicitly embed the (x, y) coordinates, by concatenating the i-coordinate channel, I_i, and the j-coordinate channel, I_j, together with the RGB input, where $I_i, I_j \in \mathbb{R}^{H \times W}$. The entries of the i-th column of I_i are all equal to i, and those of the j-th row of I_j are all equal to j, where $i = 0, \ldots, W-1, j = 0, \ldots, H-1$, and W and H are the original width and height of a local skin patch, respectively. I_i and I_j are normalized to the range of $[-1, 1]$, and they will be resized to the accepted input size of the network by bilinear interpolation. Third, we slightly increase the accepted input size in order to find the optimal input size of the network. For a fair comparison, a global average pooling layer is added at the end such that the network always produces a 128-D feature. As a trade-off between computational complexity and accuracy, we consider input size of 42×42 pixels. By combining the above changes, our final model (*i.e.*, PoreNet) achieves an FPR95 of 10.37%. We do not observe over-fitting problem as the FPR95 on the test set decrease gradually during training.

3.3. Dataset

We use the author-released pre-trained model of HardNet [11] as our preliminary local descriptor, called HardNet++, which was trained on the union of Brown [16] and HPatches [17] datasets. For each pore features, a local patch size of $48\sigma \times 48\sigma$ pixels is considered, where σ is the scale of the keypoint. Then, the grayscale patch is resized to 32×32 pixels, which is the required input size of the HardNet++.

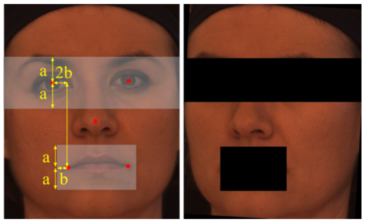

Fig. 4: Illustration of the measurement of the occluded regions (left), and occluded regions of a query image (right).

We select faces from 4 poses (10°, 20°, 30°, and 45°) from all the 105 subjects in the Bosphorus dataset [18]. The original resolution is used and the two eyes in each image are horizontally aligned. We consider those keypoints inside a tightly cropped facial bounding box only. Matched keypoints are established between a pair of face images of adjacent poses, and they are the nearest neighbor in both images. Then, outliers are discarded by using Grid-based Motion Statistics (GMS) [19]. Those correspondences that are consistently matched across all the 4 poses form our dataset. Although there are some incorrect correspondences, as shown in Fig. 3, most of them are visually convincing.

We split the database into a test set and a training set, which contain the correspondences from the first 30 subjects and the remaining 75 subjects, respectively. Therefore, the subjects in the training set and the test set are mutually exclusive. For the evaluation protocol, we follow the design of the Brown benchmark [16], where the test set contains a set of matching pairs and a set of non-matching pairs. To reflect the performance properly, each of the non-matching keypoints is the hardest one, which is randomly chosen from 1,000 non-matching keypoints. Finally, the training set consists of 62,524 classes (250,096 patches), and the test set consists of 19,495 matching pairs and 19,495 non-matching pairs.

3.4. Pore-keypoint matching and outlier rejection

In [6], Random Sample Consensus (RANSAC) [20] was employed to perform outlier rejection from a geometric point of view. However, the human face is highly non-rigid, and the transformation between two faces cannot be modeled by a homography matrix. In our proposed method, we employ GMS [19] as the outlier rejection scheme. The main idea of GMS is to consider the local statistical information, rather geometric information. This is because we should expect that the supported matches around an inlier (*i.e.*, a true positive) should be statistically more than that around an outlier (*i.e.*, a false positive). Therefore, a match with statistically fewer supported matches in its surroundings can be considered an outlier and discarded. The details of GMS are beyond the scope of this paper, and we use it as an off-the-shelf outlier rejection method. In all of our experiments, we use the official implementation[1] with the authors suggested parameter

[1] https://github.com/JiawangBian/GMS-Feature-Matcher (commit:2c8ff5f)

Method	Descriptor	P-R10	P-R20	P-R30	P-R45	P-L45	P-All	E-Ha	E-Su	E-Fe	E-Sa	E-An	E-Di	E-All	T-S0	T-S1	T-S2	T-S3	T-All
Li *et al.* [7]	PSFIT	1.49	3.37	6.88*	12.01	11.14	7.58	1.65	0.7	1.7	1.58	0.76	2.06	1.05	0.55	1.2	1.11	1.21	0.91
	PPCASIFT	1	1.9	2.77	5.82	9.28	4.64	2.02	1.01	2.4	3.12	1.51	3.92	2.25	0.47	1.18	1.74	2.24	1.29
Bian *et al.* [19]	ORB	0.25	2.86	7.4	30.02	26.42	15.7	0.94	**0**	**0**	**0.22**	**0.14**	2.9	1.08	**0**	4.65	6.98	7.22	5.28
Ours	HardNet++	0.99	2.85	3.81	5.71	3.81	7.07	0.94	1.41	1.55	1.52	2.82	2.9	1.77	0.08	0.21	**0.78**	0.07	**0.19**
	HardNet baseline	0.15	0.95	**1.9**	**1.9**	2.86	**2.86**	0.12	0.24	1.43	1.23	1.49	**1.45**	0.96	0.1	**0.02**	**0.78**	**0.01**	**0.19**
	PoreNet	**0.05**	**0.77**	1.99	2	**1.35**	**2.86**	**0.05**	0.21	1.43	0.73	1.41	**1.45**	**0.88**	**0**	0.78	1.31	0.16	0.58

Table 2: Equal error rate (%) of different approaches. The result marked with an * is estimated from Fig. 8 in [7].

Method	Descriptor	P-R10	P-R20	P-R30	P-R45	P-L45	P-All	E-Ha	E-Su	E-Fe	E-Sa	E-An	E-Di	E-All	T-S0	T-S1	T-S2	T-S3	T-All
SphereFace [2]	-	7.12	8.39	13.44	18.6	21.44	15.54	5.14	14.08	10.24	9.09	8.88	8.44	9.39	2.33	6.57	8.07	7	6.23
Bian *et al.* [19]	ORB	1.9	14.95	31.29	44.15	43.23	30.61	0.94	1.41	4.27	1.52	1.29	4.27	2.43	**0**	4.65	6.98	7.22	5.28
Ours	HardNet++	0.95	2.86	6.51	15.03	21.75	12.19	0.59	0.43	2.86	1.52	2.82	**1.45**	1.32	**0**	**0.78**	**0.78**	1.55	0.78
	HardNet baseline	0.01	0.57	**1.9**	3.28	4.61	3.22	0.05	0.04	**1.43**	0.79	1.41	**1.45**	**0.07**	**0**	**0.78**	**0.78**	**0.78**	**0.63**
	PoreNet	**0**	**0.16**	**1.9**	**1.9**	**3.81**	**2.86**	**0.01**	**0.03**	**1.43**	**0.2**	**0.18**	**1.45**	0.48	0.04	**0.78**	1.55	1.55	1.16

Table 3: Equal error rate (%) of different approaches when query faces are manually occluded.

settings. Same as mentioned in Section 3.3, those keypoints whose descriptors are the nearest neighbor to each other in a pair of images are selected.

4. EXPERIMENTS

We have conducted extensive experiments on two HR face datasets: Bosphorus [18] and Multi-PIE [21], to evaluate the performance and robustness of our proposed method for HR face recognition. We follow the protocol used in [7]. For the Bosphorus database, the gallery set is formed based on the first frontal-view faces from all the 105 subjects. Then, all the face under the 5 pose variations: R10°, R20°, R30°, R45°, and L45° (prefixed with P-), and 6 facial-expression variations: happy, surprise, fear, sadness, anger, and disgust (prefixed with E-), form the query set. Original images, whose resolution is about 1,400×1,200 pixels, are used. For the Multi-PIE database, only the 129 subjects that appeared in all the four sessions are selected. All faces with neutral expression in Session 0 are used to form the gallery set, and all the remaining faces in Sessions 0 to 3 (prefixed with T-) form the query set. The face images are loosely cropped and downsampled to about 900×700 pixels. All the faces in both datasets are slightly rotated so that the two eyes are horizontally aligned. Each of the query faces is matched with all the gallery faces. Matching results based on the different methods are shown in Fig. 1, where the subjects were selected from the test set. Furthermore, we have conducted an additional experiment to mimic the occluded face recognition problem. The eyes and mouth regions of the query images are manually occluded, as illustrated in Fig. 4.

First, our method, based on the 3 different descriptors, is compared to Li *et al.* [7] (results are directly cited from the paper), and Bian *et al.* [19]. In fact, [19] can be regarded as our method with the keypoint detector and descriptor replaced by a weaker one (*i.e.*, ORB [22] features with the maximum feature number of 10,000). As shown in Table 2, our method is robust to pose, expression, and age variations, and it outperforms the other methods. This reveals the importance of using a high performance local descriptor. In the Bosphorus dataset, PoreNet consistently outperforms the pre-trained and trained-from-scratch HardNet models, denoted as HardNet++ and HardNet baseline, respectively. It is worth noting that the HardNet baseline and PoreNet are trained with the same training set, and the same setting. However, this advantage cannot be generalized to the Multi-PIE dataset. The reason for this is that the color tones and lighting conditions of the two datasets are visually distinct.

Second, our method is compared to Bian *et al.*, and a state-of-the-art CNN-based face recognition method, namely SphereFace [2], for the occluded face recognition problem. We use the author-released 20-layer SphereFace model and follow the pre-processing steps. SphereFace achieves a near perfect performance on both datasets when the faces are not occluded. However, as shown in Table 3, the performance of SphereFace is degraded significantly when occlusion happens. With our proposed method, the performance can be retained under heavy occlusion, and consistently outperforms other methods.

5. CONCLUSION

In this paper, a novel local descriptor for high-resolution face recognition is proposed. Compared to other existing local descriptors, the proposed descriptor is specifically designed for extracting the feature representations for the local facial-skin pore patterns. We have demonstrated the face recognition performance with the proposed local descriptor. More importantly, the accuracy of our proposed method can be retained when face images are under heavy occlusion, and it does not require any facial landmark labels. For our future work, we will consider high-resolution face recognition in the wild.

Acknowledgement: The work described in this paper was supported by the GRF Grant PolyU 152765/16E (project code: B-Q55J) of the Hong Kong SAR Government.

6. REFERENCES

[1] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2015.

[2] W. Liu et al., "Sphereface: Deep hypersphere embedding for face recognition," in *Proc. CVPR*, 2017.

[3] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, July 2018.

[4] H. Wang et al., "Cosface: Large margin cosine loss for deep face recognition," in *Proc. CVPR*, 2018.

[5] 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," Tech. Rep. 07-49, University of Massachusetts, Amherst, Oct 2007.

[6] D. Li and K. M. Lam, "Design and learn distinctive features from pore-scale facial keypoints," *Pattern Recognition*, vol. 48, no. 3, pp. 732–745, Mar 2015.

[7] D. Li, H. L. Zhou, and K. M. Lam, "High-resolution face verification using pore-scale facial features," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2317–2327, Aug 2015.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov 2004.

[9] V. Balntas, E. Riba, D. Ponsa, and K Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks," in *Proc. BMVC*, 2016.

[10] Y. Tian, B. Fan, and F. Wu, "L2-net: Deep learning of discriminative patch descriptor in euclidean space," in *Proc. CVPR*, 2017.

[11] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas, "Working hard to know your neighbor's margins: Local descriptor learning loss," in *Proc. NIPS*, 2017.

[12] D. Lin and X. Tang, "Recognize high resolution faces: From macrocosm to microcosm," in *Proc. CVPR*, 2006.

[13] U. Park and A. K. Jain, "Face matching and retrieval using soft biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 406–415, Sep 2010.

[14] J. Jo, H. Kim, and J. Kim, "3d facial shape reconstruction using macro- and micro-level features from high resolution facial images," *Image and Vision Computing*, vol. 64, pp. 1–9, Aug 2017.

[15] R. Liu et al., "An intriguing failing of convolutional neural networks and the coordconv solution," in *Proc. NIPS*, 2018.

[16] M. Brown, G. Hua, and S. Winder, "Discriminative learning of local image descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 43–57, Jan 2011.

[17] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, "Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *Proc. CVPR*, 2017.

[18] A. Savran et al., *Bosphorus Database for 3D Face Analysis*, Biometrics Identity Management. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[19] J. Bian, W. Y. Lin, Y. Matsushita, S. K. Yeung, T. D. Nguyen, and M. M. Cheng, "Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence," in *Proc. CVPR*, 2017.

[20] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Cmmun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[21] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, May 2010.

[22] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: an efficient alternative to sift or surf," in *Proc. ICCV*, 2011.