Improving RGB-D face recognition via transfer learning from a pretrained 2D network*

Xingwang Xiong, Xu Wen, and Cheng Huang

University of Chinese Academy of Sciences
{xiongxingwang18,wenxu14,huangcheng14}@mails.ucas.edu.cn

Abstract. 2D Face recognition has been extensively studied for decades and has reached remarkable results in recent years. However, 2D Face recognition is sensitive to variations in poses, facial expressions and illuminations. Depth images provide valuable information to help model facial boundaries and understand the global facial layout and provide low frequency patterns. Intuitively, RGB-D images are more robust to external environments than RGB images. Unfortunately, RGB-D datasets are orders of magnitude smaller than 2D datasets and insufficient to train a deep CNN model as effective as RGB-based models. To tackle these challenges, we present an RGB-D ResNet50 model which can be transferred from a pretrained RGB model and takes RGB-D images as input. We achieved an accuracy of 94.64% and won the 1st place on *3D Face Recognition Algorithm Challenge*, *2019 BenchCouncil International Artificial Intelligence System Challenges*.

Keywords: Face recognition · RGB-D images · Transfer learning.

1 Introduction

Face recognition is one of the most significant topics in the field of Artificial Intelligence. In recent years, deep models (e.g. AlexNet [20], VGGNet [30], GoogLeNet [32] and ResNet [13]) based on convolutional neural network (CNN) have made great progress in face recognition. But neural network structure is just one side of the coin. To explore the potential of CNN, a dataset with abundant RGB face images is required. There are several such large scale RGB image datasets available online. After LFW (13,233 images) [15], dataset volume has grown all the way (e.g. IJB series [24], CASIA-WebFace [35] and MF2 [26]) to millions of images.

However, 2D face recognition under bad environmental illumination, large head pose or big expression variations still remains challenging. An RGB-D image contains one more channel of depth information compared with RGB images. Depth information provides clues about illumination, pose and scale, and a stabler facial texture [40]. Intuitively, RGB-D image based face recognition models are more robust and could have a better performance [4], or a better upper bound at least.

* The source code is available at https://github.com/xingwxiong/Face3D-Pytorch.

Nowadays, depth sensors such as Microsoft Kinect [39] and Intel RealSense [18] are becoming easily available and popular, and even some mobile phones are equipped with depth cameras. RGB-D data volume is growing, but still far smaller than RGB ones. For example, Lock3DFace contains 5,711 images and video clips taken by Kinect V2 [37]. Datasets like ND 2006, Bosphorus and BU-3DFE also only have thousands of 3D face images [6,19,28,36]. Those datasets are insufficient to train an efficient CNN model.

To leverage conventional RGB-based works and depth features on limited RGB-D dataset, we present an RGB-D ResNet50 model which can be transferred from a pretrained RGB model and takes RGB-D images as input. We achieved a competitive accuracy of 94.64%, outperforming RGB models taking merely RGB images as input and RGB-D models traning from scratch. And we won the 1st place on *3D Face Recognition Algorithm Challenge*, one track of *2019 BenchCouncil International Artificial Intelligence System Challenges*. To our knowledge, among all the winning teams, we are the only one who uses inter-modal transfer learning to improve RGB-D face recognition [9,34].

3D Face Recognition Algorithm Challenge's topic is derived from AIBench [7,8], one of benchmarking projects proposed by BenchCouncil. Besides AIBench, BenchCouncil also proposes several other active benchmarking projects, such as BigDataBench [33], HPC AI500 [17], AIoT Bench [23], Edge AIBench [11]. The source code of AIBench is publicly available from `http://www.benchcouncil.org/benchhub/AIBench/` (Sign up to get access).

2019 BenchCouncil International Artificial Intelligence System Challenges contains a total of 4 challenge tracks, namely *International AI System Challenge based on RISC-V* [14], *International AI System Challenge based on Cambricon Chip* [21,22], *International AI System Challenge based on X86 Platform* [2,5,12] and *International 3D Face Recognition Algorithm Challenge* [9,34], respectively.

2 Related Work

Face recongnition, one of the earliest tasks in computer vision, has been researched for decades. 2D face recognition has achieved remarkable results, while 3D face recongition gets less attention.

2.1 3D Face Datasets and 3D Face Recognition

Many organizations collect data or use simulation to create 3D face datasets. However, all of them are much smaller than that of 2D face. 3D face images are usually stored in three types: depth-image, point cloud and mesh. Among them, depth-image method is cheaper than the other ones, so it's widely used nowadays.

Traditional methods for cloud point images use distances to recongnize faces, such as Iterative Closest Point(ICP) [3] and Hausdorff distance. However, these methods are quite limited. Extracting features is a more flexible method, which can deal with all types of images. In recent years, deep learning plays an important role in 3D face recognition and has a better performance [40].

Table 1. Some typical 3D face dataset

Name	Number of persons	Number of images	Data type
Bosphorus [28]	105	4,666	point cloud
BU-3DFE [36]	100	2,500	mesh
GavabDB [25]	61	540	mesh
Lock3DFace [37]	509	5,671	depth-image
ND 2006 [6]	888	13,450	depth-image
FRGC Ver2.0 [27]	466	4,007	depth-image
FaceWarehouse [1]	150	-	depth-image
Intellifusion RGB-D dataset [7]	1,205	403,068	depth-image

2.2 RGB-D Images and Datasets

RGB-D is one kind of depth-image. It combines an ordinary RGB image with depth map, which reflects the distance between the surface of items and a given viewpoint. Although containing more information, RGB-D images can only be collected by certain devices and acquisition of such RGB-D images might take a long time, which limits the scale of RGB-D datasets. RGB-D datasets are orders of magnitude smaller than 2D datasets due to the acquisition cost and insufficient to train a deep CNN model with considerable quality.

2.3 Transfer Learning on RGB-D Datasets

Now that RGB-D contains 2D images, transfer learning is a good method for RGB-D data. [16] uses transfer learning on RGB-D dataset to make action recognition. While [10] and [29] make saliency detection and object recognition respectively. Using transfer learning can solve the problems caused by limited datasets, because there are plenty of 2D datasets which can be used. Using a pre-trained model can save resources as well.

3 Transfer Learning from RGB to RGB-D

From the perspective of low-level patterns, depth images attend to have smooth variations, contrasts and borders, but lack texture information and high frequency patterns [31], which is exactly complementary to RGB images. RGB-D images can provide the model more diverse features, compared to merely RGB images. This is the very motivation for us to use RGB-D model instead of RGB model for face recognition.

As shown in Fig.1, in addition to changing the last fully connected layer, we also adjust the uppermost convolution layer of ResNet50 [13] in order to feed the model with 4-channel RGB-D images. We copy the parameters of the middle layers of the RGB model and then fine tune the entire RGB-D model on the target Intellifusion RGB-D dataset.

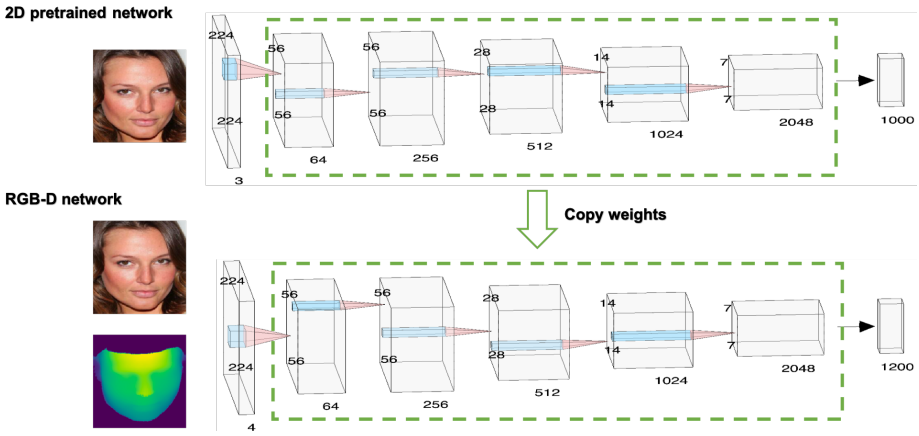

Fig. 1. Transfer learning from a pretrained RGB model.

4 Experiments

We conduct three experiments to compare the effects of depth images and inter-modal transfer learning on face recognition accuracy (see Table 2). The first experiment only uses RGB images and train the model on a pre-trained model. The second uses RGB-D images to train the RGB-D ResNet model from scratch. The third trains our RGB-D model on the target 3D dataset based on the 2D pretrained model.

Preprocessing The faces and their landmarks in images are detected and aligned by MTCNN [38]. The images are horizontally flipped with a probability of 0.5 for data augmentation. Both RGB images and depth images are normalized to 0-1 range.

Train/test split The Intellifusion RGB-D face dataset contains 403,068 images of 1,205 people. We divide the dataset into a training set and a test set in a ratio of approximately 9:1. Categories with no more than 10 samples are removed. After discarding images in which no face is detected, there are 361,799 face images in the training set of 1,200 people and 40,809 images in the test set. Dataset is split under closed-set settings, which means identities in the testing set must appear in the training set.

Training configuration We train the models with batch size of 16 on 4 Nvidia Titan Xp GPUs. We use SGD as our optimizer with a momentum of 0.9. The learning rate is initialized to 0.001 and decayed by a factor of 0.1 every 7 epochs. Following the last fully connected layer is a softmax layer for classification and identification.

Evaluation metric We report the average recognition precision in test set of 1200 people.

Table 2. Comparisons on Intellifusion RGB-D face dataset in accuracy (%).

Method	RGB images	Depth images	CNN models	Accuracy (%)
Pretrained on Imagenet	✓	✗	RGB ResNet50	94.47
Train from scratch	✓	✓	RGB-D ResNet50	88.36
Pretrained on Imagenet	✓	✓	RGB-D ResNet50	**94.64**

As shown in Table 2, the RGB-D model transfered from a 2D network has the best accuracy, compared to the RGB model taking merely RGB images as input and the RGB-D model traning from scratch. What's more, the inference speed of our RGB-D ResNet50 model is about 262.64 fps. And it shows that inter-modal transfer learning outperforms RGB models which take merely RGB images as input and RGB-D models training from scratch.

5 Conclusion

Compared to RGB images, RGB-D images contain more information about the global layout and enable the images to have stereoscopic effects. From a perspective of intuition, our RGB-D model is able to make more use of the information provided by depth images. But due to the limitation of the size of RGB-D datasets, it is difficult to develop a RGB-D model as efficient as RGB models. Using transfer learning from RGB to RGB-D can solve this problem. In this paper, we present an RGB-D ResNet50 model transferred from a pretrained RGB model and achieve an accuracy of 94.64%.

References

1. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics **20**(3), 413–425 (2013)
2. Chen, M., Chen, T., Chen, Q.: An efficient implementation of the als-wr algorithm on x86 cpus. In: International Symposium on Benchmarking, Measuring and Optimization (Bench 19). Springer (2019)
3. Cheng, S., Marras, I., Zafeiriou, S., Pantic, M.: Statistical non-rigid icp algorithm and its application to 3d face alignment. Image and Vision Computing **58**, 3–12 (2017)
4. Cui, J., Zhang, H., Han, H., Shan, S., Chen, X.: Improving 2d face recognition via discriminative face depth estimation. In: 2018 International Conference on Biometrics (ICB). pp. 140–147. IEEE (2018)
5. Deng, W., Wang, P., Wang, J., Li, C., Guo, M.: Psl: Exploiting parallelism, sparsity and locality to accelerate matrix factorization on x86 platforms. In: International Symposium on Benchmarking, Measuring and Optimization (Bench 19). Springer (2019)

6. Faltemier, T.C., Bowyer, K.W., Flynn, P.J.: Using a multi-instance enrollment representation to improve 3d face recognition. In: 2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems. pp. 1–6. IEEE (2007)
7. Gao, W., Luo, C., Wang, L., Xiong, X., Chen, J., Hao, T., Jiang, Z., Fan, F., Du, M., Huang, Y., et al.: Aibench: towards scalable and comprehensive datacenter ai benchmarking. In: International Symposium on Benchmarking, Measuring and Optimization. pp. 3–9. Springer (2018)
8. Gao, W., Tang, F., Wang, L., Zhan, J., Lan, C., Luo, C., Huang, Y., Zheng, C., Dai, J., Cao, Z., Tang, H., Zhan, K., Wang, B., Kong, D., Wu, T., Yu, M., Tan, C., Li, H., Tian, X., Li, Y., Lu, G., Shao, J., Wang, Z., Wang, X., Ye, H.: Aibench: An industry standard internet service ai benchmark suite. arXiv preprint arXiv:1908.08998 (2019)
9. Gong, T., Huiqian, N.: An implementation of resnet on the classification of rgb-d images. In: International Symposium on Benchmarking, Measuring and Optimization (Bench 19). Springer (2019)
10. Han, J., Chen, H., Liu, N., Yan, C., Li, X.: Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. IEEE transactions on cybernetics **48**(11), 3171–3183 (2017)
11. Hao, T., Huang, Y., Wen, X., Gao, W., Zhang, F., Zheng, C., Wang, L., Ye, H., Hwang, K., Ren, Z., Zhan, J.: Edge aibench: Towards comprehensive end-to-end edge computing benchmarking. 2018 BenchCouncil International Symposium on Benchmarking, Measuring and Optimizing (Bench18) (2018)
12. Hao, T., Zheng, Z.: The implementation and optimization of matrix decomposition based collaborative filtering task on x86 platform. In: International Symposium on Benchmarking, Measuring and Optimization (Bench 19). Springer (2019)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
14. Hou, P., Yu, J., Miao, Y., Tai, Y., Wu, Y., Zhao, C.: Rvtensor: A light-weight neural network inference framework based on the risc-v architecture. In: International Symposium on Benchmarking, Measuring and Optimization (Bench 19). Springer (2019)
15. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments (2008)
16. Jia, C., Kong, Y., Ding, Z., Fu, Y.R.: Latent tensor transfer learning for rgb-d action recognition. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 87–96. ACM (2014)
17. Jiang, Z., Gao, W., Wang, L., Xiong, X., Zhang, Y., Wen, X., Luo, C., Ye, H., Lu, X., Zhang, Y., Feng, S., Li, K., Xu, W., Zhan, J.: Hpc ai500: A benchmark suite for hpc ai systems. 2018 BenchCouncil International Symposium on Benchmarking, Measuring and Optimizing (Bench18) (2018)
18. Keselman, L., Woodfill, J.I., Grunnet-Jepsen, A., Bhowmik, A.: Intel (r) realsense (tm) stereoscopic depth cameras. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1267–1276. IEEE (2017)
19. Kim, D., Hernandez, M., Choi, J., Medioni, G.: Deep 3d face identification. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). pp. 133–142. IEEE (2017)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)

21. Li, G., Wang, X., Ma, X., Liu, L., Feng, X.: Xdn: Towards efficient inference of residual neural networks on cambricon chips. In: International Symposium on Benchmarking, Measuring and Optimization (Bench 19). Springer (2019)
22. Li, J., Jiang, Z.: Performance analysis of cambricon mlu100. In: International Symposium on Benchmarking, Measuring and Optimization (Bench 19). Springer (2019)
23. Luo, C., Zhang, F., Huang, C., Xiong, X., Chen, J., Wang, L., Gao, W., Ye, H., Wu, T., Zhou, R., Zhan, J.: Aiot bench: Towards comprehensive benchmarking mobile and embedded device intelligence. 2018 BenchCouncil International Symposium on Benchmarking, Measuring and Optimizing (Bench18) (2018)
24. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., et al.: Iarpa janus benchmark-c: Face dataset and protocol. In: 2018 International Conference on Biometrics (ICB). pp. 158–165. IEEE (2018)
25. Moreno, A.: Gavabdb: a 3d face database. In: Proc. 2nd COST275 Workshop on Biometrics on the Internet, 2004. pp. 75–80 (2004)
26. Nech, A., Kemelmacher-Shlizerman, I.: Level playing field for million scale face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7044–7053 (2017)
27. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). vol. 1, pp. 947–954. IEEE (2005)
28. Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L.: Bosphorus database for 3d face analysis. In: European Workshop on Biometrics and Identity Management. pp. 47–56. Springer (2008)
29. Schwarz, M., Schulz, H., Behnke, S.: Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features. In: 2015 IEEE international conference on robotics and automation (ICRA). pp. 1329–1335. IEEE (2015)
30. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
31. Song, X., Herranz, L., Jiang, S.: Depth cnns for rgb-d scene recognition: learning from scratch better than transferring from rgb-cnns. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
32. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
33. Wang, L., Zhan, J., Luo, C., Zhu, Y., Yang, Q., He, Y., Gao, W., Jia, Z., Shi, Y., Zhang, S., et al.: Bigdatabench: A big data benchmark suite from internet services. In: 2014 IEEE 20th International Symposium on High Performance Computer Architecture (HPCA). pp. 488–499. IEEE (2014)
34. Wang, Y., Zeng, C., Li, C.: Exploring the performance bound of cambricon accelerator in end-to-end inference scenario. In: International Symposium on Benchmarking, Measuring and Optimization (Bench 19). Springer (2019)
35. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
36. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3d facial expression database for facial behavior research. In: 7th international conference on automatic face and gesture recognition (FGR06). pp. 211–216. IEEE (2006)

37. Zhang, J., Huang, D., Wang, Y., Sun, J.: Lock3dface: a large-scale database of low-cost kinect 3d faces. In: 2016 International Conference on Biometrics (ICB). pp. 1–8. IEEE (2016)
38. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
39. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE multimedia **19**(2), 4–10 (2012)
40. Zulqarnain Gilani, S., Mian, A.: Learning from millions of 3d scans for large-scale 3d face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1896–1905 (2018)