RGB-D Face Recognition: A Comparative Study of Representative Fusion Schemes

Jiyun Cui[1,2], Hu Han[1(✉)], Shiguang Shan[1,2], and Xilin Chen[1,2]

[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
jiyun.cui@vipl.ict.ac.cn, {hanhu,sgshan,xlchen}@ict.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. RGB-D face recognition (FR) has drawn increasing attention in recent years with the advances of new RGB-D sensing technologies, and the decrease in sensor price. While a number of multi-modality fusion methods are available in face recognition, there is not known conclusion how the RGB and depth should be fused. We provide a comparative study of four representative fusion schemes in RGB-D face recognition, covering signal-level, feature-level, score-level fusions, and a hybrid fusion we designed for RGB-D face recognition. The proposed method achieves state-of-the-art performance on two large RGB-D datasets. A number of insights are provided based on the experimental evaluations.

Keywords: RGB-D face recognition · Signal-level fusion
Feature-level fusion · Score-level fusion · Hybrid fusion

1 Introduction

While significant progress has been made for visible light face recognition in past five years, face recognition under bad environmental illumination, large head pose and big expression variations remains challenging using only visible light face images. With the popularity of RGB-D sensors such as RealSense and Kinect, apart from visible light face image, it becomes easy to obtain near-infrared (NIR) and depth information of human face. While visible face images represent texture information, depth modality provides face space information such as shape and surface normal. Multi-modal face recognition using both color and depth images has been found to be more robust in unconstrained environment. Studies on RGB-D face recognition aims to design representation and fusion approaches which can explore the complementary information from both modalities as much as possible [10,21].

In review of published modality fusion methods, the fusion schemes can be grouped into two main categories: feature-level fusion methods and score-level fusion methods. Feature-level fusion methods usually learn modality-specific features first, which were fused to form a combined feature representation. Score-level fusion methods compute per-modality similarity scores first, and then fuse

J. Zhou et al. (Eds.): CCBR 2018, LNCS 10996, pp. 358–366, 2018.
https://doi.org/10.1007/978-3-319-97909-0_39

the scores via particular rules, e.g., a sum rule. In this paper, we provide a comparative study of four fusion scheme using the RGB-D two modalities information, apart from the above two fusion strategies, we also consider the signal-level fusion methods which combine the raw color and depth images and the hybrid fusion strategy consisting of two or more fusion schemes.

The contributions of the paper are two-fold: (i) four representative fusion strategies are summarized in RGB-D face recognition covering signal-level, feature-level, score-level, and hybrid fusions; (ii) individual fusion schemes are fully evaluated on two large-scale RGB-D datasets (Lock3DFace and our dataset) and a number of insights are provided.

2 Related Work

RGB-D multi-modal face recognition has been studied for many years. [1] proposed to extract HOG features from entropy/saliency maps calculated from RGB and depth images and then trained a Random Decision Forest (RDF) classifier for matching. [2] fused both the entropy map based match score and attribute based match scores of depth image for face recognition. [3] built a 12-layer Deep Convolutional Neural Network (CNN) consisting of six modules, in which three loss modules were added after the second, fourth and sixth network modules, respectively. In each loss module, in addition to using softmax loss for identification, contrastive loss was utilized for verification purpose. [10] proposed an approach to learn complementary features and common features from RGB-D face images during the training phase. During testing, this method extracted modal-specific features for per-modality matching, and used a score-level fusion to compute the final matching score.

Besides the fusion schemes used in RGB-D face recognitioin, there are also a number of studies on how to fuse the RGB and depth in other tasks, such as object recognition [4,6,8], scene recognition [5,9], and person re-identification [7]. [4] proposed a pair of deep residual networks for RGB and depth data to explore the sharable and modal-specific features. The input data of depth modal is surface normals instead of depth image. [5] proposed an approach that combines RGB and depth modalities from multiple sources, the depth modality has two streams networks which are transfered from RGB pre-trained modal and direct training. The extracted features of multiple modalities were added as a fusion features for recognition. [7] proposed a RGB-D based approach for person re-identification, in which, the anthropometric feature vectors were extracted in a fusion layer consisting of the depth-specific part, the sharable part and the RGB-specific part. [8] proposed an approach that combines convolutional and recursive neural networks (RNN) in which RNN worked as a fusion part to get the final features from RGB and depth. [9] proposed a novel discriminative multi-modal feature fusion method which also used two CNN steams for handling color and depth, respectively.

3 Fusion Schemes in RGB-D Face Recognition

We group the fusion schemes for RGB-D face recognition into four categories: signal-level fusion, feature-level fusion, score-level fusion and hybrid fusion. The details of each scheme are summarized below.

3.1 Signal-Level Fuison

The signal-level fusion method operates directly on the raw RGB and depth images. Since RGB images are treated as three-channel input data of the network, we can also concatenate RGB and depth into a four-channel input data for signal-level fusion. Apart from such a four-channel fusion method, we also explore other signal-level RGB-D fusion methods such as sum or average of the corresponding pixels in RGB and depth images. In these methods, the depth modality data is copied to a three-channel format to keep consistent with the RGB three-channel format. Figure 1(a) shows a general diagram of the signal-level fusion for RGB-D face recognition.

3.2 Feature-Level Fusion

The feature-level fusion is to fuse the features extracted from RGB and depth modality network and then fed into the classification layers. In feature-level fusion, the fusion can be in the form of feature concatenated or sum or average of the feature map, or projections from two modalities' feature maps. In this paper, our experiments are all based on ResNet [15] modal, the feature vector of the first fully connected layer is used as the features. With feature-level fusion, we usually get a single feature vector representing the RGB-D face images. Figure 1(b) shows a general diagram of the feature-level fusion for RGB-D face recognition.

3.3 Score-Level Fusion

Score-level fusion was thoroughly discussed in [13] for multi-biometric systems, i.e., face, fingerprint, etc. Here, we focus on the fusion of the matching scores by RGB and depth, respectively. A straightforward strategy is to calculate the average score of the two modalities' scores. Such an average fusion can be formulated in a more general format, i.e., a weighted sum of the two score, $S = \omega \cdot S_{RGB} + (1 - \omega) \cdot S_D$ where ω and $1 - \omega$ are the weights for the RGB and depth, respectively, which can be determined empirically based on the performance of each modality ($\omega = 0.5$ in our experiments). Figure 1(c) shows a general diagram of the score-level fusion for RGB-D face recognition.

3.4 Hybrid Fusion

In real applications, multiple fusion strategies might be jointly used, i.e., a hybrid fusion consisting of more than two of the above fusion schemes. We propose a

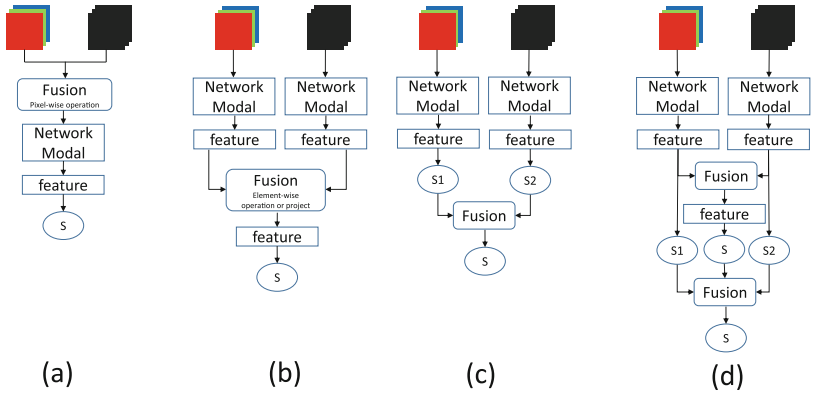

Fig. 1. The diagrams of the four representative fusion schemes for RGB-D face recognition: (a) signal-level fusion, (b) feature-level fusion, (c) score-level fusion, (d) hybrid fusion.

hybrid fusion consisting of both feature-level fusion and score-level fusion (see Fig. 1(d)). In particular, the feature-level fusion part aims to learn a joint feature representation from both RGB and depth. The score-level fusion part takes into account three matching scores obtained using RGB feature, depth feature, and the joint feature. Finally, a score-level fusion is applied with a weighed sum rule $S = \frac{\alpha \cdot S_{RGB} + \beta \cdot S_D + \gamma \cdot S_C}{\alpha + \beta + \gamma}$, where α, β and γ balance the importances of individual features, and are determined empirically ($\alpha = \beta = \gamma = 1$ in our experiments).

4 Implementation Details

4.1 Network Training

We use a ResNet-80 network as the backbone network[1] for our RGB-D face recognition experiments (see Fig. 2). Our ResNet-80 was pre-trained using a large RGB face dataset compiled from multiple public-domain datasets, such as MS Celeb [14], etc. We then finetune the pre-trained model under various RGB-D fusion schemes. The learning rate is set to 0.001.

In signal-level fusion (Fig. 1(a)), the fusion operation was performed ahead of the first of convolution layer of ResNet-80. We consider pixel-wise operations such as pixel-wise sum and average, which comes up a three-channel data. The concatenation operation generates a six-channel input data. In this situation, we revise the first convolution layer and train this layer from scratch. In score-level fusion, two ResNet-80 modals are separately trained using RGB and depth modalities, respectively. In feature-level fusion, a fully connect layer is used to

[1] We also tried AlexNet [18], GoogLeNet [17], and VGG-16 [19], but the best performance of the three model for RGB and depth fusion is 96.8%, which is lower than our ResNet-80 (98.7%). So we only report the results using our ResNet-80.

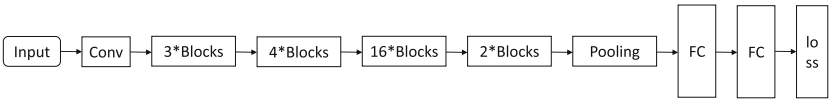

Fig. 2. An overview of the ResNet-80 network used in our experiments. Conv. and FC denote the convolutional and fully convolution layers, blocks are explained in [15].

get a concatenation of the RGB and depth features, followed by a single loss function. The proposed hybrid fusion contains an joint loss for the feature fusion layer, and two losses for RGB and depth respectively. Thus, three features (RGB, depth and RGB-D) are extracted. The final score is computed via a score-level fusion of the three scores.

4.2 The Preprocessing of the RGB-D Images

We use an open-source face recognition engine to detect the face and keypoint landmark[2] from RGB, and normalize the detected faces to 256×256. For depth image, we follow a preprocessing pipeline in [10]. We also use a bilateral filter [12] to suppress the noises in depth.

5 Experiments

5.1 Datasets and Evaluation Metrics

We provide experiments on two large-scale RGB-D datasets: Lock3DFace [11] and our RGB-D dataset [10]. Lock3DFace consists of 5,711 video sequences of 509 subjects with variations in head pose, expression, occlusion, etc. We extract 33,780 RGB-D images from the video sequences and randomly select 22,798 RGB-D images of 340 subjects for training, the remaining of 169 subjects for testing. Our RGB-D dataset contains about 845K RGB-D images of 742 subjects captured by RealSense II with variable of head pose and illumination. About 580K RGB-D images of 500 subjects are randomly selected for training and the remaining 280K RGB-D images of 242 subjects are used for testing.

We perform face identification on the two databases, and report the rank-1 identification accuracy. For each subject in the testing dataset, one frontal RGB-D image is used as gallery, and the remaining RGB-D images are used as probe. Cosine distance is used to mesure the similarity between two features.

5.2 Overall Performance and Analysis

We report the rank-1 identification accuracy by the four representative fusion schemes in Table 1. As a comparison, we also provide the unimodal face recognition accuracy, i.e., using RGB alone and using depth alone as the baselines. On

[2] https://github.com/seetaface/SeetaFaceEngine.

Table 1. The overall performance of the representative RGB-D fusion schemes in terms of the rank-1 identification accuracy and the average feature extraction time per RGB-D image.

Fusion schemes	Method	Datasets		Time (s)
		Our dataset	Lock3DFace	
Unimodality	RGB alone	97.14%	97.30%	0.012
	Depth alone	94.33%	74.45%	0.012
Signal-level fusion	Pixel-wise max	92.27%	90.35%	0.021
	Pixel-wise sum	96.54%	97.38%	0.013
	Concatenation	92.65%	79.32%	0.013
Feature-level fusion	Element-wise max	97.66%	**97.86%**	0.030
	Element-wise sum	98.90%	97.60%	0.033
	Concatenation	99.03%	97.74%	0.026
Score-level fusion	Weighted sum	98.67%	97.43%	0.024
Hybrid fusion	Use at least 2 of the above fusions	**99.05%**	97.53%	0.033

our RGB-D dataset collected by RealSense II, all the fusion methods except for the signal-level fusion can improve the face recognition by a large margin. The hybrid fusion method achieves the highest recognition accuracy among the four types of fusion methods. However, on Lock3DFace, the improvement by hybrid fusion becomes slightly smaller than using feature-level fusion. We think the main reason is that the depth captured by Kinect II is much worse than using RealSense II (see Table 1). Among the three methods in signal-level fusion, the pixel-wise sum operation is better than the other two signal-level fusion methods, but still does not show improved performance than the baseline results using RGB alone. The possible reason is that the three signal-level operations are not able to properly make use of the information in the color and depth images. Feature-level, score-level and hybrid fusions all are found to be helpful in improving face recognition accuracy than unimodality face recognition, and feature-level and hybrid fusions are slightly better than score-level fusion. In addition, the feature-level fusion and hybrid fusion report very similar rank-1 accuracies on our dataset (see an example of the top-5 matched gallery images in Fig. 3). This is consistent with previous studies where multiple feature descriptors are fused for improving classification accuracies. These results suggest that RGB and depth may not share the same network weights during feature learning. The observations on the Lock3DFace dataset is similar, except that the feature-level fusion becomes slightly better than hybrid fusion. The main reason is that each video in Lock3DFace was recorded with almost a still subject, leading to near-duplicated video frames, and thus reducing the effective training data samples during network learning. We also provide evaluations on EURECOM [16], but since this dataset is very small, we directly use the model trained

on our dataset. The rank-1 identification accuracies of using RGB and depth alone are 94.41% and 58.74%, respectively, and the fusion does not improve the accuracy. The main reasons are: (i) the race difference between our dataset and EURECOM and (ii) the device difference between RealSense and Kinect.

We also profile the average feature extraction time of individual fusion schemes on a Titan Xp GPU. While signal-level fusion does not incur additional computation cost, the other three fusion schemes have 2–3 times higher computation cost in feature extraction. This is understandable because RGB and depth are handled via separate subnetworks by the other three fusion schemes. Still, they are able to process images in real time (30fps), and should meet the requirement of general applications.

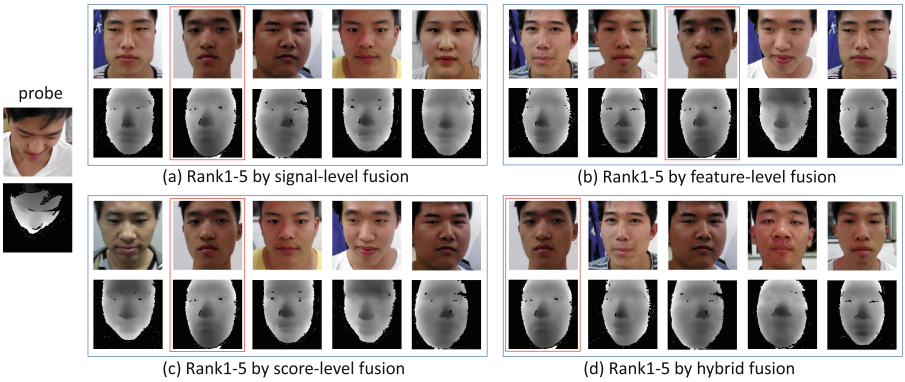

Fig. 3. The top1-5 matches by the four representative fusion methods. (a) top-5 matched gallery using signal-level fusion; (b) top-5 matched gallery using feature-level fusion; (c) top-5 matched gallery using score-level fusion; (d) top-5 matched gallery using hybrid fusion. The gallery images marked with red boxes are the correct mated gallery images for the probe.

6 Conclusions

We provide a comparative study of four representative fusion strategies covering signal-level fusion, feature-level fusion, score-level fusion, and hybrid fusion on two large-scale RGB-D databases. While signal-level fusion should retain the most amount of information in theory, the four-channel or pixel-wise fusion of RGB and depth signals does not show better performance than the feature-level or score-level fusion. Furthermore, the proposed fusion approach, a hybrid fusion scheme, achieves the best accuracy on our RGB-D dataset which is more challenging than Lock3DFace in terms of pose and illumination vairations. These motivate us to investigate new network architectures so that the signal-level fusion could better leverage the information of RGB-D to improve face recognition accuracy. In addition, we are going to study 3D reconstruction based method [20] for RGB-D face recognition.

Acknowledgement. This research was supported in part by the Natural Science Foundation of China (grants 61732004, and 61672496), External Cooperation Program of Chinese Academy of Sciences (CAS) (grant GJHZ1843), and Youth Innovation Promotion Association CAS (2018135).

References

1. Goswami, G., Bharadwaj, S., Vatsa, M., Singh, R.: On RGB-D face recognition using kinect. In: Proceedings of BTAS, pp. 1–6 (2013)
2. Goswami, G., Vatsa, M., Singh, R.: RGB-D face recognition with texture and attribute features. IEEE Trans. Inf. Forensics Secur. **9**(10), 1629–1640 (2014)
3. Lee, Y., Chen, J., Tseng, C., Lai, S.: Accurate and robust face recognition from RGB-D images with a deep learning approach. In: Proceedings of BMVC, pp. 123.1–123.14 (2016)
4. Wang, Z., Lu, J., Lin, R., Feng, J., Zhou, J.: Correlated and individual multi-modal deep learning for RGB-D object recognition, in arXiv:1604.01655 (2016)
5. Song, X., Jiang, S., Herranz, L.: Combining models from multiple sources for RGB-D scene recognition. In: Proceedings of IJCAI, pp. 4523–4529 (2017)
6. Eitel, A., Springenberg, J., Spinello, L., Riedmiller, M., Burgard, W.: Multimodal deep learning for robust RGB-D object recognition. In: Proceedings of IROS, pp. 681–687 (2015)
7. Ren, L., Lu, J., Feng, J., Zhou, J.: Multi-modal uniform deep learning for RGB-D person re-identificaiton. Pattern Recogn. **72**(12), 446–457 (2017)
8. Socher, R., Huval, B., Bath, B.: Convolutional-recursive deep learning for 3D object classification. In: Proceedings of NIPS, pp. 665–673 (2012)
9. Zhu, H., Weibel, J., Lu, S.: Discriminative multi-modal feature fusion for RGBD indoor scene recognition. In: Proceedings of CVPR, pp. 2969–2976 (2016)
10. Zhang, H., Han, H., Cui, J., Shan, S., Chen, X.: RGB-D face recognition via deep complementary and common feature learning. In: Proceedings of FG, pp. 1–8 (2018)
11. Zhang, J., Huang, D., Wang, Y., Sun, J.: Lock3DFace: a large-scale database of low-cost kinect 3D faces. In: Proceedings of ICB, pp. 1–8 (2016)
12. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Proceedings of ICCV, pp. 839–846 (1998)
13. Jain, A.K., Nandakumar, K., Ross, A.: Score normalization in multimodal biometric systems. Pattern Recogn. **38**(12), 2270–2285 (2005)
14. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: a dataset and benchmark for large-scale face recognition. In: Proceedings of ECCV (2016)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition, in arXiv:1512.03385 (2015)
16. Min, R., Kose, N., Dugelay, J.: KinectFaceDB: a kinect database for face recognition. IEEE Trans. SMC Syst. **44**(11), 1534–1548 (2014)
17. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of ICML, pp. 448–456 (2015)
18. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: Proceedings of NIPS, pp. 1097–1105 (2012)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition, in arXiv:1409.1556 (2015)

20. Han, H., Jain, A.K.: 3D face texture modeling from uncalibrated frontal and profile images. In: Proceedings of BTAS, pp. 223–230 (2012)
21. Cui, J., Zhang, H., Han, H., Shan, S., Chen, X.: Improving 2D face recognition via discriminative face depth estimation. In: Proceedings of ICB, pp. 1–8 (2018)