Heterogeneous face recognition: Multi-model fusion and Siamese Network

Miaoran Chen
Beihang University
No. 37 Xueyuan Road
Haidian District, Beijing
86-18811218639
miaoran9819@163.com

Duojia Yang
Beijing University of Posts and Telecommunications
No.10 Xitucheng Road
Haidian District, Beijing
86-13683050910
ydj980316@126.com

Ziyang Yuan
Beijing University of Posts and Telecommunications
No.10 Xitucheng Road
Haidian District, Beijing
86-18810778687
y_zy@bupt.edu.cn

Yang Huang
Beijing jiaotong University
No.3 Shuangyuancun
Haidian District, Beijing
86-13261611722
m13210889156@outlook.com

Shunchang Liu
Beihang University
No. 37 Xueyuan Road
Haidian District, Beijing
86-18810220881
liusc@buaa.edu.cn

ABSTRACT

Heterogeneous face recognition remains a challenging problem because of distortion, exaggeration and simplification. Key points annotation, manually extraction are effective in capturing common features of images, however, not efficient enough. In order to solve this problem, we designed an end-to-end network with fused neural networks to extract features. Besides, we used Pseudo- Siamese network andintegrated contrastive loss and binary cross entropy loss to complete the identification work. Caricature-Visual dataset[20], which have images of many different styles, is used to train our network. Our model not only has good performance on identifying unseen images of people in the dataset, but also works well on identifying images of people not included in the dataset.

Keywords

Heterogeneous face recognition, Model fusion, Deep learning, Siamese network

1. INTRODUCTION

Heterogeneous face recognition has becoming increasingly important. Sketch [4,15] recognition is used to track down criminals. Infra-red face recognition is used in monitoring and public security. Among them, caricature face recognition is a representative question of heterogeneous face recognition.

Firstly, the diversity of caricature styles are much larger than the diversity of sketches or infrared images. (Actually these images are in the same style). However, caricatures of different styles would have different kinds of distortion, exaggeration and color. Some caricature datasets even contain sketches and infra-red pictures. Secondly, the difference between caricatures of different people is relatively small, especially when the caricatures have same style. The cosine similarity of Figure1 is 69%, and they apparently are describing different person. Moreover, caricatures of different styles of the same person would vary a lot. The similarity of Figure2 is only 60%, although they belong to the same person.

The most important part of heterogeneous face recognition is cross-modal. The representations of same feature could have huge difference under different visual domains. [28] demonstrated that caricature could even enhance the performance of face recognition. Cause caricatures emphasize the unique characteristics of each person, making their face more recognizable. It turns out that we need to find the common features of caricatures and visual images and filter irrelative ones if we want to achieve cross-modal recognition.

While key points annotation and manually extraction are common on finding features, they require lots of energy and time to do the preparation work. Moreover, they are not robust cause the unseen images have to be either annotated or aligned automatically, additional error could be produced in this process.

To solve problems stated above, we designed an end-to-end network to achieve heterogeneous face recognition. Overall, the purposed work makes the following contributions:

We used fused neural networks (DenseNet121 and Xception) to extract features of caricatures and visual images. Maxpooling is adopted to the output of DenseNet121 and Xception , in order to reduce parameters and enhance generalization ability of model. The 2 outputs are then concatenated and then put into 2 dense layer and 1 dropout layer.

We used Pseudo-Siamese network, contrastive loss and classification function to verificate whether 2 images belong to the same person. Usually, identification and verification are combined to enhance the accuracy[5,20].However, these models could not work well on recognizing unseen identities.

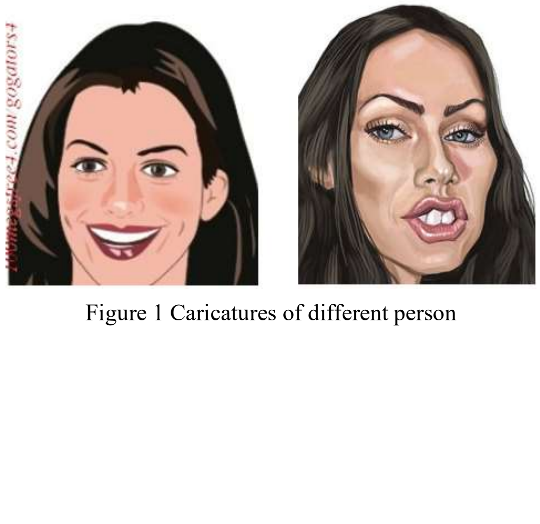

Figure 1 Caricatures of different person

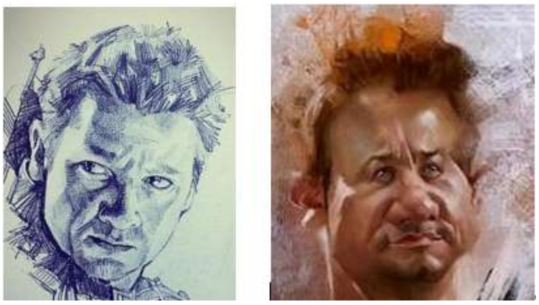

Figure 2 Caricatures of same person

2. Related Work

Although face recognition has been improved greatly and applied to a wide range, heterogeneous face recognition (HFR) problems including caricature face recognition remain challenging because of cross-modal gap and image gap between different visual domain[15]. Caricature face recognition could be viewed as a specific question of heterogeneous face recognition and the most important part of this problem is feature extraction. Early research of HFR mainly focused on subspace algorithms[29,30] and manifold learning[31,32,33]. [29] used PCA to retain the information and reduce redundancy as much as possible. [31,32,33] project image features into low-dimension space with nonlinear transformation based on the fact that the distribution of human face features in high dimensional space is non-linear. These methods highly depend on the quality of dataset, and do not have good generalization ability.

Recent years, manually extraction[11,14], key-points annotation[1,6,13] and neural networks[1,5,19] are three main approaches to extract features under caricature face recognition problem. Klare et al [11] proposed a set of qualitative features which could encode the appearance of both caricatures and photographs, then generated a similarity score with several machine learning methods including logistic regression and SVM between a caricature and a photograph. Klare trained models based on small dataset which have 196 pairs of caricatures and photographs. Abaci[14] enhanced Klare's work by adopting a genetic algorithm and logistic regression[5], extracting 32 features to identify a subject also based on a small dataset. However, WebCaricature[1], a dataset proposed by Huo, Jing, contains 6,042 cartoon images of 252 celebrities and 5,974 face images. Every image is annotated with key points produced by Face++[1]. Dong Yi et al[6] also extracted Gabor features at some localized facial points, then employed Restricted Boltzmann Machines (RBMs) to learn a shared representation locally to remove the heterogeneity around each facial point.

Besides, extracting features through neural networks is also popular. Yi Sun designed deep convolutional networks (DeepID2) [5] to extract features and use both face identification and verification signals as supervision. Saxena [19] used migration study, demonstrated CNN pre-trained on visible spectrum face images can be used to perform heterogeneous face recognition. [20] integrated previous work, used VGG to extract features and Siamese Network to verificate and identify identities. Particularly, Wenbin Li et al integrated neural networks and key points annotation[10], Li and his group designed different networks to extract features on different parts. Neural networks are becoming deeper to extract features effectively. However, over-fitting and gradient vanishing are still common in training models. [35] did a practice to fuse models and had great performance.

Many methods to identify the similarity of 2 images remain inflexible. Most of them need to set a threshold manually. [21] used cosine similarity to match face descriptors. [6] also used cosine descriptors to match RBM representations. Fingerprint information also has wide application. 2 images are normalized and a sequence is generated to describe it. The similarity of images relate to the number of same digits of 2 sequences. It is hard to determine the result of matching problem only with similarity. Siamese network[27], firstly used to judge whether signatures belong to the same person, also played an important role in the area of face recognition[34]. However, with the shared parameters, it is not flexible in dealing with some problems.

3. Dataset

We used the dataset produced by [20], which have 5091 caricatures and 6427 visual images of 205 different identities. CaVI dataset does not create the one-to-one correspondence of each identities, the paired data is generated by randomly match the caricature and visual images of a same person which is more robust than [11,14]. Besides, the style of caricatures in CaVI dataset has more diversity compared to [8,11,14], having not only different kinds of distortion and exaggeration of different facial parts, but also have various artistic styles including sketch, watercolor, oil painting and pixel picture. Moreover, the background of visual images is not homogeneous, it is very complex and different, which made the heterogeneous face recognition even more challenging giving the problems of cross-modal. However, it is more in line with actual situation.

The limited dataset also makes the caricature-visual recognition more challenging. To avoid over-fitting, we did data-augmentation to the dataset. Including image reversal, scale translation and color jittering.

4. Methodology

4.1 Extracting features with models fused network

We fused neural networks including ResNet, Xception, MobileNet, DenseNet to extract features. After experiments, the combination of Xcpetion and DenseNet121 turned out to be most effective one.

Since layers are densely connected in DenseNet[22], the input of each layer depends on the output of all previous layers, it could enhance the spread of features. Besides, features are reused based on connected channels, making the extraction more efficiently. In spite of deep network, DenseNet resolved the problem of vanishing gradient and reduced the number of model parameters, even could prevent over-fitting on small dataset to some extent, which is desirable to caricature recognition problem. However, the number of parameters increase as the depth of network increases. It is necessary for us to add bottleneck block to DenseNet to control the dimension of output of each layer. Moreover, transition block is adopted to reduce the dimension of feature map. Instead of using Relu in [22], we used LeakyRelu which is more suitable for adjusting parameters. Average pooling is used to downsample in [22], we used maxpooling in order to extract edge features. These enhancements are demonstrated to be useful in our experiments.

Xception[22] is an improvement to Inception_V3. The inception module in Inception_V3 was displaced by depthwise separable convolution in Xception. Residual learning is also employed to accelerate rate of convergence.

Model fusion is the integration of different models and their results. Majority vote fusion, weighted vote fusion, average result fusion and stack&blending. It has been demonstrated that as the number of individual learners increases, the error rate of integrated learner would decrease exponentially[24], and round towards zero. Besides, the performance of integrated learner is related to the performance and diversity of individual learners.As shown in figure 3, the integration of several individual learners could produce a strong learner. The relationship between integrated learner and individual learner is shown below.

$$E = \bar{E} - \bar{A} \qquad (1)$$

E refers to the weighted mean error after fusion. $\bar{E}$ refers to the weighted mean error of the individual learner. $\bar{A}$ refers to the diversity of models.

We chose to use DenseNet and Xception to extract features, concatenated the output representations, and added 2 dense layers and 1 dropout layer to fuse the representations. The architecture of features extraction module is shown in figure 4.

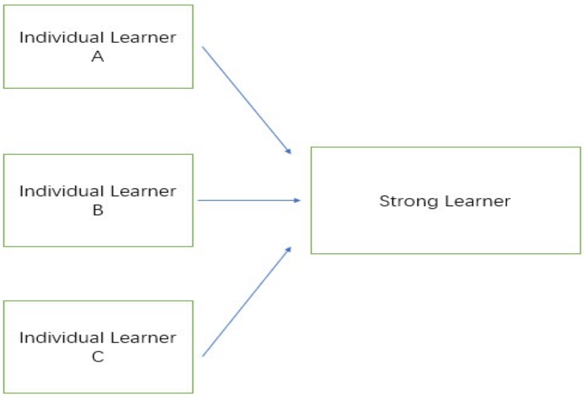

Figure 3 Relationship between individual learner and integrated learner

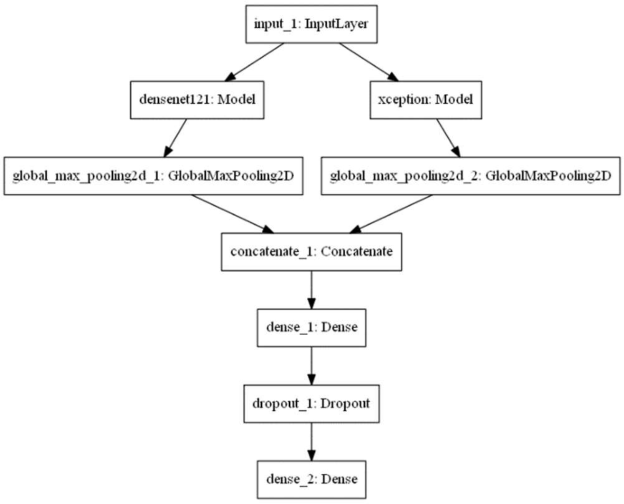
input_1: InputLayer
densenet121: Model
xception: Model
global_max_pooling2d_1: GlobalMaxPooling2D
global_max_pooling2d_2: GlobalMaxPooling2D
concatenate_1: Concatenate
dense_1: Dense
dropout_1: Dropout
dense_2: Dense

Figure 4 Architecture of extraction module

Multi-model fusion network

4.2 Find common and unique features with Pseudo-Siamese network

Inspired by [5,20], we used Pseudo-Siamese network to integrate the features of caricatures and visual images. We did not make the parameters shared in our work. Siamese network is not flexible under this question, which features extracted from caricatures of same person could vary a lot. Basically, we employed the concept of shared vision model, which first used to identify the number in the images of Mnist dataset. Like Garg's work, we designed 3 separated modules to get the common features of images, and the unique features of each image. The architecture is shown in figure 5. Contrastive loss receives 2 inputs while binary cross entropy loss only receives the concatenated feature as input.

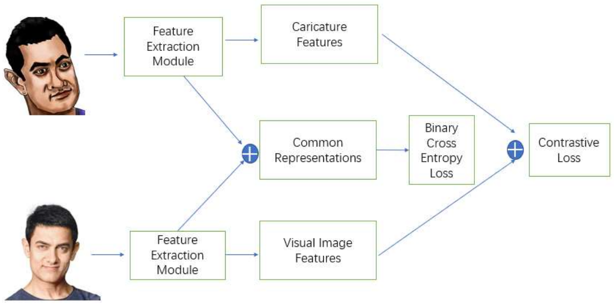

Figure 5 Architecture of verification module

Pseudo-Siamese network

Inspired by [5,20], 3 sets of transformations s, s_v, s_c are used to project the project the common features into a common subspace, extract modality specific features of caricature and visual image. s and s_v (also s and s_c) are constrained to be orthogonal to each other to minimize the redundancy.

$$s_c^T s = 0 \text{ and } s_v^T s = 0 \qquad (2)$$

Although we also projected the common features into same subspace, we did not concatenate the common features and specific features as Garg did. We used common features to do the identification work, which is regarded as a classification problem and the loss is represented as binary cross entropy loss. And the features of caricatures and images are projected to the same dimension respectively. They are combined to calculate the contrastive loss which is suitable for paied data.

4.3 Integrate contrastive loss and binary cross entropy loss

Since the input of this model is paired data, if the caricature and visual image belong to the same person, the label would be 1, if not, the label would be 0, to be specific. The contrastive loss is mainly used in feature extraction, indicating that the resemble samples still resemble with each other after feature extraction, vice versa. With the use of contrastive loss, we could make the photos of the same person close enough in the feature space and different people far enough apart in the feature space until they exceed a certain threshold. Facial images of different person do not vary a lot, caricature face images with same style especially. We could regard the heterogeneous face recognition problem as fine-grained identification problem, in which increasing inter-class differences and decreasing intra-class differences are vital.

Defined the features of caricature as a_n, features of visual image as b_n. The Euclidean distance between them could be defined as $d = ||a_n - b_n||^2$. Then the contrastive loss [26] is:

$$L = \frac{1}{2N} \sum_{n=1}^{N} yd^2 + (1 - y) \max(margin - d, 0)^2 \qquad (3)$$

y refers to whether the 2 images are matched. (If matched, y=1, otherwise, $y = 0$).

If y=1 (sample matched), the loss function is:

$$\Sigma yd^2 \tag{4}$$

Else if y=0 (sample not matched), the lose function is:

$$\Sigma(1-y)\max(margin-d,0)^2 \tag{5}$$

In other words, when the samples are not similar, and the Euclidean distance of the characteristic space is small, the loss value will increase, which is exactly what we need.

Input images are all normalized to size of $224 \times 224 \times 3$. Bottleneck block and transition block are added to DenseNet to control the dimension of features. The output of DenseNet and Xception are downsampled by a maxpooling layer, then concatenated. The following dense layers and dropout layer are used to fuse the features.

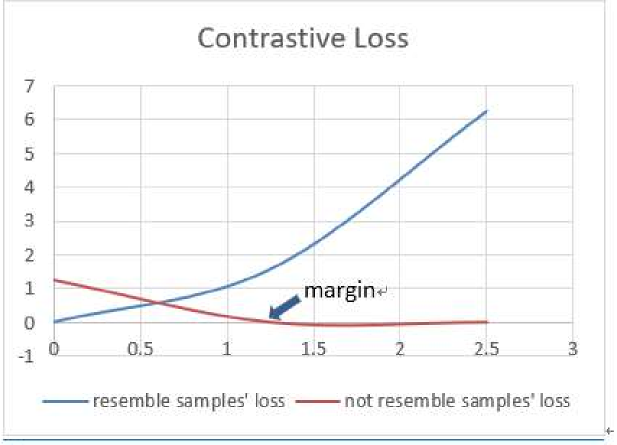

Figure 6 Contrastive Loss

We used stochastic gradient descent as optimizer, set learning rate as $l = 0.001$, decay=0, batch size=16, epoch=2. Data augmentation and enhancement are adopted to avoid over-fitting and make the network more robust.

5. EXPERIMENTS

We adopted pretrained DenseNet121 and Xception to extract features, pseudo-siamese network to complete identification and verification work. We trained on 2857 images and used 513 images to validate. It turned out that verification accuracy reached to 93% at epoch1 and 98.45% at epoch2. Besides, we tested the identification ability of model on seen images. However, due to the scarce of training data. The identification accuracy of caricature and visual image did not turn out to be well. More experiments are needed to be completed. However, we also tested the performance of model on identifying whether the unseen caricature and visual image belong to the same person. This model presented 75% accuracy on this problem which is better than the result of CaVINet.

Table 1 Experiment Results

	Verification Accuracy	Identification Accuracy of Caricature	Identification Accuracy of Visual Image
Multi-Modal Fusion (DenseNet+Xception)	98.45%	54%	56%
Multi-Modal Fusion (Inception_V3+Xception)	97.68%	54%	55%
CaVINet	97.24%	62%	59%

6. SUMMARY

This paper presents an end-to-end network to recognize heterogeneous facial images. Fused model (DenseNet121 and Xception) are adopted to extract features more effectively. Pseudo-siamese networks are used to capture the shared representations and unique features of caricature and visual image respectively. To enhance the performance on recognizing unseen faces, we integrate contrastive loss and binary cross entropy loss as our loss function.

7. REFERENCES

[1]Huo, Jing & li, Wenbin & Shi, Yinghuan & Gao, Yang & Yin, Hujun. (2017). WebCaricature: a benchmark for caricature face recognition.

[2]Jin, Yi & Lu, Jiwen & Ruan, Qiuqi. (2015). Coupled Discriminative Feature Learning for Heterogeneous Face Recognition. IEEE Transactions on Information Forensics and Security. 10. 640-652. 10.1109/TIFS.2015.2390414.

[3]Crowley, Elliot & Parkhi, Omkar & Zisserman, Andrew. (2015). Face Painting: querying art with photos. 65.1-65.13. 10.5244/C.29.65.

[4]Ouyang, Shuxin & Hospedales, Timothy & Song, Yi-Zhe & Li, Xueming. (2014). Cross-Modal Face Matching: Beyond Viewed Sketches. 10.1007/978-3-319-16808-1_15.

[5]Sun, Yi & Wang, Xiaogang & Tang, Xiaoou. (2014). Deep Learning Face Representation by Joint Identification-Verification. Proc. NIPS. 27.

[6]Yi, Dong & Lei, Zhen & Liao, Shengcai & Li, Stan. (2014). Shared Representation Learning for Heterogeneous Face Recognition. 10.1109/FG.2015.7163093.

[7]Sharma, Amit & Devale, Prakash. (2012). Face Photo-Sketch Synthesis and Recognition. International Journal of Applied Information Systems. 1. 46-52. 10.5120/ijais12-450192.

[8]Huang, Gary & Mattar, Marwan & Berg, Tamara & Learned-Miller, Eric. (2008). Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. Tech. rep.

[9]Huo, Jing & Gao, Yang & Shi, Yinghuan & Yin, Hujun. (2017). Variation Robust Cross-Modal Metric Learning for Caricature Recognition. 340-348. 10.1145/3126686.3126736.

[10]li, Wenbin & Huo, Jing & Shi, Yinghuan & Gao, Yang & Wang, Lei & Luo, Jiebo. (2019). A Joint Local and Global Deep Metric Learning Method for Caricature Recognition. 10.1007/978-3-030-20870-7_15.

[11]Klare, Brendan & Bucak, Serhat & Jain, Anil & Akgül, Tayfun. (2012). Towards automated caricature recognition. Proceedings - 2012 5th IAPR International Conference on Biometrics, ICB 2012. 139-146. 10.1109/ICB.2012.6199771.

[12]Masi, Iacopo & Wu, Yue & Hassner, Tal & Natarajan, Prem. (2018). Deep Face Recognition: A Survey. 471-478. 10.1109/SIBGRAPI.2018.00067.

[13]Lowe, David. (2004). Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision. 60. 91-110. 10.1023/B%3AVISI.0000029664.99615.94.

[14]Abacı, Bahri & Akgül, Tayfun. (2015). Matching caricatures to photographs. Signal Image and Video Processing. 9. 1-9. 10.1007/s11760-015-0819-8.

[15]Ouyang, Shuxin & Hospedales, Timothy & Song, Yi-Zhe & Li, Xueming. (2014). A Survey on Heterogeneous Face Recognition:

Sketch, Infra-red, 3D and Low-resolution. Image and Vision Computing. 56. 10.1016/j.imavis.2016.09.001.

[16]Gong, Dihong & Li, Zhifeng & Huang, Weilin & Li, Xuelong & Tao, Dacheng. (2017). Heterogeneous Face Recognition: A Common Encoding Feature Discriminant Approach. IEEE Transactions on Image Processing. PP. 1-1. 10.1109/TIP.2017.2651380.

[17]Pereira, Tiago & Anjos, André & Marcel, Sébastien. (2018). Heterogeneous Face Recognition Using Domain Specific Units. IEEE Transactions on Information Forensics and Security. PP. 1-1. 10.1109/TIFS.2018.2885284.

[18]He, Ran & Cao, Jie & Song, Lingxiao & Sun, Zhenjun & Tan, Tieniu. (2019). Cross-spectral Face Completion for NIR-VIS Heterogeneous Face Recognition.

[19]Saxena, Shreyas & Verbeek, Jakob. (2016). Heterogeneous Face Recognition with CNNs. 483-491. 10.1007/978-3-319-49409-8_40.

[20]Garg, Jatin & Peri, Skand & Tolani, Himanshu & Krishnan, Narayanan. (2018). Deep Cross Modal Learning for Caricature Verification and Identification (CaVINet). 1101-1109. 10.1145/3240508.3240658.

[21]Sarfraz, M. & Stiefelhagen, Rainer. (2015). Deep Perceptual Mapping for Thermal to Visible Face Recogntion. 10.5244/C.29.9.

[22]You, Fucheng & Zhao, Yangze. (2019). Attention Image Caption with DenseNet. Journal of Physics: Conference Series. 1302. 032048. 10.1088/1742-6596/1302/3/032048.

[23]Chollet, Francois. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. 1800-1807. 10.1109/CVPR.2017.195.

[24]Zhihua Zhou, Machine Learning[M]

[25]Sun, Yi & Wang, Xiaogang & Tang, Xiaoou. (2014). Deeply learned face representations are sparse, selective, and robust. 10.1109/CVPR.2015.7298907.

[26]Hadsell, Raia & Chopra, Sumit & Lecun, Yann. (2006). Dimensionality Reduction by Learning an Invariant Mapping. 1735 - 1742. 10.1109/CVPR.2006.100.

[27]Jane Bromley,Isabelle Guyon, Yann LeCun, Eduard Sackinger and Roopak Shah(1994). Signature Verification using a “Siamese Time Delay Neural Network”.738-744.NIPS.

[28] Mauro, R. & Kubovy, M. Memory & Cognition (1992) 20:433. https://doi.org/10.3758/BF03210927

[29] Matthew Turk,Alex Pentland. Eigenfaces for recognition. 1991, Journal of Cognitive Neuroscience.

[30] Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. Peter N Belhumeur J P Hespanha David Kriegman. 1997 IEEE Transactions on Pattern Analysis and Machine Intelligence.

[31] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using laplacianfaces. IEEE Trans. Pattern Anal. Mach. Intell.,27(3):328–340, 2005.

[32] S. Yan, D. Xu, B. Zhang, and H.-J. Zhang. Graph embedding: A general framework for dimensionality reduction. Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, 2:830–837, 2005.

[33] W. Deng, J. Hu, J. Guo, H. Zhang, and C. Zhang. Comments on “globally maximizing, locally minimizing: Unsupervised discriminant projection with applications to face and palm biometrics”. IEEE Trans. Pattern Anal. Mach. Intell., 30(8):1503–1504, 2008.

[34] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 539–546. IEEE, 2005.

[35] H. Chen, Y. Li and D. Su, "M3Net: Multi-scale multi-path multi-modal fusion network and example application to RGB-D salient object detection," 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, 2017, pp. 4911-4916.

AUTHORS’ BACKGROUND

Your Name	Position	Email	Research Field	Personal website
Miaoran Chen	Undergraduate Student	miaoran9819@163.com	Machine Learning Computer Vision	
Duojia Yang	Undergraduate Student	ydj980316@126.com	Robotics Software Engineering	
Ziyang Yuan	Undergraduate Student	y_zy@bupt.edu.cn	Mathematics	
Yang Huang	Undergraduate Student	m13210889156@outlook.com	Deep Learning Computer Vision	
Shunchang Liu	Undergraduate Student	liusc@buaa.edu.cn	Deep Learning Computer Vision	