Feature Level Information Fusion Based Deep Learning

Kejun Wang, Xuesen Hao and Xianglei Xing

1 Introduction

Feature level information fusion has been attracted widely attention. It can be considered that we extract feature then combine them. Feature level fusion can retain features and decrease amount of calculation. It can realize real time processing. In early years, they detect key points from images. Then they calculate the distance between two images. Burt PJ proposed to make fusion by Laplace pyramid. In 1995, Li H proposed wavelet method [1]. As a promising direction, Linas and Waltz analysis fusion technology delicately. Additionally, information fusion used to solve robot obstacle avoidance problem.

In recent years, facial recognition has two main methods. One of them is extract feature vector, another one is PCA(Principal Component Analysis) method [8]. These two methods base on features. Classification and identification based combine characteristic vector. Similarly, feature fusion methods also widely used in gait recognition [11], face recognition [12] and people identification [13]. It has a merit that if one tensor had problem or poor quality it would lead to low accuracy. From this point of view fusion theory and fuzzy neural network has satisfactory result [4]. It has stronger anti-interference skills. With the development of neural network [9], it is widely used to solve problems [10]. As for many other computer vision tasks, in the last few years significant performance gains have been achieved thanks to approaches based on deep networks [2, 5–7]. In 2005, Yan Lecun firstly proposed verify facial based Siamese [3]. It is different from common network. It has more than one channel as input in Siamese. It is significant to design a stable and effective system.

K. Wang · X. Hao · X. Xing (✉)
College of Automation, Harbin Engineering University,
No.145 Nantong Street, Harbin City, China
e-mail: xingxl@hrbeu.edu.cn

Z. Deng (ed.), *Proceedings of 2017 Chinese Intelligent Automation Conference*,
Lecture Notes in Electrical Engineering 458,
https://doi.org/10.1007/978-981-10-6445-6_55

Given the observations above, this paper introduces an approach based deep learning to realize feature level fusion. Our method is inspired by Siamese network. We add another channel to make fusion in facial recognition. We improve Siamese network in facial comparison. Experimental results on datasets, demonstrate the advantages of our approach over previous methods. Improved Siamese network has been proved to be useful in other dataset. In facial recognition and comparison, our method has good generalization.

Our main contributions:

1. Firstly,we establish a new dataset. We add samples to extend dataset. We set one image as a basic image and compare with remain images. If there is a result shows that they belong to the same person. We consider it as a positive sample. By contrast, we consider it as a negative sample. There are 6,848,920 positive samples and 9,910,668 negative samples.
2. Then, we crop regions of eye, nose and mouth regions. This purpose is increasing proportion of feature in an image. We train these features and face region together, and hence it is able to utilize the information of given features and improve recognition performance. We verify the effectiveness on several datasets and achieve state-of-the-art performance.
3. Moreover, an improved Siamese network is proposed to compare two images. We analyze both traditional Siamese network and improved Siamese network. Specifically, we add Spatial Transformer Network to Siamese work. We transfer single branch to seven channels.
4. Without training again when you want to compare two images. You can select two images, it will give you result.

2 The Proposed Approach

In this section we present to proposed network. We first provide an overview of our approach and we describe in details the architectures we design to realize fusion. In this paper, we propose a new deep learning framework with multiply channels as shown in Fig. 1. In traditional methods, there adapt to send all features once a time. This will lead high dimension and not unified of each feature vector. To solve these

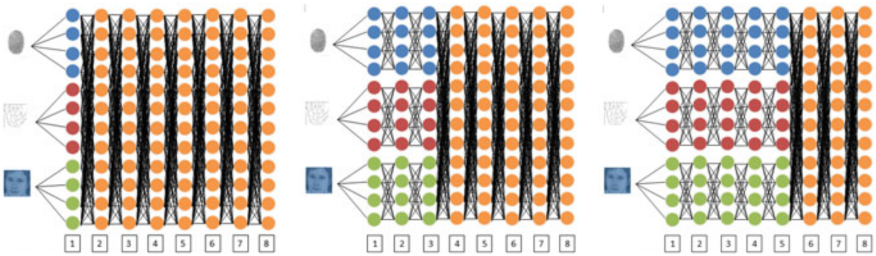

Fig. 1 Setting three channels and making fusion at different layers

problems, we adapt to set three channels as input. We try to make fusion at different layers to evaluate effectiveness of our method.

The most expensive part in terms of facial recognition is to detect the features. Despite significant progress in the past few years, facial recognition and comparison is still challenging due to the following two unanswered questions. The first one is face region has low proportion in an image. Secondly, there are more than one faces in an image. Solving these two difficulties will bring performance gain over traditional methods.

To solve the first problem, we crop face, eyes, nose and mouth regions. The process is illustrated in Fig. 2. We crop face regions by Haar algorithm. Furthermore, we detect key points by SDM algorithm. At last, we crop the other regions. The aim of crop regions is to avoid missing face region.

2.1 Feature Level Information Fusion Based Deep Learning Test on Facial Recognition

In traditional methods, they determine recognition by only one image. We propose to increase feature proportion. When we get eyes, nose and mouth regions, these features carry information, while there is no shelter. However, when sheltered in

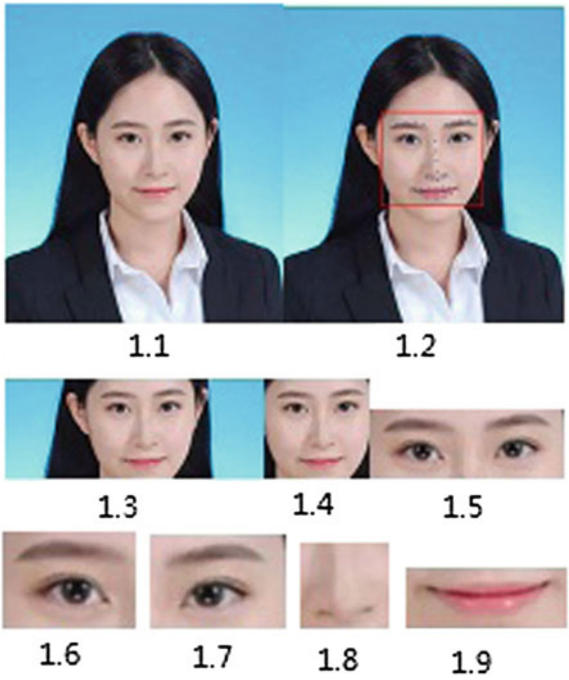

Fig. 2 Process of crop regions (Original image 1.1, detect face region 1.3, crop face region 1.4, detect eye region 1.6 and 1.7, crop nose region 1.8, crop mouth region 1.9, 1.2 and 1.5 are produced in processing)

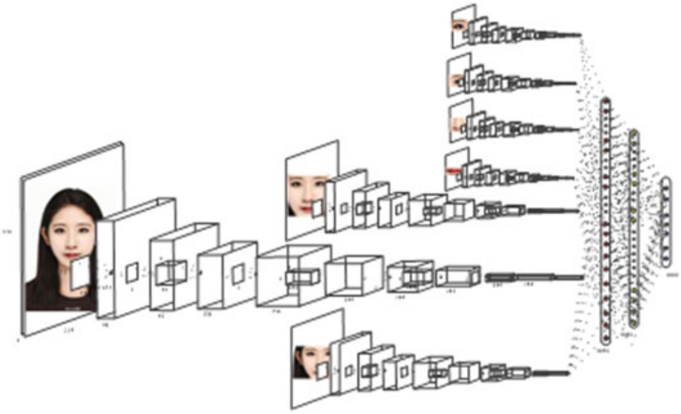

Fig. 3 Proposed fusion method for one channel

some regions, it cannot recognize efficiently. Without face region, there will be fluctuation. Furthermore, to improve accuracy, we combine face and the other parts to make fusion. It is more appropriate to achieve better result and generalization as shown in Fig. 3.

Inspired by previous works demonstrating the importance of considering feature level information in facial recognition, we propose to add another channel as shown in Fig. 4. This is specifically designed to perform facial recognition by adding another channel. If one image has low quality, we can get features from another image. In the network, it will combine these features which from two images. This will improve rate of recognition and generalization.

2.2 Feature Level Information Fusion Based Deep Learning Test on Facial Comparison

Facial comparison also called facial similarity comparison. As shown in Fig. 5, this method utilizes two channels as input. We can see the two images are the same person. The first step is detecting face region. The second step is extracting features. At last, we compare them and give the result.

Traditional neural network is widely used. However, there are problems such as low recognition and convergent slowly. These problems have effect on accuracy in practice. Through detailed analysis, we demonstrate how two channels can benefit from this network to overcome these problems in experiments. We adapt Siamese loss function in the proposed network. It can be calculated by the following formula (1):

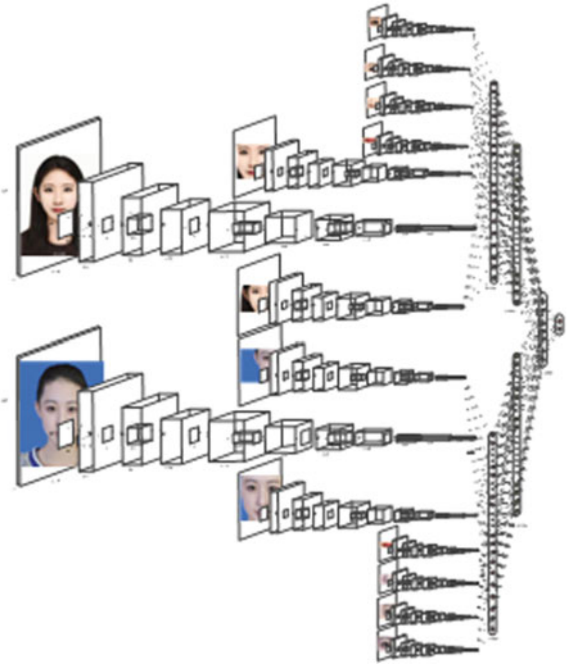

Fig. 4 Overview of our framework. Two channels with the same structure are used in facial recognition and comparison

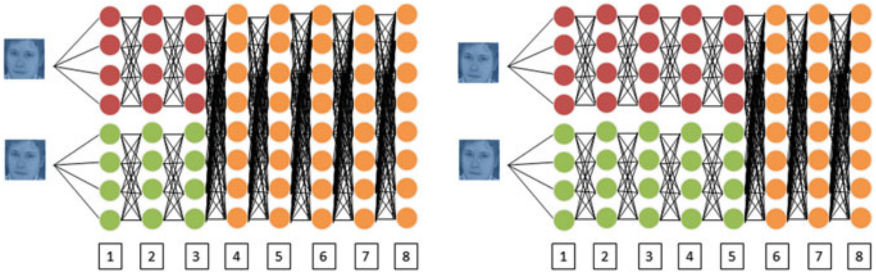

Fig. 5 Process of facial comparison. Two images belong to one person. We make fusion at different layers

$$E_W(X_1, X_2) = \|G_W(X_1) - G_W(X_2)\| \tag{1}$$

We propose to split single branch to seven channels in the middle of network as shown in Fig. 4. These features are connected in series one by one up. Previous works have not considered invariance in network. In order to have better performance, we add Spatial Transformer Networks. It has robust to translation and rotation.

3 Experiments

In this section, we present experimental evaluations and in-depth analysis of the proposed method on the new dataset. Firstly, we introduce our dataset. Then we compare our framework with the state-of-art method on LWF dataset in Table 2. Our framework is implemented under digits, and our evaluation is conducted on a NVIDIA TeslaK40 GPU. In the experiments, we show the effectiveness of our proposed method. At last, we present the result on an interface.

3.1 Prepare Datasets

Before delving into our experiments, we describe our dataset. We combine some datasets and add new samples to build new dataset. It contains LFW and CASIA-maxpy-clean dataset and our new samples. LFW dataset contains 5749 persons (13,233 images). CASIA-maxpy-clean dataset contains 10,575 persons. In this dataset, each person has 100–769 images. We add samples to our datasets. Firstly, we select 790 persons as basic images. Second, we compare each person with remain images in this dataset. If there are two images belong to one person, we regard it as a positive sample. By contrast, we regard it as a negative sample. And so on, we get 14,582 positive samples, 598,096 negative samples. Considering image size has effect on recognition. Therefore, we change images to 28*28, 56*56, 128*128 and 256*256. Then we introduce CelebFaces dataset to increase capacity. In total, we have 12,000 persons, 390,000 images. We get 6,848,920 positive samples and 9,910,668 negative samples.

3.2 Test on Facial Recognition and Analysis

It is different from usual deep learning network because of we add another input to our framework. As shown in Fig. 1. In fully connection layer, class number equal to neuron number. We compare shallow network and deep network to evaluate effective of deep network.

We evaluate the performance of four types of images and four types of network. Table 1 shows the result of our comparison. From the table, it is clear that in deep network with 256*256 images outperforms, confirming the fact that deep framework improves the recognition accuracy.

Table 1 Comparison of performace based on different network

Pixel of image	Shallow network (image undisposed) (%)	Shallow network (image preprocess) (%)	Deep network (image undisposed) (%)	Deep network (image preprocess) (%)
28*28	60.19	69.13	66.25	70.75
56*56	78.14	82.24	80.01	86.49
128*128	85.90	93.58	92.37	97.64
256*256	90.94	96.58	93.39	98.94

Table 2 Comparisons of detecting performance on LWF dataset

Method	LBP	Joint Bayesian	GMM	Original network	Single input	Our proposed
Accuracy (%)	95.17	96.33	96.52	95.71	97.13	97.98

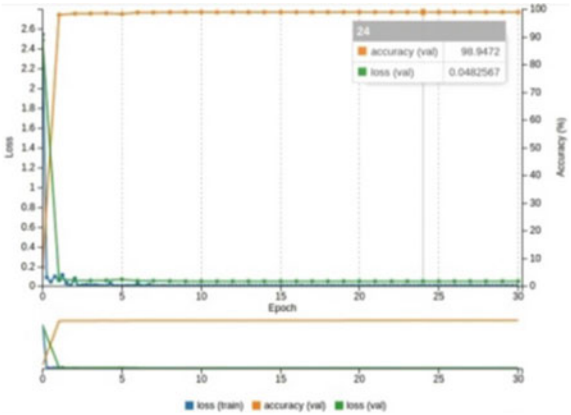

Fig. 6 Result of our proposed method on the new dataset (*blue line* is loss value of training, *orange line* is accuracy value)

We compare our approach with conventional methods. The results are summarized in Table 2. On LWF dataset, our approach outperforms all of the compared approaches. It is remarkable that our method achieves 97.98% accuracy. As shown in Table 2, it is easy to observe that different detectors affect the performance significantly. We directly using a detector may not be a good choice when applying existing method in the real world. Otherwise the detector may lose some valuable data when there is complex background.

Observing Fig. 6, we notice that it convergent quickly and stability. We can see the accuracy reach to 98.94%. In training process, it needs 18 h on NVIDIA Tesla K40 GPU. Then we evaluate 20,000 images base on this network, it needs 1 h and 20 min on CPU(Inteli76700).

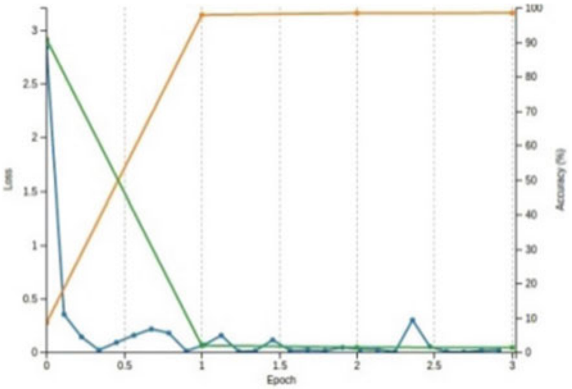

Fig. 7 Result of our approach on another dataset we select about 10,000 images to test our approach

To further demonstrate that the performance with the proposed network is not simply suit for only one dataset, we test this network by another dataset as shown in Fig. 7. It is clear that it also convergent in short time and has high accuracy. We analyze the performance of our approach on the other dataset. It assumes that this network has strong generalization ability.

3.3 Test on Facial Comparison and Analysis

Two kinds of network result are visualized in Fig. 8. We respectively evaluate the effect of improved Siamese network. It is clear that there is fluctuate in Fig. 8a. From Fig. 8a, we can see there are failed results. Compared with traditional Siamese network, the improved Siamese network has good stability and convergence as shown in Fig. 8b. Because of traditional Siamese network has few layers, and also has bad robust performance. However, it is not enough if we only add layers to network. In addition, we make the network complicated, it achieved by improved Siamese network.

We train our dataset base on improved Siamese network. We can find that it has high accuracy and stability as shown in Fig. 9. Bringing Spatial Transformer Networks, it also has effect on stability.

Furthermore, we develop interface based our proposed methods as shown in Fig. 10. These are planted to C# and Winform platform. You can select two images from your own datasets at random. It will detect regions of face and eyes. At last, there output the result of similarity without train network again. This method not only has stability but also useful in practice.

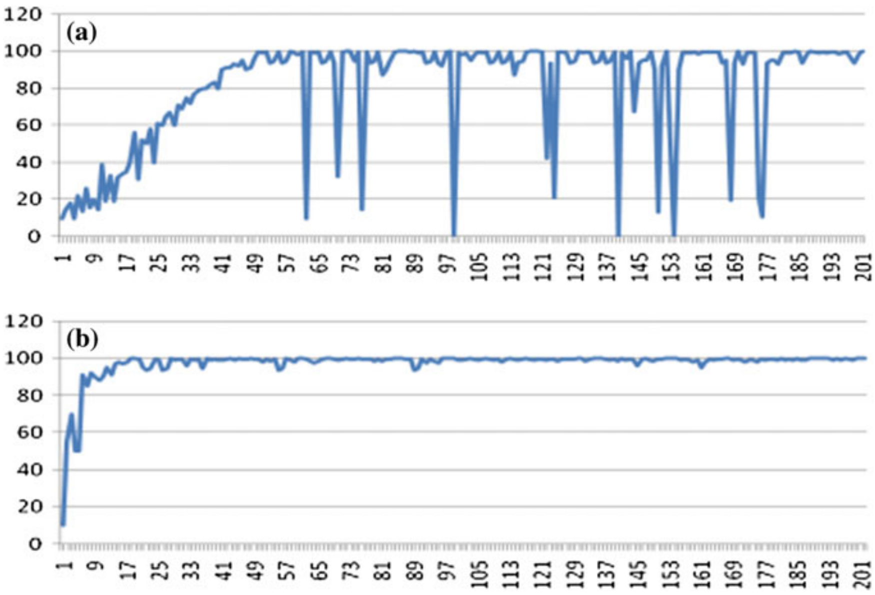

Fig. 8 Comparison of training result between Siamese and improved Siamese network. The *blue line* is result of accuracy

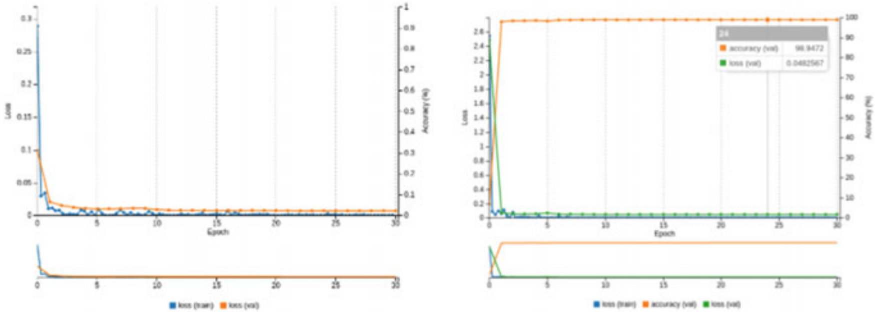

Fig. 9 Loss function and accuracy of improved Siamese network

4 Conclusion

In this paper, we propose to add channel to network. Our experiments show the proposed method can achieve satisfactory performance. Commonly used hand craft features, as they do not have good robustness. Differently from previous methods, the proposed method is possible to learn features from the improved network. We show that by increasing feature proportion and adding another input to network, it is possible to improve rate of recognition. An improved Siamese network is proposed

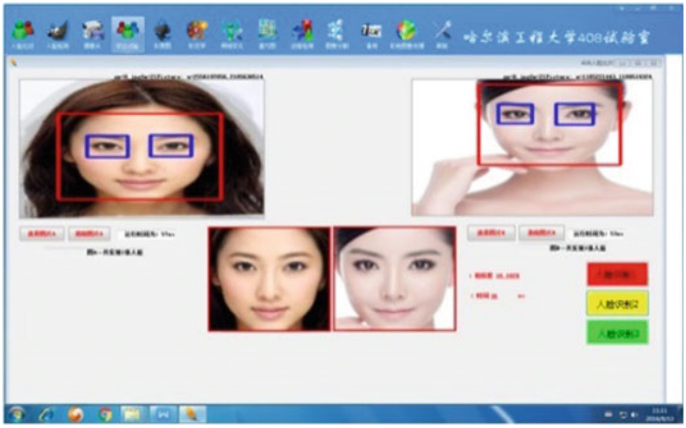

Fig. 10 Interface of facial detection and comparison result

by adding Spatial Transformer Networks. It is validated through series of experiments that our method has generalization ability. Hence, relevant application areas and topics with potential for further research.

Acknowledgement This work was supported by the Fundamental Research Funds for the Central Universities of China, Natural Science Foundation of China, and Natural Science Fund of Heilongjiang Province of China under Grand No HEUCFJ170404, 61573114, 61703119, F2015033 and QC20170702.

References

1. Li H, Nozaki T (1995) Wavelet analysis for the plane turbulent jet: analysis of large eddy structure. Jsme Int J 38(4):525–531
2. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
3. Chopra S, Hadsell R, LeCun Y (2005) Learning a similarity metric discriminatively, with application to face verification. IEEE computer society conference on computer vision and pattern recognition, vol 1, pp 539–546
4. Yuan X, Zhu QD, Lan H (2006) Multi-sensor information fusion based on rough set theory. J Harbin Inst Tech 38(10):1669–1672
5. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. International conference on neural information processing systems, vol 25. Curran Associates Inc., pp 1097–1105
6. Zbontar J, LeCun Y. (2014) Computing the stereo matching cost with a convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1592–1599
7. Ahmed E, Jones M, Marks TK (2015) An improved deep learning architecture for person re-identification. Computer vision and pattern recognition, pp 3908–3916
8. Moore B (2003) Principal component analysis in linear systems: controllability, observability, and model reduction. IEEE Trans Autom Contr 26(1):17–32

9. Sun Y, Wang X, Tang X (2015) Deeply learned face representations are sparse, selective, and robust. Computer vision and pattern recognition, pp 2892–2900
10. Zhang C-L, Zhang H, Wei X-S, Wu J (2016) Deep bimodal regression for apparent personality analysis. J Eur Conf Comput Vision
11. Xing X, Wang K, Yan T, Lv Z (2016) Complete canonical correlation analysis with application to multi-view gait recognition. Pattern Recognit 50:107–117
12. Xing X, Wang K (2016) Couple manifold discriminant analysis with bipartite graph embedding for low-resolution face recognition. Sig Process 125:329–335
13. Xing X, Wang K, Yan T, Lv Z (2015) Fusion of gait and facial features using coupled projections for people identification at a distance. IEEE Sig Process Lett 22(12):2349–2353