Vis Comput (2018) 34:1691–1699
https://doi.org/10.1007/s00371-017-1443-0

ORIGINAL ARTICLE

Deeper cascaded peak-piloted network for weak expression recognition

Zhenbo Yu[1] · **Qinshan Liu**[1] · **Guangcan Liu**[1]

Published online: 22 September 2017

Abstract Facial expression recognition is in general a challenging problem, especially in the presence of weak expression. Most recently, deep neural networks have been emerging as a powerful tool for expression recognition. However, due to the lack of training samples, existing deep network-based methods cannot fully capture the critical and subtle details of weak expression, resulting in unsatisfactory results. In this paper, we propose Deeper Cascaded Peak-piloted Network (DCPN) for weak expression recognition. The technique of DCPN has three main aspects: (1) Peak-piloted feature transformation, which utilizes the peak expression (easy samples) to supervise the non-peak expression (hard samples) of the same type and subject; (2) the back-propagation algorithm is specially designed such that the intermediate-layer feature maps of non-peak expression are close to those of the corresponding peak expression; and (3) an novel integration training method, cascaded fine-tune, is proposed to prevent the network from overfitting. Experimental results on two popular facial expression databases, CK+ and Oulu-CASIA, show the superiority of the proposed DCPN over state-of-the-art methods.

Keywords Facial expression recognition · Peak-piloted · Deep network · Cascaded fine-tune

✉ Zhenbo Yu
zbyu@nuist.edu.cn; 369415386@qq.com

Qinshan Liu
qsliu@nuist.edu.cn

Guangcan Liu
gcliu@nuist.edu.cn

[1] B-DAT, School of Information and Control, Nanjing University of Information Science and Technology, 219 Ningliu Road, Nanjing 210044, Jiangsu, China

1 Introduction

Facial expression recognition, which aims to predict six basic facial expressions including disgust, angry, fear, happy, sadness and surprise, is a classic problem in the field of computer vision. In the past few years, expression recognition has drawn considerable attention [1,4,6,13], as it can facilitate many other face-related tasks such as face recognition [17] and alignment [34]. Among various methods, deep neural networks, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have demonstrated outstanding performance in expression recognition [10,14,15,36].

However, most existing methods focus on recognizing the strong expressions of clearly separable but ignore the weak expressions that are ubiquitous in the normal communication of daily life. Due to the lack of salient features, weak expression is more difficult to recognize compared to strong expression (see Fig. 1). There has been sparse research in the direction of recognizing weak expression, e.g., [11,21,35,36]. In particular, the so-called Peak-piloted Deep Network (PPDN) proposed in [36] is specially designed for coping with weak expressions. The key idea of PPDN is about a peak-piloted feature transformation, which utilizes the intermediate-layer feature maps of peak expressions to supervise those of the corresponding non-peak expressions. This method improves the capability to capture critical and subtle details of weak expressions, and thereby it outperforms the state-of-the-art methods in facial expression recognition.

Despite these significant progress, weak expression recognition is still a very challenging problem due to two main difficulty. First, as a special task of expression recognition, weak expression recognition suffers from the common difficulty that different subjects may exhibit the same expression with diverse visual appearances and facial intensities. Sec-

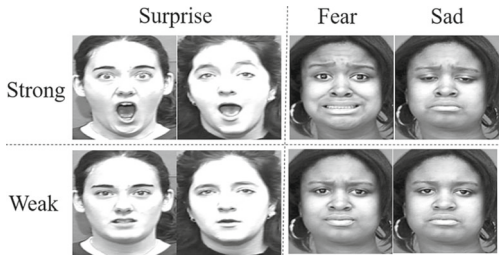

Fig. 1 Examples of strong and weak expression

ond, very often the weak expressions may not bring about dramatic changes in visual appearance, there is a great similarity between different weak expressions. For instance, as shown in Fig. 1, the weak expressions of fear and sadness are quite similar to each other [36].

In this work, we propose a novel method termed Deeper Cascaded Peak-piloted Network (DCPN) for weak expression recognition. The same as PPDN, our DCPN uses the intrinsic correlations between weak and strong expressions to magnify the critical and subtle details of weak expressions. In order to capture the critical and subtle details of weak expressions more precisely, the proposed DCPN uses a deeper, larger network architecture compared to the network used in PPDN. Furthermore, to prevent the enlarged network architecture from overfitting, we propose a new integration training method called cascaded fine-tuning.

The training process of DCPN contains three main stages as demonstrated in Fig. 2. In the first stage, the basic network of DCPN is firstly pre-trained on the ImageNet dataset and then fine-tuned for facial expression recognition. In the second stage, for every frame in each sequence, the fine-tuned network is used to generate the prediction score of the corresponding expression label. The frame with the highest score is taken as the peak expression image (e.g. the strong expression), while the others are considered as non-peak expression images. The peak expression image is often the most easily recognizable expression in each sequence, and it tends to be the last frame in the sequence which begins with a neural emotion and ends with a peak of the emotion. In the last stage, an image pair, consisting of a peak and a non-peak expression of the same type and subject, serves as an input to the network. The image pair passes through several intermediate layers to generate feature maps for each expression image. The L2-norm of the difference between the feature maps of the image pair is then minimized. This network utilizes a different back-propagation algorithm named peak gradient suppression (PGS) [36], which encourages the feature maps of a non-peak expression toward those of the corresponding peak expression. Stochastic gradient descent (SGD) [5] drives the feature maps of the image pair to be close to each other, and PGS drives the non-peak expression images toward the corresponding peak expression image.

Overall, this work is to establish a refined version of PPDN [36] in the purpose of improving the ability to recognize weak expressions so as to fundamentally improving the accuracy of facial expression recognition. Our main contributions are summarized as follows:

- Compared with the network adopted by PPDN, our DCPN uses a deeper and larger network, which can capture the subtle details of expressions more precisely and thus shows better performance in weak expression recognition.
- To prevent the enlarged network architecture from overfitting, we propose a new integration training method called cascaded fine-tuning. The experiments demonstrated on several popular facial expression recognition databases show that our method distinctly outperforms PPDN and can achieve state-of-the-art performance.

2 Related work

Deep learning algorithms have shown excellent performance of facial expression recognition in latest significant conferences [10,14,15,36] and competitions [2,7,8,30]. These methods can be divided into two categories: sequence-based and still image approach. In the first category, Liu et al. [18] propose a method called 3D CNN-DAP, which first applies a 3D convolutional network (C3D) to facial expression recognition. Jung et al. [15] propose a method called DTAGN, which integrates with a C3D and a fully connected DNN. Jaiswal et al. [14] is the first to use the CNN in combination with Bi-directional Long Short-Term Memory (BLSTM) for facial expression recognition, which outperforms the winner of the FERA 2015 challenge [30]. Fan et al. [8] propose a novel hybrid network combining RNN and C3D which wins the EmotiW2016 [7] facial expression recognition competition. In the second category, Yu et al. [32] utilize an ensemble of multiple CNNs. Bargal et al. [2] propose a hybrid network containing VGG16 [25],a modified VGG (13 layers) and Residual Network [12]. Yao et al. [31] propose a new deeper and wider network structure than inception [27]. Zhao et al. [36] propose a novel peak-piloted feature transformation, which can be utilized on all layers of the network to help to recognize the weak expression.

From above methods, using a deeper and wider network structure (more powerful capability of extracting features) and methods (e.g. multiple networks integration [8],fine-tune [10],joint fine-tune [15], etc.) of preventing the enlarged network from overfitting have become the mainstream of facial expression recognition. Sequence-based methods utilize both appearances and dynamic motions to exploit the correlations between different facial expression intensities in each sequence from the same subject, and still image meth-

ods are more generic, recognizing facial expressions in both sequences and still images.

In contrast to sequenced-based and still image methods, PPDN takes a sequence of images as a input in the training phase to take dynamics into account, and it takes one testing image as the input in the process of testing. It combines the advantages of sequence-based and still image methods. In this paper, the proposed DCPN designs a new deeper and wider network structure on the basis of PPDN and a more powerful integration training method to prevent the enlarged network from overfitting, and it has a more powerful capability to recognize weak expressions.

3 Deeper cascaded peak-piloted network

In this section, we will introduce the DCPN framework, which improves the capability to recognize the weak expression on the basis of PPDN.

3.1 Overall framework

The overall pipeline of the proposed DCPN is shown in Fig. 2. During the training process, DCPN takes an image sequence as a input. This image sequence passes through the network, which is firstly pre-trained on the ImageNet dataset and then fine-tuned for facial expression recognition. After the network, it is divided into two parts: a peak expression and non-peak expressions. This two parts are used as the new input to fine-tune the network again. This integration training method of cascaded fine-tune integrates two identical networks which shares the same parameters, and then fine-tunes the parameters of the network two times.

3.2 Basic network architecture

To further improve the performance of weak expression recognition, we design a new network inception-w as the basic network architecture on the basis of PPDN (see Fig. 3). Compared with GoogLeNet [27], which is used as the basic network architecture in PPDN, inception-w factorizes convolutions and aggressive dimension reductions to reduce much computational cost, then computational and memory savings can be used to increase both the width and depth of the network to improve the ability to capture critical and subtle details to recognize the weak expression. In contrast to Inception-v4 [26], some inception structures in Inception-w are removed due to the limited facial expression training databases.

Inception-w utilizes three different inception structures: Inception-A, Inception-B and Inception-C, which are firstly proposed in Inception-v3 [28] and deeper and wider than the traditional inception structure in GoogLeNet. And Inception-w also uses two reduction modules: Reduction-A and Reduction-B, which are firstly proposed in Inception-v4. Furthermore, Inception-w substitutes the fully convolutional

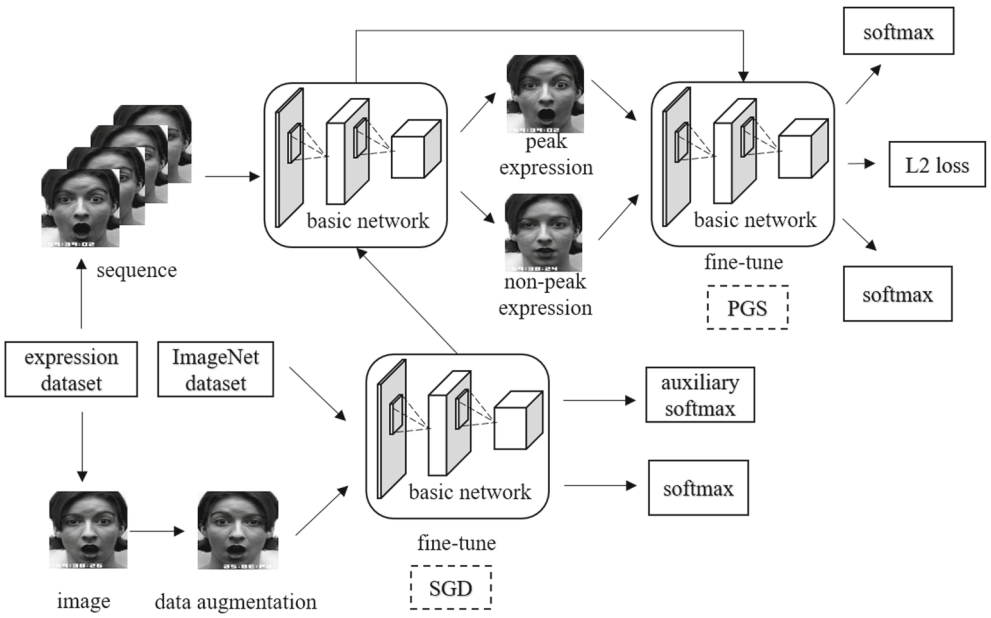

Fig. 2 Illustration of three training stages of DCPN. In the first stage, the pre-trained basic network is fine-tuned with data augmentation to get a better initialization. In the second stage, we use the basic network to choose a peak expression and non-peak expressions from each sequence. In the last stage, the resulting network is fine-tuned with peak-piloted feature transformation. During the back-propagation process, the stochastic gradient decent is used in the first stage, and peak gradient suppression is used in the last stage

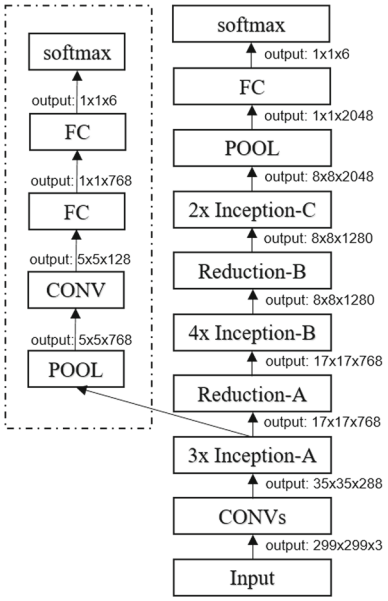

Fig. 3 Basic network architecture. The output size of each module is the input size of the next one, and the network is 42 layers deep, but the computation cost is only about 2.5 higher than that of GoogLeNet

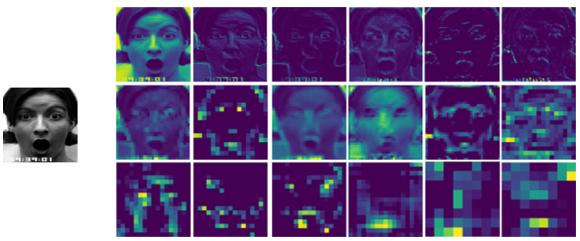

Fig. 4 Example visualizations of the different layers

(FC) layers for the fully connected layers to reduce the number of the parameters. In total, Inception-w implements three different inception structures and two reduction modules after five convolutional layers and two max pooling layers. We denote the five convolutional layers and two max pooling layers as CONVs, and the structure of the CONVs is the same as that in Inception-v3. After that, the first FC layer generates the intermediate features with 2048 dimensions, and the second FC layer generates the logit values of label predictions for six basic expressions. The auxiliary classifier is on the top of the last 17×17 layer as is shown in the dashed box of Fig. 3. The utility of auxiliary classifier is introduced to improve the convergence of the deep network. The different layers of the DCPN architecture produce feature maps as can be seen in Fig. 4

3.3 Three-stage cascaded framework

The three-stage cascaded framework can be introduced explicitly in the following:

Stage 1 Facial expression databases, e.g. CK+ [22] and Oulu-CASIA [29], provide only thousands of images. However, a typical deep network has many parameters, and this will make a deep network prone to overfitting. To overcome this problem, various data augmentation techniques are required. However, some traditional data augmentations, such as rotation, translation and random clipping, may bring noise to the facial expression databases. In this stage, each image passes three different linear transformations before being sent to the deep network. Those transformations are random horizontal flip and random changes of the brightness and saturation. Most of the models pre-trained using the ImageNet dataset outperform the model without any pre-trained due to a good initialization provided by the pre-trained models; therefore, the basic network is pre-trained using the ImageNet dataset. To combat the vanishing gradient problem in very deep network and improve the convergence during the training, the loss function is defined as the summation of an auxiliary classifier loss and a cross-entropy loss:

$$\begin{aligned} J &= \frac{1}{N}\left(J_1 + J_2 + \lambda \sum_{i=1}^{N} \|W\|^2\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i; W)) + \frac{1}{N}\mu \sum_{i=1}^{N} L(y_i, f_{\text{aux}}(x_i; W)) \\ &\quad + \lambda \|W\|^2, \end{aligned} \tag{1}$$

which is an extended version of that in PPDN [36]. Here, J_1 and J_2 indicate the auxiliary classifier loss and the cross-entropy loss, respectively. x_i is a activation feature of a batch with N training images, $y_i \in \{0, 1, 2, 3, 4, 5\}$ is a vector storing the ground truth labels. $f(x_i; W)$ and $f_{\text{aux}}(x_i; W)$ are defined as the logit values of the deep network and the auxiliary classifier. W are parameters of the deep network. L is cross-entropy loss function between the logit values of expression labels and the corresponding ground truth labels. The final regularization term is used to penalize the complexity of network parameters W. A stochastic gradient decent method is utilized for fine-tuning the network.

Stage 2 The peak frame is not known a priori in the real-world videos or some facial expression recognition databases. The image sequence from the same subject is taken as the input of the deep network, which is fine-tuned in stage 1. Then the frame with the highest prediction score in each sequence is treated as a peak expression image, while the others are treated as the non-peak expression images. This training stage is more applicable to videos where the information of the peak expression is not available.

Stage 3 During this stage, the fine-tuned deep network takes an image pair with a peak and a non-peak expression of the same type and subject as an input. This image pair

passes through the intermediate layers of the deep network to generate feature maps for each expression image. The L2-norm of the difference between the feature maps of non-peak and peak expression images is then minimized, to embed the evolution from non-peak to peak expressions into the DCPN framework. To supervise the feature maps of the non-peak expression image with those of the peak expression image, the deep network is learned by a loss function that contains the L2-norm of the difference between the feature maps of peak and non-peak expressions. We do not need the auxiliary classifier to improve the convergence of the deep network in this stage because the fine-tuned deep network has already roughly converged. Following PPDN [36], we fine-tune the deep network for the second time with a loss function defined as follows:

$$\begin{aligned} J' &= \frac{1}{N}\left(J'_3 + J'_1 + J'_2 + \lambda \sum_{i=1}^{N} \|W\|^2\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} L\left(y_i^p, f\left(x_i^p; W\right)\right) + \frac{1}{N}\sum_{i=1}^{N} L\left(y_i^n, f\left(x_i^n; W\right)\right) \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}\sum_{j\in\Omega} \| f_j\left(x_i^p; W\right) - f_j\left(x_i^n; W\right)\|^2 + \lambda\|W\|^2, \end{aligned} \tag{2}$$

where J'_3, J'_1 and J'_2 indicate the L2-norm of the difference between the feature maps of each expression image pair and two cross-entropy losses for recognition, respectively. Ω is set of layers that exploit the peak-piloted transformation, and f_j is feature map in the j-th layer. x_i^n denotes a face with non-peak expression and x_i^p denotes a face with the corresponding peak expression. To drive the intermediate-layer feature maps of non-peak expressions toward those of the corresponding peak expression, we adopt instead a special-purpose back-propagation algorithm which is based on peak gradient suppression (PGS) [36]:

$$\begin{aligned} W^+ &= W - \frac{\gamma}{N}\frac{\partial J'_3}{\partial f_j(W; x_i^n)} \times \frac{\partial f_j(W; x_i^n)}{\partial W} \\ &\quad - \frac{1}{N}\gamma\nabla_W(J'_1) - \frac{1}{N}\gamma\nabla_W(J'_2) - 2\gamma W, \end{aligned} \tag{3}$$

where γ is learning rate. The difference between SGD and PGS is that the gradients due to the feature responses of the peak expression image $-\frac{\gamma}{N}\frac{\partial J'_3}{\partial f_j(W;x_i^p)} \times \frac{\partial f_j(W;x_i^p)}{\partial W}$ are suppressed in 3. In this way, PGS supervises the feature maps of non-peak expressions toward those of peak expressions instead of making the peak and non-peak expression images get close to each other. The process of optimizing the loss function 2 in stage 3 by PGS can be seen as peak-piloted feature transformation, which is first proposed in PPDN.

3.4 Cascaded fine-tune method

Due to the large number of model parameters and small training set, the deep network is prone to overfitting. To alleviate this problem, we propose a novel integration training method called cascaded fine-tune. First, we fine-tune pre-trained deep network on the ImageNet dataset with data augmentation, this provides a good initialization of the deep network. To speed up the convergence of the network, we add an auxiliary classifier loss to our loss function. Then, we fine-tune the resulting networks by adding peak-piloted feature supervision on various layers, and this drives the feature maps of non-peak expressions toward those of peak expressions to improve the performance of weak expression recognition.

4 Experiments

Although DCPN shows great performance in weak expression recognition, we still conduct extensive experiments on two popular facial expression databases instead of micro-expression databases due to the fact that we need the strong expression to supervise the weak expressions during the training. Weak expressions must be hard samples for traditional facial expression recognition, so the improvement of the ability to recognize the weak expression can obviously improve the performance of facial expression recognition.

4.1 Data pre-processing

Most of the papers on the subject use face crop techniques to throw off the useless information for high accuracy. We utilize Multi-task Cascaded Convolutional Network (MTCNN) [33], which achieves superior accuracy over the state-of-the-art techniques for face detection and alignment, to crop faces from each dataset. In each sequence, the position of the face regions on each image cropped by MTCNN must be slightly different from each other. Face regions consist of four facial landmarks, including the coordinates of the top left corner and the lower right corner. To reduce the noise caused by resizing non-aligned face regions and the effects of scale variability caused by minimizing the L2-norm of the difference between the peak expression and non-peak expression, we choose the minimum coordinates of the top left corner and the maximum coordinates of the lower right corner to produce new face regions in each sequence to ensure that the facial frames from the same subject will be aligned with each other. Then we crop the central region of the face image with an area containing 87.5% of the resulting face region, and resize it to a size of 299 × 299 by bilinear interpolation algorithm. As can be seen from Fig. 5, some details that have nothing to do with facial expression recognition have

Fig. 5 Illustration of a standard pre-processing results in CK+, which involves face detection, central cropping and bilinear interpolation algorithm

Fig. 6 Illustration of a standard pre-processing results in oulu-CASIA, which involves face detection, zero padding and bilinear interpolation algorithm

been dropped after the data pre-processing, such as freckles, mustaches and breakouts.

As can be seen from Fig. 6, the process of data pre-processing in oulu-CASIA is a little different from that in CK+. Due to the fact that oulu-CASIA is more challenging than CK+, so we need to minimize the face regions as much as possible. We crop faces utilizing the coordinates of two eyes provided by MTCNN rather than using those of face regions directly, then determine the final rectangular face by keeping the distance between two eyes equal. Finally, we turn rectangle into square through zero padding and resize it to a size of 299 × 299.

4.2 Description of the databases

Facial expression recognition databases usually provide video sequence. We conduct all experiments on two popular databases, CK+ and Oulu-CASIA database. CK+ is a representative database for facial expression recognition. It contains six basic facial expressions and one non basic expression (contempt). It is composed of 593 sequences from 123 subjects, of which only 309 are annotated with six basic expression labels and 18 are annotated with one non basic expression label. There are 118 subjects which are divided into ten groups. Nine subsets are used for training, and the remaining subset is used for testing. In this database, each sequence starts with a neural expression and ends with a peak expression. Oulu-CASIA contains 480 image sequences of six basic facial expressions under normal illumination conditions. There are 80 subjects, and 10-fold cross-validation is performed in the same way as in the case of CK+. Similar to the CK+ database, the facial expression evolves from a neural to a peak expression.

4.3 Experimental setting

DCPN uses Inception-w as the basic network architecture. The pre-processed face regions are resized to 299 × 299.

In the first stage, the convolutional layer weights are initialized with those of the pre-trained model on the ImageNet dataset. We first fine-tune only the last two FC layers by setting the learning rate as 0.01 for 20,000 iterations, and then fine-tune all the layers by setting the learning rate as 0.0001 for 10,000 iterations. We set $\mu = 0.4$ in Eq.1. All models are trained using a weight decay λ of 0.00004 and a batch size of 32 images or image pairs. The stochastic gradient descent with a momentum of 0.9 is also used for training the network.

In the last stage, the peak-piloted feature transformation is only employed in the last two FC layers which shows the better performance than used on whole or partial layers of the network [36]. The main reason is that the peak-piloted feature transformation is more useful for supervising the highly semantic features extracted by the deep network than fine-grained ones extracted by the shallow network. The final network is fine-tuned by setting the learning rate as 0.000001 for 20,000 iterations. Following the standard setting of [23], we use 10-fold subject-independent cross-validation for evaluation in all experiments.

4.4 Evaluation on facial expression recognition

As is shown in Table 1, to evaluate the effects of various aspects of our approach and compare the performance of our approach to that of other existing approaches fairly, we divide the databases under the standard setting and conduct two sets of experiments. One for evaluating the performance of recognizing weak expressions, peak expressions and combined expressions of our method, the other for comparing the performance of our method with other existing methods on the same database.

Table 1 shows the results of data segmentation under the standard setting [36], where Weak is the number of weak expressions consisting of the 7th to 9th frames in each sequence, Strong is the number of strong expressions con-

Table 1 Standard partition of the dataset

dataset	Train	Test	Weak	Strong	Peak
CK+	3419	382	911	927	309
Oulu	6750	751	1440	1440	480

Table 2 Average accuracy on CK+ database

Method	Weak (%)	Peak (%)	combined (%)
PPDN	83.36	99.30	95.33
Inception-w	88.13	99.52	97.79
DCPN	92.48	99.60	98.28

Table 3 Average accuracy on Oulu-CASIA database

Method	Weak (%)	Peak (%)	combined (%)
PPDN	67.95	84.59	74.99
Inception-w	69.75	85.41	76.42
DCPN	72.22	86.23	77.28

Table 4 Performance comparisons of still image methods on CK+ database

Method	Average accuracy (%)
AdaGabor[3]	93.3
LBPSVM[24]	95.1
BDBN[20]	96.7
PPDN[36]	97.3
Inception-w	97.1
DCPN	98.6

sisting of the last one to three frames in each sequence, Peak is the number of peak expressions consisting of the frames with the highest prediction score in each sequence.

The main advantage of DCPN is its improved ability to capture the critical and subtle details, and it can obviously improve the performance of recognizing weak expressions. To test this, we evaluate on three different test sets, including "Peak", "Weak" and "combined". The average accuracy of 10-fold cross-validation is shown in Tables 2 and 3. "Inception-w" shows the average accuracy in the first stage of DCPN. It is obviously that our approach results in the first stage outperforms "PPDN", and the most substantial improvements are obtained on the test set of the weak expression, 92.48 and 72.22% of DCPN vs 88.13 and 69.75% of "Inception-w" on CK+ and Oulu-CASIA, respectively. This is evidence in support of the great performance of recognizing weak expressions of DCPN. And the improved performance of the weak expression recognition also facilitates the ability to facial expression recognition, DCPN outperforms "PPDN" on the combined sets, where both peak and non-peak expressions are evaluated.

Table 4 compares the DCPN to still image-based approaches on CK+, under the standard setting which uses the strong expression (e.g. the last one to three frames) for training and testing.

Table 5 Performance comparisons of sequence-based methods on CK+ database

Method	Average accuracy (%)
3DCNN-DAP[18]	92.4
STM-ExpLet[19]	94.2
DTAGN(Joint)[15]	97.3
PPDN[36]	99.3
Inception-w	99.5
DCPN	99.6

Table 6 Performance comparisons of sequence-based methods on Oulu-CASIA database

Method	Average Accuracy
HOG 3D[16]	70.63%
AdaLBP[29]	73.54%
Atlases[9]	75.52%
STM-ExpLet[19]	74.59%
DTAGN(Joint)[15]	81.46%
PPDN[36]	84.59%
Inception-w	85.41%
DCPN	86.23%

Tables 5 and 6 compare DCPN to sequence-based approaches on CK+ and Oulu-CASIA. Unlike the still-based approaches, sequence-based approaches use the image sequences for training and testing. So given an image sequence in the test phases, we use DCPN to choose the peak expression and then test the average accuracy of the peak expression. DCPN achieves better performance of facial expression recognition than other state-of-the-art methods. On the CK+ database, it has gains of 2.2% and 0.3% over "DTAGN(Joint)"[15] and "PPDN"[36]. On the Oulu-CASIA database it achieves 86.23% vs. the 81.46% of "DTAGN(Joint)"[15] and the 84.58% of "PPDN"[36].

5 Conclusions

In this paper, we propose a novel Deeper Cascaded Peak-piloted Network for weak expression recognition. We design a deeper network Inception-w and utilize the peak-piloted feature transformation to improve the performance of the weak expression recognition and then we also present a integration training method called cascaded fine-tune to prevent the deep network from overfitting. The proposed DCPN shows its improved ability to recognize the weak expression and achieve the state-of-the-art performance on two popular facial expression recognition databases.

Acknowledgements The work of Qingshan Liu is supported by National Natural Science Foundation of China (NSFC) under Grant 61532009. The work of Guangcan Liu is supported in part by NSFC under Grant 61622305 and Grant 61502238, and in part by the Natural Science Foundation of Jiangsu Province of China (NSFJPC) under Grant BK20160040.

References

1. Agarwal, S., Santra, B., Mukherjee, D.P.: Anubhav : recognizing emotions through facial expression. Vis. Comput. 1–15 (2016)
2. Bargal, S.A., Barsoum, E., Ferrer, C.C., Zhang, C.: Emotion recognition in the wild from videos using images. In: ACM International Conference on Multimodal Interaction, pp. 433–436 (2016)
3. Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Recognizing facial expression: Machine learning and application to spontaneous behavior. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005. vol. 2, pp. 568–573 (2005)
4. Chi, J., Tu, C., Zhang, C.: Dynamic 3d facial expression modeling using Laplacian smooth and multi-scale mesh matching. Vis. Comput. **30**(6–8), 649–659 (2014)
5. Chopra, S., Hadsell, R., Lecun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005. vol. 1, pp. 539–546 (2005)
6. Danelakis, A., Theoharis, T., Pratikakis, I.: A spatio-temporal wavelet-based descriptor for dynamic 3d facial expression retrieval and recognition. Vis. Comput. **32**(6–8), 1–11 (2016)
7. Dhall, A., Goecke, R., Joshi, J., Hoey, J., Gedeon, T.: Emotiw 2016: video and group-level emotion recognition challenges. In: ACM International Conference on Multimodal Interaction, pp. 427–432 (2016)
8. Fan, Y., Lu, X., Li, D., Liu, Y.: Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In: ACM International Conference on Multimodal Interaction, pp. 445–450 (2016)
9. Guo, Y., Zhao, G., Pietikainen, M.: Dynamic Facial Expression Recognition Using Longitudinal Facial Expression Atlases. Springer, Berlin (2012)
10. Han, S., Meng, Z., KHAN, A.S., Tong, Y.: Incremental boosting convolutional neural network for facial action unit recognition. Adv. Neural Inf. Process. Syst. **29**, 109–117 (2016)
11. He, J., Hu, J.F., Lu, X., Zheng, W.S.: Multi-task mid-level feature learning for micro-expression recognition. Pattern Recognit. **66**, 44–52 (2016)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
13. Hung, A.P., Wu, T., Hunter, P., Mithraratne, K.: A framework for generating anatomically detailed subject-specific human facial models for biomechanical simulations. Vis. Comput. **31**(5), 527–539 (2015)
14. Jaiswal, S., Valstar, M.: Deep learning the dynamic appearance and shape of facial action units. In: Winter Applications in Computer Vision, pp. 1–8 (2016)
15. Jung, H., Lee, S., Yim, J., Park, S.: Joint fine-tuning in deep neural networks for facial expression recognition. In: IEEE International Conference on Computer Vision, pp. 2983–2991 (2015)
16. Klaser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: British Machine Vision Conference 2008, Leeds, September (2008)
17. Li, X., Mori, G., Zhang, H.: Expression-invariant face recognition with expression classification. In: The Canadian Conference on Computer and Robot Vision, p. 77 (2006)
18. Liu, M., Li, S., Shan, S., Wang, R., Chen, X.: Deeply Learning Deformable Facial Action Parts Model for Dynamic Expression Analysis. Springer International Publishing, Berlin (2014)
19. Liu, M., Shan, S., Wang, R., Chen, X.: Learning expression lets on spatio-temporal manifold for dynamic facial expression recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1749–1756 (2014)
20. Liu, P., Han, S., Meng, Z., Tong, Y.: Facial expression recognition via a boosted deep belief network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1805–1812 (2014)
21. Liu, Y.J., Zhang, J.K., Yan, W.J., Wang, S.J., Zhao, G., Fu, X.: A main directional mean optical flow feature for spontaneous micro-expression recognition. IEEE Trans. Affect. Comput. **7**(4), 1–1 (2016)
22. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J.: The extended Cohn–Kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: Computer Vision and Pattern Recognition Workshops, pp. 94–101 (2010)
23. Metaxas, D.N., Huang, J., Liu, B., Yang, P., Liu, Q., Zhong, L.: Learning active facial patches for expression analysis. In: Computer Vision and Pattern Recognition, pp. 2562–2569 (2012)
24. Shan, C., Gong, S., Mcowan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis. Comput. **27**(6), 803–816 (2009)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv:1409.1556
26. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Workshop Track International Conference on Learning Representations, pp. 1–12 (2016)
27. Szegedy, C., Liu, W., Jia, Y., Sermanet, P.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–13 (2014)
28. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
29. Taini, M., Zhao, G., Li, S.Z., Pietikainen, M.: Facial expression recognition from near-infrared video sequences. In: International Conference on Pattern Recognition, pp. 1–4 (2011)
30. Valstar, M.F., Almaev, T., Girard, J.M., Mckeown, G.: Fera 2015 second facial expression recognition and analysis challenge. In: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, pp. 1–8 (2015)
31. Yao, A., Cai, D., Hu, P., Wang, S., Sha, L., Chen, Y.: Holonet: towards robust emotion recognition in the wild. In: The ACM International Conference, pp. 472–478 (2016)
32. Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: ACM on International Conference on Multimodal Interaction, pp. 435–442 (2015)
33. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multi-task cascaded convolutional networks. IEEE J. Solid State Circuits **23**(99), 1161–1173 (2016)
34. Zhang, Z., Luo, P., Chen, C.L., Tang, X.: Facial landmark detection by deep multi-task learning. In: European Conference on Computer Vision, pp. 94–108 (2014)
35. Zhao, R., Gan, Q., Wang, S., Ji, Q.: Facial expression intensity estimation using ordinal information. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3466–3474 (2016)
36. Zhao, X., Liang, X., Liu, L., Li, T., Han, Y., Vasconcelos, N., Yan, S.: Peak-piloted deep network for facial expression recognition. In: European Conference on Computer Vision, pp. 425–442 (2016)

Zhenbo Yu received his bachelor degree from the school of Information and Control, Nanjing University of Information Science and Technology, Nanjing, China, in 2016, where he is pursing the master degree. He took second place in 2015 and first place in 2016 in one major category of the ImageNet challenge. His research interest is facial expression analysis.

Qinshan Liu is a Professor with the School of Information and Control Engineering, Nanjing University of Information Science and Technology, Nanjing, China. He received the Ph.D. degree from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China, in 2003 and the M.S. degree from Southeast University, Nanjing, China, in 2000. He was an Assistant Research Professor with the department of Computer Science, Computational Biomedicine Imaging and Modeling Center (CBIM), Rutgers University of New Jersey, Piscataway, NJ, USA. Before joining Rutgers University, from 2010 to 2011. Before he joined Rutgers University, he was an Associate Professor with the National Laboratory of Pattern Recognition. He was a recipient of the President Scholarship of the Chinese Academy of Sciences in 2003. His research interests include image and vision analysis, including face image analysis, graph- and hypergraph-based image and video understanding, medical image analysis, and event-based video analysis.

Guangcan Liu received the bachelor's degree in mathematics and the Ph.D. degree in computer science and engineering from Shanghai Jiao Tong University, Shanghai, China, in 2004 and 2010, respectively. He was a Post-Doctoral Researcher with the National University of Singapore, Singapore, from 2011 to 2012, the University of Illinois at Urbana-Champaign, Champaign, IL, USA, from 2012 to 2013, Cornell University, Ithaca, NY, USA, in 2014. Since 2014, he has been a Professor with the School of Information and Control, Nanjing University of Information Science and Technology, Nanjing, China. His research interests touch on the areas of machine learning, computer vision, and image processing.