Journal of Systems Engineering and Electronics
Vol. 28, No. 4, August 2017, pp.784 – 792

Identity-aware convolutional neural networks for facial expression recognition

Chongsheng Zhang[1], Pengyou Wang[1], Ke Chen[2,*], and Joni-Kristian Kämäräinen[2]

1. The Big Data Research Center, Henan University, Kaifeng 475001, China;
2. Department of Signal Processing, Tampere University of Technology, Tampere 33720, Finland

Abstract: Facial expression recognition is a hot topic in computer vision, but it remains challenging due to the feature inconsistency caused by person-specific characteristics of facial expressions. To address such a challenge, and inspired by the recent success of deep identity network (DeepID-Net) for face identification, this paper proposes a novel deep learning based framework for recognising human expressions with facial images. Compared to the existing deep learning methods, our proposed framework, which is based on multi-scale global images and local facial patches, can significantly achieve a better performance on facial expression recognition. Finally, we verify the effectiveness of our proposed framework through experiments on the public benchmarking datasets JAFFE and extended Cohn-Kanade (CK+).

Keywords: facial expression recognition, deep learning, classification, identity-aware.

DOI: 10.21629/JSEE.2017.04.18

1. Introduction

Facial expressions are important cues for affective computing, i.e. recognising, interpreting, processing, and simulating human affects, hence it is very important for human-computer interaction. In the light of this, the problem of recognising human expressions, given a facial image or a set of images (termed as facial expression recognition), is one of the active research topics in computer vision and has both academic and social significances. In a typical protocol for the facial expression recognition, the face is first detected, next imagery features are extracted, then these features will be classified (using a classification model) to pre-defined expression classes. The problem is therefore formulated into a classification framework to learn a discriminative classification function from a low-level imagery representation to express class labels (e.g., anger, disgust, fear, happy, sadness, surprise, neutral).

In the existing frameworks for the facial expression recognition, robust and invariant features have a very important influence on the recognition performance; but after the rise of deep learning in the last 3 – 5 years, the end-to-end learning approach by deep convolutional neural networks (CNNs [1 – 3]) is considered as the state-of-the-art approach which automatically discovers the best imagery features for the classification/pattern recognition. In a number of visual recognition problems, such as object categorisation [1,4] and scene understanding [5 – 7], the whole image regions are usually fed into the deep learning framework with the corresponding class labels during training. In this work, we will show that, using a person specific local part detection as a pre-stage prior to deep the CNN training, we can achieve the improved performance. We name our approach as identity-inspired CNN (I^2CNN) which owes its strength to the fact that the current facial image databases contain a too small number of examples for CNNs to learn the subtle person-specific details that are characteristics for the facial expression recognition.

CNN has also been verified its effectiveness on face verification [8 – 12] to cope with the expression variation (from neutral to non-neutral expressions). Intuitively, the success of deep convolutional networks for face verification indicates that deep learning models can capture person-specific characteristics across identities under the neutral expression. The concept of this work is rather simple: inspired by the deep identity nework (DeepID-Net), this paper adopts a similar design principle, by employing a number of global and local image patches generated from the detected facial images to train a more discriminative deep model than the original CNN that only uses the whole facial images. The rational of such a setting can be explained by the fact that the generalisation capability of the deep learning model can be significantly improved when increasing the variety of training samples (more global and

Manuscript received July 03, 2016.
*Corresponding author.
This work was supported by the Academy of Finland (267581), the D2I SHOK Project from Digile Oy as well as Nokia Technologies (Tampere, Finland).

local image patches than only the whole images in our case). In other words, facial patches used in our proposed framework can provide both local and global identity evidences to mitigate the negative influence of inter-person variations on the facial expression recognition.

The novelties and contributions of this work are summarized as follows.

(i) Our work adopts the state-of-the-art approach, CNN, for the problem of the facial expression recognition.

(ii) Different from the conventional deep CNN methodology, we propose the concept of I^2CNN which, through the detection of local face parts, can better capture and exploit inter-identity variations of expressions.

(iii) Our approach achieves a superior performance on the two popular benchmarks of facial expression recognition: the JAFFE and CK+ datasets.

2. Related work

2.1 Facial expression recognition

Facial expression recognition is usually formulated into a multi-class classification problem, i.e., classify the facial images or frames into independent expression categories [2,3,13 – 22]. Beyond the expression recognition on videos [13 – 16], there are many researches focusing on recognising facial expressions from still images [2,3,17,19,21,22]. The existing methods can be categorised into shallow methods [17,18,21,22] and deep architecture based methods [2,3,20,23].

Before deep learning methods rising in [1], the pipeline for the facial expression recognition is usually as follows: imagery feature extraction including geometry based and appearance based methods, e.g., spatially-pooling local binary patterns (LBP) [22], then classification function learning with, e.g., neural networks [24 – 26], support vector machines [22,27] or Bayesian network [28]. The method in [22] obtains a better performance through using boosting LBP features to enhance the performance of the facial expression recognition. Recently, deep belief networks (DBNs) have been applied to this problem [2,3,23], which achieves significant improvements in the recognition performance.

In [18], texture and shape features are used for facial expression recognition. Reference [3] detected facial components by a face parse detector, which was trained via DBNs. Reference [2] proposed combining the deep discriminative features of DBNs and the multi-layer perceptron (MLP) classification method, which outperformed state-of-the-art methods. Reference [19] used two deep networks: one extracted the temporal appearance features, the other obtained the temporal geometry features based on the facial landmark points, these features were then combined using an integration approach to improve the performance of the facial expression recognition. In order to address the problem of the non-frontal facial expression recognition, [20] first extracted the scale invariant feature transform (SIFT) features of facial images, then fed the extracted SIFT features to deep neural networks to learn an optimal set of discriminative feature vectors. While the facial expression in the wild is still a difficult problem, [29] achieved a better performance with deep neural networks. Besides still images, a few approaches have been proposed to understand image sequences, where many of these methods employ dynamic Bayesian networks [14,30,31].

The common spirit shared with the frameworks in [2,3,19,20,23] and our proposed I^2CNN method lies in the local image patch based processing, which can capture fine-grained details and dissimilarity across person identities. Nevertheless, the existing deep learning models i.e. (DBNs) [2,3,19,20,23] still depend on manually-designed, "engineered" features, which motivates us to introduce an end-to-end deep learning framework based on the recent convolutional neural network architectures.

2.2 Convolutional neural networks

Convolutional neural networks have attracted wide attention during the last few years and have demonstrated their effectiveness on a number of vision applications such as face recognition [8 – 12], face detection [32 – 35], pedestrian detection [36 – 41], object categorisation [1,4], scene understanding [5 – 7] and other tasks [42 – 50]. Specifically, deep models are designed as the end-to-end learning manner that exploits the powerful hierarchical architecture and thousands or millions of network parameters to learn features and the classification function. Nowadays, neural networks are becoming deeper and deeper, e.g., VGGNets [11] have 16 or 19 layers, GoogleNets [51] have 22 layers, while ResNets [52] have 50,101,152 layers. These approaches can generally achieve better performances on many pattern recognition tasks.

Inspired by the success of DeepID-Net for the face classification, we propose a so-called I^2CNN approach to address the facial expression recognition problem, which will be elaborated in the following section.

3. Methodology

The biggest challenge in the facial expression recognition lies in the large visual feature variations caused by person-specific characteristics of expressions and variations from extrinsic conditions such as illumination and the view point. As shown in Fig. 1, "fear" is represented with large appearance variations. For instance, some people have evident wrinkles around their mouth (top row), while others have clearly dilated eyes (bottom row).

Fig. 1 Different people expressing the same facial expressions differently

The difference between our proposed I^{2}CNN approach and the CNN methods lies in the imagery input of deep models. Specifically, not only the whole image regions of faces, but also the facial components in the form of local patches are employed in our proposed I^{2}CNN model. The work flow of our proposal in comparison with the conventional approach is illustrated in Fig. 2.

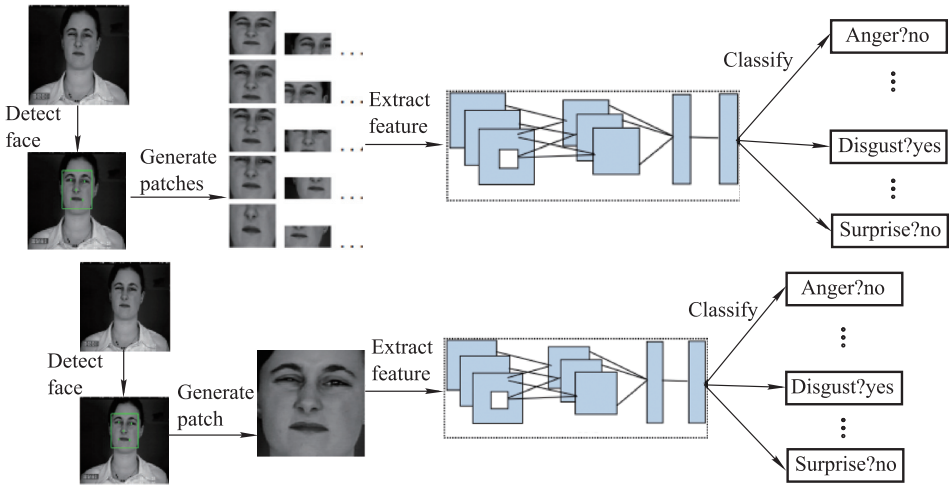

Fig. 2 Identity-aware (top) versus the conventional CNN approach for facial expression recognition

Our proposed framework consists of the following steps:

Step 1 Given the training facial images and their expression labels, we first localize the face foreground area from the image background. Then, we generate a number of face part patches which contain both the local and global personal identity information.

Step 2 We feed the generated image patches into CNN for deep model training. The same network is shared by all patches for computational convenience.

Step 3 The second to the last layer outputs are concatenated as features and a strong classifier, support vector machine (SVM), is trained upon these features for predicting the expression classes.

For testing, we first adopt a face detector to a new image/instance, then feed the detected face areas to the trained I^{2}CNN model to produce the classification outcome.

3.1 Face localisation and facial patches generation

To generate patches that are robust to identity-specific deformable face movements due to the expressions, we need a method to detect the facial foreground area and face specific landmark areas — in this sense, our approach adopts a part-based facial representation.

For the problem of face and facial parts detection, we utilize the recent CNN-based approach by [42]. The method is employed to localize the facial foreground and feature points (i.e., the centers of two eyes, the nose tip and

the corners of the mouth). Specifically, to utilize a coarse-to-fine framework, we perform the detection at multiple scales, for which different network structures are adopted.

The motivation of generating a number of patches is to capture the subtle changes of both local isolated and global correlated facial components for specific expression. For image patch generation, we also adopt the method in [8]. In details, given a cropped facial foreground and detected facial feature points, we generate ten image patches with three scales (see Fig. 3). For generalisation, we horizontally flip all the image patches, thus obtain a total number of 60 patches for each face image.

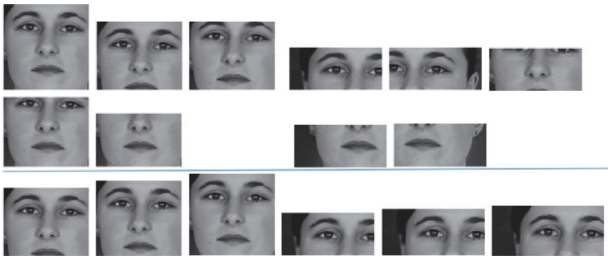

Fig. 3 Example patches produced by the first stage of our approach: the patch detection CNN by [8]

3.2 Convolutional network structure

For fair comparisons, we adopt the original structure of the AlexNet network [1] for which public implementations are available. As shown in Fig. 4, the network consists of five convolutional layers, three fully-connected layers and pooling layers. The length, width and height of each cubiod denote the number of the feature map and the dimensions of the feature map. The square in cubiods denotes the kernel size. The number which is marked under the last three fully connected layers is the number of neurons of each layer. The final layer is the classification layer for which the classification error is back-propagated to the network. The input of the network is 227×227, and the dimension of the final layer varies according to the number of classes it predicts.

The important operations of CNN are convolution, pooling and ReLU ($y = \max(0, x)$), etc. The shape of the output feature map is calculated by

$$h\_o = (h\_i - kernel\_size\_h + 2 \times pad\_h)/\text{stride\_}h + 1$$
$$w\_o = (w\_i - kernel\_size\_w + 2 \times pad\_w)/\text{stride\_}w + 1. \tag{1}$$

In (1), h_i, w_i specify the height and width of the input feature map respectively, likewise, h_o, w_o denote the height and width of the output feature map, $kernel\_size\_h$, $kernel\_size\_w$ denote the shapes of weight filter, pad_h, pad_w specify the numbers of pixels which are added to each side of the input feature map. $stride\_h$, $stride\_w$ denote the intervals between filter regions which are applied to the input feature map. The convolution operation of the CNN is expressed in (2).

$$y_j = \max(0, \sum_i (w_{i,j} * x_i) + b_j) \tag{2}$$

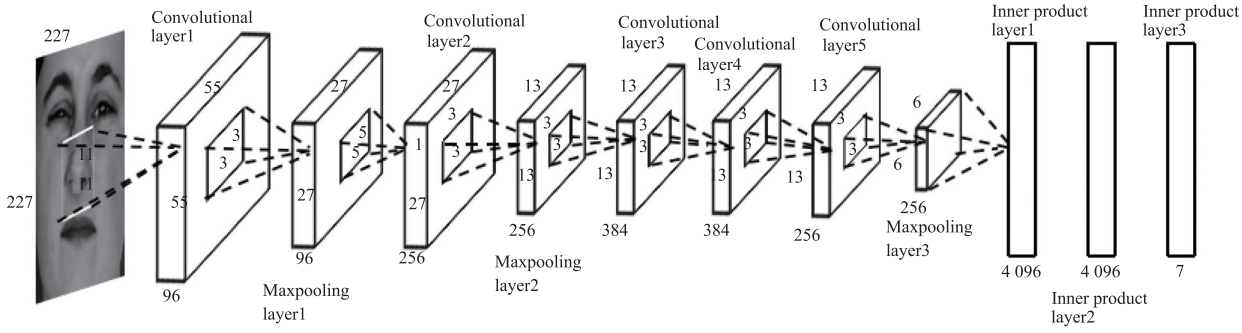

Fig. 4 Structure of our I^{2}CNN approach

In (2), $*$ denotes convolution, x_i and y_j specify the ith input feature map and jth output feature map, respectively. b_j is the bias of the jth output feature map. $w_{i,j}$ denotes the weight filter between the ith input feature map and the jth output feature map. The weight filter is locally shared with the same values to learn different features in different regions for neurons in the same output feature map.

The final layer uses the softmax function to denote the probability distribution of different classes. The function is given in (3).

$$y_j = \frac{\exp(y'_j)}{\sum_i (\exp(y_i))} \tag{3}$$

In (3), $y'_j = \sum_i^{4\,096} (w_{i,j} * x_i) + b_j$, that is, the input is the linear combination of the $4\,096$ dimensions as one feature, the output is y_j for each neuron in the final layer. The sample will be predicted to be the class which $\max(y_j)$ stands for.

The configurations of the layers and the network structure have strong effect on the classification performance. Having the ReLU layer in the network, we employ the outputs of the penultimate layer as 4 096-dimensional deep feature representation extracted for each image patch. For the 60 patches (Section 3.1), the final size of the feature vector for facial expression classification is $60\times4\,096$. The feature vector is fed to the SVM classifier which outputs the final classification. We use the feature vector of the train data and validation data to train an SVM model, then classify the test data using the models in each fold. We use the average value of the 10-fold validations as the final accuracy.

3.3 Implementation details

During training, after the 60 patches have been detected, they are rescaled to the size of 256×256, from which the CNN inputs are randomly cropped into patches of 227×227. During testing, the same procedure is performed, except that the 227×227 cropped patches are taken from the centre of each detected patch. We use the training data and validation data to train a deep model according to the above criteria. After that, we use the Matlab Interface of Caffe [53] and the trained deep model to extract the deep feature for the training data, validation data, test data, respectively.

For the SVM classifier, we adopt 10-fold cross-validation to tune the training parameters. In deep model learning, the learning rate is initially set to 10^{-2}, then decreased after each 50 000 epocs by a factor of 10, until the accuracy is not increased any longer. We use the weight-decay strategy and the dropout (using for the first two fully-connected layers) method to avoid over-fitting. The former was initially set to 5×10^{-4}, while the ratio of the dropout layer is initially set to 0.5.

4. Experiments

4.1 Datasets and settings

For the facial expression recognition evaluation, we use the popular benchmark datasets: JAFFE [54] and CK+ [55]. For the JAFFE dataset, we select all the images for our experiments belonging to 10 subjects, having 213 images in total. Each sequence has seven basic emotions (i.e., neutral, anger, disgust, fear, happy, sadness, surprise). Illustrative images of the JAFFE dataset are shown in Fig. 5.

Fig. 5 Illustrative examples from the JAFFE dataset

CK+ dataset contains the face images of 593 image sequences belonging to 123 persons captured with varying appearance and under the controlled indoor environment. The illustrative images of the CK+ dataset are shown in Fig. 6. For the CK+ dataset, we select 309 sequences for our experiments, the selection criterion is that the sequence can be labelled as one of the six basic non-neutral emotional expressions (anger, disgust, fear, happy, sadness, surprise). Following the setting in [22,23], for each sequence, we select the first frame and the last three peak frames for our experiments.

Fig. 6 Illustrative examples from the CK+ dataset

4.2 Comparisons with state-of-the-art methods

We conduct two experiments to compare with the state-of-the-art methods for the facial expression recognition on the JAFFE and CK+ datasets, whose results are shown in Tables 1 and 2. On one hand, we achieve better classification performance than the state-of-the-art for the first experiment on the JAFFE dataset. On the other hand, the superior performance of our proposed I^{2}CNN method on the CK+ dataset can be achieved for both six and seven emotions recognition, whose results are shown in Table 2. For both cases, our accuracies for each emotion class are very high, which can be seen in the confusion matrices in Fig. 7. In Fig. 8, we depict the results from [22] in confusion matrices. Reference [22] used LBP features of images and the SVM clasifier with RBF kernel. Compared to Fig. 8, we obtain a higher accuracy rate for almost all the expressions, except fear and sadness. The improvements in the performance of the facial expression recognition are depicted by the colors in the confusion matrices.

Table 1 Comparative evaluation on JAFFE dataset

Method	Accuracy/%
I^{2}CNN (our)	75.280 6
SVM (RBF)+Boosted-LBP ([22])	81.0
SVM (linear)+Boosted-LBP ([22])	79.8
SVM (polynomial)+Boosted-LBP ([22])	79.8

Table 2 Comparative evaluation on the CK+ dataset with six and seven emotion classes

Method	6 emo	7 emo
I^{2}CNN (our)	98.3	96.2
BDBN ([23])	96.7	-
SVM(RBF) + Boosted-LBP ([22])	95.1	91.4
FP + SAE ([3])	-	91.1
SVM(RBF) + LBP ([22])	92.6	88.9
SVM(RBF) + Gabor ([22])	89.8	86.8
CSPL+SVM ([56])	-	89.9
LDP+template matching ([57])	-	86.9
ITBN ([58])	-	86.3
Geometric representation + Gaussian three-augmented navie Bayes classifiers ([13])	73.2	-

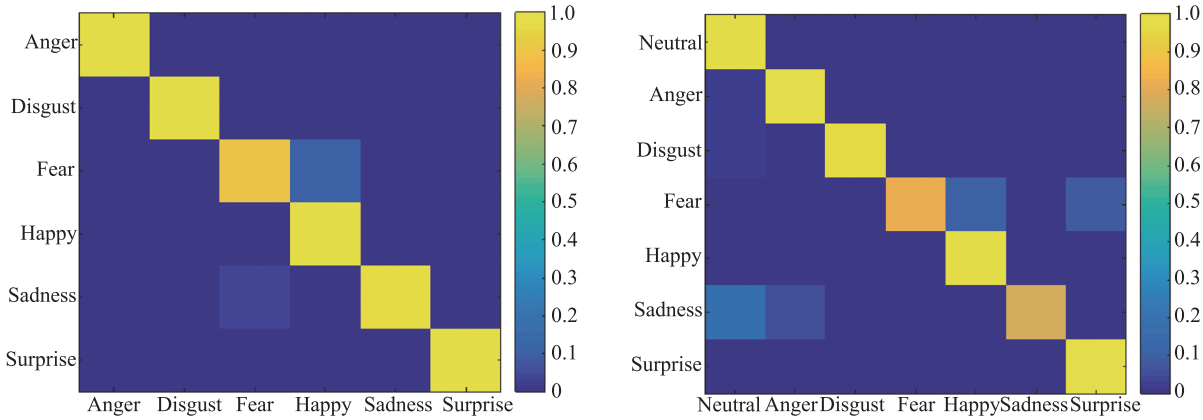

Fig. 7 I^{2}CNN confusion matrices for CK+ six (left) and seven (right) emotion classifications

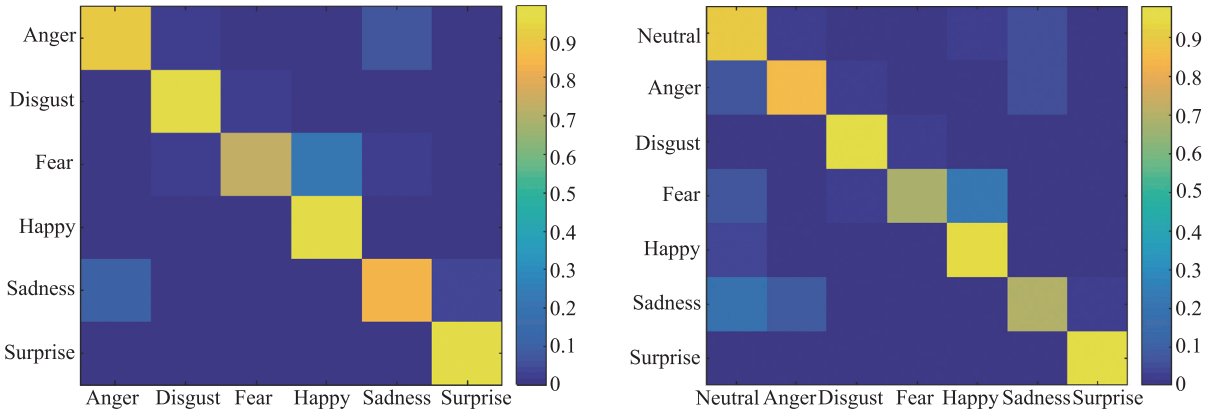

Fig. 8 SVM(RBF) + LBP ([22]) confusion matrices for CK+ six (left) and seven (right) emotion classifications

4.3 Identity-inspired versus conventional convolutional networks

As shown in Table 3, we evaluate the proposed I^{2}CNN and the convolutional CNN [1], respectively. We can observe that, better classification performances of the proposed I^{2}CNN can be obtained on both datasets, with a clear margin that can only be explained by the identity-inspired parts. Experimental results on the CK+ dataset show that the classification accuracy of six emotions recognition and seven emotions recognition can be increased to 98.3% and 96.2% compared to the conventional CNN approach. The accuracy of seven emotions on JAFFE can be increased by 19.76%. In Section 4.4, we evaluate the influence of the number of image patches on the performance of recognition.

Table 3 I^{2}CNN versus conventional CNN

Method	Accuracy/%		
	JAFFE	CK+ (6)	CK+ (7)
I^{2}CNN	75.280 6	98.3	96.2
CNN	62.860 2	92.2	89.7

4.4 Evaluation on the number of image patches

We evaluate the effect of image patches to the performance of I^{2}CNN. We try different combinations of patches for k = 1,5,60 patches, the results are given in Table 4. k = 1 means that we only use the one global patch of the image, k = 5 means that we only use the five local patches based on the detected five facial feature points, while k = 60 denotes that we use all the patches of each image. The performance of seven emotions on the JAFFE dataset is improved from 62.860 2% to 75.280 6% compared to the single patch. Obviously, we can achieve a better performance on the facial expression recognition by improving the patches of each image. Similarly, using all patches we can improve the performance by 6.62% and 7.25%, compared to the single patch for six emotions and seven emotions on the CK+ dataset, respectively. Table 4 shows that we can improve the performance by using more facial patches.

Table 4 Evaluation of the number of patches used in I^{2}CNN

Method	Accuracy/%		
	JAFFE	CK+ (6)	CK+ (7)
I^{2}CNN-60	75.280 6	98.3	96.2
I^{2}CNN-5	70.982 5	94.7	93.5
I^{2}CNN-1	62.860 2	92.2	89.7

5. Conclusions

Facial expression recognition has many real-world applications, including sentiment analyses, human-computer interaction (HCI), and automatic driving solution where we can monitor the mood/emotions of a driver so as to avoid potential accidents. The main issue in the facial expression recognition is how to extract key facial features that can better differentiate different emotions, where people can utilize the traditional, engineered feature extraction methods such as LBP [59], Gabor [60], and SIFT [61]. However, in recent years, CNN has been proved to be able to automatically extract features from low to high levels and outperform state-of-the-art in recognition accuracy in many pattern recognition tasks.

In this work, we propose a novel I^{2}CNN approach to mitigate the negative effect of inter-person expression variations on the face recognition. Compared to the existing conventional CNN methods, our proposal is based on multi-scale global images and local facial patches, it can achieve significant performance improvement on the JAFFE and CK+ datasets.

However, there are still many open challenges. Firstly, the conditions of illumination, the pose of human faces, the difference of races, have significant influences on the facial

expression recognition. Secondly, tuning the best parameters for CNN is very time-consuming. To solve this problem, in the future work we will design a simple but more effective CNN network for the facial expression recognition.

Acknowledgement

The authors would like to acknowledge CSC-IT Center for Science, Finland for generous computational resources, and the donations of Tesla K40 GPU from NVIDIA for supporting our academic research.

References

[1] A. Krizhevsky, I. Sutskever, G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 2012: 1097 – 1105.

[2] X. Zhao, X. Shi, S. Zhang. Facial expression recognition via deep learning. *IETE Technical Review*, 2015: 1 – 9.

[3] Y. Lv, Z. Feng, C. Xu. Facial expression recognition via deep learning. *Proc. of the International Conference on Smart Computing*, 2014: 303 – 308.

[4] K. He, X. Zhang, S. Ren, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. *Proc. of the European Conference on Computer Vision*, 2014: 346 – 361.

[5] R. Girshick, J. Donahue, T. Darrell, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 580 – 587.

[6] S. Gupta, P. Arbeláez, R. Girshick, et al. Indoor scene understanding with RGB-D images: bottom-up segmentation, object detection and semantic segmentation. *International Journal of Computer Vision*, 2015, 112(2): 133 – 149.

[7] H. Jung, M. K. Choi, K. Soon, et al. End-to-end pedestrian collision warning system based on a convolutional neural network with semantic segmentation, *arXiv*: 1612. 06, 2016.

[8] Y. Sun, X. Wang, X. Tang. Deep learning face representation from predicting 10,000 classes. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 1891 – 1898.

[9] Y. Sun, X. Wang. Hybrid deep learning for face verification. *Proc. of the IEEE International Conference on Computer Vision*, 2013: 1489 – 1496.

[10] F. Schroff, D. Kalenichenko, J. Philbin. Facenet: a unified embedding for face recognition and clustering. *Computer Vision and Pattern Recognition*, 2015: 815 – 823.

[11] O. M. Parkhi, A. Vedaldi, A. Zisserman. Deep face recognition. *Proc. of the British Machine Vision Conference*, 2015: 1 – 12.

[12] W. Ouyang, X. Wang, X. Zeng, et al. Deepid-net: deformable deep convolutional neural networks for object detection. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015: 2403 – 2412.

[13] I. Cohen, N. Sebe, A. Garg, et al. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, 2003: 160 – 187.

[14] R. E. Kaliouby, P. Robinson. Real-time inference of complex mental states from facial expressions and head gestures. *Real-time Vision for Human-Computer Interaction*, DOI: 10. 10710-387-27890-7_11.

[15] I. Kotsia, I. Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Trans. on Image Processing*, 2007: 172 – 187.

[16] M. Pantic, I. Patras. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Trans. on Systems, Man, and Cybernetics-Part B*, 2006, 36(2): 433 – 499.

[17] P. K. Manglik, U. Misra, H. B. Maringanti, et al. Facial expression recognition. *Proc. of the IEEE International Conference on Systems, Man and Cybernetics*, 2004: 2220 – 2224.

[18] W. Zheng, C. Liu. Facial expression recognition based on texture and shape. *Proc. of the Wireless and Optical Communication Conference*, 2016: 1 – 5.

[19] H. Jung, S. Lee, J. Yim, et al. Joint fine-tuning in deep neural networks for facial expression recognition. *Proc. of the IEEE International Conference on Computer Vision*, 2015: 2983 – 2991.

[20] T. Zhang, W. Zheng, Z. Cui, et al. A deep neural network driven feature learning method for multi-view facial expression recognition. *IEEE Trans. on Multimedia*, 2016, 18(12): 2528 – 2536.

[21] D. Ghimire, J. Lee. Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *Sensors*, 2016: 7714 – 7734.

[22] C. Shan, S. Gong, P. W. McOwan. Facial expression recognition based on local binary patterns: a comprehensive study. *Image and Vision Computing*, 2009: 803 – 816.

[23] P. Liu, S. Han, Z. Meng, et al. Facial expression recognition via a boosted deep belief network. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 1805 – 1812.

[24] C. Padgett, G. W. Cottrell. Representing face images for emotion classification. *Advances in Neural Information Processing Systems*, 1997: 894 – 900.

[25] Y. L. Tian. Evaluation of face resolution for expression analysis. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2004: 82 – 82.

[26] Z. Zhang, M. Lyons, M. Schuster, et al. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. *Proc. of the IEEE International Conference on Automatic Face and Gesture Recognition*, 1998: 454 – 459.

[27] M. S. Bartlett, G. Littlewort, M. Frank, et al. Recognizing facial expression: machine learning and application to spontaneous behavior. *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005: 568 – 573.

[28] I. Cohen, N. Sebe, A. Garg, et al. Facial expression recognition from video sequences. *Proc. of the IEEE International Conference on Multimedia and Expo*, 2002: 121 – 124.

[29] A. Mollahosseini, B. Hassani, M. J. Salvador, et al. Facial expression recognition from world wild web. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016: 1509 – 1516.

[30] J. Hoey, J. J. Little. Value directed learning of gestures and facial displays. *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004: 1026 – 1033.

[31] Y. Zhang, Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2005, 27(5): 699 – 714.

[32] M. Szarvas, A. Yoshizawa, M. Yamamoto, et al. Multi-view face detection using deep convolutional neural networks. *Proc. of the ACM International Conference on Multimedia Retrieval*, 2015: 224 – 229.

[33] S. Yang, P. Luo, C. L. Chen, et al. Wider face: a face detection benchmark. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 5525 – 5533.

[34] H. Li, Z. Lin, X. Shen, et al. A convolutional neural network cascade for face detection. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015: 5325 – 5334.

[35] Y. Zheng, C. Zhu, K. Lu, et al. Towards a deep learning framework for unconstrained face detection. *Proc. of the 8th IEEE*

International Conference on Biometrics: Theory, Applications and Systems, 2016: 1 – 8.
[36] X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012: 3258 – 3265.
[37] W. Ouyang, X. Zeng, X. Wang. Modeling mutual visibility relationship in pedestrian detection. *Computer Vision and Pattern Recognition*, 2013: 3222 – 3229.
[38] X. Zeng, W. Ouyang, X. Wang. Multi-stage contextual deep learning for pedestrian detection. *Proc. of the IEEE Conference on Gomputer Vision*, 2013: 121 – 128.
[39] P. Luo, X. Wang, X. Tang. Pedestrian parsing via deep decompositional network. *Proc. of the IEEE International Conference on Computer Vision*, 2013: 2648 – 2655.
[40] P. Luo, Y. Tian, X. Wang, et al. Switchable deep network for pedestrian detection. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 899 – 906.
[41] X. Zeng, W. Ouyang, M. Wang, et al. Deep learning of scene-specific classifier for pedestrian detection. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 472 – 487.
[42] Y. Sun, X. Wang, X. Tang. Deep convolutional network cascade for facial point detection. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013: 3476 – 3483.
[43] P. Sermanet, D. Eigen, X. Zhang, et al. Overfeat: integrated recognition, localization and detection using convolutional networks. *arXiv*: 1312.6229, 2014.
[44] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*: 1409. 1556, 2014.
[45] K. He, X. Zhang, S. Ren, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2015, 37(9): 1904 – 1916.
[46] W. Ouyang, X. Chu, X. Wang. Multi-source deep learning for human pose estimation. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 2337 – 2344.
[47] J. Zhang, S. Shan, M. Kan, et al. Coarse-to-fine auto-encoder networks for real-time face alignment. *Proc. of the European Conference on Computer Vision*, 2014: 1 – 16.
[48] K. Simonyan, A. Vedaldi, A. Zisserman. Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv*: 1312. 6034, 2014.
[49] R. C. Malli, M. Aygun, H. K. Ekenel. Apparent age estimation using ensemble of deep learning models. *arXiv*: 1606. 02909, 2016.
[50] X. Tang, X. Wang, P. Luo. Hierarchical face parsing via deep learning. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012: 2480 – 2487.
[51] C. Szegedy, W. Liu, Y. Jia, et al. Going deeper with convolutions. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 1 – 9.
[52] K. He, X. Zhang, S. Ren, et al. Deep residual learning for image recognition. *arXiv*: 1512. 03385, 2015.
[53] Jia, Yangqing, Shelhamer, et al. Caffe: convolutional architecture for fast feature embedding. *arXiv, Computer Science*, 2014: 675 – 678.
[54] M. Lyons, S. Akamatsu, M. Kamachi, et al. Coding facial expressions with gabor wavelets. *Proc. of the IEEE International Conference on Automatic Face and Gesture Recognition*, 1998: 200 – 205.
[55] P. Lucey, J. F. Cohn, T. Kanade, et al. The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2010: 94 – 101.
[56] L. Zhong, Q. Liu, P. Yang, et al. Learning active facial patches for expression analysis. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012: 2562 – 2569.
[57] T. Jabid, M. H. Kabir, O. Chae. Robust facial expression recognition based on local directional pattern. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, 51: 784 – 794.
[58] Z. Wang, S. Wang, Q. Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013: 3422 – 3429.
[59] T. Ojala, M. Pietikainen, T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2002: 971 – 987.
[60] L. Wiskott, J. M. Fellous, N. Kuiger, et al. Face recognition by elastic bunch graph matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1997, 19: 775 – 779.
[61] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, 60: 91 – 110.

Biographies

Chongsheng Zhang was born in 1982. He is a full professor of Henan University, China, where he is also the director of the big data research. He received his Ph.D. degree at INRIA, France. He has published more than 20 papers in peer-reviewed conferences and journals, including IEEE ICDM, PAKDD. He has (co-) authored three books and holds three patents. His research interests include data classification, data stream mining and deep learning.
E-mail: Chongsheng.zhang@yahoo.com

Pengyou Wang was born in 1990. He is currently a second-year master student at Henan University, China. He has lead or participated in many projects, including knowledge graph based intelligent job matching system, property management system, facial expression recognition and insect sound recognition. He has co-authored one book which is about deep learning and the usage of Caffe. His research interests are deep learning and pattern recognition.
E-mail: 1204287950@qq.com

Ke Chen was born in 1985. He received his Ph.D degree in computer vision at the School of Electronic Engineering and Computer Science, Queen Mary, University of London, UK. He is currently the Academy of Finland post-doctoral research fellow at the Department of Signal Processing, Tampere University of Technology. His research interests include computer vision, pattern recognition, neural dynamic modelling, and robotic inverse kinematics.
E-mail: ke.chen@tut.fi

Joni-Kristian Kämäräinen was born in 1974. He is an associate professor of signal processing at the Department of Signal Processing, Tampere University of Technology, Finland. He received his M.S. and Ph.D. degrees from Lappeenranta University of Technology in 1999 and 2003, respectively. He leads the Computer Vision Group and his research focuses on 2D and 3D scene analysis, object detection and recognition, signal processing and machine intelligence.
E-mail: joni.kamarainen@tut.fi