Facial Expression Recognition Method Based on Cascade Convolution Neural Network

Weida Liu
Department of Electrical Engineering
Jilin Engineering Normal University
Jilin, China
e-mail: liuweida2021@126.com

Jian Fang*
Department of Electrical Engineering
Jilin Engineering Normal University
Jilin, China
e-mail: fangj@jlenu.edu.cn
*corresponding author

2021 International Wireless Communications and Mobile Computing (IWCMC) | 978-1-7281-8616-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/IWCMC51323.2021.9498621

***Abstract*—In view of the problem that the convolution neural network research of facial expression recognition ignores the internal relevance of the key links, which leads to the low accuracy and speed of facial expression recognition, and can't meet the recognition requirements, a series cascade algorithm model for expression recognition of educational robot is constructed and enables the educational robot to recognize multiple students' facial expressions simultaneously, quickly and accurately in the process of movement, in the balance of the accuracy, rapidity and stability of the algorithm, based on the cascade convolution neural network model. Through the CK+ and Oulu-CASIA expression recognition database, the expression recognition experiments of this algorithm are compared with the commonly used STM-ExpLet and FN2EN cascade network algorithms. The results show that the accuracy of the expression recognition method is more than 90%. Compared with the other two commonly used cascade convolution neural network methods, the accuracy of expression recognition is significantly improved.**

Keywords—Convolution Neural Network, Facial Expression Recognition, Cascade Algorithm

I. INTRODUCTION

The deep convolution neural network model developed by Professor Hinton's student Alex et al won the championship in the Image Net visual recognition competition in 2012, greatly surpassing the traditional methods, and reduced the error rate of image classification from 26.2% to 15.3% [1, 2]. It is proved that convolution neural network (CNN) enhance figure classification accuracy [3] and be an ideal facial expression recognition method. Since then, the research of convolution neural network in facial expression recognition has been developed rapidly. this paper studies the indispensable links of facial expression recognition, such as face contour detection [4, 5], face marker location [6-8] and face key point recognition [9, 10].

However, these studies ignore the internal relationship of these links, resulting in low accuracy and speed of facial expression recognition, which can't meet the requirements of facial expression recognition. With further expression recognition improvement, Ni Zhuang, Hao xiang Li, Kai peng Zhang and others proposed a cascaded multi-channel CNN (CMC-CNN) structure, Several CNNs are put together, facial key points can be located from coarse to fine, and facial expression changes can be recognized accurately. Ni Zhuang proposed a multi-task learning method about cascaded convolution neural network [11], which is called MCFA (cascaded convolutional neural network method), to predict multiple face attributes at the same time. This method trains three tasks corresponding to three cascade sub networks from coarse to fine at the same time, and makes use of the inherent dependence of the three tasks to effectively improve face attribute classification. The performance of this method is better than other facial expression recognition methods on challenging CelebA and LFWA data sets.

A cascade structure of three convolutional neural networks (CNN) in series is designed to help speed up the CNN cascade and obtain high-quality facial feature localization by Haoxiang Li [12], containing 6-CNN-cascade, in which the number of CNN network of binary classification of face and non-face is three, the number of that of bounding box calibration is also three, the calibration is adopted for a multi-class classification of discrete displacement patterns. Each calibration level is used to measure the detection window position so that it can be input to the next level. This method runs 14FPS on a single CPU core of VGA resolution image and 100FPS on GPU, and achieves the most advanced detection level. Kaipeng Zhang on the basis of this model, the number of adoption layers of each layer convolution is reduced from 5 × 5 to 3 × 3 [13], which is convenient for the light miniaturization of the convolution model and saves operation time.

According to cascade CNN model, a series cascade algorithm model for educational robot expression recognition is constructed, which enables the educational robot to recognize multiple students' facial expressions simultaneously, quickly and accurately in the process of movement, in the balance of the accuracy, rapidity and stability of the algorithm. Wireless communication lays the foundation for the rapid acquisition of expressions [14-21].

II. Principle of Expression Recognition Model Based on Cascade Convolution Neural Network

Based on the cascade neural network composed of face detection network (Detection net, D-net), face local region selection network (Landmark net, L-net) and face key point output network (Key point net, K-net), this section designs the cascade convolution neural network multi-task learning model, and compares the expression recognition experiment of educational robot in the expression recognition database and the real classroom environment. Verify the rapidity, accuracy and stability of the cascade model.

A. The Overall Structure of the Model

Because of the great visual changes caused by posture, expression and illumination, a face bounding box calibration link based on CNN is introduced into the cascade structure of convolution neural network (CNN) to help speed up the CNN cascade and obtain high quality facial feature localization. the cascade structure of convolution neural network (CNN) is shown in figure 1, which consists of three parts. The order is face detection network (Detection net, D-net), face logo region selection network (Landmark net, L-net), and face key points output network (Key point net, K-net).

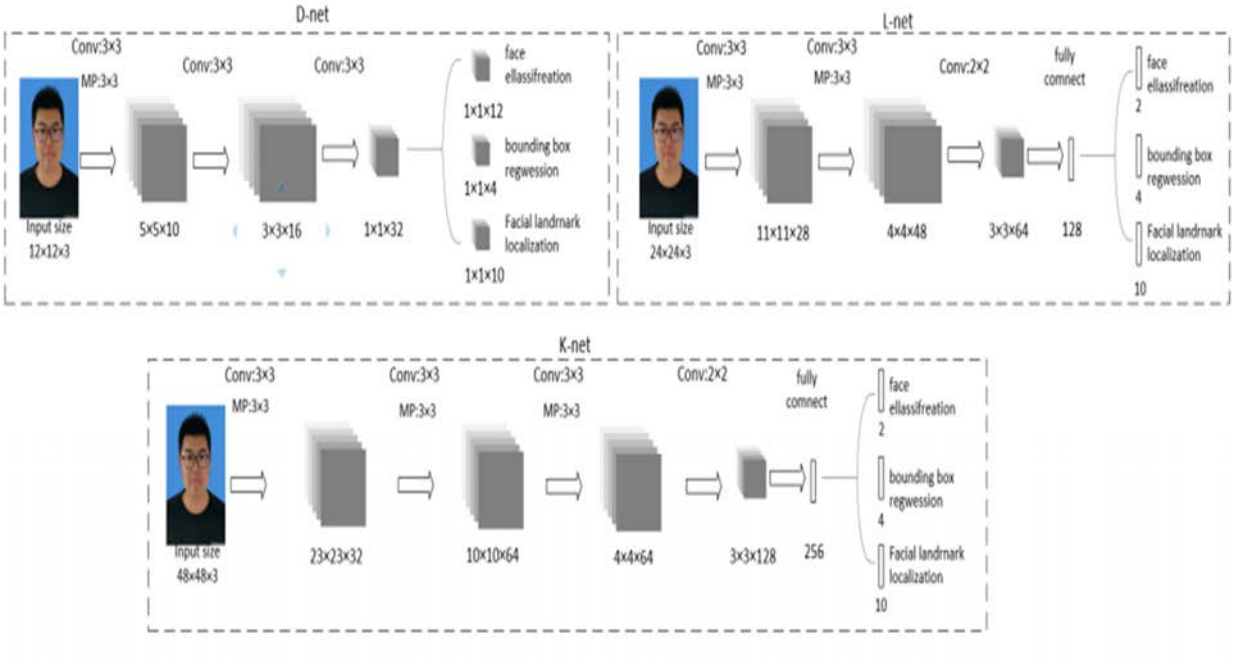

Fig. 1. Expression recognition model of cascaded convolution neural network

D-net: The various scales scanning is adopted in entire face image to quickly remove more than 90% irrelevant image information. The other images may be processed into 12 × 12 images in turn to modify their angle and position to obtain inner facial features.

L-net: The NMS (non-maximum suppression) calculation is used for further eliminate the remaining false information, and the relevant face image is cut and resized to 24 × 24 image to extract fine input facial expression figure features. The facial expression image in L-net is transferred to K-net, and a lot of false facial image information is further eliminated, which is used to test the algorithm model.

K-net: The fine input image features are extracted, and the input image size is adjusted to 48 × 48 to achieve more supervised face region recognition. the locations of five facial landmarks is output by the network.

In the cascade, there are six CNNs, face and non-face binary classification uses three CNNs, bounding box calibration also uses the same number CNNs. The adjustment of position of the detection window is achieved by calibration level. The light and small CNN model in the 12-layer and 24-layer network model does not have enough differentiation to solve the problem of identifying errors. For this reason, after the 12-layer and 24-layer network model, the NMS is used to iteratively select the detection window with the highest confidence and is applied to all detection windows on different scales after the 48-layer network model, which makes the most accurate detection window correct and avoids the redundant 48-layer network model evaluation.

The following factors can drop out CNN network performance: 1) some filters may influence distinguish ability because of the lack of convolution layer diversity; 2) is the binary face detection classification task require fewer filters than multi-class detection tasks. For this reason, the number of filters is reduced from 5 × 5 to 3 × 3, which is usually selected by cascade networks, in order to reduce the amount of calculation and get better performance with less running time. After convolution and full connection layer (except output layer), we use ReLU function as nonlinear activation function.

The specific steps of face recognition are as follows:

Step 1: use D-net to get the candidate's face window and bounding box regression vector. Then, the estimated bounding box regression vector calibrates candidate. After that, use NMS to merge highly overlapping candidates.

Step 2: use L-net to remove false image information, using bounding box regression to calibrate, and carry out NMS (non-maximum suppression), so as to extract the face logo region.

Step 3: use K-net to recognize the key points of human face, especially the position of the five key points.

Step 4: Online hard sample mining task is carried out to finish the training process in the face / non-face classification.

B. Model Training

We use three tasks to train our CNN face detector: face / non-face classification, bounding box regression and face signature location.

a) face classification: the learning goal is expressed as two kinds of classification problems. For each sample Xi, we use cross entropy loss as

$$L_i^{\det} = -\left(y_i^{\det}\log\left(p_i\right) + \left(1 - y_i^{\det}\right)\left(1 - \log\left(p_i\right)\right)\right) \tag{1}$$

pi: The probability that the sample feature quantity Xi is a facial feature. The symbol $y_i^{\det} \in \{0,1\}$ denotes a factual variable.

b) Bounding box regression: for each candidate *window*, we predict the offset between it and the nearest ground reality (that is, the left, top, height and width of the bounding box). The learning goal is expressed as a regression problem, and we use Euclidean loss for each sample Xi.

$$L_i^{box} = \left\| \bar{y}_i^{box} - y_i^{box} \right\| \tag{2}$$

$\bar{\mathbf{y}}_\mathbf{i}^{\mathbf{box}}$ represents the regression target value of the network model, $\mathbf{y}_\mathbf{i}^{\mathbf{box}}$ are real coordinate values. There are four coordinate variables, including the upper left corner, height, and width, $\mathbf{y}_\mathbf{i}^{\mathbf{box}} \in \mathbf{R}^4$.

c) Facial landmark location: similar to the bounding box regression task, facial landmark detection is described as a regression problem, and we minimize the Euclidean loss to

$$L_i^{landmark} = \left\| \bar{y}_i^{landmark} - y_i^{landmark} \right\| \tag{3}$$

$\bar{\mathbf{y}}_\mathbf{i}^{\mathbf{landmark}}$ represents the regression target value of the network model, $\mathbf{y}_\mathbf{i}^{\mathbf{landmark}}$ are real coordinate values. There are ten coordinate variables, including the upper left corner, height and width, $\mathbf{y}_\mathbf{i}^{\mathbf{landmark}} \in \mathbf{R}^{10}$.

d) Multi-source training: face-training, non-human face-training and partially aligned face-training. The background region can be done directly through the indication of the sample type. Then, the overall learning goal can be expressed as

$$\min \sum_{i=1}^{N} \sum \alpha_j \beta_i^j L_i^j \tag{4}$$

N :the number of training samples, α j: the task importance. We use (α det=1, α box=0.5, α landmark= 0.5) in P net and R net, and (α det=1, α box=0.5, α landmark = 1) in output network (O net) to locate human face landmarks more accurately. β j I ∈ {0jue 1} is a sample type indicator. In this case, it is natural to use random gradient descent to train these CNN.

III. EXPERIMENT

This section compares the expression recognition experiments of this algorithm with the commonly used and representative cascade network algorithms of STM-ExpLet and FN2EN through CK+ and Oulu-CASIA expression recognition database to verify the accuracy and rapidity of the cascade model. Results as shown in Table 1, the accuracy of our method can reach more than 90%. The recognition speed of the cascade structure in this paper is up to 100FPS, which is better than the other two methods. Compared with the other two commonly used cascade convolution neural network methods, the accuracy of facial expression recognition is significantly improved.

TABLE I. EXPERIMENTAL RESULTS OF CK+ AND OULU-CASIA DATABASES

Method	CK+	Oulu-Casia	Speed
STM-ExpLet	84.62%	85.33%	20 FPS
FN2EN	81.99%	82.02%	90 FPS
ours	90.12%	90.07%	100 FPS

IV. CONCLUSION

A three-level cascade algorithm model for educational robot expression recognition is designed. Through the CK+ and Oulu-CASIA expression recognition database, the expression recognition experiments of this algorithm and the commonly used STM-ExpLet and FN2EN cascade network algorithms are compared[21]. The proposed facial recognition method has an accuracy of more than 90% and a speed of 100 FPS. Compared with the other two commonly

used cascade convolution neural network methods, it significantly improves the accuracy and speed of facial expression recognition. The future work will be in the real classroom environment, the use of educational robots for dynamic, real-time student expression recognition, to further verify the feasibility of this algorithm.

ACKNOWLEDGMENTS

This work was supported by Science and Technology Development project of Jilin Province: Patent information analysis and strategic Research of educational Robot.

REFERENCES

[1] Russakovsky O, Deng J, Su H, et al:Imagenet large scale visual recognition challenge. International Journal of Computer Vision115(3),211-252 (2015).

[2] Krizhevsky A, Sutskever I and Hinton G E: Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems1097-1105(2012).

[3] Le Cun Y, Bottou L, Bengio Y, et al: Gradient based learning applied to document recognition. In Proc. IEEE2278-2324(1998).

[4] André Teixeira Lopes, Edilson de Aguiar, Alberto F,etc: Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order . Pattern Recognition (61)610–628(2017).

[5] Kuan Li, Yi Jin, Muhammad Waqar Akram, etc: Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy. The Visual Computer1-14(2019).

[6] Y.S. Gan a, Sze-Teng Liong , Wei-Chuen Yau,etc: OFF-ApexNet on micro-expression recognition system. Signal Processing: Image Communication (74),129-139(2019).

[7] Yan Shi, Zijun Zhang, Kaining Huang,etc: Human-computer interaction based on face feature localization.Vis. Commun. Image R(70),102740-45(2020).

[8] Gilderlane Ribeiro Alexandre, José Marques Soares, George André Pereira: Thé. Systematic review of 3D facial expression recognition methods. Pattern Recognition (100):107108-107124(2020).

[9] S Zhe Sun a, Raymond Chiong b, Zheng-ping Hua: self-adaptive feature learning based on a priori knowledge for facial expression recognition. Knowledge-Based Systems 204,106124-32(2020).

[10] G. E. Hinton, R. R. Salak hutdinov: Reducing the Dimensionality of Data with Neural Networks. REPORTS313,504-507(2016).

[11] Ni Zhuang, Yan Yan, Si Chen, etc: Multi-task Learning of Cascaded CNN for Facial Attribute Classification. 24th International Conference on Pattern Recognition 2069-2074(2018).

[12] Haoxiang Li, Zhe Linz, Xiaohui Shen, etc: A Convolutional Neural Network Cascade for Face Detection.CvF conference,5325-5334(2018).

[13] Kaipeng Zhang,Zhanpeng Zhang, Zhifeng Li, etc: Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE SIGNAL PROCESSING LETTERS23 (10), 1499-1503(2016).

[14] N. Zhang, N. Cheng, A. T. Gamage, K. Zhang, J. W. Mark and X. Shen, "Cloud assisted HetNets toward 5G wireless networks," in IEEE Communications Magazine, vol. 53, no. 6, pp. 59-65, June 2015

[15] B. Rong, Y. Qian, K. Lu, H. Chen and M. Guizani, "Call Admission Control Optimization in WiMAX Networks," IEEE Transactions on Vehicular Technology, vol. 57, no. 4, pp. 2509-2522, July 2008.

[16] S. Sun, M. Kadoch, L. Gong and B. Rong, "Integrating network function virtualization with SDR and SDN for 4G/5G networks," IEEE Network, vol. 29, no. 3, pp. 54-59, May-June 2015.

[17] S. Sun, L. Gong, B. Rong and K. Lu, "An intelligent SDN framework for 5G heterogeneous networks," in IEEE Communications Magazine, vol. 53, no. 11, pp. 142-147, November 2015.

[18] Y. Wu, B. Rong, K. Salehian and G. Gagnon, "Cloud Transmission: A New Spectrum-Reuse Friendly Digital Terrestrial Broadcasting Transmission System," IEEE Transactions on Broadcasting, vol. 58, no. 3, pp. 329-337, Sept. 2012.

[19] Ankarali Z E, Pek¨oz B, Arslan H. Flexible radio access beyond 5G: A future projection on waveform, numerology, and frame design principles. IEEE Access, 2017, 5: 18295–18309.

[20] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, “Toward 6G networks: Use cases and technologies,” IEEE Commun. Mag., vol. 58, no. 3, pp. 55–61, Mar. 2020.

[21] Z. Zhang et al., “6G wireless networks: Vision, requirements, architecture, and key technologies,” IEEE Veh. Technol. Mag., vol. 14, no. 3, pp. 28–41, Sep. 2019.