Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 107 (2017) 135 – 140

Procedia Computer Science

International Congress of Information and Communication Technology (ICICT 2017)

Facial Expression Recognition with Faster R-CNN

Jiaxing Li[a], Dexiang Zhang[a], Jingjing Zhang[a], Jun Zhang[a],

Teng Li[a], Yi Xia[a], Qing Yan[a], and Lina Xun[a]

[a] *The School of Electrical Engineering and Automation Anhui University, Hefei 230601, China*

** Corresponding author: zdxdzxy@126.com Tel.: 086-0551-63861094*

Abstract

In order to avoid the complex explicit feature extraction process and the problem of low-level data operation involved in traditional facial expression recognition, we proposed a method of Faster R-CNN (Faster Regions with Convolutional Neural Network Features) for facial expression recognition in this paper. Firstly, the facial expression image is normalized and the implicit features are extracted by using the trainable convolution kernel. Then, the maximum pooling is used to reduce the dimensions of the extracted implicit features. After that, RPNs (Region Proposal Networks) is used to generate high-quality region proposals, which are used by Faster R-CNN for detection. Finally, the Softmax classifier and regression layer is used to classify the facial expressions and predict boundary box of the test sample, respectively. The dataset is provided by Chinese Linguistic Data Consortium (CLDC), which is composed of multimodal emotional audio and video data. Experimental results show the performance and the generalization ability of the Faster R-CNN for facial expression recognition. The value of the mAP is around 0.82.

Keywords: Facial expression recognition; Faster R-CNN; deep learning; graphics processing unit

1. Introduction

Facial expression is a kind of effective way of human communication. Facial expression recognition is the key technology for realizing human-computer interaction to be the emotional computing system. Facial expression has a broad application prospect in many research fields, such as virtual reality, video conference, customer satisfaction survey and other fields.

Despite of great encouraging progress have been made in this research field, there are still many problems existing. On one hand, the traditional feature extraction methods are completely relied on human experience, which is still too complicated for real application. Therefore, the traditional methods are very difficult to extract useful features comprehensively and effectively. On the other hand, the traditional methods cannot dispose the big data and achieve better performance. So it is not easy to meet the real application requirement. In most situations, this kind of method cannot be employed effectively.

Peer-review under responsibility of the scientific committee of the 7th International Congress of Information and Communication Technology
doi:10.1016/j.procs.2017.03.069

To address the above issues, we propose an end to end recognition method based on Faster R-CNN [1]. The proposed method can be used to solve the existing problems. First, the Region Proposal Networks (RPNs) [1] are used to predict efficient and accurate region proposal [4]. The pipeline of the proposed method just use one convolutional neural network (CNN) for all purpose. So, the region proposal is nearly cost free by sharing convolutional features of the down- flow detection network. Second, the RPN also improved the regional proposal [4] quality and the accuracy of the overall target detection.

2. Faster R-CNN algorithm

Faster R-CNN [1] can be simply regarded as the system consisting of regional proposal network and Fast Regions with Convolutional Neural Network Features (Fast R-CNN). The regional proposal network is used to instead Selective Search algorithm [4] of Fast R-CNN. The proposed method focuses on solving three problems: 1) how to design a regional proposal network; 2) how to make proposal network region; 3)how to share feature extraction network.

2.1. Candidate Region (anchors).

Characteristics can be seen as a 256 channel image with a scale of 51×39, for each position of the image. The method considers the nine possible candidate windows, which are three areas of {128^2, 256^2, 512^2} multiplied by three ratios of {1:1, 1:2, 2:1}. These candidate windows are said to be as "anchors". Fig. 1 shows the anchor 51×39 center, as well as 9 anchor examples.

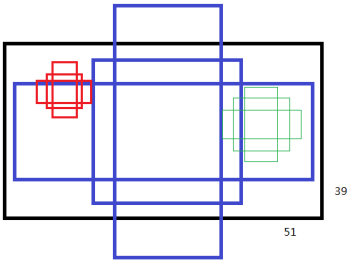

Fig. 1. The 51×39 anchor centers as well as 9 anchor examples

Each position of the output of classification layer [6] (cls_score) shows the probability that the 9 anchors belong to the foreground and background. And each position of the output of regression layer [6] (bbox_pred) shows that the corresponding window of the 9 anchors should be translated to scale parameters. For each location, the classification layer outputs the probability of the foreground and the background from the 256 dimension, while the regression layer outputs 4 translation scaling parameters. From a local respective, the two layers are the whole connection network, while from a global respective, since the network share same parameters in all positions (51×39), the network used in the present study is actually with a size of 1×1.

2.2 Sharing feature.

Region proposal network (RPN) and Fast R-CNN require an original feature extraction network. The ImageNet Classification Library is used to train the initial parameter ω_0 of the network. Then the network is fine-tuned by the specified dataset. The proposed method provides three methods: 1) train RPN to extract the anchors on the training set from ω_0 . 2) train Fast R-CNN by using the anchors from ω_0 , and the parameter is denoted as ω_1 ; 3) train RPN from ω_1 . Fig. 2 shows the detail steps of the shared feature.

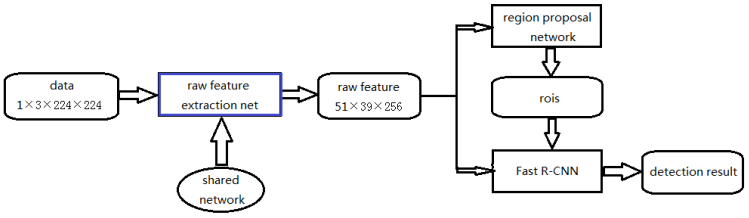

Fig. 2. The steps of the shared feature

With these definitions, we minimize an objective function following the multi-task loss in Fast R-CNN [2]. For an anchor box i, its loss function is defined as:

$$\mathrm{L}(p_i,t_i) = L_{cls}(p_i,p_i^*) + \lambda p_i^* L_{reg}(t_i,t_i^*) \tag{1}$$

Here p_i is the anchor *i* being an object [1] of the probability of prediction. If the anchor label is positive, p_i^* is 1, if it is negative, it is 0. t_i = { t_x , t_y , t_ω , t_h }$_i$ represents the predicted bounding box of 4 parameterized coordinates [6], and t_i^* = { t_x^* , t_y^* , t_ω^* , t_h^* }$_i$ indicates that the ground-truth box [6] is associated with a positive anchor. The classification loss L_{reg} is the softmax loss of two classes [5]. The details of our facial expression recognition method is illustrated in Fig. 3.

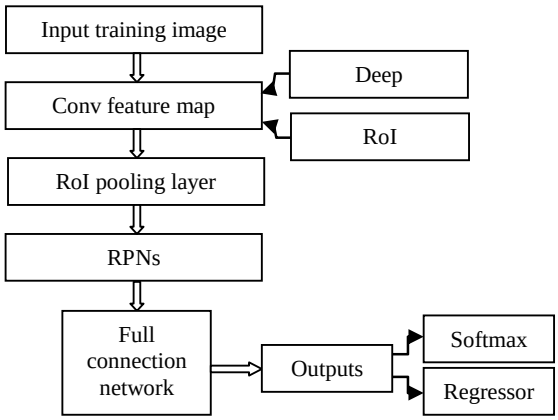

Fig. 3: Flow chart of our facial expression recognition method

3. Experiments

3.1. Dataset and Features

The dataset is provided by Chinese Linguistic Data Consortium(CLDC) [9], which is composed of multimodal emotional audio and video data. Total eight expression are collected from TV series or movies. Some examples are shown in fig. 4.

Fig. 4. The picture of the data from CLDC (From left to right: worried, angry, disgust, surprise, anxious, happy, sad and neural)

In the experiment, the data set consists of 66486 pictures with eight categories. Among them are 6174 worried pictures, 10862 angry pictures, 1687 disgust pictures, 2574 surprise pictures, 12019 anxious pictures, 9867 happy pictures , 18326 sad pictures and 4977 neural pictures. The ratio of training, validation and testing data is 8:1:1. Since Faster RCNN was used to detect face directly, the background was considered as one class. So, total 9 categories were used in this research.

3.2. Data label making

Since Faster R-CNN is used, the region of interest (ROI) of each image must be marked first. The software can be employed to achieve the coordinates of ROI. Then the coordinates are transfer into xml format. The ROI of some images is shown in Fig. 5.

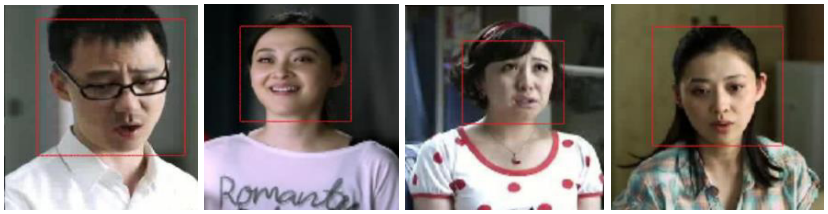

Fig. 5. Data label making

3.3. Model design

We use the three pascal_voc.model provided by the py-faster-rcnn, respectively, VGG_CNN_M_1024 [8], ZF [7] and VGG16 [3], which are called the faster_rcnn_alt_opt network to fine-turn imagenet model. The depth of the three networks is increasing. So we can compare the experimental results to further analyze the accuracy of the experiment through the three networks.

3.4. Parameter setting.

The solver parameters of the network are set as follows: *Stage1_fast_rcnn_train.pt*: base_lr *(faster):* 0.001, lr_policy *(faster):* "step", stepsize *(faster):* 30000, display *(faster):* 20, average_loss *(faster):* 100, momentum

(faster): 0.9, weight_decay *(faster):* 0.0005. *Stage1_rpn_train.pt*: base_lr *(faster):* 0.001, lr_policy *(faster):* "step", stepsize *(faster):* 60000, display *(faster):* 20, average_loss *(faster):* 100, momentum *(faster):* 0.9, weight_decay *(faster):* 0.0005. *Stage2_fast_rcnn_train.pt*: base_lr *(faster):* 0.001, lr_policy *(faster):* "step", stepsize *(faster):* 30000, display *(faster):* 20, average_loss *(faster):* 100, momentum *(faster):* 0.9, weight_decay *(faster):* 0.0005. *Stage2_rpn_train.pt*: base_lr *(faster):* 0.001, lr_policy *(faster):* "step", stepsize *(faster):* 60000, display *(faster):* 20, average_loss *(faster):* 100, momentum *(faster):* 0.9, weight_decay *(faster):* 0.0005. The Partial parameters of the network are set as follows: *data_param_str_num_classes*: 9, *cls_score_num_output*: 9, *bbox_pred_num_output*: 36.

3.5. Training and testing

We treat over 0.8 IoU overlap for all region proposals with a ground-truth box as positives and the rest as negatives for box's class. We bias the sampling towards positive windows because they are extremely rare compared to background.

4. Experimental result

In order to improve the reliability of the experimental results, three different depth of the network is used to train and test data, that is, VGG_CNN_M_1024 [8], ZF [7] and VGG16 [3]. The training results of the three kinds of networks are shown in Table 1.

As we can see from table 1, with the increase of the depth of the network, the mAP value of the training detection is improved. It can be clear to find that some types with large amount data, such as the type of anxious, the recognition rate is not very high. On the contrary, some types with small training data, such as the type of disgust, but the recognition rate is relatively high. That's because it is not balanced for the training data of each type, it leads to the mutual interference of some kinds of data in training. But it, this problem, has not much impact on the results of the training data, we can still obtain the final recognition rate and recognition results. In the future, we will make the appropriate judgment and correction for the imbalance of the data, so that the problem will not affect the results of the experiment, so as to improve the readability and accuracy of the data.

From the table 1, the recognition rate of the neural type is very low make the mAP value is not very high for all test data. The neural expression is very hard to identify because there are too many human factors to determine neural type. When label images, some other class expression are very easily determined to be a neural type. In the future, we can improve the recognition rate by increasing the training data of the neural type.

Table1. The training results of the three kinds of networks

pascal_voc.model	worried	angry	disgust	surprise	anxious	happy	sad	neural	mAP
VGG_CNN_M_1024	0.7922	0.8998	0.9798	0.8879	0.8806	0.8513	0.9078	0.3604	0.8200
ZF	0.8015	0.8998	0.9739	0.8979	0.8799	0.8666	0.9083	0.3344	0.8203
VGG16	0.8315	0.9016	0.9782	0.8973	0.8857	0.8812	0.9091	0.3646	0.8312

Some example detections using Faster R-CNN on CLDC are shown in fig. 6. We can see from the fig. 5, it is very good for recognition effect of 6 kinds of types (worried, angry, disgust, anxious, happy and sad), and some recognition rate even reached 100%, such as the types of disgust and angry. It can be seen that the experimental results have reached the expected goal.

However, there are not very good identification for the types of surprise and neural. First, the type of surprise have many similarities with the types of angry and happy, which does not have the very good recognition feature. So the type of surprise are identified as angry or happy in testing data. The type of neural is also due to the human element when it is used as a label. In later experiments, we can increase the weight of the eyes and mouth from the image to recognize the feature of images accurately, because it is mainly through the behavior characteristics of the

eyes and mouth for the facial expression recognition . At the same time, it can increase the training data of each type and enhance the recognition of the data to increase the recognition rate of the type.

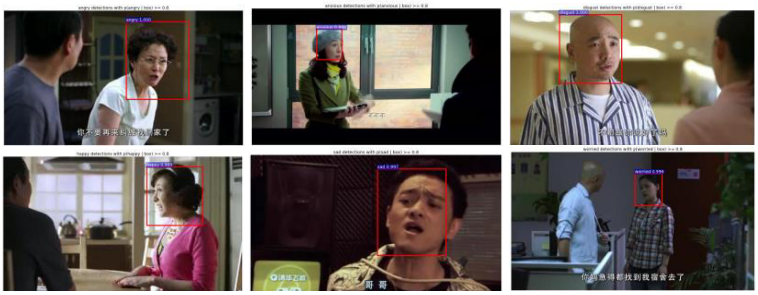

Fig. 6: Example detections using Faster R-CNN on CLDC

5. Conclusion

In this paper, Faster R-CNN was used to identify facial expression. There are the following advantages used in facial expression recognition: the original image was used as the whole network input. The process of feature extraction in the traditional facial expression recognition is avoided. The features are extracted by network from training dataset automatically. The Region Proposal Networks (RPNs) was used to generate a efficient and a accurate region proposal. In each image, the proposed method locate the face region and recognize the expression directly. The experimental results show that the proposed method achieved better recognition performance.

Acknowledgements

This work was financially supported by the Chinese National Science Foundation Grant (No. 61272025, No. 61402004, No. 61602002, No. 61300056, No. 61572029 and No. 61271098), and is supported by Anhui Provincial Natural Science Foundation (No. 1608085MF136 and No. 1408085QF118). This paper is partially supported by Science and Technology Project of Anhui Province (No. 1501b042207 and No. 1604d0802019).

References

1. Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J]. Computer Science, 2015.
2. R. Girshick. Fast R-CNN. arXiv: 1504.08083, 2015.
3. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.
4. J. R. Uijlings, K. E. vandeSande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. IJCV, 2013
5. R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.
6. C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In ECCV, 2014.
7. M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional neural networks. In ECCV, 2014.
8. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.
9. Chinese Linguistic Data Consortium, Chinese Conference on Pattern Recognition (CCPR) , 2016