Research on Face Expression Detection Based on Improved Faster R-CNN

Weiran Hua
College of Computer and Information Engineering
Hubei Normal University
Huangshi 435002, Hubei, China

Qiang Tong
College of Computer and Information Engineering
Hubei Normal University
Huangshi 435002, Hubei, China

***Abstract*—Because facial expression is easy to be confused, and is easily affected by environment, Angle and other factors, this paper proposes an improved Faster R-CNN based facial expression detection method. In this method, histogram equalization and adaptive histogram equalization are preprocessed for SFEW 2.0 of the facial expression data set, and the facial expression data is enhanced and expanded. Then the repetitive experimental optimization of the hyperparameters is carried out to improve the training and learning effect of the model and improve the detection accuracy. In the end, based on the regularization model structure optimization, Soft-max cross entropy classification loss function and L1 Smooth regression loss function with parameter constraint term were proposed. The regularization method was used to optimize parameter weight, improve detection accuracy, and an improved Faster R-CNN model adapted to face expression characteristics was obtained.**

Keywords—Faster R-CNN, Facial expression, Target detection, Deep learning, Convolutional neural network

I. INTRODUCTION

In People's Daily life, communication is very important, in addition to language, movement of communication, sometimes often a facial expression can let the other side know what you are thinking. Psychologist Mehrabian proposed [1] that emotional expression through facial expressions accounts for 55% of information. Facial expressions can reflect not only a person's current mood, but also their personality, cognitive and even physiological states. In 1971, the famous psychologists Ekman and Friesen pointed out [2] that the main emotions of human beings can be roughly divided into six categories: anger, happiness, sadness, surprise, disgust and fear. Since then, research on facial expressions has been continuously carried out. In recent years, with the developing of artificial intelligence, human-computer interaction intelligent has become increasingly important, the study of facial expressions can let the machine of artificial intelligence for more human emotional information, and analysis to understand human mental emotional status and demand, improve the level of human interaction, let artificial intelligence better service for the human, improve the level of people's life. In this paper, based on Faster R-CNN deep learning network algorithm, Faster R-CNN model was improved and optimized to improve the recognition and detection performance of facial expressions.

II. FACE EXPRESSION DATA SET

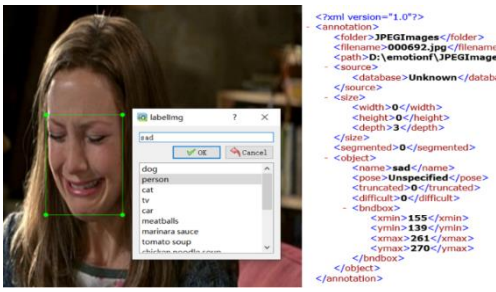
This experimental data set of Facial Expression detection adopts the image data set SFEW 2.0 (Static Facial Expression Recognition in The Wild 2.0) used in The EmotiW 2015 (The Emotion Recognition in The Wild 2015) competition [3]. This data set is a total of 1765 screenshots containing expressions taken from various films, each of which is 720×576 in size. There are 6 basic expressions and 1 neutral expression, which are respectively angry, disgust, fear, happy, sad, surprise and neutral. Sample samples of SFEW 2.0 data set are shown in Figure 1.

Fig. 1. Sample of SFEW 2.0 dataset

A. Annotate and Partition Data Sets

To train the data set on Faster R-CNN, it is necessary to first convert the data set into the corresponding PASCAL VOC [4] data format, which consists of three parts: JPEGImages, Annotations and ImageSets, which are respectively used to store data image samples, data image annotation information and data image index. Perform the following operations on the SFEW 2.0 data set:

- according to PASCAL VOC format, raw images of various facial expressions in SFEW 2.0 data set were stored in JPEGImages directory.
- label the facial expression area of the image to obtain the category and region coordinate information of the expression. In the face identity detection of the actual system, because the face only occupies a small part in the image, we need to label the image, so that the algorithm can get the location and category of facial expressions in the imported image. The open source annotation tool labelImg is used to annotate face expressions to get a more accurate expression unit target window. The annotation labels of face expressions are saved in xml file format, including file name, file source, image size, object category, bound-box coordinate and other information, and then the annotated xml files are stored in the Annotations directory. The annotation and specific format of the xml file of the open source annotation tool labelImg are shown in Figure 2, where (a) is an example of the annotation expression through labelImg and (b) is an example of the specific format information of the xml file.



(a) labelImg mark (b) xml file format

Fig. 2. Labeling of Facial Expression

- will each kind of expression SFEW 2.0 data set according to the proportion of 8:1:1 randomly divided into training set, validation set and test set three parts, the training set of facial expression of 8, validation and test sets the facial expression of the remaining 1 into the training set, validation set and test set corresponding facial expression image indexes are stored in the train.txt, val.txt and test.txt file.

B. Data Preprocessing and Data Expansion

Facial expression imaging is easily affected by environmental factors such as light and the occlusion of objects such as hair and glasses will also affect the recognition and detection of facial expression. At the same time, the detailed information changes of facial expressions are quite complex, such as the imaging Angle, the blink of an eye or the opening and closing of the mouth, etc. will produce a variety of different performances under the same expression, these subtle differences will bring difficulties to the detection work. Through image preprocessing, the influence of environmental factors such as light can be effectively reduced, facial expression texture details can be enhanced, and the effect of deep learning convolutional neural network to extract facial expression features can be improved.

- Image Graying. For the face expression detection work studied in this paper, whether the image is color has little impact on the face expression. The gray-scale operation of the image can save computing space and reduce the dimension of the input image. The graying image is shown in Figure 3, and the image turns gray. It can be seen that the graying image does not lose facial expression details, and the dimension of the image is also reduced to save storage space.

Fig. 3. Image Graying

- Histogram Equalization. Values of pixels in an image histogram equalization is statistics, displayed by the histogram, and distribution of histogram equalization and balance, and increase the number of less number and reduce the number of pixels, the shallow part, a fuzzy image into a deeper, more clear image, enhancing image contrast, make the overall image grayscale more balanced, local contrast enhancement image, to strengthen the details of the facial expression characteristic information, The image after histogram equalization is shown in Figure 4.

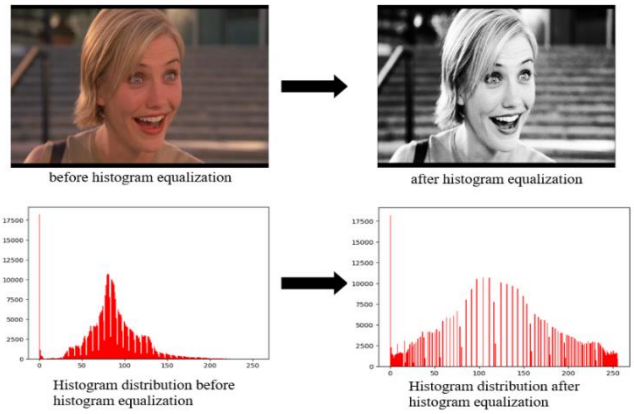


Fig. 4. Histogram Equalization

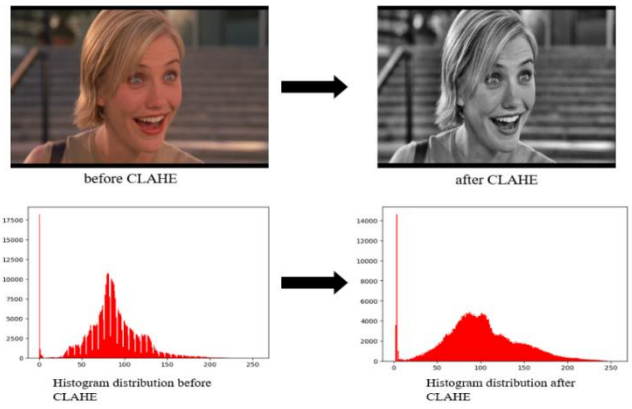


Fig. 5. Contrast Limited Adaptive Histogram Equalization

- Contrast Limited Adaptive Histogram Equalization. The restricted contrast adaptive histogram equalization (CLAHE) clipped the histogram and distributed the clipped part evenly on the whole interval, which improved the local contrast of the image and overcame

the problem of excessive image distortion and noise. The image processed by CLAHE is shown in Figure 5. It can be seen CLAHE enhances the local contrast of the image, strengthens the detailed information of facial expression, and avoids image distortion.

- Data expansion. If a neural network has too many weight parameters and not enough sample size, it will produce overfitting, that is, it is too close to the characteristics of the training data, making the generalization ability of the model insufficient. Due to the limited ability of the author, the SFEW 2.0 data set selected in this experiment only contains 1765 facial expression images, which is a small data sample set. It is inevitable that there will be overfitting problems in the training of multi-layer network structure such as deep learning. Therefore, it is necessary to expand the data sample.
- Common methods of data expansion include rollover, rotation, translation and scaling. The original data set and the pre-processed data set are flipped. The image after Flip transformation is shown in Figure 6.

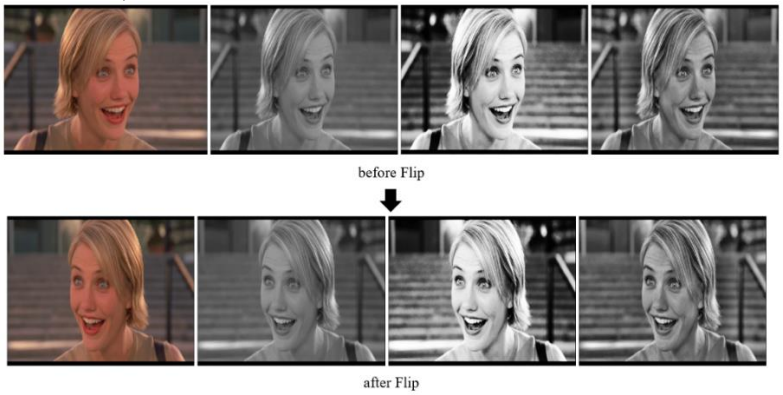


Fig. 6. Flip

III. Hyper Parameter Optimization

Hyper Parameter are parameters entered manually before training, usually determined by experience, and used to help estimate model parameters during training. For deep learning network models with different network structures, network layers and training sample sizes, adjusting and optimizing different hyper Parameter can achieve better model training effect.

In order to evaluate the effects of learning rate, batch size and iteration times on facial expression detection accuracy, and get the algorithm model has a better training effect of hyper parameter values, based on the repetitive experiments, the test accuracy was verified when the learning rate was 0.001 and 0.002, the batch size was 64, 128 and 256, and the number of iterations was 8000 and 10000. The experimental verification results of the learning rate, batch size and iteration times on the detection accuracy of facial expression are shown in Table I.

TABLE I. Influence of Learning rate, Batch size and Iteration on Detection Accuracy

Learning rate	Batch size	Iterations	mAP (%)
0.001	64	8000	0.6466
		10000	0.7994
0.001	128	8000	0.7695
		10000	0.8056
0.001	256	8000	0.6536
		10000	0.8018
0.001	512	8000	0.6932
		10000	0.7622
0.002	64	8000	0.5060
		10000	0.5360
0.002	128	8000	0.6000
		10000	0.6797
0.002	256	8000	0.7116
		10000	0.7649
0.002	512	8000	0.7146
		10000	0.7642

It can be seen when the learning rate is 0.001, the batch size is 128, and the number of iterations is 10,000, the mAP value is the highest, reaching 80.56%. Therefore, in this chapter, the hyper Parameter of facial expression detection based on improved Faster R-CNN were set as learning rate equal to 0.001, batch size equal to 128, and iteration times equal to 10,000.

IV. Loss Function Optimization

The loss function, in simple terms, is the degree of error between the predicted data and the actual data. If the predicted data is far from the actual data, the value of the loss function will be large. on the contrary, if the predicted data is very close to the actual data, the value of the loss function will be small. Without considering overfitting, the smaller the loss function is, the higher the fitting effect of the model will be. There are various types of loss functions, and different loss functions have their own advantages and disadvantages and applicable scope. For different models and different target tasks, corresponding loss functions should be selected to achieve better optimization effect.

Faster R-CNN is composed of RPN and Fast R-CNN, it adopts the multi-task loss function to simultaneously calculate the loss function of target classification and bounding-box regression, and then adds them together. The formula is shown in (1).

$$\mathrm{L}(\mathrm{p},\mathrm{u},\mathrm{t}^{\mathrm{u}},\mathrm{v}) = \mathrm{L}_{\mathrm{cls}}(\mathrm{p},\mathrm{u}) + \lambda[\mathrm{u} \geq 1]\mathrm{l}_{\mathrm{loc}}(\mathrm{t}^{\mathrm{u}},\mathrm{v}) \qquad (1)$$

L_{cls} is the classification loss function. l_{loc} is a regression loss function. λ is the equilibrium factor. $[u \geq 1]$ is the indicator function, and when $u \geq 1$, it is the object class, and the function value is 1, and there is regression loss. When $u = 0$, it is the background class, and the function value is 0, and there is no regression loss.

A. Classification Loss Function Optimization

Sigmoid cross entropy loss function is often used to predict the probability distribution of the target, and its formula is shown in (2).

$$L = \frac{-1}{n}\sum_{n=1}^{N}[p_n \log \hat{p}_n + (1-p_n)\log(1-\hat{p}_n)] \quad (2)$$

This is the loss function used in the classification of Faster R-CNN, and its form is shown in (3).

$$L_{cls}(p_i, p_i^*) = -\log[p_i^* p_i + (1-p_i^*)(1-p_i)] \quad (3)$$

i is the index of the mapping anchor. p_i is the prediction probability of the first i anchor. When the IoU of anchor and callout box is greater than 0.7, the probability of callout box p_i^* is denoted as 1. when the IoU of anchor and callout box is less than 0.3, p_i^* is denoted as 0.

On this basis, the Soft-max cross entropy loss function is introduced in this paper. The cross entropy can be combined with Soft-max to form the Soft-max cross entropy loss function. The formula is as follows.

$$L = \frac{-1}{N}\sum_{n=1}^{N}\log\left(\hat{p}_{n,l_n}\right) \quad (4)$$

$$P_{nk} = \frac{e^{x_{nk}}}{\sum_{k'} e^{x_{nk'}}} \quad (5)$$

x is the prediction confidence score of each category. P_{nk} is the mapping probability distribution of score x.

Therefore, Soft-max cross entropy loss function L_{cls} classifies loss formula as shown in (6).

$$L_{cls}(p, u) = -\log p_u \quad (6)$$

the probability distribution of each RoI is $p = (p_0, p_1, \cdots, p_k)$, which is one background and K target respectively, and p is calculated by Soft-max function. u is the category of Ground truth.

B. Regression Loss Function Optimization

In the regression process of Faster R-CNN, the loss function based on Smooth L1 was adopted. The regression loss formula of l_{loc} for Bounding box is shown as follows.

$$l_{loc}(t^u, v) = \sum_{i\in\{x,y,w,h\}} smooth_{L1}(t_i^u - v_i) \quad (7)$$

$$smooth_{L1}(x) = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & otherwise \end{cases} \quad (8)$$

where $x = t_i^u - v_i$. Deviation target $v = (v_x, v_y, v_w, v_h)$, represents the Bounding box parameterized coordinates of Ground truth. The predicted deviation $t^u = (t_x^u, t_y^u, t_w^u, t_h^u)$, represents the parameterized coordinates of the predicted Bounding box.

On this basis, this paper proposes a Smooth L1 loss function with parameter constraint term, whose formula is as follows.

$$smooth_{L1}(x) = \begin{cases} x^2 * (0.5 * \sigma^2), & |x| < 1/\sigma^2 \\ |x| - (0.5/\sigma^2), & otherwise \end{cases} \quad (9)$$

where, σ is the parameter constraint term.

The parameter constraint term is usually obtained by experience or experiments. In this paper, the influence of on improving detection accuracy of Faster R-CNN is obtained based on repeated experiments, and the results are shown in Table II.

TABLE II. INFLUENCE OF σ ON DETECTION ACCURACY

iterations	10000	10000	10000	10000	10000
$\boldsymbol{\sigma}$	0.5	1	2	3	4
mAP(%)	80.01	80.56	81.17	81.33	81.05

It can be seen from Table II that when the σ value of is 3, the improved Faster R-CNN average detection accuracy average mAP is the highest, reaching 81.33%, so the value of parameter constraint item is 3.

V. EXPERIMENTAL RESULTS AND ANALYSIS

The system environment and hardware environment of the facial expression detection experiment based on improved Faster R-CNN are shown in Table III.

TABLE III. HARDWARE AND SOFTWARE CONFIGURATION OF SYSTEM

CPU	Inter(R) Core(TM) i7-7700HQ 2.80GHz
Memory	8.00 GB
GPU	NVIDIA GeForce GTX 1050 Ti
Operating System	Win10
Python	3.6
Deep learning framework	Tensorflow

The SFEW 2.0 data sets after image preprocessing and expansion passed the learning rate equal to 0.001, the batch size equal to 128, and the number of iterations equal to 10000, after introducing the improved Faster R-CNN training of the classification loss function of Soft-max cross entropy and the Smooth L1 regression loss function with parameter constraint term, the detection results on the test set are shown in Table IV.

TABLE IV. THE EXPERIMENT RESULTS OF EACH EXPRESSION

	angry	**disgust**	**fear**	**happy**	**neutral**	**sad**	**surprise**	**mAP**
AP	90.12	61.84	84.87	92.43	88.84	81.28	69.93	81.33

It can be seen from Table IV that the average precision mAP of facial expression detection based on the improved Faster R-CNN reaches 81.33%, indicating that the improved Faster R-CNN has better overall performance in the detection of facial expression and can better realize the detection and classification of facial expression. The actual detection effect of this facial expression detection is shown in Figure 7 (a) ~ (g).

(a) angry (b) disgust (c) fear (d) happy

(e) neutral (f) sad (g) surprise

Fig. 7. Experimental Effect

As shown in Figure 7, the improved Faster R-CNN facial expression detection algorithm successfully detects the facial expression in the image, selects the region of the facial expression, and classifies it, and marks the category result and confidence score of the category on the box. It can be seen that (a) angry, (c) fear, (d) happy and other facial expressions with obvious features can be well recognized and detected. However, as the disgust expression itself is relatively diverse and complex, it is easy to be confused with other expressions. In (b), the algorithm detects the disgust error as happy. In general, the facial expression detection mAP based on the improved Faster R-CNN reached 81.33%, which could better satisfy the detection and classification of facial expressions and detect and recognize most facial expressions.

VI. Conclusion

In this paper, Faster R-CNN was improved and optimized, histogram equalization and adaptive histogram equalization were preprocessed for the facial expression data set, and the facial expression feature details were improved while the dimension was reduced. The image data was enhanced and expanded by Flip transformation. Repeated experiments were conducted on the common hyperparameter learning rate, batch size and iterations, and the hyperparameter value with better detection effect was obtained, and the training effect of the model was optimized. Finally, based on the regularization model structure optimization, Soft-max cross entropy classification loss function and L1 Smooth regression loss function with parameter constraint term were introduced, and the value of constraint term parameter was determined through experiments, which improved the detection accuracy. The experimental results showed that the improved Faster R-CNN could better adapt to the features of facial expressions, and the mAP reached 81.33%. However, there was still room for improvement for the easily confused facial expressions such as disgust.

References

[1] A. Mehrabian, “Communication without words”, *Psychological today*, vol. 2, no. 4, pp. 53-55, 1968.

[2] H. J. Visser, R. J. M. Vullers, “RF Energy Harvesting and Transport for Wireless Sensor Network Applications: Principles and Requirements”, *Proceedings of the IEEE*, vol. 101, no. 6, pp. 1410-1423, 2013.

[3] A. Dhall, O. V. Ramana Murthy, R. Goecke, J. Joshi and T. Gedeon, “Video and Image based Emotion Recognition Challenges in the Wild: EmotiW 2015, ACM International Conference on Multimodal Interaction (ICMI), 2015.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, et al., “The PASCAL Visual Object Classes(VOC) challenge. International Journal of Computer Vision, vol. 88, no. 2, pp. 303-338, 2010.

[5] A. Lumini, L. Nanni, A. Codogno and F. Berno (2019). Learning morphological operators for skin detection. Journal of Artificial Intelligence and Systems, 1, 60–76.

[6] M. Saravanan and A. Priya (2019). An Algorithm for Security Enhancement in Image Transmission Using Steganography. Journal of the Institute of Electronics and Computer, 1, 1-8.