Design and application of learner emotion recognition for classroom

Chao Su¹, Guozhong Wang¹

¹School of Electronic and Electrical Engineering, Shanghai University of Engineering and Science, Shanghai, 201600, China

Chao Su e-mail: 1311810748@qq.com
Guozhong Wang e-mail: wanggz@sues.edu.cn

Abstract. The emotion recognition of learners in the classroom plays an important role in improving classroom efficiency. At present, the recognition methods based on traditional image processing generally have problems such as low recognition accuracy, difficulty in feature extraction. In order to effectively solve the above problems, a deep learning-based method is proposed to perform emotion recognition for learners. This method first replaces the DarkNet-53 of YOLO v3 by introducing the convolution structure of MobileNet, which makes the model lighter and reduces the parameters. Then improve the loss function of the model by using GIoU loss. Finally, K-means clustering algorithm is used to obtain an anchor suitable for the emotion recognition scene on the self-made dataset. Thus, a high-accuracy and lightweight learner emotion recognition model ER_YOLO is obtained. Experimental results show that the improved model mAP increased by 4%, F1 score increased by 3.2%, detection time reduced by 1/3, and parameters reduced by 1/10.

1. Introduction

Psychological research shows that emotion is a very important non-intellectual factor in teaching, which can promote or prevent memory, help reasoning operations and problem solving, and play a key role in perception, memory, thinking, executive control and decision-making [1]. Research further shows that positive emotions can help stimulate learners’ learning motivation, cultivate learners’ interest in learning, and promote learners’ cognitive process, while negative emotions can affect their patience, attention and hinder their cognition process [2]. Therefore, the accurate recognition of learners’ emotions is of great significance for improving classroom efficiency.

At present, the methods of emotion recognition mainly focus on research based on physiological signal analysis, psychometric data analysis, cognitive evaluation methods, and behavior analysis methods [3]. Among them, methods based on behavior analysis have been the research hotspots in recent years, mainly including: methods based on facial expression recognition, methods based on speech signal recognition, methods based on text symbol recognition, and methods based on body gesture recognition. American scholar Melabian [4] (Albert Mehrabian) pointed out that emotional expression = 7% of words + 38% of voice + 55% of facial expressions. Psychologist Ekman [5] research also shows that the accuracy of mapping from facial expressions to a single emotional state is 88%. Therefore, most of the current research on emotion recognition is based on facial expressions.

In recent years, domestic and foreign scholars have put forward many researches on emotion recognition based on facial expressions. For example, Tingting Wang et al. [6] proposed a learning
fatigue recognition and intervention method based on facial expression recognition in response to the physical or psychological "learning fatigue" state of online learners, and defined concentration, fatigue and normal 3 kinds of expressions related to learning. Bo Sun et al. [7] believed that there is an important relationship between emotion and cognition, and proposed an emotion analysis framework based on facial expressions in a smart learning environment, they believe that in the process of facial expression recognition, the facial expression characteristics related to the individual should be separated from the expression characteristics irrelevant to the individual. Zehui Zhan [8] based on the existing teaching agents that lack emotional interaction and weak cognitive inference functions, and proposed a remote learner emotion and cognitive recognition model that combines facial expression recognition and eye tracking technology. Renjia Wei et al. [9] proposed an online learning system model with an emotion recognition module based on the current "emotional lack" phenomenon in online learning, combined with "constructivism" and "postmodernism" education theories. Jingjie Yan et al. [10] believe that multi-modal emotion recognition is an important content in the current emotional computing field, and proposed a dual-modal emotion recognition method based on bilateral sparse partial least squares of expression and posture. Zhenguo Xu et al. [11] adopted a method different from traditional learner emotion recognition, proposed a learner emotion recognition method based on convolutional neural network, and designed a 7-layer convolutional neural network, which effectively solved the feature extraction Difficult question.

In summary, it is not difficult to find that most of the previous studies on learner emotion recognition are based on traditional recognition methods, which mainly include image pre-processing, face detection, feature extraction, expression classification and other processes [12], as shown in Figure 1 below. However, in traditional emotion recognition methods, the feature extraction relies on manual labeling and selection, which not only requires a lot of work, but the feature extraction is greatly affected by subjective factors to a certain extent, thus increasing the difficulty of recognition. Therefore, in order to solve the problems of difficulty in feature proposal, low accuracy, and application to actual scenes, and obtain a learner emotion recognition model suitable for strict real-time requirements in the classroom, this paper proposes a learner emotion recognition method based on improved YOLOv3, using The MobileNet [13] convolution structure with deep separable convolution as the core replaces the DarkNet53 convolution structure of YOLOv3. At the same time, GIoU [14] loss is used to improve the loss function and the data constructed independently by the K-means [15] algorithm Re-clustering on the dataset to obtain a suitable anchor value, thereby obtaining a lightweight emotion recognition model ER_YOLO.

![Figure 1. Flow chart of traditional emotion recognition](image)

2. Model design

2.1. Overview

Aiming at the problems of low recognition rate and difficulty in feature extraction in traditional methods for learner emotion recognition, a learner emotion recognition method based on improved YOLOv3 is proposed. The overall flow chart is shown in Figure 2 below, which is divided into two parts, namely the emotion recognition part and the decision fusion part. First, perform emotion recognition by improving the model ER_YOLO obtained by YOLOv3. Then analyze the classroom state for the seven emotions, which are divided into five classroom states: listening, understanding, doubting, disdain, and uncertainty. Finally, the final classroom evaluation is conducted through 5 kinds of classroom state analysis, that is, learners' classroom participation, classroom deviation, and uncertainty.
2.2. YOLOv3

YOLOv3 is a single-stage target detection algorithm, improved from YOLOv1 and YOLOv2. Compared with YOLOv2, YOLOv3 uses up-sampling and fusion methods to fuse 13x13, 26x26, and 52x52 features at three different scales. By detecting on the fused feature maps of multiple scales, YOLOv3 is more effective for small objects. The detection accuracy rate is greatly improved.

2.2.1. Network structure. The YOLOv3 network structure is shown in Figure 3 below. The input image is processed by the Darknet-53 convolutional network and divided into three YOLO layers of different scales. The detection is performed on the fusion feature maps of multiple scales, so that YOLOv3 has a better detection effect than YOLOv2.

2.2.2. Loss function. The loss function of YOLOv3 mainly contains three parts: target positioning offset loss, target confidence loss and target classification loss. The YOLOv3 loss function is shown in the following formula (1), where $\lambda_1, \lambda_2, \lambda_3$ is the balance coefficient of each loss.

$$L(O, o, C, c, l, g) = \lambda_1 L_{conf}(o, c) + \lambda_2 L_{cls}(O, C) + \lambda_3 L_{loc}(l, g)$$  \hspace{1cm} (1)$$

The target positioning offset loss uses the mean square error of the true deviation value and the predicted deviation value, as shown in the following formula (2):

$$L_{loc}(l, g) = \sum_{i \in \text{pos \ in\{x, y, w, h\}}} (l_i - \hat{g}_i)^2$$  \hspace{1cm} (2)$$
Among them:
\[
\begin{align*}
\hat{l}^i &= b^i - c^i, \quad \hat{l}^i = b^i - c^i, \\
\hat{g}^i &= g^i - c^i, \quad \hat{g}^i = g^i - c^i \\
\hat{l}^w &= \log(b^w / p^w), \hat{l}^h &= \log(b^h / p^h), \\
\hat{g}^w &= \log(g^w / p^w), \hat{g}^h &= \log(g^h / p^h)
\end{align*}
\]

The above formula (2) \( \hat{l} \) represents the coordinate offset of the predicted rectangular frame, \( \hat{g} \) represents the coordinate offset between the matched GTbox and the default frame, \((b^x, b^y, b^w, b^h)\) represents the predicted target rectangular frame parameter, \((c^x, c^y, p^w, p^h)\) represents the default rectangular frame parameter, and \((g^x, g^y, g^w, g^h)\) represents the matched real target rectangle parameters.

The target confidence indicates the probability of the detection target in the predicted target rectangle. The target confidence loss \( L_{conf}(o, c) \) uses binary cross-entropy loss, as shown in the following formula (3):
\[
L_{conf}(o, c) = -\sum (o \ln(\hat{c}) + (1 - o) \ln(1 - \hat{c})) \\
\hat{c} = \text{Sigmoid}(c)
\]

Among them, \( o \in \{0,1\} \) indicates whether the target actually exists in the predicted target bounding box \( i \), 0 indicates that it does not exist, and 1 indicates that it exists. \( \hat{c} \) represents the \text{Sigmoid} probability of predicting whether there is a target in the target rectangle \( i \).

The target category loss \( L_{cla}(O, C) \) also uses the binary cross-entropy loss, as shown in the following formula (4):
\[
L_{cla}(O, C) = -\sum_{i \in \text{pos}} \sum_{j \in \text{cla}} (O_{ij} \ln(\hat{C}_{ij}) + (1 - O_{ij}) \ln(1 - \hat{C}_{ij})) \\
\hat{C}_{ij} = \text{Sigmoid}(C_{ij})
\]

Where \( O_{ij} \in \{0,1\} \) indicates whether the \( j \) target actually exists in the predicted target bounding box \( i \), 0 indicates that it does not exist, and 1 indicates that it exists. \( \hat{C}_{ij} \) represents the \text{Sigmoid} probability of the \( j \) target in the bounding box \( i \) of the network predicted target.

2.3. ER_YOLO

Although YOLOv3 occupies a large position in engineering applications. However, the YOLOv3 convolutional network has too deep layers and too many parameters, and the training of the model has relatively high requirements for hardware. In actual scenes, there are few hardware devices equipped with GPU, so the real-time performance is not strong.

Aiming at the above shortcomings of YOLOv3, this article will improve YOLOv3 from the three aspects of convolutional network structure and loss function and anchor box clustering, thus, obtain a lightweight model—ER_YOLO, and apply it to the learner's emotion recognition.

2.3.1. Improvement of convolution structure. Since the convolutional neural network was proposed, due to the relatively large amount of calculation of the standard convolution and the parameters of the convolution kernel, after continuous optimization, such as Group convolution and Dilated convolution have been produced. As well as various forms of convolution such as Depthwise separable
convolution [16]. Among them, the deep separable convolution [13] is the core of the mobile-side lightweight network representing MobileNet.

Depth separable convolution is a combination of two parts: Depthwise Convolution (DW) and Pointwise Convolution (PW) [17] to extract features. Compared with standard convolution, its parameter Less, lower computational cost. The structure of standard convolution and depth separable convolution is shown in Figure 4 below.

Assuming that the size of the input feature map is $D_i \times D_j \times M$, the size of the convolution kernel is $D_k \times D_k \times M$, and the size of the output feature map is $D_o \times D_o \times N$, the parameters of the standard convolution using Figure (a) are:

$$W_{\text{std}} = (D_k \times D_k \times M) \times N$$  \hspace{1cm} (5)

And the size of the depth convolution in Figure (b) is $(D_k, D_k, 1)$, there are $M$ in total, and the size of point convolution in Figure (c) is $(1,1, M)$, there are $N$ in total, then the parameters of depth convolution and point convolution are:

$$\begin{cases} W_{\text{depthwise}} = (D_k \times D_k \times 1) \times M \\ W_{\text{pointwise}} = (1 \times 1 \times M) \times N \end{cases}$$  \hspace{1cm} (6)

Among them: $W_{\text{depthwise}}$ represents the depth convolution parameter quantity, and $W_{\text{pointwise}}$ represents the point convolution parameter quantity. Therefore, the ratio of the depth separable convolution parameter to the standard convolution parameter is:

$$\eta = \frac{W_{\text{depthwise}} + W_{\text{pointwise}}}{W_{\text{std}}} = \frac{(D_k \times D_k \times 1) \times M + (1 \times 1 \times M) \times N}{(D_k \times D_k \times M) \times N} = \frac{1}{N} + \frac{1}{D_k}$$  \hspace{1cm} (7)

Therefore, from the above formula (7), the parameter amount of the depth separable convolution is the $\frac{1}{N} + \frac{1}{D_k}$ of the standard convolution.

This article uses the Mobilenet network structure with deep separable convolution as the core to replace the network structure Darknet-53 in YOLOv3. Compared with the original YOLOv3, the improved model parameters are greatly reduced.
2.3.2. Improvement of loss function. IoU (Intersection and Union Ratio) has scale invariance, so it can reflect the detection effect of the predicted detection frame and the real detection frame. However, if there is no intersection between the real frame and the predicted frame, then IoU=0. At this time, the overlap between the two cannot be reflected, and learning and training cannot be performed.

In response to the above shortcomings, Hamid et al. proposed the concept of Generalized Intersection over Union (GIOU) [18].

\[ GIOU = IoU - \frac{A_c - U}{|A_c|} \]  

(8)

Among them: \( A_c \) represents the minimum area of the closed area that contains both the real box and the predicted box, \( U \) represents the union of the real box and the predicted box, \( \frac{|A_c - U|}{|A_c|} \) represents the proportion of the closed area that does not belong to the two boxes in the closed area.

Therefore, on the basis of replacing the convolutional structure Darknet53 with MobileNet, this paper uses GIOU as a part of the loss function to improve the target positioning offset loss \( \hat{L}_{loc}(l,g) \), then the improved target positioning offset loss is:

\[ \hat{L}_{loc}(l,g) = \sum_{i \in \text{pos}} \sum (1 - GIOU) \]  

(9)

Therefore, the improved YOLO v3 loss function is:

\[ L(O,o,C,c) = \lambda_{conf}L_{conf}(o,c) + \lambda_{cls}L_{cls}(O,C) + \lambda_{loc}\hat{L}_{loc}(l,g) \]  

(10)

2.3.3. K-means anchor box clustering. YOLOv3 uses a priori box to predict the bounding box coordinates, and uses the K-means clustering algorithm [15] to get the number and size of anchors. However, in the scene of learner emotion recognition in the classroom, the a priori box obtained by clustering the COCO data set is not suitable. Therefore, this paper uses the K-means algorithm to re-cluster the self-made data set to obtain a suitable anchor. The anchor values obtained by re-clustering are: 86x101, 110x172, 135x107, 147x125, 162x145, 175x122, 187x256, 204x159, 228x356. Used to replace the original YOLOv3 anchor values: 10x13, 16x30, 33x23, 30x61, 62x45, 59x119, 116x90, 156x198, 373x326. The experimental results are shown in Figure 5 below.

![Figure 5. K-means clustering result graph](image)

3. Experimental results and analysis

3.1. Dataset
There are many types of expression databases, among which the most commonly used are the Kaggle Facial Expression Recognition Challenge database Fer-2013, Cohn-kanade (CK) dataset and CK+, etc.
Based on the Kaggle database, this article selects images with obvious facial expressions, and combines the data collected by the web crawler to construct its own data set. There are a total of 14,194 pictures, of which 12,194 are used as training sets and 2,000 are test sets, which are divided into 7 expressions: surprised, anxious, Disgust, happiness, sadness, angry and neutral.

3.2. Model evaluation norm
In the target detection task, to evaluate the pros and cons of a model effect, indicators such as Precision, Recall, and mean average precision (mAP) are often used, as shown in the following formula (11).

\[
\begin{align*}
\text{Recall} &= \frac{TP}{N} \\
\text{Precision} &= \frac{TP}{TP + FP}
\end{align*}
\]

(11)

Among them: TP stands for True Positive, FP stands for False Positive, and N stands for the total number of targets. In general, the higher the accuracy rate, the better, but the accuracy rate and the recall rate affect each other, high accuracy rate and low recall rate, or low accuracy rate and high recall rate. In order to better balance the relationship between the two indicators, one way is to draw the accuracy and recall rate curve, that is, the PR curve, and the other way is to calculate the score \( F_\beta \), as shown in Equation (12).

\[
F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot (\text{Precision} + \text{Recall})}
\]

(12)

Among: when \( \beta = 1 \) is called \( F_1 \) score.

3.3. Model comparison
This article compares ER_YOLO with YOLOv3, YOLOv3-spp, YOLOv3-tiny to illustrate the effectiveness of the model. As shown in Table 1 below. The test results of each model are shown in Figure 6 below.

| Network model | mAP@0.5 | Recall | Precision | F1    | Time/s | Params          |
|---------------|---------|--------|-----------|-------|--------|-----------------|
| YOLOv3        | 0.67    | 0.72   | 0.68      | 0.695 | 0.0322 | 6.15614e+07     |
| YOLOv3-spp    | 0.78    | 0.80   | 0.65      | 0.717 | 0.1053 | 6.2611e+07      |
| YOLOv3-tiny   | 0.53    | 0.65   | 0.58      | 0.613 | 0.0091 | 8.68605e+06     |
| ER_YOLO       | 0.71    | 0.78   | 0.68      | 0.727 | 0.0117 | 8.68836e+06     |

From Table 2 we can see that the improved model ER_YOLO has the highest F1 score, which is an increase of 3.2%, 1.0% and 11.4% compared to YOLOv3, YOLOv3-spp and YOLOv3-tiny, respectively. The detection time is only 1/3 and 1/10 of the detection time of YOLOv3 and YOLOv3-spp, but compared to YOLOv3-tiny, the detection time is increased by 0.26%. To a certain extent, it can be considered that the detection time of the two is almost the same. At the same time, the parameters of the improved model ER_YOLO are reduced a lot compared to YOLOv3 and YOLOv3-spp, only 1/7 and 1/7 of YOLOv3 and YOLOv3-spp respectively. But compared to YOLOv3-tiny, the improved model parameters increased by 1/5, which is almost the same.
4. Application of learner emotion recognition
The traditional teaching model is a kind of "intellectualism", which only pays attention to the cognitive level of learners, and rarely considers the role of non-intellectual factors such as human emotion, personality, hobbies, etc. in learning activities. The learner emotion recognition method proposed in this paper can quickly, accurately, and real-time recognize learner emotions, which is of great significance for emotional interaction in the remote teaching environment, and it also helps to improve classroom efficiency.

5. Conclusion
Learner emotion recognition in a smart environment is one of the important indicators of classroom observation and evaluation, and it is also a hot issue in the field of emotional computing research in recent years. The article proposes a learner emotion recognition model based on deep learning. The experimental results show that the improved model parameters are greatly reduced and the detection speed is faster, which is suitable for actual engineering projects. However, the research in this article also has many shortcomings: (1) Facial expressions are easily affected by the external environment. For example, too dark or too bright light will affect the detection accuracy, and some facial expressions are subtle, making recognition difficult. (2) The influence of learner's posture on facial expression is not considered.

Acknowledgments
The author would like to thank the classmates for their technical support for this article, and thank the tutor for his opinion on the author's writing.

References
[1] Y Liu, Q.F Fu, X.L Fu. The interaction between cognition and emotion[J]. Science Bulletin, 2009, 54: 2783-2796.
[2] Z.L Meng. Emotional Psychology [M]. Beijing: Peking University Press, 2005.
[3] Z.G Xu. Emotion recognition and its application in learning images in smart learning environment [D]. Shandong: Shandong Normal University, 2019.
[4] MEHRABIAN A. Communication without words[J]. Psychology today, 1968, 2(4): 53-56.
[5] EKMAN P, FRIESEN W V. Facial action coding system: a technique for the measurement of facial movement[M]. Palo Alto: Consulting Psychologists Press, 1978.

[6] T.T Wang, Y.W Wu, X.Y Ai. Recognition and intervention method of learning fatigue based on facial expression recognition[J]. Computer Engineering and Design, 2010, 31(08): 1764-1767+1778.

[7] B Sun, Y.N Liu, J.B Chen, etc. Emotion analysis based on facial expressions in a smart learning environment[J]. Modern Distance Education Research, 2015, 2:96-103.

[8] Z.H Zhan. Emotion and Cognition Recognition Model of Distance Learners Based on Intelligent Agent—Coupling of Eye Tracking and Facial Expression Recognition Technology[J]. Modern Distance Education Research, 2013, (05):100-105.

[9] R.J Wei, Y.Z Ding, L Zhang, etc. Design and implementation of emotion recognition module in online learning system [J]. Modern Educational Technology, 2014, 24(03): 115-122.

[10] J.J Yan, W.M Zheng, M.M Xin, etc. Bimodal Emotion Recognition of Expression and Posture[J]. Chinese Journal of Image and Graphics. 2013, 18(9): 1101-1106.

[11] Z.G Xu, G.W Zhang, X.Z Meng, etc., learner emotion recognition and application based on deep learning [J]. Audio-visual Education Research, 2019, 02: 87-93.

[12] H.Y Li. Multi-pose face correction and recognition based on affine transformation [D]. Jiangsu: Soochow University, 2013.

[13] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861,2017.

[14] H Rezatofighi, N Tsoi, J Y Gwak, et al. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. Computer Vision and Pattern Recognition, arXiv:1902.09630.

[15] Hartigan J A, Wong M A. Algorithm AS 136: A K-Means Clustering Algorithm[J]. Journal of the Royal Statistical Society,1979,28(1):100-108.

[16] M Gu. Research on Tibetan independence flag detection algorithm based on improved convolutional neural network and selective retrieval [D]. Gansu: Lanzhou University, 2019.

[17] H.L Jiang. Research on shared electric vehicle collision detection and early warning based on deep learning [D]. Jiangsu: Southeast University, 2018.

[18] Z.H He, Z.X Wang. Based on improved YOLO v2 body-in-white solder joint detection method [J]. Computer Engineering, 2020.