Computer Methods and Programs in Biomedicine 215 (2022) 106621

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Facial expression recognition based on deep learning

Huilin Ge, Zhiyu Zhu*, Yuewei Dai*, Biao Wang, Xuedong Wu

School of Electronic Information, Jiangsu University of Science and Technology, Zhenjiang 212003, China

ARTICLE INFO

Article history:
Received 17 September 2021
Revised 22 December 2021
Accepted 3 January 2022

Keywords:
3D face depth information
Deep learning
Facial expression recognition
Target detection
Convolutional neural network

ABSTRACT

Background and objective: Facial expression recognition technology will play an increasingly important role in our daily life. Autonomous driving, virtual reality and all kinds of robots integrated into our life depend on the development of facial expression recognition technology. Many tasks in the field of computer vision are based on deep learning technology and convolutional neural network. The paper proposes an occluded expression recognition model based on the generated countermeasure network. The model is divided into two modules, namely, occluded face image restoration and face recognition.
Methods: Firstly, this paper summarizes the research status of deep facial expression recognition methods in recent ten years and the development of related facial expression database. Then, the current facial expression recognition methods based on deep learning are divided into two categories: Static facial expression recognition and dynamic facial expression recognition. The two methodswill be introduced and summarized respectively. Aiming at the advanced deep expression recognition algorithms in the field, the performance of these algorithms on common expression databases is compared, and the strengths and weaknesses of these algorithms are analyzed in detail.
Discussion and results: As the task of facial expression recognition is gradually transferred from the controlled laboratory environment to the challenging real-world environment, with the rapid development of deep learning technology, deep neural network can learn discriminative features, and is gradually applied to automatic facial expression recognition task. The current deep facial expression recognition system is committed to solve the following two problems: (1) Overfitting due to lack of sufficient training data; (2) In the real world environment, other variables that have nothing to do with expression bring interference problems.
Conclusion: From the perspective of algorithm, combining other expression models, such as facial action unit model and pleasure arousal dimension model, as well as other multimodal models, such as audio mode, 3D face depth information and human physiological information, can make expression recognition more practical.

1. Introduction

Image recognition is a technology that uses computer to process, analyze and understand images to identify different patterns of targets and objects. It is a main research direction in the field of computer vision and plays an important role in intelligent data acquisition and processing based on image. Image recognition technology can efficiently complete the detection and recognition of specific target objects (such as handwritten characters, products or faces), image classification and marking, and subjective image quality evaluation. At present, image recognition technology has a broad commercial market and optimistic application prospects in Internet application products such as image retrieval, commodity recommendation, user behavior analysis and face recognition. Moreover, it has long-term development potential in high-tech industries such as UAV, autonomous driving and intelligent robot, as well as many fields such as geology, medicine and biology. Early image recognition systems mainly used directional gradient histogram [1] and scale invariant feature transformation [2], and then input the extracted features into the classifier for classification and recognition. These functions are basically manually designed. For different recognition problems, the extracted features will directly affect the performance of the system. Therefore, researchers need to further study the unsolved problem areas in order to design better adaptive features to improve the performance of the system. The image recognition system in this period is often for a specific recognition task, and the data scale is small, the generalization ability is poor, so it is not easy to achieve the ideal recognition effect in daily application.

* Corresponding authors.
E-mail addresses: zhuzy@just.edu.cn (Z. Zhu), dywjust@163.com (Y. Dai).

https://doi.org/10.1016/j.cmpb.2022.106621
0169-2607/© 2022 Elsevier B.V. All rights reserved.

Deep learning is a branch of representation learning based on data of artificial neural network. Deep learning includes supervised learning, semi-supervised learning and unsupervised learning. In deep learning, deep neural network, deep belief network and recurrent neural network have been widely used in speech recognition, computer vision, audio recognition, unmanned driving and natural language processing. Rina dechter first introduced deep learning in 1986. In addition, Igor aizenberg introduced the concept of artificial neural network in 2000. In fact, Alexey ivakhnenko and Lapa proposed supervised feedforward learning network as early as 1965. In 1986, Geoffrey Hinton proposed the back-propagation algorithm of multilayer perceptron (MLS), and used the SIGMOD activation function for nonlinear transformation, so as to solve the problems of nonlinear learning and classification. In 2006, Geoffrey Hinton et al. Pointed out that the pre-training weight should be used to initialize the model and fine tune the model according to the supervised training. With the further development of deep learning, the proposal of lenet [3] in 1998 marked the emergence of convolutional neural network (CNN).

However, due to the backward hardware at that time, convolutional neural network is not in an advantage compared with other machine learning methods (such as SVM [4]). With the further development of computing devices, the Alex net that is proposed by Hinton et al. has made significant achievements in the computer vision competition ILSVRC 2012. Convolutional neural networks have made rapid progress in recent decades.

2. Related research

2.1. Review of face recognition technology

Face recognition has the characteristics of easy access, easy operation and diversified features. In the last century, many scholars studied face recognition. However, due to the underdeveloped network, limited resources of face images and poor quality of photos, many scholars mostly studied it from the perspective of algorithm, but the recognition accuracy is low, far from the human eye recognition effect. With the gradual maturity of machine learning technology, there are many powerful algorithms, such as genetic algorithm, Bayesian classifier, support vector machines. The application of these algorithms in face recognition technology improves the accuracy of face recognition to a certain extent, but its feature extraction is complex and single, which is greatly affected by human factors, so these methods are not widely used. With the development of science and technology, a large number of face image resources can be obtained, and the rapid development of deep learning provides new vitality for face recognition, and its recognition accuracy is greatly improved.

Sun et al. proposed the DeepId [5] face recognition method, which uses deep learning to learn a group of advanced features to represent face information. After learning the face feature information, only classifiers are used for classification, and the recognition rate is better on LFW data. Zhu et al. embedded the improved DeepId model into the embedded system, and achieved good recognition results and speed. Compared with DeepId2 [6] the recognition accuracy of DeepId2 has been greatly improved. The main reason is that the recognition of feature extraction not only takes into account the classification accuracy, but also takes into account the gap between classes, which are used as recognition signal and verification signal respectively, and then the two signals are weighted and combined for common supervision training. Compared with DeepId2, DeepId2+ [7] improves the robustness of occlusion. Schroff et al. [8] believe that although the field of face recognition has made great progress (DeepFace [9], DeepId2), the effective implementation of large-scale face verification and recognition is still facing severe challenges. This paper proposes a compact Euclidean space metric mapping method for end-to-end learning image coding into Euclidean space, and then face recognition based on this coding. At the same time, Facenet uses the triple loss function in the training process and calculates the distance between the same class of images and different classes of images through the learned feature vectors, so that the distance between the same class decreases and the distance between different classes increases. Coincidentally, Wen et al. [10] proposed a central loss function, using the central loss function and the traditional softmax loss joint supervision training can effectively reduce the distance between the same category, and then improve the classification and recognition ability.

In order to enable the features learned from the network to be more distinguishable, many scholars have improved on the basis of softmax loss, which not only simplifies the training effect, but also achieves a better recognition effect. For example L-softmax loss [11] A-softmax loss [12] and Cosface [13] add constraints in the aspect of classification angle, change the classification interval of the original classification function, improve the learning ability of the network for features, make features aggregate within classes, disperse between classes, and achieve very good recognition effect. Chen et al. [14] proposed a very effective lightweight face recognition network model, the model parameters are less than 1 million, the recognition accuracy is higher than MobileNetV2 [15] and the speed is also greatly improved. At the same time, weighted global security local average pooling is used in the last layer of the network. Although this method increases the training parameters of the model to a certain extent, it can effectively reduce the impact of reduced accuracy due to the common pooling.

2.2. Research status of target detection at home and abroad

Target detection is a very important problem in the field of computer vision. In recent years, many detectors based on convolutional neural network have been proposed to increase the speed and accuracy of target detection. However, the scale change in target detection is still an important problem for all detectors. There are basically two kinds of target detection methods, namely single-stage method and two-stage method.

The two-stage method divides the target detection task into two stages: extracting the region of interest, and then starting classification and regression in the region of interest. R-CNN [16] uses the region generated from the edge box [17] or selective search [18]. It is suggested to generate region based features from the pre-trained convolutional neural network and use SVMs for classification. Fast R-CNN [19] introduces end-to-end classification and position regression loss to train convolutional neural network. Fast R-CNN [20] is replaced by network (RPN) with regional recommendation selective search. RPN is used to generate candidate bounding boxes (anchor boxes) and filtrate the background area, and then another small network is used for classification and bounding box position regression based on these suggestions. R-FCN [[30]] uses position tactful region of interest pool (PSROI) instead of the region of interest pooling in speed fast RCNN. This method can increase the speed and accuracy of the detector. At present, deformable convolution and deformable PSROI are proposed in the variable roll in network [[31]], which can continue to increase the accuracy of R-FCN.

The single-stage method eliminates the ROI extraction process and directly classifies and regresses the candidate anchor frames. YOLO [[32]] separates the input image into several grids and locates and classifies each section of the image. Using this concept, YOLO can detect the target quickly, but the accuracy is not ideal. YOLOv2 [[33]] improves on the YOLO network and proposes a joint training strategy of classification and detection, which allows YOLOv2 to predict the detection of object classes without labeled detection data. At the same time, a multi-scale training method is used to make the same model run in different sizes of images, which provides a proper trade-off between speed and accuracy. YOLOv3 [21] uses darknet-53 as the backbone network, and the residual structure can retain the original features as much as possible when extracting new features, and can eliminate the problem of gradient disappearance to a certain extent. At the same time, in order to reduce the negative effect of gradient caused by pooling, the convolution step size of 2 is used instead of pooling to achieve downsampling. In order to increase the detection effect of the algorithm for small targets, YOLOv3 adopts the upsampling fusion method similar to FPN [[34]] increases the semantic information of the feature map, and adopts the multi-scale detection method.

In order to support multi-label objects, we use logistic instead of Softmax to predict categories. Through these improvements, the speed and accuracy have been greatly improved. SSD [[35]] is another effective single-stage target detector. In order to detect objects of different scales, a feature pyramid is constructed. In this way, large feature maps can detect small targets and small feature maps can detect large targets. At the same time, different from the last layer of YOLO, SSD uses convolution feature map to detect. SSD uses the anchor frame design concept of Fast R-CNN for reference, and sets up preselection frames with different length-width ratios in each pixel, which can reduce the difficulty of training to a certain extent. Compared with the two-stage detection, SSD has higher detection accuracy and real-time efficiency.

However, SSD still has the problem of class imbalance, so it can not effectively detect small targets. According to these shortcomings, some scholars have made improvements. Through the upsampling method, the deep feature map and shallow feature map are fused, so as to enrich the semantic information of the feature map and make the detection of small targets more convenient. RFBnet [[36]] adds hole convolution to the structure of Inception [[37]] to simulate the human visual receptive field to improve the feature extraction ability of the network, and improves the detection accuracy under the condition of ensuring the speed. Focal loss [[38]] believes that there are many pre selection boxes in a picture, but only a few of them contain targets, which makes the number of minus samples too large, accounting for the best of the gross loss, and most of them are amiable in samples classification, resulting in the trouble of category imbalance. Focal loss solves the troublesome of category imbalance by cutting down the weight of easily classified samples. At the same time, a simple and dense network is designed to ensure accuracy and speed.

3. Methods

3.1. Depth target detection algorithm

In the realm of computer vision, object recognition technology is an algorithm that can detect sample objects in videos and photos. Recent target recognition algorithms mostly rely on high-performance GPU chips based on multilayer neural network and deep learning software framework and built-in thousands of stream processors. Therefore, it is also known as deep object detection (Deep OD). Object detection has always been a vital issue in the realm of computer vision. Before deep learning is applied to object detection, traditional object detection techniques are based on manual feature collection. Traditional target detection methods are generally divided into three stages: region selection, feature extraction and regression classification. However, there are some problems: in the region selection, the strategy effect is poor and the time complexity is high, and the robustness of the features extracted by hand is not very good. With the application of deep learning in the field of target detection, target detection in view of deep learning can be differentiated into two types: two-stage target detection and single-stage target detection.

Two-stage target detection, as the name suggests, is that the whole target detection process is divided into two stages: first, propose possible regions and extract features at the same time, then classify and regress. Two-stage target detection methods are based on region location. The most common is the R-CNN(Region-CNN) series.

Single-stage target detection, as the name suggests, is to complete the task of detection and classification in one training without step-by-step training. On the condition of assuring the correctness of classification, the performance is greatly improved. Common single-stage target detection algorithms contain YOLO, SSD, YOLOv2 and YOLOv3.

From the view of the application, target detection contains significant target and general target detection. Inside these, general target detection contains object recognition, vehicle detection, pedestrian detection, face detection and so on. Significant target detection should imitate the effect of extracting important targets from images in human vision. Representative target detection algorithms include R-CNN, FastR-CNN, Fast R-CNN SSD (Single Shot MultiBox Defender) and Yolo (You Only Look Once), etc. These algorithms continuously improve the recognition rate and performance on standard open datasets by improving the topology of neural network.

3.2. Joint optimization method

This paper proposes a joint optimization method of Softmax loss and improved Island loss, which integrates the expression classification task based on Softmax loss and the metric learning task based on improved Island loss into the AlexNet-Emotion network. Softmax loss can make the model learn expression features, and the improved Island loss can make the expression features learned by the model more discriminative and compact. Through multi-task learning, the advantages of the two tasks are brought into play, so as to improve the discrimination capacity and robustness of the pattern.

In this paper, an easy and available Island loss function is selected, and its shortcomings are improved. The effect is proved to be better than the original function. The main function of island loss is to push the samples to the corresponding class centers and keep the class centers away from each other. When counting the range between the sample and the homologous class center, the square difference loss function is used to count the loss and backpropagation of all samples. So as to deal with the troublesome that the loss of all samples causes a large number of calculation loss and does not pay attention to the more difficult to classify samples. Based on the island loss, this paper adds online difficult sample mining technology, which is called improved Island loss. By only calculating the loss of the relatively difficult samples which exceed the radius of the class center, the model can further adjust the relatively difficult samples and speed up the convergence rate of the model. The calculation formula of the center distance L_c between the sample and the homologous class is as follows:

$$L_c = \frac{1}{2}\sum_{i=1}^{m} max\left(||x_i - c_{y_i}||^2 - \tau,\ 0\right) \tag{1}$$

Where m is the number of batch training samples, $y_i$is the real label of the sample $x^{(i)}$, c_{y_i} is the class center corresponding to the sample $x^{(i)}$, and $x_{\ i}$ is the feature vector of FC7 layer of sample $x^{(i)}$, τ is the threshold.

Cosine similarity is used to calculate the distance between class centers. The smaller the value is, the larger the distance between class centers is. The improved formula of Island Loss function is as follows:

$$L_{IL} = L_c + {}_1 \sum_{c_j \in K} \sum_{\substack{c_k \in K \\ c_{k \neq c_j}}} \left(\frac{c_k c_j}{||c_k||_2 \left|\left|c_j\right|\right|_2} + 1 \right) \tag{2}$$

Where k is the expression type, • is the vector dot multiplication, $||c_j||_2$ is the L2 norm of c_j, λ_1 is the loss weight of the class center, $c_k$is the class center of class K.

Therefore, the joint optimization method of Softmax loss and improved Island loss is put forward to drill the AlexNet-Emotion network by using the advantages of Softmax loss and improved Island loss, which can effectively solve the problem of small inter-class variance and large intra-class variance. The overall loss function of the model is as follows:

$$L = L_s + L_{IL} \tag{3}$$

$$L_s = -\frac{1}{m} \sum_{i=1}^{m} \log \frac{e^{z_{yi}}}{\sum_{j=1}^{k} e^{z_j}} = -\frac{1}{m} \sum_{i=1}^{m} \log \left(\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{k} e^{W_j^T x_i + b_j}} \right) \tag{4}$$

Among them, L_s is the Softmax loss, λ is the weight of the improved Island loss, y_i is the real category label of the sample $x^{(i)}$, $x_{\,i}$is the expression feature vector of FC7 layer of sample F, m and K are the account of drilling samples and expression categories, $W_j$and b_j are the weight and bias of category j, and z_j is the score of category j.

The expression class center is calculated by the feature vector of the average sample, which will introduce noise interference, so the random gradient descent (SGD) way is applied to update the class center. The updating formula of the j-type expression center is as follows:

$$\Delta c_j = \frac{\sum_{i=1}^{m} \delta(y_i, j)\left(c_j - x_i\right)}{1 + \sum_{i=1}^{m} \delta(y_i, j)} + \frac{\lambda_1}{|K| - 1} \sum_{\substack{c_k \in N \\ c_k \neq c_j}} \frac{c_k}{\|c_k\|_2 \|c_k\|_2} - \left(\frac{c_k \cdot c_j}{c_{k2} \left|\left|c_j\right|\right|_2^3} \right) c_j \tag{5}$$

$$c_j^{t+1} = c_j^t - a c_j^t \tag{6}$$

Where a is the class center update learning rate and t is the tth iteration; When y_i=j, $\delta(y_i$,j)=1; $y_i \neq$ j, $\delta(y_i$,j)=0.

In order to effectively detect the change process of expression intensity in video frames, this paper uses IRNN to extract the temporal information in video.

Specifically, a video is converted into N frames of images, and the spatial features of each frame are extracted by using the AlexNet-Emotion network to obtain n 128 dimensional expression feature vectors, which are sent to the IRNN network. The IRNN learns the dependency of each frame in the sequence through the hidden layer, so as to pick up the temporal features of the expression, the hidden layer representation of each moment contains the expression change information of the previous moment.

To deal with the troublesome that it is uneasy to integrate the expression features between frames effectively, attention module is introduced to distinguish the importance of each frame. By comparing the hidden layer representation of the current frame with other frames, the importance of the current frame can be obtained. The time complexity of any two video frames in the video is O(n^2).

Inspired by the introduction of attention mechanism in machine translation, so as to cut down the comparison time, this paper uses attention mechanism to embed IRNN network. As shown in Fig. 4.1, the IRNN network generates the video coding vector c, gives weight to the hidden layer representation of each frame by computing the similarity between the video coding vector c and the hidden layer representation $h_{\,i}$, and carries out linear weighting to obtain the video level expression features, with time complexity of O(n). The calculation formula is as follows:

$$a_i = c^T V h_i \tag{7}$$

$$_i = \frac{\exp(a_i)}{\sum_{t=1}^{T} a_t} \tag{8}$$

$$f_v = \sum_{I=1}^{T} \xi_i h_i \tag{9}$$

Where t is the number of input IRNN video frames, f_v is the video level expression feature, ξ_i is the weight of the ith hidden layer representation, and V is the learning parameter.

3.3. Occluded face recognition model based on dual discrimination countermeasure network

The paper proposes an occluded expression recognition model based on the generated countermeasure network. The model is divided into two modules, namely, occluded face image restoration and face recognition, as shown in Fig. 5.

4. Results

4.1. Deep facial expression recognition based on static image

Due to the convenience and availability of network static data processing, a large number of researches are based on static images without considering time information for expression recognition. Direct training of deep network on relatively small facial expression databases will inevitably result in overfitting problem. In order to solve this problem, many related researches use additional auxiliary data to pre-train and build their own network, or directly fine tune based on an effective pre-training network such as AlexNet (Krizhevsky et al., 2012), visualgeometry group (VGG, [27]), VGG-Face (Parkhi et al.,2015) and GoogLeNet [[37]]. Large face recognition databases such as CASIA WebFace (Yi et al., 2014), CFW (cellular face in the wild) (Zhang et al., 2012) and FaceScrub dataset (Ng and Winkler, 2014), and relatively large facial expression databases such as FER2013 (Goodfellow et al., 2013) and TFD (the Toronto to face database) are suitable for auxiliary training. Kaya et al. (2017) pointed out that VGG face model pre-trained on face data is more suitable for facial expression recognition task than Imagenet model pre-trained on object data. [28] also pointed out that pre-training on large face databases and further fine-tuning on additional expression databases can effectively improve the expression recognition rate. In addition, [29] proposed a multi-stage fine-tuning strategy: in the first stage, additional facial expression database FE R 2013 was used to fine-tune the existing pre-training model; In the second stage, the known training set of (EmotiW) is used to fine-tune the model to make it more suitable for the target database. Ding et al. (2017) proposed the Face et2ExpNet framework to eliminate the interference of face information retained in the pre-training model on the expression recognition task.

4.2. Deep facial expression recognition network based on dynamic image sequence

The temporal correlation information between consecutive frames in an input sequence is beneficial to facial expression recognition. This section focuses on the deep spatiotemporal expression recognition network based on dynamic sequence. The network takes a certain range of frames in a period of time window

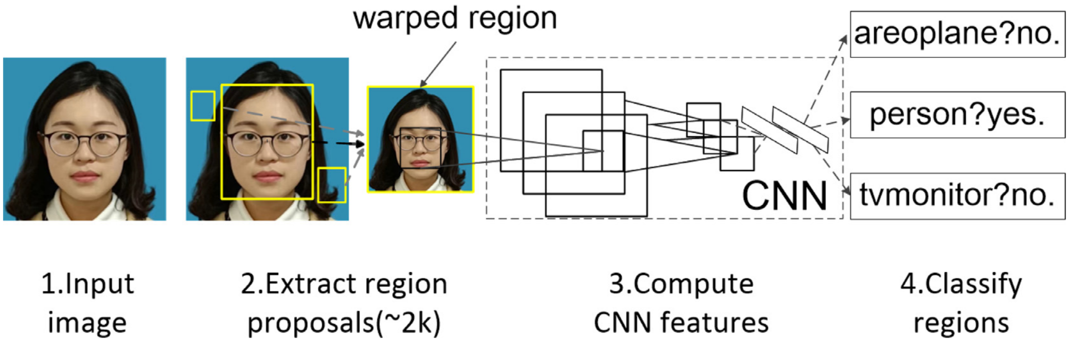


Fig. 1. Architeture of R-CNN.

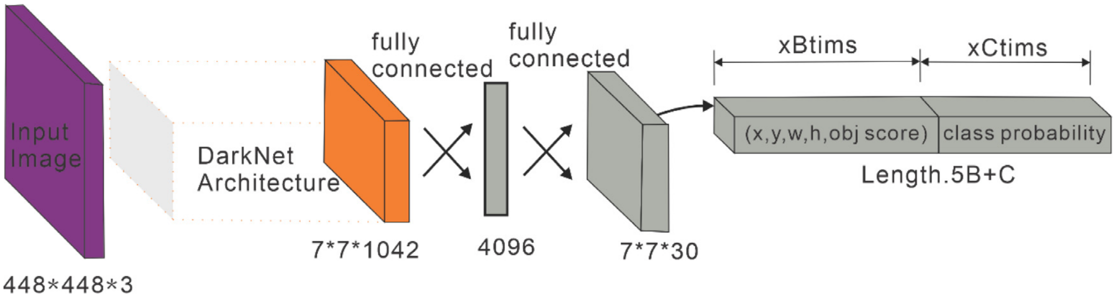


Fig. 2. Architeture of YOLO.

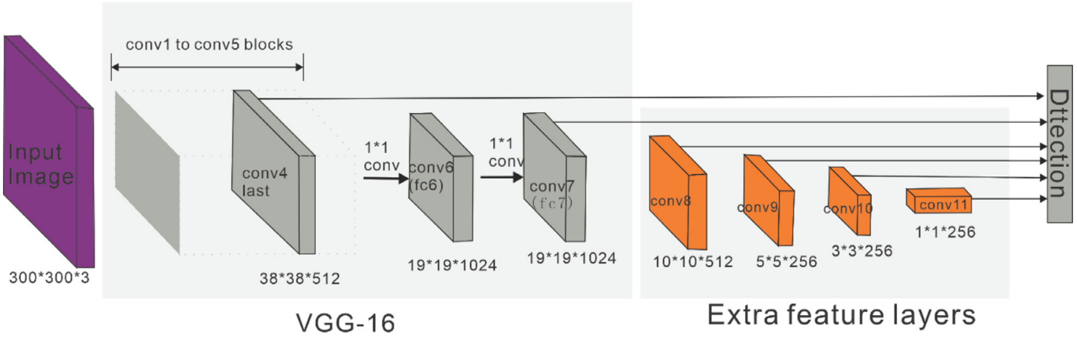


Fig. 3. Architeture of SSD.

as a separate input, considers the spatiotemporal motion mode in video sequence, and uses spatial and temporal information to capture more subtle expressions. Dynamic facial expression recognition based on video is to recognize the expression of the person in the video and classify the video into basic expression. Each video frame contains expression changes. How to integrate the features of each frame into video level representation effectively using temporal information is a difficult problem in video classification. In the early work, CNN was used to extract spatial information and generate video-level expression features [22] through average aggregation.

In recent years, RNN propagates information on sequences through continuous hidden layer representation, and makes effective use of temporal information. Some scholars further propose CNN-RNN spatiotemporal model, which uses the last-minute hidden layer representation as video level expression features. However, the above mentioned model cannot effectively use the expression change information to generate video level expression features. To solve this problem, a dynamic facial expression recognition model based on attention mechanism is proposlINed on the basis of AlexNet-Emotion network. Firstly, the spatial information of video frame is extracted by using AlexNet-Emotion network. Then, the cyclic neural network is used to extract the time information of the video frame. Secondly, the attention mechanism is used to give more weight to some video frames with obvious expression to generate video level expression features. Finally, the classification is realized through the full connection layer [23].

4.3. Summary of target detection based on deep learning

Whether it is two-stage target detection or single-stage target detection, the performance of target detection algorithm is usually measured by detection speed and detection accuracy. Both two-stage target detection and single-stage target detection try to find a balance between speed and accuracy. However, the emphasis of the two methods is different. The single-stage target detection algorithm pays more attention to the detection speed, and the two-stage target detection algorithm pays more attention to the detection accuracy [24]. Moreover, in the development of the two kinds of target detection algorithms, they learn from each other and fuse, and achieve better detection results. The two types of target detection algorithms have different application scenarios. For example, real-time target detection, single-stage target detection is better. When the image scene is more complex, the accuracy of two-stage target detection algorithm is needed.

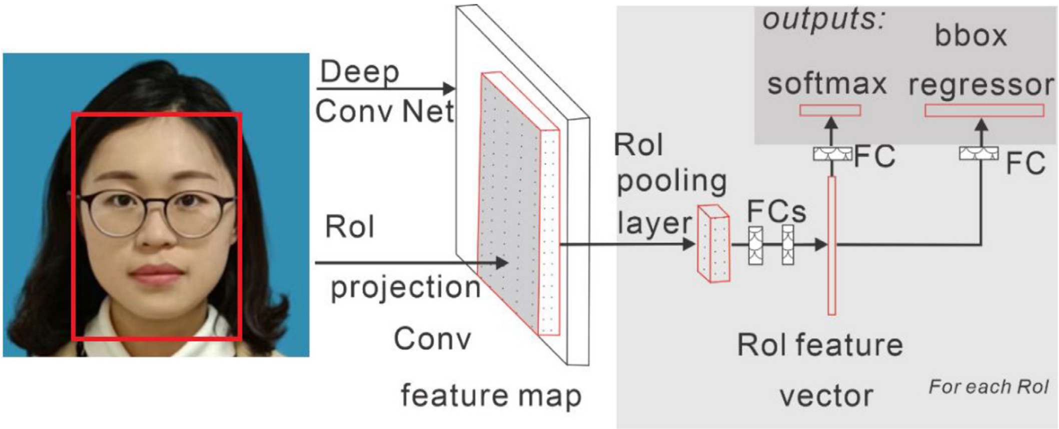


Fig. 4. Architeture of Fast R-CNN.

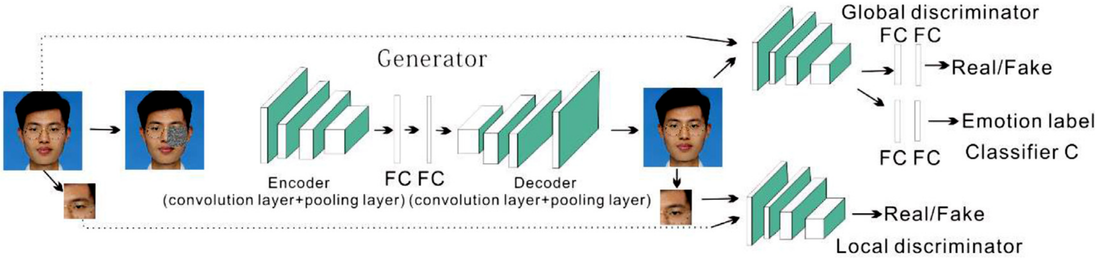


Fig. 5. Occlusion expression recognition model based on double discrimination network.

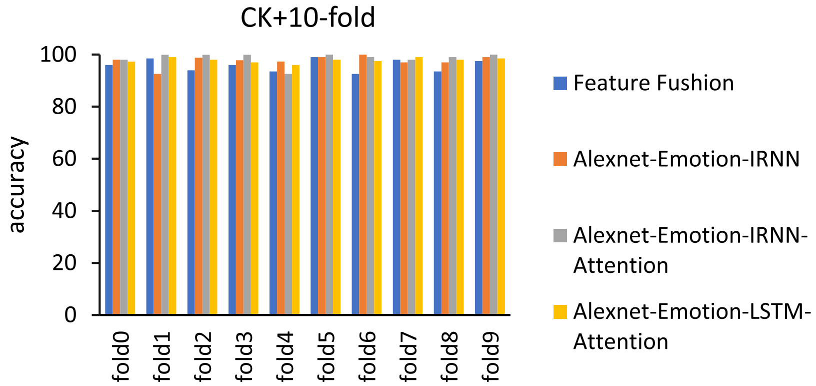
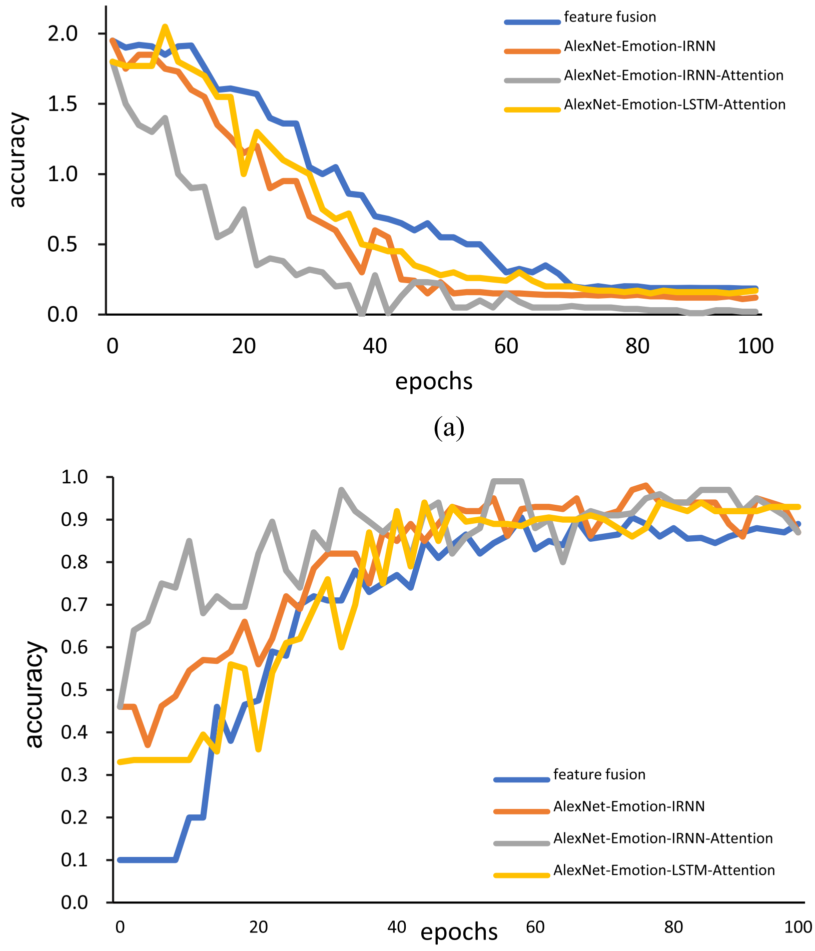


Fig. 6. Ten fold cross validation accuracy bar graph of four groups of experiments based on CK + dataset.

Table 1
Comparison results between the proposed algorithm and other algorithms.

Algorithm	CK+	MMI
AUDN	93.7%	75.85%
DeRL	96.57%	72.67%
IL-CNN	94.39%	70.67%
Inception	93.2%	77.6%
IACNN	95.37%	71.55%
AlexNet-Emotion Iinit	94.41%	65.88%
AlexNet-Emotion Softmax loss	96.45%	76.89%
AlexNet-Emotion w/o Angle Variance	93.73%	71.84%
AlexNet-Emotion Pre	97.14%	78.68%

In order to further provide the validity of the algorithm proposed in this paper, compared with the current popular algorithm, the comparison results are shown in Table 1. Liu et al. analysed action unit inspired deep networks (AUDN). Firstly, micro action pattern (MAP) is learned through convolution layer and pooling layer. Then feature grouping is used to simulate larger receptive field through adaptive combination with related maps, so as to generate more abstract middle layer semantics, Finally, in each receptive field, a multi-layer learner is used to construct a grouping sub network to further generate high-level representation, which is conducive to expression analysis. [26] proposed the de expression residue learning (DeR L) algorithm. In the process of generating neutral expression, the expression information will be filtered, and the expression information is still retained in the middle layer. The residual information in the middle layer is used to generate expression features. Cai et al. proposed using Island loss optimization [25] model. Mollahoseini et al. proposed a model consisting of two convolution layers, a maximum pooling layer and four inception layers. In this paper, the AlexNet-Emotion network is proposed to provide the auxiliary information of face key point angle change feature map. The comparison results in Table 1 show that the correctness of the algorithms proposed in this paper is higher



(a)

(b)

Fig. 7. (a) Training loss of four groups of experiments (b) accuracy of test set.

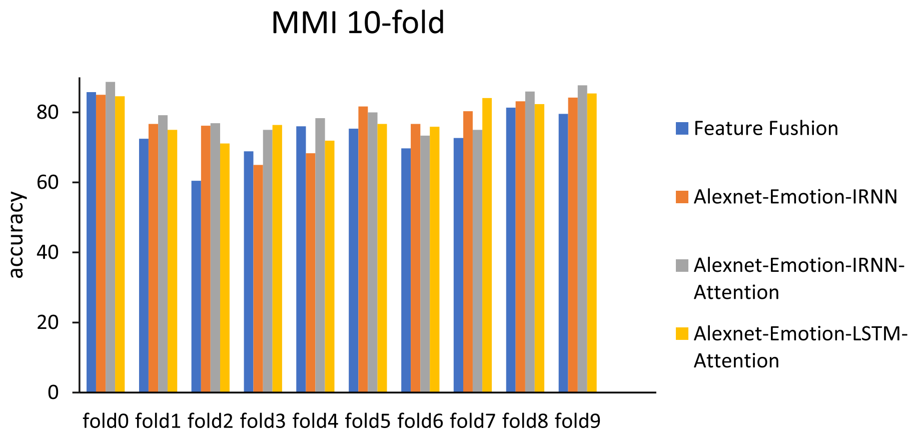


Fig. 8. Ten fold cross validation accuracy bar graph of four groups of experiments on MMI dataset.

than that of the above algorithms, which verifies the effectiveness of the AlexNet-Emotion model.

Ten fold cross validation accuracy bar graph of four groups of experiments on CK + dataset (Fig. 6) is presented. Furthermore, we also illustrate the training loss of four groups of experiments as well as the accuracy of test set (Fig. 7). Finally, the Ten fold cross validation accuracy bar graph of four groups of experiments based on the MMI dataset is performed (Fig. 8).

5. Discussion

This paper firstly recommends the background knowledge of facial expression recognition, and summarizes the evolution and development of database and algorithm in the field of facial expression recognition. It points out that deep learning has become the mainstream framework in this field. Then, the expression recognition algorithms based on deep learning are divided into two categories (static expression recognition network and dynamic expression recognition network). By comparing the performance of a large number of algorithms based on the current common expression database, it is found that dynamic expression network can often achieve better recognition results than static expression network. In view of the fact that the expression is in a dynamic process, and the deep expression recognition algorithm based on dynamic sequence can capture the effective temporal information and make use of the identity information of the characters in the same sequence, the space-time deep network will become a major trend in the field of expression recognition.

From the perspective of dataset, collecting a large variety of training data with accurate expression tags can fundamentally improve the expression recognition rate. From the perspective of algorithm, combining other expression models, such as facial action unit model and pleasure arousal dimension model, as well as other multimodal models, such as audio mode, 3D face depth information and human physiological information, can make expression recognition more practical.

6. Conclusions

In the field of computer vision, based on the research status at home and abroad, facial expression recognition technology has made great progress and development. However, there are still many challenges and difficulties waiting for researchers to solve. For example, the research on facial expression recognition in real scenes, and there is a certain confusion between different expressions. The expression of facial emotion may vary with region, culture, and environment. There are differences in personality. Therefore, facial expression recognition technology needs to be improved from different aspects to achieve better results. Deep learning iprovides a very strong research direction. Using convolution layer, pooling layer and full connection layer of convolution neural network, we can let this network structure study and pick up relevant features by itself, and put them to use. This feature applies plenty of convenience for several researches, and can leave out the very difficult modeling process in the past. In addition, deep learning has attained great improvement in image classification, object detection, pose estimation and image segmentation. On the one hand, deep learning has a wide sphere of applications and strong versatility, so we could continue to try hard to expand it to other application realms. On the other hand, deep learning still has a lot of potential issues to explore and discover (Figs. 1–3, Eqs. (1)–(9)).

Declaration of Competing Interest

The authors declare that there is no conflict of interest in this paper.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 62006102).

References

[1] C. He, Overview of face recognition technology, Intell. Comput. Appl. 6 (5) (2016) 112–114.

[2] M.H. Zhu, S.T. Li, Y. Hua, Facial expression recognition method based on sparse representation, Pattern Recognit. Artif. Intell. 27 (8) (2014) 708–712.

[3] Y.L. Xue, X. Mao, C. Catalin-Daniel, Robust facial expression recognition method under occlusion conditions, J. Beijing Univ. Aeronaut. Astronaut. 36 (4) (2010) 429–433.

[4] Y. Chenglin, P. Hailong, Face recognition framework based on effective computing and adversarial neural network and its implementation in machine vision for social robots, Comput. Electr. Eng. 92 (2021) 92.

[5] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, 2014, pp. 1891–1898.

[6] Y. Sun, Y. Chen, X. Wang, et al., Deep learning face representation by joint identification verification, in: Proceedings of the Advances in Neural Information Processing Systems, Barcelona, SPAIN, 2014, pp. 1988–1996.

[7] Y. Sun, X. Wang, X. Tang, Deeply learned face representations are sparse, selective, and robust, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, 2015, pp. 2892–2900.

[8] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: a unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, 2015, pp. 815–823.

[9] Y. Taigman, M. Yang, M.A. Ranzato, et al., Deepface: closing the gap to human-level performance in face verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, 2014, pp. 1701–1708.

[10] Y. Wen, K. Zhang, Z. Li, et al., A discriminative feature learning approach for deep face recognition, in: European Conference on Computer Vision, Springer, Cham, Berlin, 2016, pp. 499–515.

[11] W. Liu, Y. Wen, Z. Yu, et al., Large-margin softmax loss for convolutional neural networks, in: Proceedings of the International Conference on Machine Learning, 2, ICML, Stockholm., 2016, p. 7.

[12] F. Wang, J. Cheng, W. Liu, et al., Additive margin softmax for face verification, in: IEEE Signal Processing Letters, 25, PISCATAWAY, USA, 2018, pp. 926–930.

[13] H. Wang, Y. Wang, Z. Zhou, et al., CosFace: large margin cosine loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, 2018, pp. 5265–5274.

[14] S. Chen, Y. Liu, X. Gao, et al., MobileFaceNets: efficient CNNs for accurate real–time face verification on mobile devices, in: Chinese Conference on Biometric Recognition, Springer, Cham, Berlin, 2018, pp. 428–438.

[15] M. Sandler, A. Howard, M. Zhu, et al., MobileNetV2: inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, 2018, pp. 4510–4520.

[16] R. Girshick, J. Donahue, T. Darrell, et al., R-CNN for object detection, in: Proceedings of the IEEE Conference, 2014.

[17] Zhu G., Porikli F., Li H. Tracking randomly moving objects on edge box proposals. arXiv preprint arXiv: 1507.08085, 2015.

[18] J.R.R. Uijlings, K.E.A. Van De Sande, T. Gevers, et al., Selective search for object recognition, Int. J. Comput. Vis. 104 (2) (2013) 154–171.

[19] C. Lee, J.H. Kim, K.W. Oh, Comparison of faster R-CNN models for object detection[C]//2016 16th international conference on control, in: Automation and systems (iccas), IEEE, 2016, pp. 107–110.

[20] S. Ren, K. He, R. Girshick, et al., Faster R-CNN: towards real-time object detection with region proposal networks, IEEE transactions on pattern analysis and machine intelligence 39 (6) (2016) 1137–1149.

[21] X.X. Zhang, X. Zhu, Moving vehicle detection in aerial infrared image sequences via fast image registration and improved YOLOv3 network, Int. J. Remote Sens. 41 (11) (2020) 4312–4335.

[22] Nguyen H.T. Contributions to facial feature extraction for face recognition. université de grenoble, 2014.

[23] Z. Tang, G. Zhao, T. Ouyang, Two-phase deep learning model for short-term wind direction forecasting, Renew. Energy 173 (2021) 1005–1016, doi:10.1016/j.renene.2021.04.041.

[24] C. Lunerti, R.M. Guest, R. Blanco-Gonzalo, et al., Environmental effects on face recognition in smartphones, in: Proceedings of the 51st IEEE International Carnahan Conference on Security Technology, IEEE, 2017.

[25] Z.H. Tang, Y.Y. Li, X.Y. Chai, H.Y. Zhang, S.X. Cao, Adaptive nonlinear model predictive control of NOx emissions under load constraints in power plant boilers, J. Chem. Eng. Jpn. 53 (1) (2020) 36–44, doi:10.1252/jcej.19we142.

[26] L.C. Yan, B. Yoshua, H. Geoffrey, Deep learning, Nature 521 (7553) (2015) 436–444.

[27] Simonyan K., Zisserman A., Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[28] Knyazev B., Shvetsov R., Efremova N., et al. Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. arXiv preprint arXiv:1711.04598, 2017.

[29] Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 4694-4702.

[30] Gabrielsson S. A moral endeavour in a demoralizing context: Psychiatric inpatient care from the perspective of professional caregivers[D]. Luleå tekniska universitet, 2015.

[31] Nazir D, Hashmi KA, Pagani A, et al. Title. Preprints 2021, 1, 0[C]//Conference on Document Analysis and Recognition (ICDAR). IEEE. s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affil-iations. 1 Department of Computer Science, Technical University of Kaiserslautern, 67663 Kaiserslautern, Germany; dn8034@gmail. com (DN); khurram_azeem. hashmi@ dfki. de (KAH); muhammad_zeshan. afzal@ dfki. de (MZA); alain. pagani@ dfki. de (AP); didier. stricker@ dfki. de (DS); 2 Mindgarage, Technical University of Kaiserslautern, 67663 Kaiserslautern, Germany 3 German Research Institute for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany, 4 Department of Computer Science, Luleå University of Technology, 971 87 Luleå, Sweden; marcus. liwicki@ ltu. se (ML);, 2017, 1: 1417-1422.

[32] W. Chen, M. Fang, Y.H. Liu, et al., in: Monocular semantic SLAM in dynamic street scene based on multiple object tracking[C]//2017, IEEE, 2017, pp. 599–604.

[33] Xu X., Dou P., Le H.A., et al. When 3D-Aided 2D Face Recognition Meets Deep Learning: An extended UR2D for Pose-Invariant Face Recognition[J]. arXiv preprint arXiv:1709.06532, 2017.

[34] J. Wang, J. Ding, H. Guo, et al., Mask OBB: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images[J], Remote Sensing 11 (24) (2019) 2930.

[35] Felsberg M., Five years after the Deep Learning revolution of computer vision: State of the art methods for online image and video analysis[J]. 2017.

[36] Rezaei M, Azarmi M. Deepsocial: Social distancing monitoring and infection risk assessment in covid-19 pandemic[J]. Applied Sciences, 2020, 10(21): 7514.

[37] S. Paul, L. Singh, A review on advances in deep learning[C]//2015 IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions (WCI), IEEE (2015) 1–6.

[38] H Ge, Y Dai, Z Zhu, et al., A Robust Face Recognition Algorithm Based on an Improved Generative Confrontation Network[J], Applied Sciences 11 (24) (2021) 11588.