SIN-CHN SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2025, 22(1), 1017.
https://doi.org/10.62617/mcb1017

Article

Biomechanically-informed emotion recognition algorithm of sports athletes based on deep neural network

Weiwei Zhou[1], Zheng Yang[2,*]

[1] Ministry of Sports, Faculty of Disaster Prevention Science and Technology, Langfang 065201, China
[2] College of Humanities, Hebei Oriental University, Langfang 065001, China
*** Corresponding author:** Zheng Yang, yangz202404@163.com

CITATION

Zhou W, Yang Z. Biomechanically-informed emotion recognition algorithm of sports athletes based on deep neural network. Molecular & Cellular Biomechanics. 2025; 22(1): 1017.
https://doi.org/10.62617/mcb1017

ARTICLE INFO

Received: 5 December 2024
Accepted: 23 December 2024
Available online: 6 January 2025

Abstract: The environment of sports competition is changing rapidly, and there is a certain relationship between athletes' decision-making and executive functions and athletes' emotions. Positive emotions can enhance reaction inhibition, while negative emotions will damage the inhibition function. Therefore, identifying the emotions of athletes in sports competitions can help coaches quickly grasp the emotional state of athletes, so as to make targeted decisions. With the advent of the era of big data and the continuous in-depth development of deep neural networks, the emergence of various networks and network models has not only made rational use of a large amount of data, but also promoted the continuous development of emotion recognition. This paper takes the research on the emotion recognition algorithm of sports athletes as the object, uses the faster CNN network to recognize the facial emotion, modifies the backbone network model and loss function parameters in the network, selects the best performing network through comparative experiments, and applies it to the research field of emotion recognition algorithm of sports athletes. While understanding the emotional state of athletes in sports competitions, it lays a solid foundation for the follow-up study of athletes' emotional recognition algorithm in sports competitions. The main research contents of this paper are as follows: firstly, this paper selects data sets with different characteristics, classifies them according to the status of athletes in sports competitions, and labels the data sets. Secondly, Fast Region-based Convolutional Neural Networks (R-CNN) is used to train the labeled data set and obtain the model, and compare the accuracy in different model and loss function parameter conditions. Finally, according to the experimental comparison results, the network with the highest accuracy is selected and applied to the research of athletes' emotion recognition algorithm in sports competitions.

Keywords: deep neural network; emotion recognition; sports athletes; Fast R-CNN network

1. Introduction

In the current era of rapid development of science and technology, as a country, its strength is not only reflected by military, economic and other hard power, but also needs strong enough soft power. Only when hard power and soft power complement each other, can it fully reflect the strength of a country. Sports athletes are the part that can best reflect the soft power of a country. If sports athletes want to achieve good results in the competition, they often cannot do without emotional changes [1]. At the same time, emotion has always been a topic of special attention. From prenatal education to adults, emotional changes deeply represent their physical and mental health. With the continuous progress of artificial intelligence and the increase of its application in the education industry, deep neural network is also gradually emerging, and through the research and continuous improvement of many researchers in this

field, deep learning network has become an extremely popular research. How to use deep neural network to identify the emotions of sports athletes has become a research problem. At the same time, in the current sports competition, the emotion of athletes is also a research problem that psychologists in the sports field pay close attention to athletes' emotions in competition and even life is closely related to the success or failure of the competition. It can be said that emotion has an important impact, so the realization of sports athletes' emotion recognition ability plays an important role in improving athletes' physical health level and competition. The research on sports athletes' emotion recognition and other related knowledge can effectively solve the current physical and mental health problems of sports athletes. With the deepening of the research in the field of emotion, emotion recognition has gradually attracted the attention of scholars, but there is a lack of research on athletes' emotion recognition algorithm in sports field.

Biomechanics, the study of the mechanical laws governing living organisms, plays a crucial role in understanding the physical expressions of emotion in athletes. It provides insights into how emotional states affect muscle tension, movement patterns, and energy expenditure, which are vital for recognizing and analyzing emotions in high-intensity sports environments. By integrating biomechanical data with deep neural network analysis, we can enhance the accuracy of emotion recognition algorithms, leading to a more comprehensive understanding of athletes' emotional states and their impact on performance.

This paper takes the research of athletes' emotion recognition algorithm in sports competition as the object, and uses the model analysis method to study the athletes' emotion recognition algorithm based on deep neural network. The main research contents are as follows:

1) This paper introduces the emotional expression data set used in this paper and the preprocessing operation of the data.
2) For the network used in this paper, its network structure and the loss function used in the network are introduced. Through the comparative experiments using different network models and the parameter values of the loss function, a more accurate network is obtained.

2. Literature review

Emotion recognition has become an increasingly prominent field of study, aiming to capture, interpret, and model the emotional states of individuals for a wide range of applications, including human-computer interaction, educational environments, and personalized services. In recent years, researchers have focused on leveraging advanced optimization strategies, deep neural network architectures, multimodal feature fusion, and cross-domain adaptation to achieve improved emotion recognition accuracy and generalization capabilities, all of which are particularly relevant when designing systems that cater to the dynamic and high-pressure scenarios of sports athletes' training and performance. By examining the most recent developments in this field, it is possible to identify trends, methodologies, and challenges that can guide the development of a Deep Neural Network (DNN)-based emotion recognition framework tailored for the sports domain.

In the literature, face-based emotion recognition has demonstrated substantial progress through the integration of deep learning architectures with enhanced optimization algorithms. For instance, Sumalakshmi and Vasuki [1] introduced an Ameliorate Grasshopper Optimization Algorithm (AGOA) to refine the feature selection process in facial emotion recognition. AGOA, which augments a conventional Grasshopper Optimization Algorithm with opposition-based learning, Lévy flights, and Gaussian mutations, works in tandem with Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) units to achieve remarkable recognition rates (93.90% accuracy) on the YALE face database. This sophisticated interplay of deep feature extraction and biologically inspired optimization underscores the potential of advanced meta-heuristics in improving model performance. Beyond facial expressions, researchers have also expanded their attention to other modalities, such as electroencephalograms (EEGs). Zhang et al. [2], a semi-supervised multi-source domain adaptation algorithm (MGFKD) is employed for cross-subject EEG-based emotion recognition, highlighting the utility of domain adaptation for new subjects and scenarios. This is particularly significant for sports contexts, where athletes may exhibit idiosyncratic emotional responses to training regimens, competition pressure, and fatigue, and where rapid model adaptation is crucial.

Audio-based emotion analysis also complements visual and physiological modalities, Na and Yong [3], where a music recognition and classification algorithm uses neural networks and gradient descent to refine audio emotion feature extraction. Achieving an accuracy of around 85%, this demonstrates the feasibility of emotion recognition based solely on audio signals. However, not all proposed algorithms in the literature maintain their scientific rigor and trustworthiness, as evident from [4], which was retracted due to compromised peer review processes. Such incidents serve as a cautionary tale, reminding the community that stringent validation and verification of experimental results are essential, especially when adapting these findings for high-stakes domains such as sports, where decision-making might impact training outcomes or athlete well-being.

Recent studies have begun to explore the integration of biomechanical signals, such as electromyography (EMG) and motion capture data, with traditional emotion recognition methods. These physiological markers offer a direct measurement of the body's response to emotional stimuli, complementing the analysis of facial expressions and providing a more nuanced understanding of athletes' emotional states. The combination of biomechanical and neural network analysis has the potential to revolutionize the field of sports psychology by offering objective, real-time data on athlete emotional states, which can inform training strategies and competitive tactics.

Beyond unimodal approaches, multimodal fusion techniques have emerged as a vital strategy to enhance the robustness and accuracy of emotion recognition systems. For instance, Xu [5] presents an improved CNN-based algorithm in online educational environments, constructing decision trees for effective classification and proposing faster model training and execution times. Although the focus is on online education, the underlying methodology—augmenting CNN performance and employing decision-level management—can translate effectively to the sports domain, where scalability and real-time inference are critical. Similarly, Cheng et al. [6] proposed a

weight-adaptive multimodal fusion approach that integrates audio and video inputs. Leveraging "time-distributed CNNs + LSTMs" for audio and "DeepID V3 + Xception" for visual data, this method outperforms single-modality baselines by nearly 4%. Such multimodal strategies could offer superior accuracy in sports contexts, especially when integrating facial expressions, vocal cues from the athlete's breathing or exertion sounds, and supplementary physiological signals (like heart rate or EEG) that reflect the athlete's stress level or mental state.

Physiological signals have gained increasing traction due to their intrinsic link to emotional states and reduced susceptibility to deceptive external expressions. Studies like Zhang et al. [7] introduce a Physiological signal-based, Mean-threshold, and Decision-level fusion algorithm (PMD) to classify emotional states. By selecting key EEG and peripheral physiological features and employing mean-threshold methods, ensemble learning models surpass traditional classifiers (such as Gaussian Naive Bayes, Linear Regression, and Support Vector Machines) in terms of accuracy [8,9]. These insights are highly relevant to the sports domain, where an athlete's physiological responses to training load, fatigue, and psychological stress are often more reliable indicators of their internal state than facial expressions or vocal cues alone. By fusing multiple data sources—EEG, heart rate, galvanic skin response—into a unified DNN-based emotion recognition architecture, it may be possible to build a more stable, generalizable, and context-aware model.

3. Emotional data processing of athletes in sports competitions based on python

3.1. Emotional expression data set

The data set used in this paper is mainly Real-world Affective Faces Database (RAF-DB) data set, and some images in ck+ data set and NVIE data set are selected to improve the practicability of the data. A total of 15,000 athlete images are selected in this data set for training and verification. The categorization of seven emotions into three groups (positive, negative, and neutral) is supported by research in sports psychology. Positive emotions such as happiness and excitement have been linked to improved reaction times and decision-making abilities in competitive environments. In contrast, negative emotions like anger, disgust, fear, and sadness are associated with decreased cognitive efficiency and impaired inhibitory control, which can adversely affect performance. Neutral emotions, which represent a baseline state, allow athletes to maintain focus and consistency. This grouping aligns with studies suggesting that emotional states influence athletic outcomes and provide a practical framework for analyzing emotions in sports settings.

The data preprocessing stage involved several key steps to ensure the robustness and diversity of the training dataset. For data augmentation, brightness and contrast adjustments were applied to account for variations in environmental lighting. Brightness was randomly adjusted within a range of ± 20% of the original image value, while contrast was varied by a factor between 0.8 and 1.2 to simulate real-world conditions. Additionally, random horizontal flipping and slight rotation (± 10 degrees) were used to enhance the dataset diversity and generalization capabilities of the model.

The image selection criteria from the three datasets (DAF-DB, CK+, NVIE) were as follows (**Table 1**):

1) DAF-DB: Images were chosen based on their clear labeling of seven emotions and diverse imaging scenes, including both natural and artificial settings, to increase variability.
2) CK+: Dynamic expressions were used to capture the progression of emotions, but only those frames with clear peak emotional expressions were included.
3) NVIE: To ensure uniformity in quality, only high-resolution images with consistent annotations were selected, excluding those with heavy occlusions or poor lighting.

Table 1. summarizes the selection criteria and distribution of images from each dataset.

Dataset	Selection Criteria	Number of Images Used
DAF-DB	Clear labeling; diverse scenes	10,000
CK+	Peak dynamic expressions only	3000
NVIE	High resolution; no occlusions	2000

3.2. Research on emotion recognition algorithm based on deep neural network

For researchers of deep neural networks, the role of fully connected neural networks is beyond doubt. The traditional neural network consists of three parts [10]. The first is the input layer, which has only one and is mainly responsible for input data; Then there is the hidden layer. The neural network with more than 2 hidden layers in the neural network is called the deep neural network. For the hidden layer, the increase of the number of layers will increase the depth and complexity of the network. The number of nodes in the hidden layer can be adjusted. If you want to make the performance of the neural network more powerful, you need to set more nodes in the hidden layer, but the number of parameters will also increase, consuming a lot of memory; Finally, the output layer, whose number of neurons depends on the task.

Due to the advantages of convolutional neural network, the number of parameters produced in the training process is less than that of neural network, so the efficiency and model accuracy are also higher. Therefore, compared with other neural networks, convolutional neural network has more research in the field of image. **Figure 1** shows the structure of convolutional neural network. With the continuous development of convolutional neural network, LeNet, AlexNet, Visual Geometry Group (VGG) and other network models have been proposed by researchers. These network models make the model more accurate by deepening the network structure, especially the ResNet model, which greatly improves the accuracy of the model by introducing residual blocks [2].

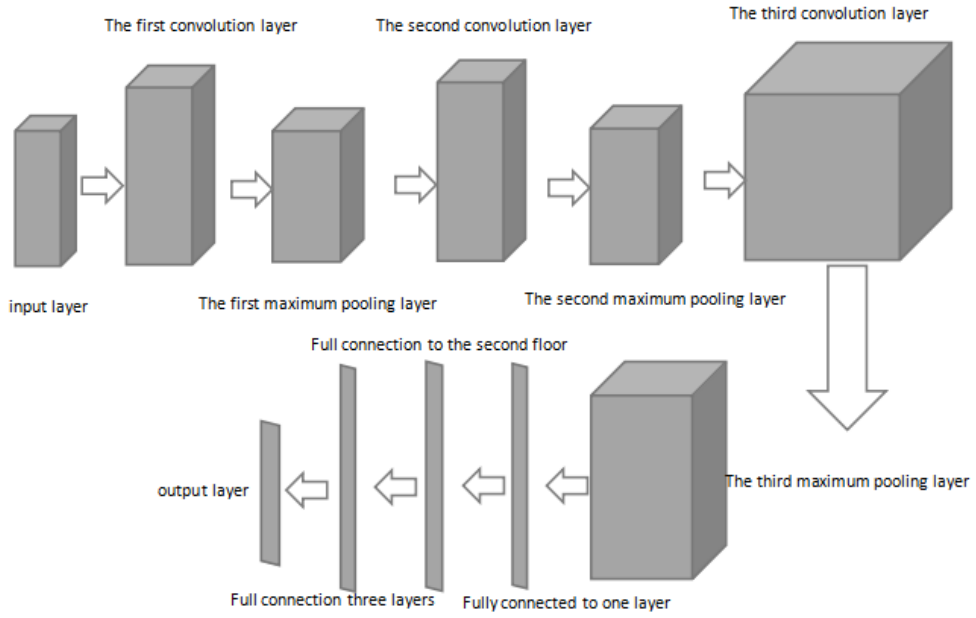

Figure 1. Structure diagram of convolutional neural network.

According to the network model mentioned above and the Fast R-CNN network used in this paper, this section will introduce VGG model, ResNet model and SPPNet model [11].

3.2.1. Model

VGG model was proposed by Simonyan and Zisserman in 2014, and its name is also derived from the initials of the group. At the same time, as the most popular VGG model, VGG16 model has been deeply loved by the majority of researchers since it was proposed. VGG16 is the network model used to extract features in Fast R-CNN network. In the 2014 competition, I got excellent results. Although I didn't get the first place, I performed better in the transfer learning task than GoogLeNet, which was at the first place at that time [12]. VGG16 network is divided into three layers. At the same time, its simple network structure also deeply attracts relevant researchers. The first layer is a thirteen layer convolution layer, the second layer is a three-layer full connection layer, in which the discarding method is used to reduce the number of parameters, and the third layer is a classification layer composed of softmax, in which the target is classified. In VGG16 model, all relu functions are used as activation functions. Usually, among many models, VGG model is the first one to extract features from images. In terms of model architecture, VGG model basically follows AlexNet's model architecture, but it is deeper than AlexNet in terms of layers, and neural network integration is also used in VGG model, so that the accuracy of VGG model can be higher and the object recognition is more accurate. **Table 2** shows the structure of VGG16 model.

Table 2. Structure of VGG16 model.

structure	size of input	structure	size of input
Input (channel = 3)	224 × 224 × 3	Conv3-256	56 × 56 × 256
Conv3-64	224 × 224 × 64	Conv3-256	56 × 56 × 256
Conv3-64	224 × 224 × 64	Maxpool	28 × 28 × 256
Maxpool	112 × 112 × 64	Conv3-512	28 × 28 × 512
Conv3-128	112 × 112 × 128	Conv3-512	28 × 28 × 512
Conv3-128	112 × 112 × 128	Conv3-512	28 × 28 × 512
Maxpool	56 × 56 × 128	Maxpool	14 × 14 × 512
Conv3-256	56 × 56 × 256	Conv3-512	14 × 14 × 512
Conv3-512	14 × 14 × 512	Conv3-512	14 × 14 × 512
Maxpool	7 × 7 × 512	FC1	1 × 1 × 4096
FC2	1 × 1 × 4096	Softm ax	1000

It can be seen from **Table 2** that VGG networks all use 3 × 3 convolution cores and 2 × 2 pooled windows to extract features. At the same time, VGG model suppresses over fitting by adding dropout in full connection. Compared with AlexNet model, VGG model uses smaller convolution kernel. This method can not only effectively reduce the number of parameters, but also make training, verification and testing more efficient. At the same time, its convolution kernel size is relatively small, which can deepen the network. This method is extremely beneficial for image classification tasks. The characteristics of VGG model, such as relatively simple structure, high efficiency and strong application, have attracted a large number of researchers. At the same time, the success of VGG model has also proved that the efficiency of learning image features can be improved by increasing the network depth, and the network structure design method of VGG model has played a great role in promoting the development of deep neural network. However, VGG16 still has some shortcomings, such as: The training time of VGG16 is too long, and it is difficult to adjust parameters and requires a large amount of storage.

3.2.2. ResNet model

ResNet (residual network) was proposed by four Chinese in Microsoft Research Institute. By introducing residual blocks, we can deepen the network structure and make the features extracted by the network more effective. In the 2015 IL, svrc competition, ResNet won the first place with an error rate of 3.57%, showing an accuracy beyond the recognition of normal human eyes [13]. Compared with VGG network, the number of parameters in ResNet model is less, so the recognition rate is higher. It can be found from the model introduced before that in the development process of deep learning, the number of layers of the model and the depth of its network structure are increasing. Therefore, it can be assumed that by deepening the structure of the network, that is, increasing the number of layers of the network model, the effect shown by the model must be better than that of the relatively shallow network. But practice shows us again that the training error will also increase with the increase of the model structure. In response to this phenomenon, He et al. Proposed a solution, that is, to use ResNet model to solve this problem. ResNet network adopts

the idea of connecting layers, adds direct channels in the network layer, maps the data to the subsequent network layer, and adds it to the calculated results. Through this jump connection, the gradient disappearance problem with the increase of the number of layers can be moderately alleviated [14]. **Figure 2** shows the basic structure of the residual network.

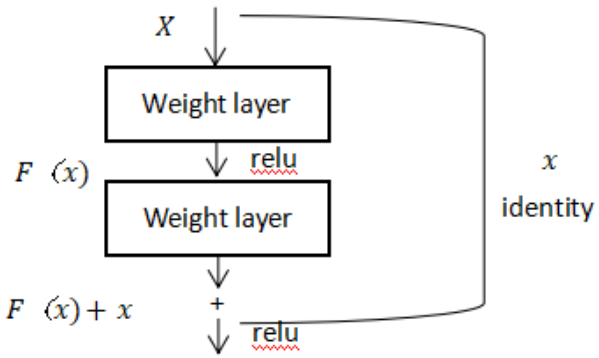

Figure 2. ResNet.

The structure shown in **Figure 2** is also called residual block, and the model proponent proposed ResNet with network layers of 34, 50, 101 and 152 through experiments. The proposed structure did not cause the effect of adding network layers to decline, on the contrary, the error rate was greatly reduced. The residual structure formula is as follows:

$$y = F(x) + x \tag{1}$$

In Equation (1), $F(x)$ is the output after convolution operation, and the second is the original input. **Figure 3** shows the specific setting of the residual block. Among them, one adopts two-layer 3 × 3 convolution networks connected in series, and the other adopts 1 × 1, 3 × 3 and 1 × 1 three-layer convolution networks.

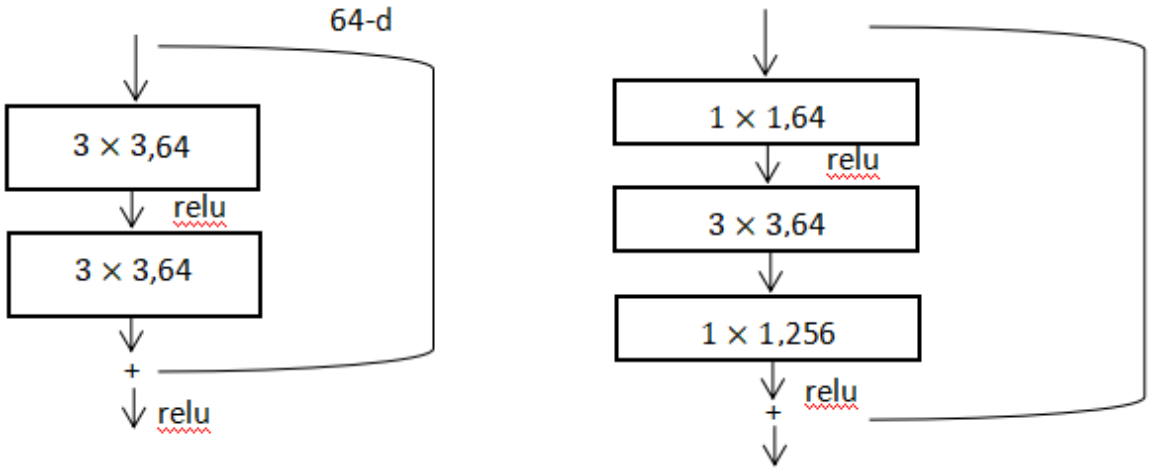

Figure 3. Specific design scheme of residual block.

Table 3 shows the network structure of ResNet-18.

Table 3. ResNet-18 network structure.

The ResNet-18 structure of the	**big or small**
Input	224 × 224 × 3
Conv	64 × 112 × 112
Maxpooling	64 × 56 × 56
	64 × 56 × 56
Conv-x	64 × 56 × 56
	64 × 56 × 56
	64 × 56 × 56
Conv-x	128 × 28 × 28
	128 × 28 × 28
	128 × 28 × 28
	128 × 28 × 28
Conv-x	256 × 14 × 14
	256 × 14 × 14
	256 × 14 × 14
	256 × 14 × 14
Conv-x	512 × 7 × 7
	512 × 7 × 7
	512 × 7 × 7
	512 × 7 × 7
AveragePooling	512 × 7 × 7
	512 × 7 × 7
FC	1000

3.2.3. SPPNet model

Because all neural networks have strict requirements for the size of the input image, some operations are needed to process the image to make the size of the input data consistent, which greatly reduces the efficiency of related tasks [15]. To solve this problem, the Kaiming proposed SPPNet model in 2014. The introduction of this model can not only make the network ignore the problem of image size proportion in training, but also improve the accuracy of network recognition [16].

Compared with the previous network model, SPPNet model makes full use of spatial relations to extract the features of the input image and fix the size of the image. The main function of SPPNet is embodied in the SPP layer, that is, after the convolution of the last layer of the network, the features are further extracted through pooling operation, and the size of the output feature map is fixed [2]. The feature map generated in this way is consistent in size and convenient for transmission to the full connection layer. **Figure 4** shows the core architecture of SSPNet network model.

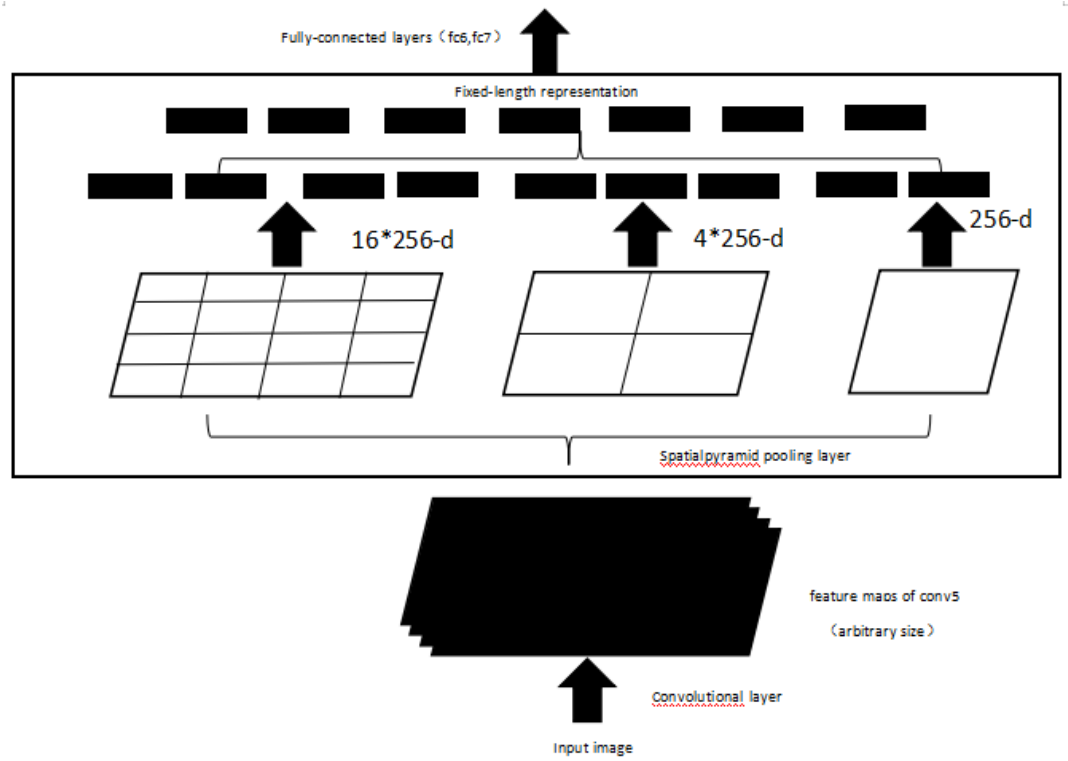

Figure 4. Core architecture of SPPNet network model [2].

As can be seen from **Figure 4**, the main idea of SPPNet model is to divide the convoluted feature map (of any size) into 16, 4 and 1 feature maps, pool these three feature maps at the same time, and splice the pooled features into fixed dimension output [2]. For the convolution layer and pooling layer in the network, the input size will not affect the calculation, but the parameters in the full connection layer have a great relationship with the size of the data [17]. SPPNet network perfectly solves the problem of the correspondence between parameters and data in the full connection layer by fixing the size of the pooled data. It can be seen in the article that the speed of SSPNet is about 100 times higher than that of R-CNN, which uses image preprocessing to change the size of the image and uses selective search algorithm to extract features. Compared with the convolution network before SPPNet, SPPNet has excellent robustness, and the network performance and accuracy using SPPNet model have been greatly improved. At the same time, the proposal of SPPNet has also greatly promoted R-CNN, and the idea of SPPNet has been applied to Fast R-CNN and Fast R-CNN after R-CNN.

To adapt the ResNet50 architecture for emotion recognition tasks, several modifications were implemented to enhance its effectiveness for facial emotion datasets. The final fully connected layer was replaced with a dense layer having three outputs corresponding to the emotion categories (positive, negative, neutral) and activated using the softmax function. This change allowed the network to directly predict the probability of each emotional state. Additionally, dropout layers with a rate of 0.5 were added before the final dense layer to mitigate overfitting, especially important given the relatively small size of the training dataset. The model also leveraged advanced data augmentation techniques, including random rotations, horizontal flips, and slight variations in brightness and contrast. These transformations

enriched the training dataset and helped the model generalize better to unseen data. Furthermore, the learning rate schedule was carefully designed, starting with an initial learning rate of 0.005, which was reduced by a factor of 10 every 20 epochs. This approach facilitated rapid convergence in the early stages while avoiding overfitting during later training phases. The choice of a batch size of 16 balanced computational efficiency with stable gradient updates. The momentum optimizer, with a decay rate of 0.9 and an L2 regularization weight of 0.001, was employed to stabilize training and reduce the risk of overfitting. These hyperparameter choices were validated through extensive cross-validation experiments, demonstrating their effectiveness in achieving optimal performance on the emotion recognition task.

3.3. Biomechanical data integration

In addition to facial expression analysis, this study also incorporates biomechanical data to enhance the emotion recognition algorithm. EMG sensors and motion tracking systems were used to collect data on muscle activity and movement patterns during training and competition. These data were then processed and integrated with the deep neural network model to identify distinct biomechanical signatures associated with various emotional states. The integration of biomechanical data provides a more holistic view of the athlete's emotional and physiological response, improving the algorithm's ability to recognize and classify emotions accurately.

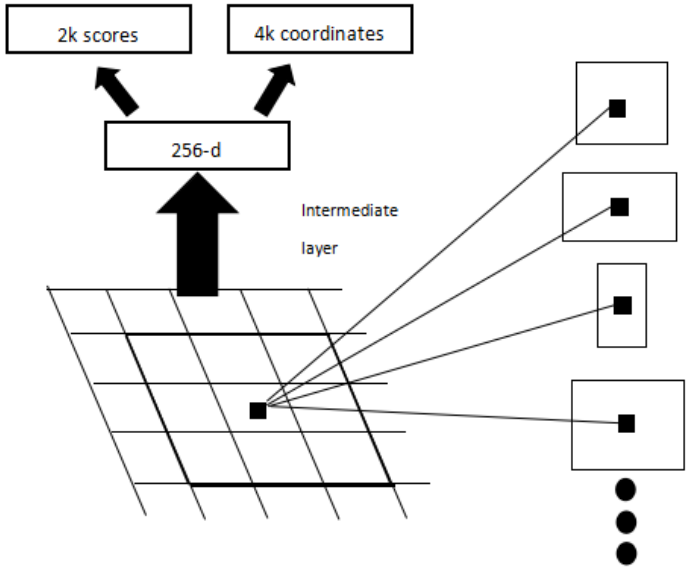

Figure 5. RPN architecture diagram.

Fast R-CNN proposes to use Region Proposal Network (RPN) to extract candidate frames and offset the position of foreground frames. The main steps of RPN network are as follows: The first step is to set the size and proportion of each pixel (also known as anchor) in the feature map obtained from the backbone network to form an anchor box (generally, 9 anchor boxes will be generated for each pixel); The second step is to input each anchor box into two network layers, one of which is used for image classification, that is, to judge whether the feature map contained in the anchor box belongs to the foreground, and the other network layer is used for position

regression, that is, to judge the offset of the anchor box from the real box [2]; Finally, summarize the calculation results of the two network layers, filter the anchor box through Intersection over Union (IOU) and non maximum suppression, and enter the filtered results into the next layer. Therefore, the essence of RPN can also be regarded as "classless target detector based on sliding window". **Figure 5** shows the RPN network structure.

3.3.1. Intersection and union ratio

The emergence of intersection union ratio (IOU) is mainly used as a measure to determine the relationship between anchor box and real box.

Equation (2) is the IOU formula [2].

$$IOU = \frac{A \cap B}{A \cup B} \tag{2}$$

In Equation (2), a and B represent the position coordinate information of anchor frame and real frame respectively. IOU can be seen from Equation (2).

In fact, it is calculating the coincidence degree between the anchor box and the real box. For example, the location coordinate information of an anchor box is (boxxl, boxyl, boxx2, boxy2), the location coordinate of the real box is (TXL, two 1, TX2, two 2, the coordinates of the intersection part of the anchor box and the real box are (max cboxxl, TXL, Max C boxyl, two 1), min (boxx2, TX2), min C boxy2, two 2). Based on these information, the area of the intersection part of the anchor box and the real box can be calculated. Thus, the IOU results of anchor box and real box are obtained.

3.3.2. Non maximum suppression

For the target detection task, it is not only necessary to calculate the relationship between the anchor box and the real box by the intersection and union ratio, but also need to filter the selected anchor box in some way, and select some of them as the prediction box [2]. This method is called non maximum suppression (NMS) by researchers. As the name suggests, non maximum suppression is to use the algorithm to screen the candidate boxes according to the confidence, and select the candidate box with the greatest confidence by calculating the intersection and union ratio of the two. Its main function is to select the target boundary box with the greatest confidence from a large number of overlapping anchor boxes, remove unnecessary anchor boxes, and improve the efficiency of the task. Non maximum suppression mainly uses the circular idea to continuously screen the preselection box, and select the preselection box with the largest local confidence to train it. The main workflow is as follows: The first step is to sort the bounding box according to the confidence level; The second step is to select the bounding box with the highest confidence, take it out of the bounding box and classify it as a candidate box; The third step is to calculate the intersection ratio of the candidate box and the remaining bounding box, and delete the bounding box that is greater than the threshold; Then repeat the second and third steps until there is no boundary box.

3.4. Loss function

In this paper, the network structure of Fast R-CNN is used for research. The

formula of the loss function of Fast R-CNN is shown in Equation (3) [2].

$$L(P_i, t_i) = \frac{1}{N_{cls}} \sum_i L_{cls}(P_i, P_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \quad (3)$$

In Equation (3), i refers to the index of the anchor frame, P_i refers to the predicted classification probability of the frame, t_i refers to the position coordinate information of the bounding box predicted by the anchor frame that is, the predicted coordinate offset, for which it refers to whether the *i* th frame is a positive sample or a negative sample, p_i^* refers to the position coordinate information of the real frame corresponding to the ith frame, that is, the actual coordinate offset, λ is the weight balance parameter, N_{cls} refers to the number of mini batch, and N_{reg} refers to the number of candidate frames [2]. Among them, the positive sample refers to the sample with the largest IOU between the prediction frame and the real frame or the value of IOU is greater than the threshold value, while the negative sample refers to the sample with the value of IOU less than the threshold value. In addition, the sample between the positive sample and the negative sample does not participate in the training. For the Fast R-CNN loss function, it is composed of classification loss and regression loss.

In Fast R-CNN, the classification loss is expressed as $\frac{1}{N_{cls}} \sum_i L_{cls}(P_i, P_i^*)$ in Equation (3). By introducing the cross moisture loss function, namely $L_{\text{cls}}(P_i, P_i^*)$, the cross moisture loss value of each anchor is calculated, and then the results are added to remove the mini batch to obtain the classification loss value. The formula of cross moisture loss function is shown in Equation (4).

$$L_{cls}(P_i, P_i^*) = -log[P_i, P_i^* + (1 - p_i)(1 - p_i^*)] \quad (4)$$

In Fast R-CNN, the regression loss is expressed as $\lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*)$ in Equation (3), where $L_{reg}(t_i, t_i^*)$ is also expressed as $R(t_i, t_i^*)$, R is the smooth $L1$ loss function, and Equation (5) is the smooth $L1$ loss function. For regression loss, first calculate the smooth $L1$ loss in the positive sample, and then add the results and divide by the number of candidate boxes to obtain the regression loss value.

$$Smooth_{L1}(x) = \begin{cases} 0.5^* x^2 & |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \quad (5)$$

4. Analysis of experimental results

Aiming at the VGG and ResNet models proposed above, this paper tests the performance of VGG16 and ResNet50 models as the backbone network model of Faster R-CNN network respectively; Then modify the parameters in the Faster R-CNN loss function and compare the performance of different parameters.

4.1. Experimental preparation

4.1.1. Experimental environment

In this paper, the windows 10 operating system is used, in which the CPU is i7-9700k, the running memory is 16 g, and the GPU is NVIDIA GeForce 20705. The language is python, of which the version of Python is 3.7 and the development

environment is pycharm2019. The third-party libraries used in the experiment are as follows:

Numpy: As an open source code, it is developed by numeric, the predecessor of numpy, in combination with numarray and other extension libraries. Is one of the many extension libraries of the python language. Numpy contains many mathematical function operations, which are not only for constants, but also high-dimensional arrays, matrices, etc.

OpenCv: As a computer vision library supporting a variety of operating systems, it was officially released at the end of 2010. At the same time, OpenCV not only supports python, but also has interfaces for C, C++, PHP and other languages. As a lightweight third-party library, its main role is in real-time visual applications, and it can complete most tasks of image and video operations through OpenCV.

4.1.2. Data preparation and processing

In this paper, the public data set is used, and LabelImg software is used to mark the data set. According to the emotions of athletes in the picture, the previous seven basic emotions are divided into three categories. Labelimg is a visual image annotation tool, which is simple and convenient to operate, and can directly generate the dataset text used by Fast R-CNN. Then, the image is processed according to the image processing method mentioned in Chapter 3. It includes random changes in brightness, contrast, etc. of the image to enhance the effect of the model results. Because Fast R-CNN doesn't have high requirements for image size, it doesn't deal with the image size specifically during preprocessing. **Table 4** describes the changes of six emotional characteristics based on neutral expression.

Table 4. Emotional shape analysis.

Mood	Characteristic
Take offence	Eyebrows wrinkled together, eyelids drawn tight and mouth closed
Detest	Low eyebrows, eyes around the cheek lift, mouth slightly open
Fear	Eyebrows wrinkled together, upper eyelids raised, mouth open, mouth pulled back
Happy	Eyebrows bend slightly down, the end of the eyes may appear wrinkled, and the mouth is open
Sad	The inside of the eyebrows wrinkled together, the eyes tail slightly lower, and the mouth pulls wide or closed
Surprised	Eyebrows raised and curved, raised lines, eyes enlarged, mouth open

4.2. Comparative experiment between VGG16 model and ResNet50 model

The experiment of this paper is to compare different models. VGG16 model and ResNet50 model are selected in this paper. The learning rate in the experiment is set to 0.0001, and the optimizer adopts momentum optimization method (momentum, regularization method is set to l2decay, kinetic energy attenuation is set to 0.9, and batch size is set to 16 for training. The experimental results are shown in **Table 5** [2]:

Table 5. Model comparison results.

Model	Accuracy rate (%)
VGG16	81.6
ResNet50	89.3

The experimental comparison results show that for the Fast R-CNN network, based on the same conditions, the accuracy of ResNet50 is 7.7% higher than that of VGG16.

While the experimental results demonstrate the superiority of the ResNet50 model compared to VGG16, it is essential to situate this work within the broader landscape of state-of-the-art emotion recognition methods. Recent advancements in convolutional neural networks (CNNs) and hybrid approaches have achieved notable success in emotion recognition tasks. For instance, the Capsule Network (CapsNet) has shown promise in capturing spatial hierarchies within data, which are often missed by traditional CNNs.

Moreover, Mini-Xception, a lightweight CNN model, has been developed to improve real-time emotion recognition efficiency. Despite its reduced complexity, Mini-Xception achieves competitive accuracy by combining residual modules with separable convolutional structures, allowing it to balance speed and accuracy effectively. Compared to these methods, the modified ResNet50 in this study demonstrated superior accuracy on the facial emotion dataset, suggesting its suitability for scenarios where accuracy is prioritized. However, further investigation could explore integrating features from EEG signals or combining spatial and temporal features to compete with hybrid methods like those using CapsNet.

To validate the stability of the ResNet50 model, five-fold cross-validation was conducted using the combined dataset. The mean accuracy across folds was 89.1%, with a standard deviation of 0.45%, demonstrating consistent performance and robustness of the model.

In addition to accuracy, inference time was also evaluated to compare the practical applicability of ResNet50 and VGG16 models in real-time scenarios. On the test set, the average inference time per image was 25 ms for ResNet50 and 45 ms for VGG16 when tested on an NVIDIA GeForce RTX 2070 GPU. The significant reduction in inference time makes ResNet50 more suitable for real-time applications, despite its higher computational complexity. **Table 6** presents the detailed comparison of performance metrics.

Table 6. Comparison of performance metrics.

Metric	VGG16	ResNet50
Accuracy (%)	81.6	89.3
Cross-validation mean (%)	81.2	89.1
Cross-validation SD (%)	0.53	0.45
Inference Time (ms)	45	25

4.3. Comparative experiment based on Fast R-CNN loss function

This experiment is based on the ResNet50 model, which performs well in the previous section. The experiment is carried out for the Fast R-CNN network loss function proposed above, and the performance of parameters under different conditions of 1, 5, 10 and 15 is compared. Since the learning rate setting in Experiment 1 is not conducive to the training rate, this experiment updates the learning rate. In the first 20 rounds, the learning rate is set to 0.005, and then the learning rate is still 0.0001. The experimental results are shown in **Table 7**:

Table 7. Comparison results of parameters in loss function.

λ accuracy rate	**accuracy rate (%)**
1	89.7
5	90.3
10	91.2
15	90.8

It can be seen from the experimental results that the performance is the best when the λ parameter value is 10.

In this chapter, firstly, the network structure of convolutional neural network and three models VGG, ResNet, and SPPNet are introduced. VGGmodel is mainly the network model used by traditional Faster R-CNN, and ResNet is mainly used for comparative experiments. Then it introduces the loss function of RPN network and Faster R-CNN network, in which RPN network is mainly used to extract features in Faster R-CNN network. Finally, a comparative experiment is carried out on the two models proposed in the first section and the loss function introduced in the fourth section. Through the experiment, it is found that when ResNet50 is used as the backbone network and the parameter value in the loss function is 10, the accuracy of the model is the best.

5. Conclusion

With the advent of the era of big data, deep neural network has made rapid progress, and its related applications are also increasing. At the same time, deep neural network technology is also deepening under the research of scholars, and its applications in voice, image, audio and other fields are also emerging in endlessly. Among them, there are many applications in the field of image, such as face recognition, emotion recognition and so on. For target detection task, although there are many researches on emotional expression, there are few applications in emotional recognition. Based on Faster R-CNN, one of the current mainstream networks, this paper studies the emotion recognition algorithm of sports athletes based on deep neural network. Through in-depth analysis of the research status, the research work is summarized as follows: 1) Through the selection of data sets with different characteristics, and its data annotation work; 2) introduce the relevant technical knowledge of this paper from point to surface, and select Faster R-CNN as the main research network of this paper by introducing the relevant knowledge. According to the prototype of Faster R-CNN, ResNet50 network model is used to replace the

original VGG16 backbone network model in Faster R-CNN, and comparative experiments are carried out. Compared with VGG16 network model, ResNet50 has a higher accuracy. At the same time, modify the parameter value of the network loss function for experimental comparison. It is concluded that when the parameter value is 10, the effect is the best, which provides a new method for the emotional recognition algorithm of sports athletes.

The application of the algorithm in the experiment demonstrates its effectiveness in accurately recognizing the emotions of sports athletes. This capability holds significant value for practical applications, particularly for coaches and sports psychologists. By integrating the system into wearable devices or real-time monitoring software, it becomes possible to assess athletes' emotional states during competitions. This real-time feedback allows coaches to make immediate adjustments to strategies or interventions, thereby optimizing performance. For instance, when heightened levels of negative emotions, such as anxiety or frustration, are detected, coaches can implement calming strategies or guide athletes to refocus their attention. Conversely, recognizing positive emotional states can support decisions to maintain or amplify successful approaches.

Despite its potential, the system faces certain limitations that must be addressed to enhance its practical usability. Real-time monitoring during competitions may be influenced by environmental variables, such as poor lighting, occlusions, or rapid head movements, which could compromise the accuracy of facial emotion recognition. Additionally, as the system predominantly relies on facial data, it may not fully capture internal emotional states that are not outwardly expressed. Future advancements could incorporate physiological data, such as heart rate or EEG signals, to provide a more comprehensive assessment of athletes' emotional states.

The integration of biomechanical data with deep neural network analysis presents a promising avenue for future research in emotion recognition for sports athletes. By leveraging the physiological responses measured through EMG and motion analysis, researchers can develop more accurate and responsive emotion recognition systems. These advancements could lead to personalized training regimens that adapt in real-time to an athlete's emotional state, optimizing performance and managing mental health. Future work in this area should focus on refining the collection and analysis of biomechanical data, as well as exploring additional physiological markers that could further enhance the emotion recognition algorithm.

Author contributions: Conceptualization, WZ; methodology, WZ; software, ZY; validation, WZ; formal analysis, WZ; investigation, ZY; resources, ZY; writing—original draft preparation, WZ; writing—review and editing, ZY; visualization, WZ; supervision, WZ. All authors have read and agreed to the published version of the manuscript.

Ethical approval: Not applicable.

Conflict of interest: The authors declare no conflict of interest.

References

1. Sumalakshmi CH, Vasuki P. Ameliorate grasshopper optimization algorithm based long short term memory classification for

face emotion recognition system. Multimedia Tools and Applications. 2024; (13): 83.
2. Zhang R, Guo H, Xu Z, et al. Mgfkd: A semi-supervised multi-source domain adaptation algorithm for cross-subject eeg emotion recognition. Brain Research Bulletin, 2024; 208.
3. Na W, Yong F. Music recognition and classification algorithm considering audio emotion. Scientific programming, 2022; 2, 3138851.1-3138851.10.
4. Weng Y, Lin F. multimodal emotion recognition algorithm for artificial intelligence information system. Wireless Communications & Mobile Computing, 2023.
5. Xu Z. Application of convolution neural network algorithm in online education emotion recognition. International journal of web-based learning and teaching technologies, 2023; 18(Pt.2), 579–591.
6. Cheng Y, Zhou D, Wang S, Wen L. Emotion-recognition algorithm based on weight-adaptive thought of audio and video. Electronics 2023; (2079-9292), 12(11).
7. Zhang Q, Zhang H, Zhou K, Zhang L. Developing a physiological signal-based, mean threshold and decision-level fusion algorithm (pmd) for emotion recognition. Journal of Tsinghua University: Natural Science Edition (English Edition), 2023; 28(4), 673–685.
8. Lu H, King S, Watts O. Combining a Vector Space Representation of Linguistic Context with a Deep Neural Network for Text-To-Speech Synthesis, 2013, 5(9):261–265.
9. Watts O, Wu Z, King S. Sentence-level control vectors for deep neural network speech synthesis. 2015, 351(1): 2217–2221.
10. Ma T, Wang F, Cheng J, et al. A Hybrid Spectral Clustering and Deep Neural Network Ensemble Algorithm for Intrusion Detection in Sensor Networks. Sensors, 2016, 16(10):1701.
11. Kong Q, Sobieraj I, Wang W, et al. Deep Neural Network Baseline for DCASE Challenge. 2018, 291(21):2581–2590.
12. Su H, Chen H. Experiments on Parallel Training of Deep Neural Network using Model Averaging. Computer Science, 2020, 5(3):86–87.
13. Bai Y, Yang K, Wei Y, et al. Automatic Image Dataset Construction from Click-through Logs Using Deep Neural Network. Acm International Conference on Multimedia. ACM, 2021, 295(9):1023–1032.
14. Lakis N, Jiménez JA, Mancini-Mar EA, et al. Neural correlates of emotional recognition memory in schizophrenia: Effects of valence and arousal. Psychiatry Research Neuroimaging, 2011, 194(3):245–256.
15. Punkanen M, Eerola T, Erkkil J. Biased emotional recognition in depression: Perception of emotions in music by depressed patients. Journal of Affective Disorders, 2011, 130(12):118–126.
16. Bollano G, Ettorre D, Esiliato A. Customizable method and system for emotional recognition: 2009, 5(5):14–1.
17. Souto T, Baptista A, Tavares D, et al. Facial emotional recognition in schizophrenia: preliminary results of the virtual reality program for facial emotional recognition. Revista De Psiquiatria Clínica, 2013, 40(4):129–134.