# *Research on Campus Physical Measurement System Based on CNN-LSTM*

Mengzhong Wei,Yuting Fu,Chao Li
Huaiyin Institute of Technology
Computer&Software Engineering
Huai'an China
1455301587@qq.com

Yuanyuan Li,Wanli Feng,Quanyin Zhu*
Huaiyin Institute of Technology
Computer&Software Engineering
Huai'an China
*Corresponding Author's email:hyitzqy@qq.com

***Abstract*— In response to the 14th Five-Year Plan and the 2035 long-term goal outline, the goals of building smart communities and smart industries are clearly defined. Starting from the campus, there are many physical test projects in colleges and universities. Problem, so the campus physical test is transformable and innovative. This article first proposes the idea of building a smart body measurement system; then, analyzes the CNN-LSTM (Convolutional Neural Network-Long short term memor)algorithm that is mainly used in the construction of the system; secondly, it analyzes the accuracy of the recognition of specific action processes and the debugging process of the actual parameters. Carry out optimization and improvement; finally, make a report on the enforceability of the campus physical test system.**

***Keywords CNN-LSTM; human pose estimation; smart application; Opencv***

## INTRODUCTION

Since the 14th Five-Year Plan and the 2035 long-term goal outline [1] was released, research on the construction of smart campuses and smart campuses have been increasing. In the traditional sense, smart campus projects such as college information file management platform, subject competition organization management platform, colleges and universities. The construction of public platforms, etc. focuses on the processing of campus information data in the context of colleges and universities. With the rapid development of the Internet of Things and artificial intelligence, under the initiative of schools and enterprises to combine production and education, many smart classrooms, smart restaurants, etc., integrate smart campus teaching equipment or The terminal operates in colleges and universities, and has obvious effects in assisting campus teaching and improving the quality of campus life. The construction of the campus physical measurement system proposed in this paper focuses on the construction of the human skeleton feature model and the accuracy of human action recognition. Analyze the mode of the action through the student's physical test video, refer to the actual physical test score, record the student's performance data, and finally get the physical test result.

## RELATED WORK

### *A. CNN-LSTM algorithm basis*

Feedforward Neural Network, also known as Multilayer Perceptrons, is the most classic and basic mode in deep learning [2]. The purpose of this algorithm is to find an optimal approximation function f, so as to obtain a mapping from input x to output y: y = f(x; θ), where θ represents the parameters of the neural network. Figure 1 is a schematic diagram of a typical one-layer fully connected forward neural network:

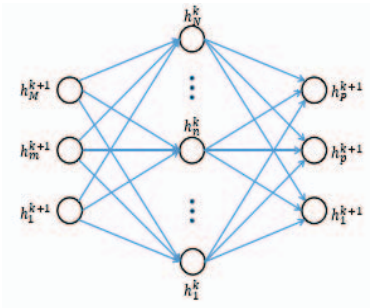

Figure 1. Model of neural network.

In the forward neural network, the information flow starts from the input X, passes through the intermediate f calculation, and outputs y. There is no feedback connection in this process, and there is no output of a certain layer as the input of the next layer. The forward neural network generally has a multi-layer structure. Assuming that l represents the number of neural network layers, and k represents the current level, the calculation mode of the forward neural network is:

$$h^k = \sigma(b^k + w^k h^{k-1}) \quad (1)$$

Among them, $h^k$ represents the output of the Kth hidden layer, $b^k$ represents the bias of the kth layer, $w^k$ represents the corresponding weight, and σ represents the activation function, which is generally a sigmoid function or a tanh function.

When the parameters of the forward network θ are fixed, we can get the output of the neural network through Algorithm(1). The problem now is how to solve the parameters of the forward network. Like ordinary machine learning problems, we can solve the optimal parameter θ by minimizing the cost function J(θ). Among them, L represents the cost function of each sample, and f(x, θ) is the output of the neural network when the input is x.

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00035

| Algorithm 2Multi-layer neural network forward propagation |
|---|
| Input: network depth l |
| Input: W(i),i ∈ {1, ..., l}, the weight of each layer of the model |
| Input: b(i),i ∈ {1, ..., l}, the bias of each layer of the model |
| Input: x, the input of the neural network |
| Output: ˆy, the output of the neural network |
| Output: J, the value of the loss function of the neural network |

*B. Convolutional Neural Network*

Convolutional Neural Network (CNN) is a type of the previous neural network. LeCun et al. proposed that it has become the most commonly used and important in deep learning model after Hinton et al. won the first place in the ImageNet competition with CNN in 2012[3].The convolutional layer is the basis of the convolutional neural network, and the convolution operation is equivalent to using a sliding window to do a weighted average on an image. If our input data is x, the convolution kernel is f, and the output is y, here

$$\mathbf{x} \in \mathbb{R}^{H \times W \times D}, \mathbf{f} \in \mathbb{R}^{H' \times W' \times D \times D''}, \mathbf{y} \in \mathbb{R}^{H'' \times W'' \times D''} \quad (2)$$

Then the convolution operation can be expressed as:

$$y_{i''j''d''} = b_{d''} + \sum_{i'}^{H'} \sum_{j'}^{W'} \sum_{d'}^{D} f_{i'j'd} \times x_{i''+i'-1,j''+j'-1,d'd''} \quad (3)$$

The role of the Pooling layer is mainly to reduce the dimensionality of the output of the convolutional layer while retaining effective features. Pooling is mainly divided into two types: Max-pooling and Sum-pooling. Max-pooling operation can be expressed as:

$$y_{i''j''d} = \max_{1 \le i' \le H', 1 \le j' \le W'} x_{i''+i'-1,j''+j'-1,d}. \quad (4)$$

Sum-pooling and Max-pooling are similar to Sum-pooling, except that Max-pooling takes the maximum response. Local Response Normalization (LRN) normalizes the different channels at each position of the input data. The specific formula is as follows:

$$y_{ijk} = x_{ijk} \left( \kappa + \alpha \sum_{t \in G(k)} x_{ijk}^2 \right)^{-\beta}, \quad (5)$$

Among them, for each channel k, G(k) ⊂ {1, 2, ..., D} corresponds to the input subset. For LRN, input data and output data have the same dimensions.

For most machine learning problems, we can usually mix multiple models to get better generalization errors, but for deep neural networks, training multiple neural networks and then mixing the models often costs too much calculation resource. Dropout can be regarded as an approximation of model mixing. The method of Dropout is to set the input neurons to 0 with a certain probability p, so these neurons will not be counted in the forward and reverse processes.

*C. Action recognition based on multi-resolution coding*

The framework of the action recognition algorithm is shown in Figure 3-1. The algorithm is mainly composed of three parts: feature extraction, feature coding and action classification. Our action recognition framework combines traditional features and deep neural network features through different The coding method encodes them separately, and finally performs feature fusion and action classification [4].

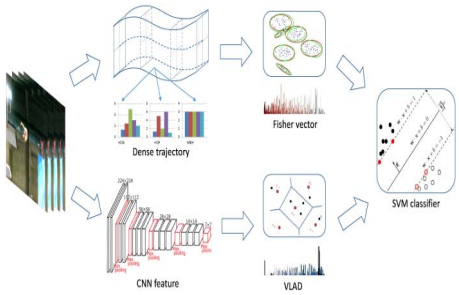


Figure 2. Framework diagram of multi-resolution coding action recognition.

LCD (Latent Concept Descriptor) is a feature extraction algorithm that encodes the features of neural networks through the traditional bag-of-words model [5]. Generally speaking, when we use CNN for feature extraction, we will use the features of the last fully connected layer as the input features. However, LCD encodes the features of the last layer of convolution, because this layer retains the input spatial information.

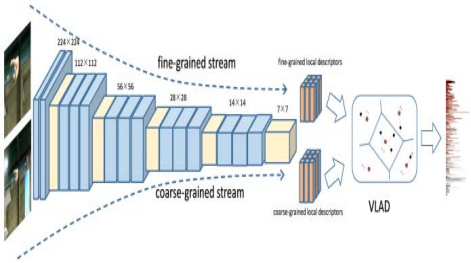


Figure 3. Multi-resolution LCD algorithm

First, we combined LCD features with Two-stream CNN. The original LCD feature only encodes one of the two-stream CNNs of Two-stream CNN, that is, a single frame image is used as the input of the neural network to extract high-level semantic information. Therefore, the original LCD can only obtain static features such as scenes in the video. In order to be able to extract the dynamic features of the video, we did the same operation on the optical flow neural network, which is to treat the

pool5 layer as a local feature and then encode it. The two CNNs in Two-stream CNN will be respectively encoded with LCD, and the extracted features will be feature fusion at the end.

In this section, we will introduce how to detect a significant area in a frame of image [6]. In identifying problems, usually, only one person is doing a certain action in a video. So, generally speaking, The most prominent area in the image includes the most important part of the frame. Our algorithm independently extracts features for this part during LCD encoding, which is equivalent to increasing the weight of the salient area in the final video feature. Given a whole input image I, we first use the EdgeBox algorithm to detect its candidate area (Object Pro-posal)[7]. The detection of object candidate regions generally refers to an algorithm for quickly screening out regions that may contain objects during object detection. After all the candidate regions are obtained, the formula for determining the final saliency region is as follows:

$$b = \frac{1}{T}\sum_{t}^{T} \Pi(w_t b_t < \alpha S)\Pi(w_t b_t > \beta S) b_t \quad (6)$$

Among them, S represents the area of the input image, and α and β are manually determined parameters. In our experiment, we set their values to 0.1 and 0.6, respectively. This formula is equivalent to screening all candidate regions according to their area, filtering out the too large and too small regions, and averaging all the remaining regions, so as to get the most significant region in the end. Of course, we can also calculate the saliency map of the input image from this[8]:

$$saliency(Pxy) = \sum_{t}^{T} \Pi(p_{xy} \in b_t) S_{xy} \quad (7)$$

Among them, (x, y) represents the position of a pixel. The algorithm for the detection of the entire salient area is shown in Figure 4, which shows the original image in turn, with the 10 candidate areas with the highest score, the detection result of the salient area, and the visible image of the image. Of course, we only need to use the saliency region in our motion detection algorithm.

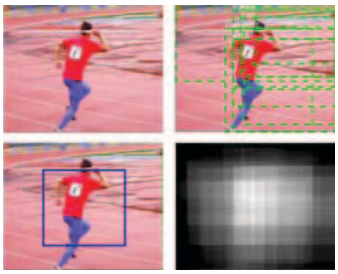

Figure 4. Significant area detection

## EASE OF USE

### A. Experimental results

In this section, we will first briefly introduce the data set used in the experiment: Olympic Sports. The Olympic Sports data set includes sports videos of various athletes collected from YouTube. The data set includes a total of 16 action categories (including weightlifting, bowling, basketball layup, diving, high jump, etc.), including a total of 783 videos . We use 649 videos as the training set, and the remaining 134 videos as the test set. Similarly, on this data set, the evaluation index is mAP.

### B. Experimental results

In the data set we use Map[9] (Mean Average Precision) as the evaluation index. For ordinary classification problems, generally we will directly use accuracy as an evaluation indicator. AP (Average Precision) can be calculated with the following formula:

$$AP(\varepsilon) = \frac{1}{N_i}\sum_{k=1}^{N_i} P(k) \quad (8)$$

Among them, P(k) represents the accuracy (Preci-sion) when k videos are retrieved from the database by setting the threshold, and mAP can be expressed as the averagevalue of all APs, namely:

$$mAP = \frac{1}{M}\sum_{i=1}^{M} AP(\varepsilon) \quad (9)$$

### C. Visualization of results

Filter out the fixed actions from Olympic Sports, use Openpose and Opencv tools to instantiate the above-mentioned action pattern recognition process, and get the action posture recognition diagram:

Figure 5. Action recognition diagram

### D. Visualization of results[10]

Starting from the 14th Five-Year Plan proposed by the country, this article starts from the direction of smart applications and uses related ML algorithms to process video streams. The more basic ones are the popularization of CNN-LSTM algorithm forward algorithm and the detection of action pattern recognition algorithms. , Gave a detailed overview of the action recognition process and

saliency area detection in multi-resolution video analysis, and finally conducted an experiment on the collected instance Olympic Sports data set, and the class roughly inferred the action mode and movement correlation. This research has good application prospects in campus physical testing, which can save a lot of manpower and material resources and achieve intelligent physical testing. But it is worth noting that in practical applications, user identification and anti-cheating measures need to be improved.

## IV. SUMMARY

Starting from the 14th Five-Year Plan proposed by the country, this article starts from the direction of smart applications and uses related ML algorithms to process video streams. The more basic ones are the popularization of CNN-LSTM algorithm forward algorithm and the detection of action pattern recognition algorithms. , Gave a detailed overview of the action recognition process and saliency area detection in multi-resolution video analysis, and finally conducted an experiment on the collected instance Olympic Sports data set, and the class roughly inferred the action mode and movement correlation. This research has good application prospects in campus physical testing, which can save a lot of manpower and material resources and achieve intelligent physical testing. But it is worth noting that in practical applications, user identification and anti-cheating measures need to be improved.

## REFERENCES

[1]AlNatour Ahlam,Gillespie Gordon Lee,Alzoubi Fatmeh. "We cannot stop smoking": Female university students' experiences and perceptions.[J]. Applied nursing research : ANR,2021,61:

[2] Yunxin Huang, Fei Chen, Shahe Lv, Xuedong Wang. Facial Expression Recognition: A Survey [J]. Symmetry, 2019,11(10).

[3] ZHANG A M, XU Y. Attention Hierarchical Bilinear Pooling Residual Network for Expression Recognition [J].Computer Engineering and Application, 2020, 56(23):161- 166.

[4]Gera D, Balasubramanian S. Landmark Guidance Independent Spatio-Channel Attention and Complementary Context Information based Facial Expression Recognition[J].Pattern Recognition Letters, 2021, 145:58-66.

[5]LI G H, YUAN Y F, Ben X Y, ZHANG J P. Spatiotemporal attention network for micro- expression recognition[J].Journal of Image and Gra-phics,2020,25(11):2380-2390.

[6]Duta I C, Liu L, Zhu F, et al.Pyramidal convolution: rethinking convolutional neural networks for visual recognition[EB/OL].(2020-6-20)[2021-5-9]

[7] Adil B, Nadjib K M, Yacine L. A novel approach for facial expression recognition[C]//2019 International Conference on Networking and Advanced Systems (ICNAS). IEEE, 2019: 1-5.

[8] Dapogny A,Bailly K,Dubuisson S. Confidence -weighted local expression predictions for occlusion handling in expression recognition and action unit detection[J]. International Journal of Computer Vision, 2018, 126(2): 255- 271.

[9] Li Y, Zeng J, Shan S, et al. Occlusion aware facial expression recognition using CNN with attention mechanism[J]. IEEE Transactions on Image Processing, 2018, 28(5): 2439-2450.

[10] Gera D, Balasubramanian S. Landmark Guidance Independent Spatio-Channel Attention and Complementary Context Information based Facial Expression Recognition[J]. Pattern Recognition Letters, 2021, 145:58-66.