applied sciences

MDPI

Article

Deep Clustering Efficient Learning Network for Motion Recognition Based on Self-Attention Mechanism

Tielin Ru [1,*] and Ziheng Zhu [2]

[1] Sports Department, Xi'an University of Science and Technology, Xi'an 710054, China
[2] College of Computer Science and Technology, Xidian University, Xi'an 710071, China
* Correspondence: rtl7608@xust.edu.cn

Abstract: Multi-person behavior event recognition has become an increasingly challenging research field in human–computer interaction. With the rapid development of deep learning and computer vision, it plays an important role in the inference and analysis of real sports events, that is, given the video frequency of sports events, when letting it analyze and judge the behavior trend of athletes, often faced with the limitations of large-scale data sets and hardware, it takes a lot of time, and the accuracy of the results is not high. Therefore, we propose a deep clustering learning network for motion recognition under the self-attention mechanism, which can efficiently solve the accuracy and efficiency problems of sports event analysis and judgment. This method can not only solve the problem of gradient disappearance and explosion in the recurrent neural network (RNN), but also capture the internal correlation between multiple people on the sports field for identification, etc., by using the long and short-term memory network (LSTM), and combine the motion coding information in the key frames with the deep embedded clustering (DEC) to better analyze and judge the complex behavior change types of athletes. In addition, by using the self-attention mechanism, we can not only analyze the whole process of the sports video macroscopically, but also focus on the specific attributes of the movement, extract the key posture features of the athletes, further enhance the features, effectively reduce the amount of parameters in the calculation process of self-attention, reduce the computational complexity, and maintain the ability to capture details. The accuracy and efficiency of reasoning and judgment are improved. Through verification on large video datasets of mainstream sports, we achieved high accuracy and improved the efficiency of inference and prediction. It is proved that the method is effective and feasible in the analysis and reasoning of sports videos.

Keywords: self-attention mechanism; deep embedded clustering; LSTM; group behavior recognition; feature enhancement

Citation: Ru, T.; Zhu, Z. Deep Clustering Efficient Learning Network for Motion Recognition Based on Self-Attention Mechanism. *Appl. Sci.* **2023**, *13*, 2996. https://doi.org/10.3390/app13052996

Academic Editors: Xin Ning, Weijun Li and Sahraoui Dhelim

Received: 12 January 2023
Revised: 17 February 2023
Accepted: 21 February 2023
Published: 26 February 2023

1. Introduction

The application of computer vision technology in human–computer interaction [1], detection and recognition of abnormal events and other fields has developed rapidly. Among them, the automatic recognition of human activities [2–4] is more important because it includes a wide range of applications in real scenes, including monitoring human behavior changes and personal health monitoring, and the application scenarios are not only limited to sports fields, military environments, hospitals, but also in social security management, crime fighting urban management, serving people's livelihood and other fields, it plays an important role, so people have a strong research interest in the identification and inference of human activities in videos. The deduction and analysis of athletes' behavior [5,6] in sports videos has become a hot research topic. Due to the large interference of surrounding background information and different scales of human behavior targets in the detection of video information in public places, it is difficult to further improve the accuracy of athletes' behavior detection. Over the years, many researchers have developed

various methods; however, it is difficult to find the expression of athletes' behavior changes that can be accurately understood.

Modern sports are mostly in the form of groups. The behavior analysis and inference methods of single athletes cannot be directly transplanted into group events. The identification of multi-player behavior events [7–9] is a challenging research problem. Usually, in the scene where many people are active, only a small part of people actually participate in the actual events. In previous studies, researchers relied on manually constructed models to detect and discriminate sports videos. They solved this problem by using local context descriptors and graphical models to model context information. Due to the complexity of the model and the inability to express deep level features [10], these effects are not satisfactory. Refs. [11,12] use correlation filters [13] to identify action instances, and extract context descriptors from humans and surrounding areas to identify group activities, which can obtain relatively low computational complexity, but its effectiveness is questionable. Refs. [14–16] conducts behavior analysis through the range of motion of human joints, that is, the 3D coordinate set of human body joints is given to identify the activities performed. Ref. [17] proposes a framework to understand and infer human social behavior in video image sequences, and infers group behavior through the single person's social behavior. Ref. [18] evaluates the performance of athletes by considering the changes in the first person perspective in basketball. This method splits the group behavior, first individuals, then individuals superimposed and combined, and finally infers the group behavior. In addition, Ref. [19] deployed a multi-mode deep learning network, first reconstructed from wearable and environmental sensor data, checked individual behavior used to classify the activities of a group of people in their daily work tasks, observed whether there was a common temporal and spatial dynamics at the group activity level, and solved the problem in a hierarchical manner. Ref. [20] proposed a semantic retention teacher student (SPTS) model for group activity recognition in videos, which aims to mine semantic retention attention to automatically find key people and discard misleading people to mine group activities. Ref. [21] proposed a new method to effectively locate the sports field from a single broadcast image of the game. By annotating some key frames and extending the location to similar images, the layout of the field was obtained and applied to football and hockey. RNN is used to represent tracking features, learn the attention weight when behavior changes, combine these features at each time, and then use another RNN to process the features of interest for behavior detection and classification [22].

At this stage, researchers have proposed several deep learning methods to construct the structure of group context, and these methods have achieved better performance than traditional methods. For example, they use recurrent neural networks (such as LSTM) to model individual behaviors and group activities sequentially, input images as a whole, classify behaviors, build effective classifiers, and conduct feature dominance, in which the maximum average pool is used to aggregate human features. Although different people should receive the same degree of attention, at a certain time, different people's contributions to the whole group's activities are different. For example, in football, the player's "free kick" action should be more worthy of attention than other players, and the captured details should also be greater so that the type of sports can be better analyzed and judged, and the efficiency of reasoning and judgment can be improved.

Based on the problem background of the above scheme, we specifically solved this problem. The contributions of this paper can be summarized as follows:

(1) Through the LSTM, this network can not only solve the problem of gradient disappearance and explosion in the recurrent neural network (RNN), but also capture the internal correlation between multiple people in the sports field for identification, etc.
(2) On the basis of (1), the DEC is added to integrate the motion coding information in key frames to improve the judgment efficiency.
(3) With the self-attention mechanism, it can not only analyze the whole process of the whole sports video macroscopically, but also focus on the specific attributes of the movement to capture more important details, extract the key posture features of

athletes, and further enhance the features, effectively reducing the parameters of the self-attention mechanism in the calculation process, reducing the computational complexity while maintaining the ability to capture details, and improving the accuracy and efficiency of reasoning and judgment. Through verification on large video datasets of mainstream sports, we achieved high accuracy and improved the efficiency of detection and recognition.

The logical structure of this paper is as follows: In Section 2, we introduce the related work, and mainly discuss the hottest group behavior detection technology at present. In Section 3, we introduce the specific methods and measures of this paper, build the overall algorithm flow chart of the full text, and elaborate the mechanisms of DEC, LSTM and self-attention in detail. In Section 4 of the experimental module, we compare and evaluate the effectiveness of DEC, and experiment with a pedestrian dataset and football dataset in the algorithm incorporating the self-attention mechanism, and compare with other mainstream network architectures. In Section 5, we discuss the full text and describe the advantages and limitations of this method. In the final conclusion, we summarize the full text and look forward to the future work direction.

2. Related Work

Compared with previous research, this method has made further improvement in clustering, the self-attention mechanism and the deep learning network.

Clustering [23] is an important topic in machine learning and data mining. In recent years, deep neural networks (DNNs) [24,25] have received extensive attention in various clustering tasks. In particular, semi-supervised clustering can significantly improve the clustering performance by introducing only a small amount of prior information into a large number of unsupervised data. However, these clustering methods ignore that the defined clustering loss may destroy the feature space, resulting in non-representative and meaningless features. Aiming at the problem that the existing semi-supervised deep clustering feature learning process has some deficiencies in maintaining the local structure, this paper proposes a deep embedded clustering algorithm, which uses an incomplete automatic encoder to maintain the internal local structure of data during feature expression learning. By combining clustering loss, pair constraint loss and reconstruction loss, the cluster label allocation and feature expression are jointly optimized. The problem of data clustering involves a wide range of scenarios. The purpose of clustering is to divide similar data into a class according to similarity measures (such as Euclidean distance). With the development of science and technology, data collection becomes more convenient, the dimensions of data generated become higher and higher, and the correlation between data becomes more complex. As data sets become larger and more diverse, existing algorithms need to be adjusted to maintain the quality and efficiency of clustering. The traditional clustering algorithm considers all dimensions of the data because the data dimension is not high and in order to obtain all information. However, in high-dimensional data, multiple dimensions of data are usually irrelevant, and these irrelevant dimensions will hide clustering in noisy data, making clustering algorithms confused. In a very high-dimensional dataset, all objects are almost equidistant from each other, thus completely concealing clustering. The feature selection method has been successfully used to improve the clustering quality to some extent. Compared with other complex neural network structures, the structure proposed in this study uses the potential representation form of self-encoder learning data, and then uses this form for clustering. This study made innovations in the potential representation form and fusion mode of extracted data. In cluster analysis, the field of unsupervised machine learning has been studied from different aspects, covering a wide range of scenarios, including deep learning [26,27], pattern recognition, image processing [28,29], biological information [30], and so on. How to choose the feature space and how to define the distance measure are the key points of the clustering algorithm. DEC clustering uses an automatic encoder to construct a nonlinear embedding. A clustering layer is stacked on the self-encoder for clustering tasks to achieve parameterized high-dimensional data clustering.

Compared with the K-means clustering algorithm, local discrimination and global fusion model, this algorithm greatly improves the effect in clustering tasks, and further enhances the stability of the model.

In the self-attention mechanism [31,32], most of the current research is to design local and global self-attention encoders, and build feature aggregation modules to extract features. The essence of the self-attention mechanism is to highlight or emphasize the important information of the target object through a series of attention weight coefficients, suppress some irrelevant details, and correlate the global information. It is flexible and can directly capture the connection between local and global information. In the self-attention module, firstly, the input features are normalized using batch standardization, the global feature information is extracted and activated using the activation function, then the intermediate fusion features are obtained by pixel addition with the original image. Secondly, cross-channel information fusion is achieved using the convolution operation, and then the image features are non-linearly activated using the activation function. Finally, we add the activated feature map and the input image pixel by pixel to construct the identity mapping condition so as to obtain the output feature map of the self-attention module with more abundant feature information.

The convolutional neural network (CNN) [33], as the leading technology of deep learning in recent years, has made remarkable achievements in image-recognition tasks, and also performs well in face detection [34,35], speech recognition, expression recognition, modal awareness recognition, gender recognition and other real computer-vision applications. However, the traditional convolutional neural network model construction has many shortcomings, such as excessive dependence on empirical knowledge, unpredictability, training difficulty, etc., which leads to the network structure and parameter settings needing to spend a lot of time on the tuning test. In order to break through the limitation that the CNN structure design and parameter setting excessively rely on empirical knowledge, we applied LSTM. Compared with other neural networks, LSTM has more potential to successfully optimize convolutional neural networks. It is of great significance for research on different classification tasks, improves the convergence speed, realizes the adaptive construction of convolutional neural networks, improves the classification accuracy of neural networks, and has strong learning and generalization capabilities, from the single-layer neural network to the two-layer neural network, and then to the multi-layer neural network; with the increase in the number of network layers and the adjustment of the activation function, its nonlinear fitting ability is constantly enhanced. In LSTM, the corresponding RNN and its corresponding RNN are time-series models with long-term memory capability. However, RNN will remember the previous state of the network during the calculation process, which greatly increases the amount of calculation and the calculation time of the model. In order to overcome this disadvantage, LSTM adds three gates (forgetting gate, input gate and output gate) on the basis of RNN to selectively remember the previous state of the network, which improves the calculation speed of the model. Among them, the forgetting gate selectively memorizes part of the time output, the input gate combines the current time input with the previous time output to change the current state, and the output gate obtains the current neural network output through the activation function after the current neural network state changes.

In the next part of the article, we will introduce the methods we used, the experimental analysis, and the final conclusions and references.

3. Method

The overall motion recognition algorithm framework of this paper is mainly divided into the following three modules, as shown in Figure 1: (1) integrating self-attention mechanism; (2) complete clustering operation; and (3) capture the internal correlation of people on the field through LSTM and finally obtain the result of inference and recognition.

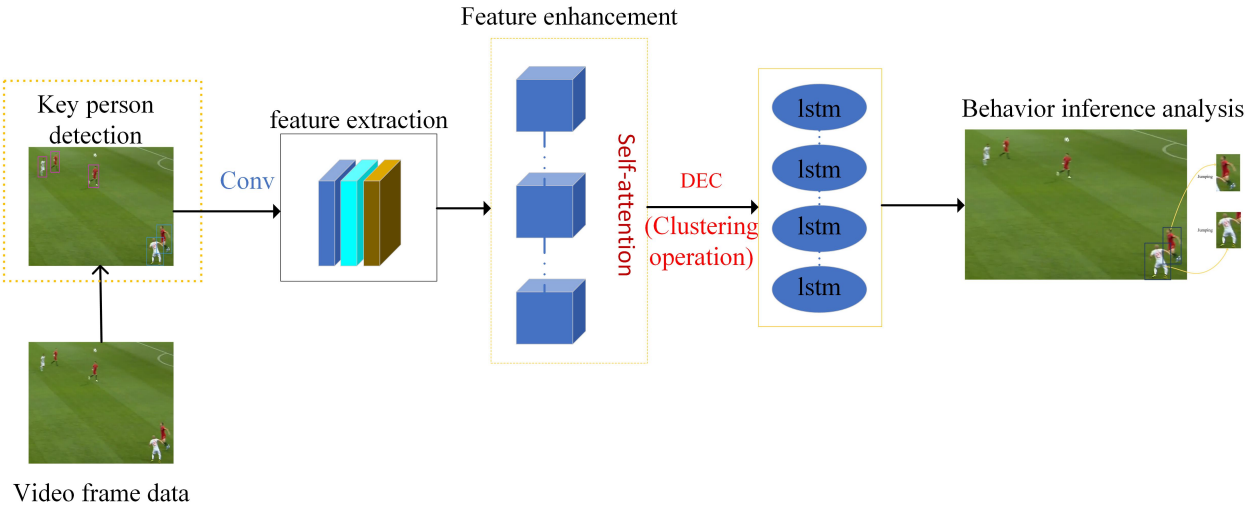

Figure 1. Overall flow chart of algorithm. First, input the video frame data to locate and detect the key people, and then extract the features through convolution. Secondly, incorporate the self attention mechanism here to enhance the features, and then cluster through DEC. Finally, conduct behavior analysis and inference through LSTM.

3.1. Deep Embedded Clustering

In order to optimize inference detection technology, we apply deep embedded clustering algorithms here. Compared with traditional clustering algorithms, their distance metrics are often limited to the original data space. When the input dimensions are high, they are often invalid. The automatic encoder can retain the local structure of the data generation distribution. In this case, the use of clustering embedding space will not change its performance. In order to ensure the effectiveness of clustering, the stacked denoising automatic encoder for preprocessing is no longer applicable. Because clustering should be performed on the characteristics of clean data, rather than using noise data in the denoising automatic encoder, the noise is directly removed in this paper. The stacked noise elimination automatic encoder degenerates into an incomplete automatic encoder. Therefore, on the one hand, DEC can reflect the essential characteristics of the distribution of the input data, and on the other hand, it can learn the feature representation and clustering allocation of the data to the potential hidden variable space and iteratively optimize the target to improve the clustering performance.

The DEC model uses the automatic encoder for data pre-training. The network parameters are initialized through the automatic encoder, and a two-layer neural network is defined as follows:

$$\tilde{x} \sim \text{Dropout}(x), h \sim b_1(W_1\tilde{x} + \omega_1) \tag{1}$$

$$\tilde{h} \sim \text{Dropout}(h), y \sim b_2(W_2\tilde{h} + \omega_2) \tag{2}$$

Dropout() indicates that the dimension of the input data of any setting part is 0, that is, the dimension of x input is 0, $\tilde{x}$ means the input layer, h means the hidden layer, and y means the output layer, b_1,b_2 is the activation function of the encoder, and W_1,W_2,ω_1,ω_2 is the model parameter.

3.2. Self-Attention Mechanism Fusion Depth Clustering Learning Network Model

3.2.1. Long Short Term Memory

Because traditional neural networks only use independent data vectors each time when dealing with prediction problems, there is no concept similar to "memory", which is used to deal with various tasks related to "memory". The proposal of the recurrent

neural network solves the above problems to a certain extent, but it is easy to have the problem of gradient disappearing when dealing with long-term dependence problems. In addition, RNN will remember the previous state of the network during the calculation process, which greatly increases the calculation amount and the calculation time of the model. Long term short-term memory network (LSTM) introduces a gating structure so that the network can not only remember the past information, but at the same time, it can selectively forget some unimportant information to achieve the modeling of long sequences. In order to further obtain the relationship between regional features, this paper uses the depth learning network model under the self-attention mechanism to obtain the context information of the image. The algorithm model structure is shown in Figure 2.

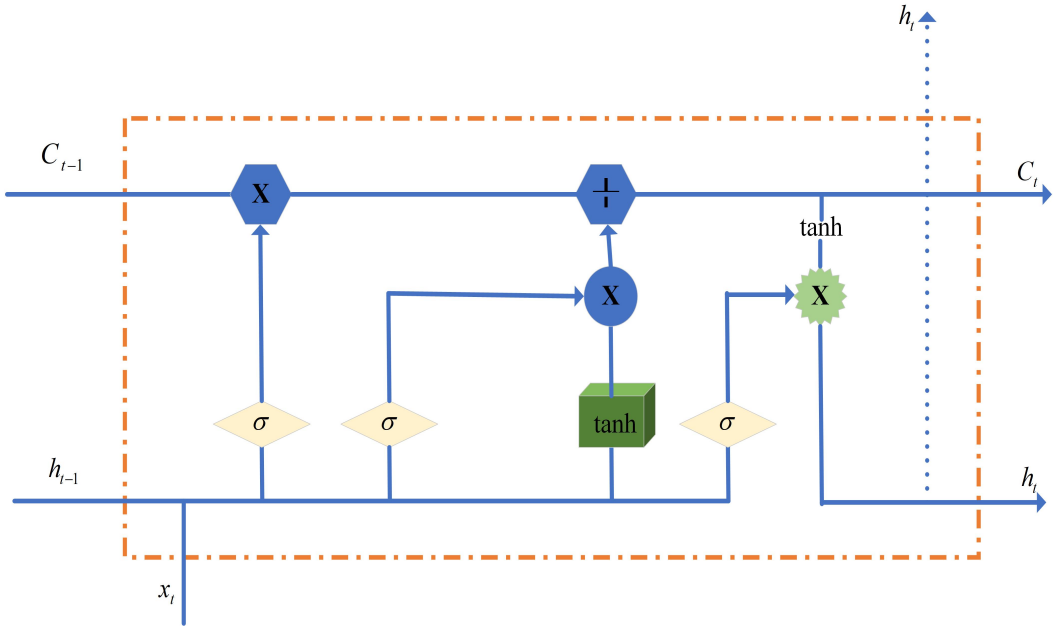

Figure 2. LSTM unit structure.

Among the two inputs in Figure 2, the input above is the memory state c_{t-1} of the unit, and the following input is the output h_{t-1} of the upper layer. In both outputs, the upper output is the memory state c_t of the next unit, and the lower output is the output h_t of the current layer. The three sigmoid activation functions of the intermediate structure represent three gate control units; since the output of the Sigmoid function is 0 1, the control valve mechanism can be realized. f_t is the forgetting gate output, which is mainly used to selectively forget the output of the previous node; i_t is the output value of the input gate, and S_t and h_{t-1} are the control, which is obtained by adding the product of the weight matrix and the output of the previous layer and the input at the current time. It is combined with the sigmoid function σ externally; output gate O_t is used to output the hidden state h_t. y_t is the output of the next time, the sum of the corresponding weight matrix and the output of the current layer plus the offset term, embedded into the activation function. x_t is the input of the current time, and Tanh is the activation function. The specific formula is

$$i_t = \sigma(\boldsymbol{W}_{xi}x_t + \boldsymbol{W}_{hi}h_{t-1} + b_i) \tag{3}$$

$$f_t = \sigma\left(\boldsymbol{W}_{xf}x_t + \boldsymbol{W}_{hf}h_{t-1} + b_f\right) \tag{4}$$

$$C_t = f_t C_{t-1} + i_t \tanh(\boldsymbol{W}_{xC}x_t + \boldsymbol{W}_{hc}h_{t-1} + b_C) \tag{5}$$

$$O_t = \sigma(\boldsymbol{W}_{x0}x_t + \boldsymbol{W}_{h0}h_{t-1} + b_o) \tag{6}$$

$$h_t = O_t \tanh(C_t) \tag{7}$$

$$y_t = \tanh\left(\boldsymbol{W}_y h_t + b_y\right) \tag{8}$$

Among them, b_i, b_f, b_C, b_o, and b_y are the bias terms corresponding to the input gate, forgetting gate, cell state, output gate and output layer, respectively. $\boldsymbol{W}_{xi}$, $\boldsymbol{W}_{hi}$, $\boldsymbol{W}_{xf}$, $\boldsymbol{W}_{hf}$, $\boldsymbol{W}_{xC}$, $\boldsymbol{W}_{hC}$, $\boldsymbol{W}_{x0}$, $\boldsymbol{W}_{h0}$, and $\boldsymbol{W}_y$ are the weight matrix corresponding to the input layer to input gate, hidden layer to input gate, input layer to forgetting gate, hidden layer to forgetting gate, input layer to cell state, hidden layer to cell state, input layer to output layer, hidden layer to output layer, and output layer.

The feature records of each athlete collected in the experiment include the posture of each limb, key touch points, individual actions, group actions and displacement values. Therefore, the input of the prediction model is a vector composed of these feature records. The vector is extracted through the convolution layer, and then the extracted features are input into the LSTM model to predict the displacement time series after the internal features of the data are extracted through the self-attention mechanism layer.

3.2.2. Self-Attention Mechanism

The self-attention mechanism has been widely used in image recognition and natural language processing, such as image classification, image detection, object detection, pedestrian recognition, etc. The self-attention mechanism simulates the key information features of things captured by the human brain when observing things, which is conducive to paying attention to the important information features of interested objects when the computer [36–39] executes image scene tasks, and this feature information is crucial for subsequent work. The traditional depth learning convolution neural network cannot extract the feature information of a specific target according to the input image, and will ignore the impact of global features on group detection. The self-attention mechanism can act on the middle layer features of the image, enrich the semantic features to help the model extract the key features with identification contained in the image, and improve the performance of the features. In addition, the self-attention mechanism is good at extracting the internal correlation of features, which can enhance the performance of network local feature extraction and global context aggregation. Common attention mainly includes spatial attention and channel attention. Spatial attention aims at improving the feature expression of key areas, transforming the spatial information in the original image into another space and retaining the key information, generating a weight mask for each area and weighting the output. Channel attention can find the correlation of different channels to the feature map, automatically obtain the weight of each channel through the neural network, and enhance the features of important channels and weaken the features of non-important channels according to the dependence of each channel.

In this paper, a self-attention mechanism is designed in LSTM, which can fuse multi-scale information. While enhancing the feature difference between the tampered area and the real area of the image, it can improve the detection performance of the tampered area of the network. The core idea of the self-attention mechanism is to highlight the part of the input data that is more relevant to the output task in the form of weight according to the probability distribution of the data and the relationship between variables. In terms of image semantic understanding, the attention weight is learned through the neural network and fused with the original feature map so as to strengthen important features and weaken irrelevant components, reduce dependence on external information, and make use of the inherent information in the features to interact with attention as much as possible. A feature map contains information, such as feature channels and positions. Some of the content in the map is of concern to us, while some of the content contributes little to the results of the detection task. Through learning, we can suppress insignificant features, enhance the expression ability of features in the network, and then improve the target detection

effect. In this paper, the self-attention mechanism used is decomposed into two parallel one-dimensional feature codes, that is, two one-dimensional vectors are obtained by averaging pooling in the horizontal and vertical directions, and then 1×1 convolution is spliced in the spatial dimension to compress channels. Through Relu activation coding, the same number of channels as the input feature map is obtained. Normalized weighting is used to adjust the features and extract the output feature map.

The self-attention mechanism adopts the multi-head attention mechanism. Its principle is to splice the calculation results of the multi-point product attention mechanism. The self-attention model can be seen as establishing the interaction between different forms of the input vector in a linear projection space. Multi-head attention is to establish different projection information in multiple different projection spaces. The input matrix is projected differently to obtain many output matrices, which are spliced together. The figure shows the schematic diagram of data splicing processing. The query vector Q (query), the key vector K (key), and the value vector V (value) are obtained from the sequence with length X through linear transformation. The same input vector X is processed by the point product attention mechanism twice, and two sets of weight matrices and two sets of Q, K, V matrices are obtained. The output of the multi-head attention mechanism model is obtained by splicing the results, as shown in Figure 3:

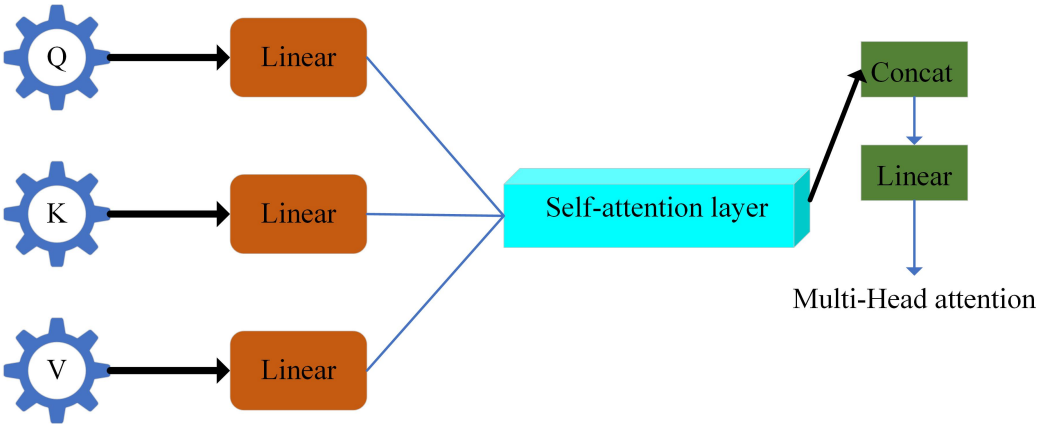

Figure 3. Data processing of multi-head attention mechanism.

It generates key, query and value through embedded transformation, calculates the weight coefficient according to the first two items, and then weights the sum of the values according to the weight to complete the adjustment of the location characteristics of the query.

The structure of the self-attention mechanism is shown in the figure. After the corresponding convolution operation, the local input feature X generates three new mapping features φ^1,φ^2 and φ^3,where $\varphi^1, \varphi^2, \varphi^3 \in \mathbb{R}^{h \times w \times c}$. After the reshaping operation, $\varphi^1, \varphi^2, \varphi^3 \in \mathbb{R}^{n \times c}$, where $n = h \times w$, then the transpose matrices of φ^1 and φ^2 are multiplied, and the sigmoid function is used as the activation function to obtain the self-attention matrix M_{sim}, where $M \in \mathbb{R}^{n \times c}$, M describes the similarity between any two positions of the tamper feature map in the spatial dimension. The relevant definition formula is shown in Formula (9):

$$M_{sim} = \frac{(\varphi^1 \varphi^2 \varphi^3)}{\sum(W_Q + W_K + W_V)} Sigmoid(\mathbb{R}) \tag{9}$$

where W_Q, W_K, and W_V are the weight parameters that can be learned, and M_{sim} represents the influence of the three location features. The more similar the location features are, the greater the relevance. The sum of the three weight parameters is combined with the embedded sigmoid function of the three mapping features to obtain the spatial dimension similarity. For W_Q, W_K, and W_V, the correlation state of the three V's is multiplied by M_{sim}, and the result is reshaped as $\mathbb{R}^{h \times w \times c}$ and multiplied by the scale parameter α. Finally,

the element summation operation is performed with the input $\mathbb{R}^{h\times w\times c}$, and the focus feature layer $E_{att} \in \mathbb{R}^{h\times w\times c}$ of self-attention is obtained, as shown in Formula (10):

$$E_{att} = \alpha\left(W_Q, W_K, W_V\right) M_{sim} + \mathbb{R}^{h\times w\times c} \tag{10}$$

where α is a learnable parameter initialized to 0. Similar to the spatial self-attention mechanism, the channel self-attention mechanism obtains the final channel attention output as $E_{att} \in \mathbb{R}^{h\times w\times c}$. In the context of the video image, the features of the context part of the image information concerned by each point are different. Therefore, the model needs to be able to distinguish the importance of different parts of the context to each point effectively so as to extract the information related to a specific point. The self-attention mechanism is used to extract information related to a particular point from the context. The specific calculation formula is as follows: Concat() represents the concatenation operation, and E^i_{att} is the context feature representation of point perception initialized by the Gaussian standard distribution and updated automatically continuously during network training, combined with the three learnable weights to obtain relevant information of specific points, as shown in Formula (11):

$$E = \left(W_Q, W_K, W_V\right) \mathrm{Concat}\left(E^1_{att}, E^2_{att}, \ldots, E^i_{att}\right) \tag{11}$$

Different from multi-classification tasks, such as semantic segmentation and target detection, image tamper localization is a two classification task, that is, the image area is tampered or non-tampered, and the prediction range of each pixel is [0 1]. At the same time, the space and channel attention matrix graph describes the similarity of two positions in the space and channel dimensions. Therefore, unlike this method, which uses the Softmax activation function, this paper uses the sigmoid activation function with a value range of [0 1], more suitable for binary tasks when calculating spatial and channel attention matrices. E can be seen as the weighted sum of context features, so the output features have more context views and selectively aggregate context information according to specific point attention.

The self-attention mechanism can fully extract features, and carry out self-attention weighting to enhance the effect of feature extraction to obtain the relationship between each action of the video image, and effectively extract the context of elements in the data set. In addition, compared with the process of feature extraction by the convolution layer, the self-attention mechanism can effectively reduce the amount of computation and the time cost of model training.

3.3. Overall Pseudocode Architecture

The overall Algorithm 1 is as follows. The purpose is to solve the problem of gradient disappearance in RNN, and introduce the gradient propagation model $c(t) = c(t-1)$ of the corresponding unit memory state c_{t-1}. Secondly, the information is loaded into the long-term memory unit through the input gate, and the problem of activation function saturation is solved through the forgetting gate. Then, we select the appropriate memory for output through the output gate to solve the problem of reducing the controllability of the gate unit. Finally, the hidden state h_t is introduced to transform the simple feedback structure of the neural network into a fuzzy historical memory structure, and the network is successfully constructed.

Algorithm 1: Display of Overall Algorithm Structure of Motion Recognition

Input: Build a deep learning network, load the video frame dataset and initialize the pre-training through the automatic encoder to optimize the detection technology.

Load dynamic motion video frame training dataset;
for *Conduct key person positioning detection* **do**
    After convolution operation;
    repeat;
    end
for *Integrate into self-attention mechanism* **do**
    Through feature enhancement;
    repeat;
    end
for *Cluster operation by integrating DEC algorithm* **do**
    The automatic encoder is used for data pre-training, the network parameters are initialized by the automatic encoder, and a two-layer neural network is defined;
    The posture of each limb, key touch points, individual actions, group actions and displacement values are used as the vector;
    end
for *After feature extraction through convolution layer again, after the internal features of the data extracted from the attention mechanism layer, the extracted features are input into the LSTM model for displacement time series prediction* **do**
    Conduct behavioral information inference;
    Get the final prediction result;
    end

4. Experiment

4.1. Experimental Platform and Setting

The operating system used in the experiment is Windows 10 and Intel processor, chip type GeForce GTX 1650, CPU main frequency 2.40 GHz, memory 8 GB. It was used for training on a server with GPU. The algorithm in this paper is based on Pytorch, Python's deep learning framework. The compilation environment is Pycharm.

In order to test the effectiveness of the algorithm in this paper, the dataset of a high-definition football match video of a certain World Cup is selected. The dataset consists of about 118.8 GB, 64 videos, 30 frames per second, and a resolution of 1920 × 1080, including the posture of players without the ball, the action stretch in the ball state, the tactical behavior transformation of the group, etc. Figure 4 is the basic flow chart of the experiment.

4.2. Experimental Results and Analysis

The experiment is conducted to verify the effectiveness of the method in this paper by comparing relevant modules. Specifically, it is mainly divided into the following parts.

4.2.1. Deep Embedded Clustering Module

First, we need to evaluate the effectiveness of the DEC model clustering. To solve this problem, we verified on the MNIST handwritten data set. The dataset consists of 70,000 handwritten digital images, each of which is normalized in size and is a 28 × 28 gray-scale image. Firstly, K-means was used for data preprocessing, cluster center was selected, encoder was used for pre-training, and the Adam optimization method was used to optimize the process and improve the model performance. The optimal parameters of the model were obtained, the learning rate was set to 0.001, parameters were updated every 15 epochal training rounds, and the output dimension was set to 15. After 100 times of

training, the hidden feature space was constructed according to the mean and variance returned by the coding layer. In order to display the clustering effect intuitively, PCA (PCA, as a basic linear dimension reduction method, has no parameter restrictions and greatly reduces the calculation cost) was used to map a small part of the sampled hidden feature space to three-dimensional space for visualization. It is convenient to observe the aggregation distribution of data in the potential space, as shown in Figure 5. The clustering effect is good, which indicates that the potential features are suitable for clustering.

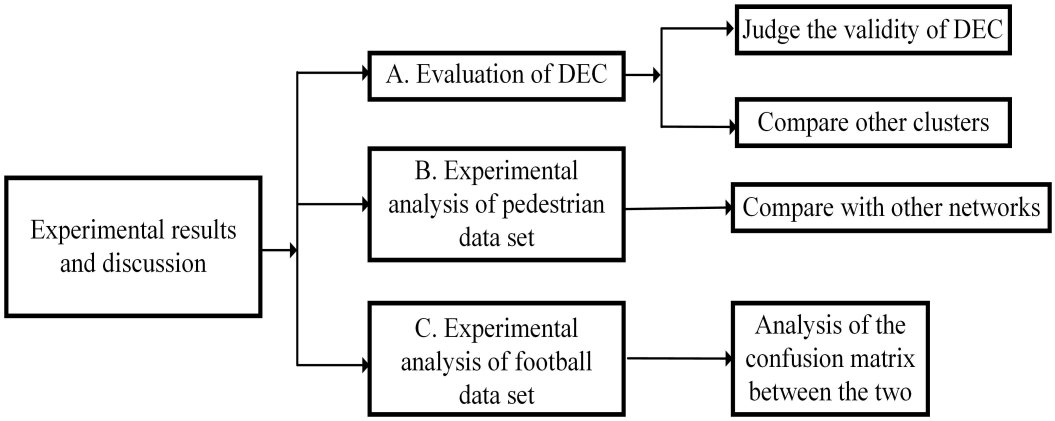

Figure 4. Basic flow chart of experiment.

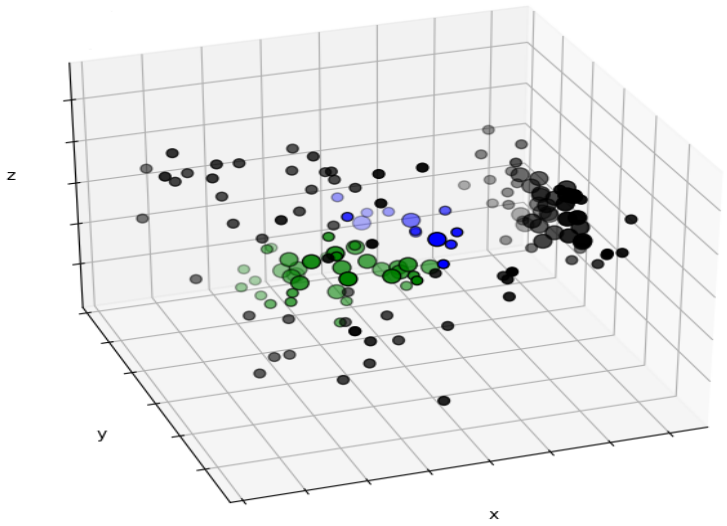

Figure 5. Data distribution of DEC potential space.

In addition, we also selected the cluster evaluation indicators to measure the effectiveness of the clusters selected in this paper, which are, respectively, accuracy (ACC), standardized mutual information (NMI) and adjusted Rand index (ARI). The closer the three values are to 1, the better the clustering effect. We tested the clustering effect with different clustering models on the MNIST dataset, and the records are shown in Table 1. According to the experimental results, the method in this paper has a significant improvement over the commonly used K-means algorithm in three evaluation indicators, up to 60.01%, and our clustering effect has also improved compared with AE and DBSCAN. This shows that the method in this paper is better than the traditional clustering algorithm and can improve the judgment efficiency.

Table 1. Indicators for comparing different clustering algorithms on MNIST dataset.

Dataset	MNIST			
	KMeans [40]	**AE [41]**	**DBSCAN [42]**	**DEC**
ACC	0.52	0.78	0.69	0.95
NMI	0.53	0.77	0.82	0.92
ARI	0.34	0.69	0.71	0.93

4.2.2. Network Model Performance Comparison Module

We have the following steps to train our network model, which incorporates the self-attention mechanism. We pretrained the self-attention modules of people and their body parts to ensure convergence. We will verify the pedestrian behavior detection in real life. The whole training process includes three steps: pretraining LSTM network, extracting behavior features through the self-attention mechanism, and finally sending the feature layer to the network. The importance of each person in each frame is generated by the self-attention mechanism, and then the behavioral characteristics of each person are summed according to the weight, the weighted features are input into the LSTM network, and the output features are classified to jointly predict the detection, individual and collective behaviors. In order to better test the pedestrian's motion information in the video, we connected the self-attention mechanism with the features extracted from the original RGB trajectory. The parameters of the network are fixed to train the network of individual behavior characteristics. In each time step, we organized the focused character features into subgroups and then input them into the network module to generate context features circularly. These features were cascaded and transmitted to the final LSTM network, followed by the softmax classification layer, which enables the entire context network to be end-to-end trained without any additional coding steps. In all experiments, we set the trade-off parameters to 1 and 2, respectively, and used the random gradient descent of the Adam optimizer to obtain the best parameters of the model. The output dimension was set to 10, and the initial learning rate was set to 0.002.

We also compared our method with the following latest research models, as shown in Table 2: The accuracy rate means that all the predicted samples are correct (for the detection of human movement behavior, the final result prediction is correct), and all the predicted experimental samples are divided. The recall rate is to predict positive samples as positive samples (that is, the detection of human movement behavior is correct) and divide the results that are originally correct.

Table 2. Comparison of pedestrian activity detection indicators.

Method	Accuracy%	Recall%
X. Li et al. [2]	91.1	91.2
X. Shu et al. [3]	91.6	91.5
M. Wang et al. [4]	92.2	92.3
H. Yu et al. [5]	93.4	93.4
S. Venugopalan et al. [6]	93.6	93.7
S. Araei et al. [13]	90.8	90.8
X. Shu et al. [16]	95.3	95.4
G. Bertasius et al. [17]	95.6	95.8
ours	**96.1**	**96.1**

We show the results of the pedestrian activity dataset in Figure 6.

The pedestrian activity dataset contains 50 video sequences of various group activities, including the location of pedestrians and their collective and individual behavior tags. Each person's behavior is identified and detected.

Table 2 shows the results of the comparison. Obviously, this method is superior to other networks listed, which shows that our framework is feasible and can improve

performance. The combination of the two further improves accuracy. This shows that DEC and the self-attention mechanism play a key role. At the same time, our results are better than the traditional deep learning model. It can be seen that important people and body parts received more attention. We also express the following confusion matrix, where walking and waiting are almost 100% recognized. On the other hand, jogging and walking are difficult to be clearly defined, which will lead to less-than-expected recognition results because they have similar visual properties.

Figure 6. Qualitative results of the collective activity dataset. The color of the bounding box represents the action label (green: walking, red: waiting).

4.2.3. Results of Football Dataset

The football dataset consists of 64 videos, including 22 players and 6 group activities. The dataset includes a long sequence and short sequence, including the angle change of different cameras, illumination change, scale change, occlusion, motion blur, fast motion, body action, low resolution, etc., under natural conditions. We divided the training set and the test set, and strictly detected each action activity. We used a time window with a length of T = 15. We used 1024 hidden units for the LSTM layer of the two networks. The hidden layer dimension was set to 1024, and different learning rates were used. The learning rate

of the sub layer gradually declined, with a decline rate of 0.95; the experimental results are shown in Figure 7. The confusion matrix is shown in Figure 8.

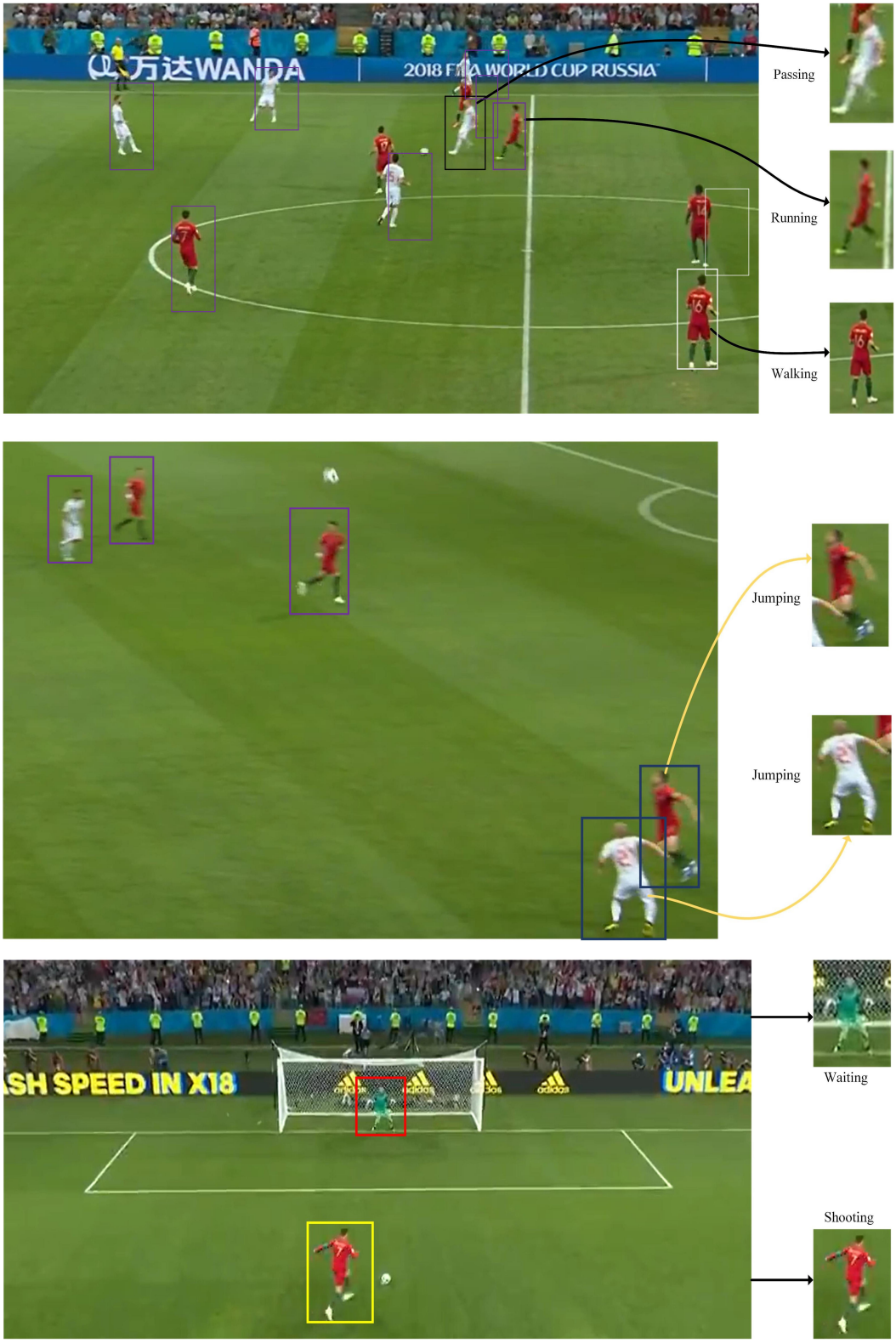

Figure 7. Qualitative results of football dataset. The color of the bounding box represents the action label (white: walking; black: passing; purple: running; blue: jumping; red: waiting; yellow: shooting).

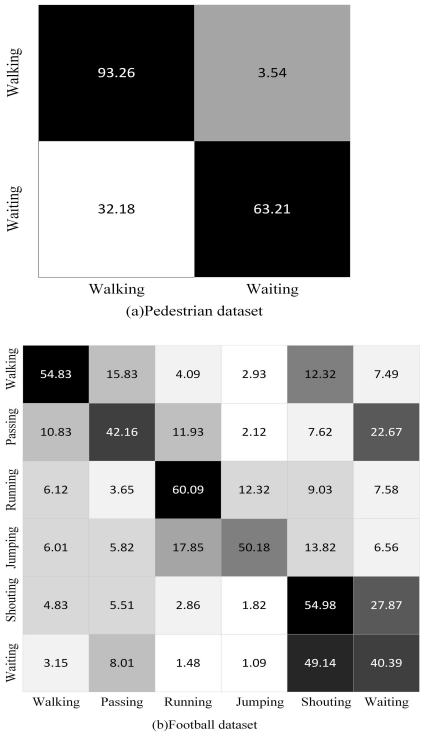

Figure 8. The confusion matrix of two data sets.

Table 3 summarizes the accuracy of our different methods on the football dataset. In general, the data stream of the video data is very stable, that is, a fixed sampling rate. In this context, as long as the number of pictures processed by the algorithm per second is greater than the number of pictures sampled per second, the algorithm is considered capable of real-time processing, with high real-time performance. It can be seen intuitively that our method is superior to a single network model, which shows that the combination of the self-attention mechanism, DEC structure and LSTM network is effective.

Table 3. Comparison results of football datasets.

Method	Accuracy%	Recall%
X. Li et al. [2]	91.2	91.2
X. Shu et al. [3]	91.7	91.5
M. Wang et al. [4]	92.3	92.3
H. Yu et al. [5]	93.4	93.2
S. Venugopalan et al. [6]	93.5	93.6
S. Araei et al. [13]	90.8	90.9
X. Shu et al. [16]	94.3	94.4
G. Bertasius et al. [17]	96.6	96.8
ours	**97.5**	**97.5**

5. Discussion

In the detection and recognition of human behavior, our method achieved good results. Previous researchers improved the YOLOv5 algorithm to detect abnormal human behavior. This method adds a shielding convolutional attention model to the original YOLOv5 backbone network. The module starts with a shielding convolutional layer, and the central area of the receptive field is covered by the nucleus. By predicting the total opening of the screen credit, the error related to the shielding information is used as the abnormal score. At the same time, the Swin CA module [43] is embedded in the detection network. By learning the features of adjacent layers, the model can better grasp the global information, thus reducing the impact of background information on the

detection results. By extracting the scale features of abnormal human behavior in different backgrounds, the complexity of the whole model calculation is reduced, and the accuracy of the model for locating the abnormal human behavior target is improved. However, there are still the following shortcomings: due to the different scenarios collected by different data sets, the characterization of human abnormal behavior is also different in different scenarios, making this method less versatile; secondly, the training of the model needs to label the detected image in advance, and the workload is large in the early stage; and finally, the method is not sensitive to the continuity of human actions in the image, which makes the judgment of human abnormalities in the detection process appear with corresponding delay or false detection and missing detection, which is worthy of our attention and improvement.

In addition, some researchers use the lightweight network MobileNet v2 to replace the original feature extraction network VGG-16, use the deformable convolution module to build a convolution layer to enhance the receptive field, and then integrate the location information into the channel attention to enhance the feature, which can capture the remote dependency between the spatial positions so as to better handle the overlapping occlusion problem. This method mainly focuses on the problems faced by the crowd abnormal behavior detection algorithm. Aiming at the high complexity of the existing model algorithm, it replaces the feature extraction network with the lightweight network MobileNet v2, thereby reducing the model parameters and improving the model running speed. In order to solve the problem of low detection accuracy in complex scenes, such as overlapping occlusion, deformable convolution is used to change the feature extraction method, and the attention mechanism is added to enhance the features. By learning the context relationship, the occlusion part is predicted, thus effectively solving the occlusion problem. Although this method is capable of application, it is particularly insufficient in the face of large-scale data sets. The method proposed by us can solves such problems and continues to be verified in future work.

However, our method also has the following shortcomings: at present, we only carried out experiments on a single football dataset, and it has not been popularized to other large-scale sports datasets. The scenarios used are also different, and the qualitative detection of athletes' behavior is also varied in different scenarios, so the universality of this method is relatively lacking. Secondly, the training of the model needs to detect and label the image video frames in advance, so the workload in the early stage is large. The method in this paper is not sensitive to the continuity of the athletes' behavior in the video images, which may lead to the corresponding missing or wrong detection of the judgment of the athletes' behavior in the recognition process, resulting in wrong results.

6. Conclusions

In this paper, we propose a deep clustering efficient learning network for motion recognition under the self-attention mechanism, which is used in group activity recognition to accurately judge the behaviors expressed by human beings. By adding DEC to LSTM, we can not only solve the problem of gradient disappearance and explosion in RNN, but also capture the internal correlation between multiple people on the sports field for identification, etc. Then, through DEC, we can integrate the motion coding information in key frames, and better analyze and judge the behavior characteristics of athletes. By incorporating the self-attention mechanism, not only can the whole process of the whole sports video be analyzed macroscopically, but also the specific attributes of the movement can be focused, and the key posture features of the athletes can be extracted to further enhance the features, effectively reducing the amount of parameters in the calculation process of self-attention, reducing the computational complexity while maintaining the ability to capture details and improving the accuracy and efficiency of reasoning and judgment. Through verification on large video datasets of mainstream sports, we achieved high accuracy, improved the efficiency of inference and prediction, and proved the superiority of the proposed framework.

In future work, we will continue to optimize our algorithm and improve the model to improve efficiency, complete the real-time classification, capture key frames and reduce the error rate when multi-person behaviors are detected simultaneously, achieve real-time monitoring and information recording, and consider conducting extensive experimental research on other sports data sets to seek greater breakthroughs.

Author Contributions: T.R. conceived and initialized the research, conceived the algorithms, and designed the experiments; Z.Z. evaluated the experiments, reviewed the paper, collected and analyzed the data; T.R. wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Shaanxi Provincial Soft Science Research Plan: "Under the Healthy China 2030 Strategy" Shaanxi Provincial Mass Sports and Health Service Industry Integration and Innovation Research, Project No. 2021MRM147.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fei, H.; Reardon, C.; Parker, L.E.; Hao, Z. Minimum uncertainty latent variable models for robot recognition of sequential human activities. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017.
2. Li, X.; Chuah, M.C. Sbgar: Semantics based group activity recognition. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2876–2885.
3. Shu, X.; Zhang, L.; Sun, Y.; Tang, J. Host–parasite: Graph lstm-in-lstm for group activity recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 663–674. [CrossRef] [PubMed]
4. Wang, M.; Ni, B.; Yang, X. Recurrent modeling of interaction context for collective activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3048–3056.
5. Yu, H.; Cheng, S.; Ni, B.; Wang, M.; Zhang, J.; Yang, X. Fine-grained video captioning for sports narrative. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6006–6015.
6. Venugopalan, S.; Rohrbach, M.; Donahue, J.; Mooney, R.; Darrell, T.; Saenko, K. Sequence to sequence-video to text. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4534–4542.
7. Kong, L.; Qin, J.; Huang, D.; Wang, Y.; Gool, L.V. Hierarchical attention and context modeling for group activity recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 1328–1332.
8. Bagautdinov, T.; Alahi, A.; Fleuret, F.; Fua, P.; Savarese, S. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4315–4324.
9. Ramanathan, V.; Huang, J.; Abu-El-Haija, S.; Gorban, A.; Murphy, K.; Li, F.-F. Detecting events and key actors in multi-person videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3043–3053.
10. Ibrahim, M.S.; Muralidharan, S.; Deng, Z.; Vahdat, A.; Mori, G. A hierarchical deep temporal model for group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1971–1980.
11. Gondal, I.; Murshed, M. Action recognition using spatio-temporal distance classifier correlation filter. In Proceedings of the 2011 International Conference on Digital Image Computing: Techniques and Applications, Noosa, QLD, Australia, 6–8 December 2011; pp. 474–479.
12. Rodriguez, M.D.; Ahmed, J.; Shah, M. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
13. Zhibin, Z.; Liping, S.; Xuan, C. Labeled box-particle cphd filter for multiple extended targets tracking. *J. Syst. Eng. Electron.* **2019**, *30*, 57–67.
14. Araei, S.; Nadian-Ghomsheh, A. Spatio-temporal 3d action recognition with hierarchical self-attention mechanism. In Proceedings of the 2021 26th International Computer Conference, Computer Society of Iran (CSICC), Tehran, Iran, 3–4 March 2021; pp. 1–5.
15. Han, J.; Shao, L.; Xu, D.; Shotton, J. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Trans. Cybern.* **2013**, *43*, 1318–1334. [PubMed]

16. Tome, D.; Russell, C.; Agapito, L. Lifting from the deep: Convolutional 3d pose estimation from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2500–2509.
17. Shu, X.; Tang, J.; Qi, G.-J.; Liu, W.; Yang, J. Hierarchical long short-term concurrent memory for human interaction recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1110–1118. [CrossRef]
18. Bertasius, G.; Park, H.S.; Yu, S.X.; Shi, J. Am i a baller? basketball performance assessment from first-person videos. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 2177–2185.
19. Rossi, S.; Capasso, R.; Acampora, G.; Staffa, M. A multimodal deep learning network for group activity recognition. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–6.
20. Tang, Y.; Wang, Z.; Li, P.; Lu, J.; Yang, M.; Zhou, J. Mining semantics-preserving attention for group activity recognition. In Proceedings of the 26th ACM international conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 1283–1291.
21. Homayounfar, N.; Fidler, S.; Urtasun, R. Sports field localization via deep structured models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2017; pp. 5212–5220.
22. Chen, Z.; Huang, J.; Ahn, H.; Ning, X. Costly features classification using monte carlo tree search. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–8.
23. Zhang, Y.-H.; Wen, C.; Zhang, M.; Xie, K.; He, J.-B. Fast 3d visualization of massive geological data based on clustering index fusion. *IEEE Access* **2022**, *10*, 28821–28831. [CrossRef]
24. Peng, H.; Zhou, S.; Weitze, S.; Li, J.; Islam, S.; Geng, T.; Li, A.; Zhang, W.; Song, M.; Xie, M.; et al. Binary complex neural network acceleration on fpga. In Proceedings of the 2021 IEEE 32nd International Conference on Application-specific Systems, Architectures and Processors (ASAP), Virtual Conference, 7–9 July 2021; pp. 85–92.
25. He, F.; Ye, Q. A bearing fault diagnosis method based on wavelet packet transform and convolutional neural network optimized by simulated annealing algorithm. *Sensors* **2022**, *22*, 1410. [CrossRef]
26. Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; Rohrbach, M. Towards vqa models that can read. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8317–8326.
27. Wang, X.; Wang, C.; Liu, B.; Zhou, X.; Zhang, L.; Zheng, J.; Bai, X. Multi-view stereo in the deep learning era: A comprehensive review. *Displays* **2021**, *70*, 102102. [CrossRef]
28. Singh, A.; Pang, G.; Toh, M.; Huang, J.; Galuba, W.; Hassner, T. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8802–8812.
29. Zeng, G.; Zhang, Y.; Zhou, Y.; Yang, X. Beyond ocr+ vqa: Involving ocr into the flow for robust and accurate textvqa. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, China, 20–24 October 2021; pp. 376–385.
30. Li, M.; Hsu, W.; Xie, X.; Cong, J.; Gao, W. Sacnn: Self-attention convolutional neural network for low-dose ct denoising with self-supervised perceptual loss network. *IEEE Trans. Med. Imaging* **2020**, *39*, 2289–2301. [CrossRef] [PubMed]
31. Zhang, X.; Sun, G.; Jia, X.; Wu, L.; Zhang, A.; Ren, J.; Fu, H.; Yao, Y. Spectral–spatial self-attention networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
32. Cheng, Z.; Yan, C.; Wu, F.; Wang, J. Drug-target interaction prediction using multi-head self-attention and graph attention network. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *19*, 2208–2218. [CrossRef] [PubMed]
33. Qiu, Z.; Yao, T.; Mei, T. Learning spatio-temporal representation with pseudo-3d residual networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5533–5541.
34. Zhang, M.; Xie, K.; Zhang, Y.-H.; Wen, C.; He, J.-B. Fine segmentation on faces with masks based on a multistep iterative segmentation algorithm. *IEEE Access* **2022**, *10*, 75742–75753. [CrossRef]
35. Ning, X.; Xu, S.; Nan, F.; Zeng, Q.; Wang, C.; Cai, W.; Li, W.; Jiang, Y. Face editing based on facial recognition features. *IEEE Trans. Cogn. Dev. Syst.* **2022**. [CrossRef]
36. Zou, Z.; Careem, M.; Dutta, A.; Thawdar, N. Joint spatio-temporal precoding for practical non-stationary wireless channels. *IEEE Trans. Commun.* **2023**. [CrossRef]
37. Zhang, Y.; Mu, L.; Shen, G.; Yu, Y.; Han, C. Fault diagnosis strategy of cnc machine tools based on cascading failure. *J. Intell. Manuf.* **2019**, *30*, 2193–2202. [CrossRef]
38. Shen, G.; Zeng, W.; Han, C.; Liu, P.; Zhang, Y. Determination of the average maintenance time of cnc machine tools based on type ii failure correlation. *Eksploat. I Niezawodn.* **2017**, *19*, 604–614. [CrossRef]
39. Shen, G.; Han, C.; Chen, B.; Dong, L.; Cao, P. Fault analysis of machine tools based on grey relational analysis and main factor analysis. *J. Physics Conf. Ser.* **2018**, *1069*, 012112. [CrossRef]
40. Chu, X.; Lei, J.; Liu, X.; Wang, Z. Kmeans algorithm clustering for massive ais data based on the spark platform. In Proceedings of the 2020 5th International Conference on Control, Robotics and Cybernetics (CRC), Wuhan, China, 16–18 October 2020; pp. 36–39.
41. Wei, R.; Garcia, C.; El-Sayed, A.; Peterson, V.; Mahmood, A. Variations in variational autoencoders-a comparative evaluation. *IEEE Access* **2020**, *8*, 153651–153670. [CrossRef]

42. Zhu, Q.; Tang, X.; Liu, Z. Revised dbscan clustering algorithm based on dual grid. In Proceedings of the 2020 Chinese Control And Decision Conference (CCDC), Hefei, China, 22–24 August 2020; pp. 3461–3466.
43. Huang, M.; Liu, Y.; Peng, Z.; Liu, C.; Lin, D.; Zhu, S.; Yuan, N.; Ding, K.; Jin, L. Swintextspotter: Scene text spotting via better synergy between text detection and text recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4593–4603.