Human Behavior Recognition Method based on Two-layer LSTM Network with Attention Mechanism

Zhijun Gao¹*, Qiaoyu Gu¹b*, Zhonghua Han¹c*

¹School of Information and Control Engineering, Shenyang Jianzhu University, Shenyang 110168, China

*gzj@sjzu.edu.cn, b*gqysyjz@163.com, c*xiaozhonghua1977@163.com

Abstract —Aiming at the problem that the exiting human skeleton-based action recognition methods cannot fully extract the relevant information before and after the action, resulting in low utilization efficiency of skeleton points, we propose a two-layer LSTM (long short term memory) network with attention mechanism. The network has two layers, the first LSTM network is used for skeleton coding and initialization of system storage units and the second LSTM network integrates attention mechanism to further process the data of the first layer network. An algorithm is designed to assign different weights to skeleton points according to the importance of human body, which greatly increases the recognition accuracy. Action classification is accomplished by multiple support vector machines. Through training and testing, the average recognition rate of 98.5% is achieved on KTH dataset. The experimental result shows that the proposed method is effective in human behavior recognition.

1. Introduction

Human behavior recognition as an important branch of artificial intelligence has been widely used in video surveillance, danger warning, human-computer interaction, virtual reality and other fields. Therefore, the research prospect of human behavior recognition is very broad. Compared with other traditional methods, using human skeleton as feature extraction has more advantages, because the skeleton data comes from the human body itself, and is not easily affected by shielding and perspective changes in the recognition process.

At present, the popular methods of human skeleton-based action recognition are based on CNN (convolutional neural network), RNN(recurrent neural network) and GCN(graph convolutional network). Scholars at home and abroad have carried out extensive research on human skeleton-based action recognition; Wang et al. [1] proposed a sequential divergence model to describe the motion state of joints by using graph convolution neural network, which realized effective extraction of spatial features. Liu et al. [2] extracted the motion features of a single joint and the relationship features of multiple joints as comprehensive features of human action recognition. This model can explore the joint features of the perspective of kinematics and space geometry. Jain et al. [3] combined the random forest algorithm with the Spark algorithm library and proposed a multi-random forest weighted large number voting algorithm, which realized the accurate identification of various actions. However, the above methods cannot fully extract the relevant information before and after the action, resulting in low utilization efficiency of skeleton points.

To solve the above problems, this paper proposes a two-layer LSTM network integrating attention mechanism. The first layer network is used to fully extract the spatio-temporal information of skeleton sequence. The data from the first layer will be further processed by the second layer network. An
algorithm is designed to assign different weights to the joints of human in this paper. Finally, Multi-classification SVMs are used to classify human behavior. The training and testing results of KTH data sets show that the proposed method has good accuracy and robustness.

2. Model Framework
The model workflow is shown in Figure 1. In the process of video frames, region locking is carried out first. Next, the spatio-temporal extraction of skeleton sequence will be extracted by the network. In order to highlight the representation information of pivotal joints, attention mechanism was incorporated in this paper to further strengthen the recognition features. Finally, the overall characteristic information will be constantly updated and classified by multiple support vector machines.

Figure 1 Network structure design diagram

3. The Specific Method

3.1. LSTM
LSTM (long short term memory) is a special RNN (recurrent neural network), and the basic unit is shown in Figure 2. Each LSTM unit includes three parts: input gate, output gate and forgetting gate.

Figure 2 The structure of LSTM

The LSTM network will be constantly updating based on the relationship between the current moment and the previous moment. It solves the problem of gradient disappearance and gradient explosion and can fully extract the spatio-temporal sequence relations hidden in the data. It runs as follows:

\[
\begin{align*}
    f_i &= \sigma(W_{fx_i} + W_{by_i}h_{i-1} + b_f) \\
    i_i &= \sigma(W_{ix_i} + W_{bh_i}h_{i-1} + b_i) \\
    C_i &= f_{c_{i-1}} + i_i \tanh(W_{cx_i} + W_{bh_i}h_{i-1} + b_c) \\
    O_i &= \sigma(W_{ox_i} + W_{bh_i}h_{i-1} + b_o) \\
    h_i &= o_i \tanh(c_i)
\end{align*}
\]

In the formula, \(i\), \(o\) and \(f\) represent the input gate, output gate and forgetting gate, \(b\) is the variable weight. At this point, the input gate determines to remember the input data, the output gate determines to output the data, and finally the forgetting gate determines when to remember or forget the value, so that the LSTM neural network can effectively improve the performance of the cyclic neural network.

3.2. LSTM With Attention Mechanism
Attention mechanism can predict the correlation between the input and the result by automatically analyzing the important feature data, which can highlight the pivotal information and improve the
accuracy of recognition. This paper designs a two-layer LSTM network on this basis. The first LSTM network is used to extract the spatio-temporal information of the skeleton, and the second LSTM network integrates the attention mechanism to adaptively assign different weights to the spatio-temporal information of the skeleton from the first layer. A global storage unit is designed to temporarily store the skeleton information of video frames. The network performs several iterations of attention mechanism to progressively improve the system's storage unit. Finally, refined global information will be used for classification. The specific network diagram is as follows:

Figure. 3 Network working diagram

In order to clearly introduce the network of this paper, we expand the two-layer network concretely, omit some of the arrows and joints and focus more on their features while trying to ignore irrelevant features, selectively focusing on the information joints as shown in Figures 4 and 5. In the first layer, the body joints in each frame from time and space are arranged into chains and fed back into the network as sequences. Each LSTM unit has a new input $x_{j,t}$, and the same joint concealment at the time before it is denoted by $(h_{j-1,t})$. The preceding joint at the same time is denoted as $(h_{j-1,t-1})$. $j \in (1,\ldots,J)$ and $t \in (1,\ldots,T)$ represent the collection of joint and frame indexes.

Figure. 4 The first layer network. Figure. 5 The whole network working diagram

The output gate of the LSTM unit is $i_{j,t}$, and the output gate is $O_{j,t}$. The units on both layers will be updated sequentially. They run as follows: $c$ and $h_t$ represent the step size of time and space. $f_{j,t}$ and $f_{j,t}$ are the spatio-temporal dimension of the forgetting gate. $x_{j,t}$ is the input and $W$ is the transformation composed of this model parameters.

$$
\begin{align*}
\begin{pmatrix}
    i_{j,t} \\
    f_{j,t}^{(T)} \\
    f_{j,t}^{(S)} \\
    o_{j,t} \\
    u_{j,t}
\end{pmatrix} &= 
\begin{pmatrix}
    \sigma \\
    \sigma \\
    \sigma \\
    \sigma \\
    \tanh
\end{pmatrix}
\begin{pmatrix}
    x_{j,t} \\
    h_{j-1,t} \\
    h_{j-1,t-1}
\end{pmatrix}
\end{align*}
$$

(6)
\[ c_{j,t} = h_{j,t} \odot u_{j,t} + f^{(S)}_{j,t} \odot c_{j, t-1} + f^{(T)}_{j,t} \odot c_{j,t-1} \]  
\[ h_{j,t} = o_{j,t} \odot \tanh(c_{j,t}) \]  

The input of the second layer is the hidden information from the first layer. If the \( h_{j,t} \) is important to the overall information, more information will be entered to update the second layer. Conversely, if the input is irrelevant, then we need to block the input gate at this step. We pay attention to the expression of important information production action sequence. We use the whole storage unit (\( \Gamma G \)) to implement the attention mechanism, the details are as follows:

\[ \Gamma G^{(0)} = \frac{1}{JT} \sum_{j=1}^{J} \sum_{t=1}^{T} h_{j,t} \]  
\[ e_{j,t} = W_{j} \tanh(W_{e2} \left( \begin{array}{c} h_{j,t} \\ \Gamma G^{(t-1)} \end{array} \right)) \]  
\[ S_{i} = \frac{\exp(e_{i,i}^{(n)})}{\sum_{i=1}^{J} \sum_{v=1}^{T} \exp(e_{i,v}^{(n)})} \]  
\[ \delta = \sum_{j=0}^{n} S_{i} h_{i} \]  

Formula (11) represents the proportion of the current hidden layer output in the storage cell. The higher the value, the greater the "attention" in the whole at the moment. This implements the transition from the current raw hidden layer state to the attention state. Finally, the result is obtained through the calculation of equation (12) and human behavior recognition will be completed by updating the global context unit.

4. Joints Weight Allocation Algorithm
In the process of recognition, all joints involved in human modeling will affect the accuracy of the results, and their contribution is different, so it is very necessary to assign different degrees of weight to joints. In the actual detection, it is found that the characteristics of hand, head, foot and other joints are obvious in the recognition process. They all have one thing in common, they are far away from the center of gravity of the body. We choose the chest joint as the center of gravity and define the line from the coordinates of the remaining joints \((x_{i}, y_{i})\) to the barycenter as \(l_{i}\). In order to be more convincing, we take the average of the sum of the distances at consecutive \((t, t + \Delta t)\) moments, and the calculation method is shown in Equation (13):

\[ l_{i} = \frac{(l'_{i} + l'_{i+\Delta t})}{2} \]  

\( l_{i} \) represents the average distance between the other bone points and the center of gravity. Larger value indicates that the joint is farther from the center of gravity, while smaller value indicates that the joint is closer to the center. It further highlights the role of pivotal points, increases the accuracy.

5. Multi-classification Support Vector Machines
The main idea of SVM is to map the input vector to the high-dimensional feature space through the selected linear and nonlinear mapping, construct the optimal decision function by using the optimization theory, and use the kernel function to replace the dot product operation of the high-dimensional space to avoid the complex calculation. Therefore, the classifier algorithm of SVM is transformed into a quadratic optimization problem and the optimal solution is obtained.
6. The Experimental Results

The experimental data set adopts the KHT data (as shown in Figures 6), which contains six types of action videos of 25 people in four different scenes: Walking, Jogging, Running, Boxing, Waving and Clapping.

In order to ensure the accuracy of the experiment, we randomly divided the KTH data set into three groups. Each time 80% was selected as the training set, 20% as the test set, and finally the average accuracy of the three groups was selected as the evaluation index. The Resize function is used to standardize the image to 224*224 when processing video frames. In terms of network parameters, the hidden layer node sets to 60 and the full connection layer node sets to 15 is the best. We set the number of iterations to 120 and the learning rate to 0.001. In order to facilitate the observation of recognition rate, confusion matrices of various behaviors are made, as shown in the following table:

| behavior       | Walking | Jogging | Running | Boxing | Waving | Clapping |
|----------------|---------|---------|---------|--------|--------|----------|
| Walking        | 98.2%   | 2%      | 0.3%    | 0      | 0      | 0        |
| Jogging        | 1%      | 98.9%   | 1%      | 0      | 0      | 0        |
| Running        | 0.8%    | 1%      | 99%     | 0      | 0      | 0        |
| Boxing         | 0       | 0       | 0       | 98.1%  | 0      | 0        |
| Waving         | 0       | 0       | 0       | 0      | 98.4%  | 0        |
| Clapping       | 0       | 0       | 0       | 1.2%   | 0      | 98.6%    |

As can be seen from the experimental results, this method can accurately identify Walking, Jogging, Running, Boxing, Waving and Clapping. For Walking and Jogging with a high degree of similarity, the recognition rate was above 98%. The average recognition rate reaches 98.5%, which proves that this paper has good recognition accuracy. The training and testing accuracy curves are as follows (the red curve represents the training process and the black curve represents the testing process):

Finally, the proposed method is compared with other joint detection methods on the KTH data, which further proves the advantages of the proposed method.

| Algorithm model                          | Average accuracy % |
|------------------------------------------|--------------------|
| Two Stream-LSTM skeleton detection\[4\]   | 96.3%              |
| LSTM+CNN skeleton detection\[5\]         | 94.2%              |
| Res-CNN skeleton detection\[6\]          | 93.6%              |
| ST-GCN skeleton detection\[7\]           | 95.8%              |
| The method adopted in this paper         | 98.5%              |
7. Conclusion
In this paper, the human behavior recognition method based on a two-layer LSTM network with attention mechanism is proposed. The network has two layers, the first LSTM network is used for skeleton coding and initialization of system storage units and the second LSTM network integrates attention mechanism to further process the data of the first layer network. The model realizes the input and output of time stream data with long time span to ensure continuous global information perception and design an algorithm to assign different weights to skeleton points. Multi-classification SVMs are used to classify human behavior. By testing the KTH data, the accuracy can arrive at 98.5% which higher than other similar human behavior recognition methods and the model also has robustness and real-time performance in practical operation. That greatly prove the method can be an efficient and strong baseline for skeleton-based action recognition. To sum up, this paper can provide a good reference for the application of human behavior recognition field and has a certain value.

References
[1] Wang J, Liu Z, Wu Y, et al. (2014) Learning Action Ensemble for 3D Human Action Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 36: 914-927.
[2] Liu J, Wang G, Duan Y, et al. (2018) Skeleton-Based Human Action Recognition with Global Context-Aware Attention LSTM Networks[J]. IEEE Transactions on Image Processing. 27: 1586-1599.
[3] Jain A, Zamir A R, Savarese S, et al. (2016) [IEEE 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Las Vegas, NV, USA] 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Structural RNN: Deep Learning on Spatio-Temporal Graphs[J]. pp: 5308-5317.
[4] Li C, Zhong Q.Y, Xie D, et al. (2017) Skeleton-based Action Recognition with Convolutional Neural Networks[C]/ 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE. pp: 597-600.
[5] Zhang P.F, Lan C.L, Xing J.L, et al. (2019) View Adaptive Neural Networks for High Performance Skeleton-based Human Action Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 41: 1963-1978.
[6] Li B, Dai Y.C, Cheng X.L, et al. (2017) Skeleton-based Action Recognition Using Translation-scale Invariant Image Mapping and Multi-scale Deep CNN[C]. Proceedings of 2017 IEEE International Conference on Multimedia & Expo Workshops, Hong Kong, China. pp: 4171-4180.
[7] Gao X, Hu W, Tang J.X, et al. (2019) Optimized Skeleton-based Action Recognition via Sparsified Graph Regression[C]. The 27th ACM International Conference on Multimedia, New York, USA. pp: 601-610.