Deep learning for skeleton-based action recognition

Lei Ye¹, Shiqing Ye²*  
¹School of Information Science and Engineering, Fudan University, Shanghai, province, 200000, China  
²School of Information Science and Engineering, Fudan University, Shanghai, province, 200000, China  
author’s e-mail: 17210720058@fudan.edu.cn

Abstract. In order to further improve the recognition rate of skeleton action recognition, and to break the limitation that most of the previous deep learning-based methods have the input content of human joint coordinates, we propose a human skeleton action recognition algorithm combining geometric features and LSTM networks. In this paper, we propose a human skeleton action recognition algorithm that combines geometric features with LSTM network. The algorithm selects geometric features based on the distances between joints and selected lines as the input of the network, and introduces a time-selective LSTM network for training. The ability to select the most recognizable temporal features using the time-selective LSTM network is demonstrated in the SBU Interaction dataset and UT Kinect dataset, achieving 99.46% and 99.30% recognition rates, respectively. The experimental results demonstrate the effectiveness of the method for human skeleton-based action recognition.

1. Introduction

Human action recognition is an important part of the computer vision field and has a wide range of applications in the fields of intelligent surveillance, human-computer interaction, and motion analysis. Traditional RGB video lacks the flexibility to achieve human positioning and scale invariance due to the lack of 3D spatial information. The human skeleton utilizes high-level representational capabilities to describe movements through the 3D coordinate positions of key joints, which are robust to changes in epiphenomena and perspectives. In addition, the maturity of skeleton real-time estimation algorithms has further advanced this technique [1].

Current research on action recognition algorithms can be divided into two categories: traditional methods [2-4] and deep learning methods [5-9]. Many traditional studies based on skeletal action recognition have attempted to build features using geometric relationships of the spatial structure of the human skeleton. Chen et al. [2] enumerated nine geometric features, including eight static features and one temporal feature. Static features encode an action pose form, and temporal features are used to represent changes in time. Vemulapalli et al. [3] used rotation and translation in 3D space to represent the 3D geometric relationships between different body parts. Shao et al. [4] proposed a class of integral invariants describing the motion trajectory to achieve an efficient and robust motion trajectory matching. Zhang et al. [10] extracted a set of geometric features, including the distance between joints and the distance from the joint to the plane formed by the joint, to describe the posture and motion.

Deep learning RNN (Recurrent Neural Networks) has a strong ability to learn the relevance and dynamics of sequential input data, and has been successfully applied to human skeleton action recognition with good results. Veeriah et al. [7] proposed a differential LSTM (Long Short-Term
Memory) network that adds a new gate inside the LSTM to track the derivatives of the memory state and thus learn the skeleton node changes between consecutive frames. Liu et al.\cite{8} proposed a two-dimensional spatio-temporal LSTM (Spatio-Temporal LSTM) framework with Trust Gates. The LSTM is extended to the spatio-temporal domain by a tree-based traversal method, and trust gates are introduced to handle inaccurate input data. Song et al.\cite{9} introduced an LSTM-based end-to-end spatial and temporal attention model that learns to selectively focus on skeletal discriminative joints in each frame of the input and gives different levels of attention to the output of different frames. Liu et al.\cite{11} proposed a Global Context-Aware Attention LSTM (GCALSTM) network that can selectively focus on the joints containing the most information in the action sequence with the help of global contextual information.

In this paper, we introduce a temporal selection network based on LSTM, which has a first layer of LSTM and a second layer of temporal pooling with the ability to select the most recognizable time period features. On this basis, inspired by Zhang et al.\cite{10}, the geometric feature that has the best recognition result among the set of tested geometric features, i.e., based on the distance between the joint and the selected line, is selected as the input to this network. The experimental results show that the algorithm has a high recognition rate, which proves the effectiveness of the method.

2. Feature Extraction

2.1. Data pre-processing

Firstly, all joint coordinates of the original human skeleton sequence data are transformed from the camera coordinate system to the human coordinate system, thus achieving viewpoint independence. The center of the hip joint is chosen as the origin of the body coordinates, and the X-axis is a 3D vector parallel to the "right shoulder" to the "left shoulder", so that the Y-axis is parallel to the 3D vector from the "center of the shoulder" to the "center of the hip". The Z-axis is then X × Y. Then the normalization is performed to scale all 3D points based on the sum of the skeleton chain distances, achieving scale independence.

2.2. Geometric feature representation

In this paper, we consider using geometric features for each frame for spatial modelling, using geometric features as input instead of joint coordinates. Based on the results of Zhang et al.'s study\cite{10}, the distance between the joint point and the selected line was chosen to model the human skeleton model for a single frame. A typical human skeleton model with 16 joints is used as an example, as shown in Figure 1, and each joint \( J \) is encoded using the coordinates \((J_X, J_Y, J_Z)\). Regarding the choice of lines, if any two joints are connected, there are 120 lines in total. To reduce the cost of calculation, only some important lines were selected from them. The line from joint point \( J_1 \) to joint point \( J_2 \) is denoted by \( L_{J1 \rightarrow J2} \). The selected lines are divided into 3 types, which are shown in Figure 2. The 3 types of lines are defined as follows:

1. The joint node \( J_1 \) and the joint point \( J_2 \) are directly adjacent in the skeleton chain. A total of 15 lines of this type.
2. Joint node \( J_1 \) and joint point \( J_2 \). One of them is at the end of the skeleton chain (left hand or right hand, left foot or right foot). The other one is two nodes away from it in the skeleton chain (head → chest, right hand → right shoulder, left hand → left shoulder, right hip → right foot, left hip → left foot). A total of 5 lines of this type.
3. Both the joint node \( J_1 \) and the joint point \( J_2 \) are at the end of the skeleton chain, and there are 10 lines of this type.

Finally, 30 lines are selected to be combined with the nodes to form the geometric features. The distance from the joint point to the selected line is defined as \( J_{L,d} \), and the distance from the line to the line is \( J_{L,d} \). The schematic diagram of \( J_{L,d} \) is shown in Figure 3, and the related calculation is shown in equation (1). Delete the features in which the line is common, such as \( J_{L,d}(J_1, L_{J1 \rightarrow J2}) \).

\[
J_{L,d}(J, L_{J1 \rightarrow J2}) = 2S_{JJ1J2} / JJ_d(J_1, J_2)
\]  
(1)
3. Time-Selected LSTM Networks

To solve the vanishing gradient problem of RNNs\(^{[12]}\), an advanced architecture LSTM was proposed. Its main idea is to introduce a gate mechanism to better control the flow of information. It compared to simple RNN neurons, LSTM neurons have a complex gate mechanism that relies on a memory mechanism controlled by input. LSTM relies on memory cells controlled by input, forgetting and output gates to store and output information. The structure of the LSTM neuron is shown in Figure 4. The definitions are as follows

\[
I_t = \sigma(W_i x_t + W_i h_{t-1} + W_i c_{t-1} + b_i),
\]

\[
f_t = \sigma(W_f x_t + W_f h_{t-1} + W_f c_{t-1} + b_f),
\]

\[
c_t = i_t \circ \text{tanh}(W_c x_t + W_c h_{t-1} + b_c) + f_t \circ c_{t-1},
\]

\[
o_t = \sigma(W_o x_t + W_o h_{t-1} + W_o c_t + b_o),
\]

\[
h_t = o_t \circ \text{tanh}(c_t),
\]

where \(i_t, f_t\) and \(o_t\) correspond to the outputs of the input gate, forgetting gate and output gate at time \(t\), respectively; \(W\) and \(b\) are the weight matrix and bias vector of the corresponding gate, respectively; \(\sigma\) denotes the sigmoid activation function. The established network model is shown in Figure 5. The first layer of the model structure is the LSTM layer, which replaces the joint coordinates with geometric features as the input \(x_t\) to the network. In this way, the output \(h_t\) at each moment can be considered as an intermediate representation of all previous frames, and since the LSTM has an oblivion gate, it can also be considered as a representation after removing irrelevant frames from previous frames. The output of all time points is combined by the following equation

\[
H = [h_1, h_2, h_3, \ldots, h_T].
\]

The second layer of the network structure is the temporal pooling layer. If the skeleton sequences are not sampled, the dimensionality of \(H\) may be different since the number of frames \(T\) may be different for different skeleton sequences. Therefore, it can be used to perform temporal maximum pooling to obtain the final fixed-length features. The pooling method is as follows

\[
q_j = \max[h_{j,1}, h_{j,2}, \ldots, h_{j,T}], \quad 1 \leq j \leq N
\]

In this way, \(q_j\) can then express the most discriminative time period features in \(H\). Finally, \(Q = [q_1, q_2, \ldots, q_N]\) is first processed using the ReLU activation function, and then sent to a fully connected layer with a softmax layer for classification.
4. Experimental operation

4.1. Style and spacing
In this paper, we validate the proposed skeleton action recognition algorithm on the SBU Interaction dataset [13] and the UT Kinect dataset [14]. Due to the different number of nodes in different datasets, the available geometric features have dimensions of 1,624 and 612 on the SBU Interaction dataset and UT Kinect dataset, respectively.

In the experiments, the negative log-likelihood loss function is used as the training loss function, and stochastic gradient descent is used to solve the optimization problem. In this paper, the learning rate is set to 0.1, the learning rate decay is set to $1 \times 10^{-5}$, and the momentum and weight decay are set to $1 \times 10^{-4}$ and $1 \times 10^{-5}$, respectively.

4.2. Experimental results
The SBU Interaction dataset is captured by Kinect and consists of 7 participants performing 8 types of two-person interactions. It has 282 skeleton sequences totaling 6,822 frames. Each skeleton has 15 joints. There are two challenges with this dataset. First, in most of the interactions, one person is acting while the other is reacting. Second, the accuracy of many of these coordinates is very low. The experiments on the SBU Interaction dataset used the standard experimental protocol provided in the literature [13], and the comparative results are shown in Table 1.

| Method                  | Recognition rate | Method                  | Recognition rate |
|-------------------------|------------------|-------------------------|------------------|
| Yun et al. [12]         | 80.30%           | ST-LSTM [8]            | 93.30%           |
| Ji et al. [13]          | 86.90%           | GCA-LSTM [11]          | 94.10%           |
| CHARM [16]              | 83.90%           | Zhang et al. [10]      | 99.02%           |
| HBRNN-L [17]            | 80.35%           | Our method             | 99.46%           |

Table 1. Skeleton-based action recognition performance comparison on SBU Interaction dataset

The UT Kinect dataset contains 200 sequences consisting of 10 action categories, obtained by capturing the actions of 10 people in different viewpoints with a fixed Kinect. Each subject performs each action twice, and each frame contains 20 skeletal joints. This dataset is challenging due to the large intra-class variability and viewpoint diversity. There are two popular protocols for this dataset. One is the protocol as in the literature [16] and the other is the protocol used in the literature [17]. In this paper, the protocol of [17] is used, and the comparison results are shown in Table 2.
Table 2. Skeleton-based action recognition performance comparison on UT-Kinect dataset

| Method                | Recognition rate | Method                | Recognition rate |
|-----------------------|------------------|-----------------------|------------------|
| Skeleton Joint Features\cite{17} | 87.90%           | Lie Group\cite{3}    | 93.60%           |
| Elastic functional coding\cite{15} | 94.90%           | GCA-LSTM\cite{11}   | 98.50%           |
| ST_LSTM\cite{8}       | 95.00%           | Zhang et al.\cite{10} | 95.96%           |
| Our method            | 99.30%           |                       |                  |

As can be seen from the two comparison tables above, the recognition rate of the model proposed in this paper is higher than previous non-deep learning approaches on both datasets and exceeds the recognition rates of most recent deep learning-based models. By comparison, it can be observed that the recognition rate on the SBU Interaction dataset is significantly better than most previous methods that used joint coordinates as network input.

This is because in this dataset, the relationship implied by the actions of two people interacting is richer than that of one person, and it is easier to find the relationship by using geometric features as input instead of joint coordinates. The model also achieves a higher recognition rate on the UT Kinect dataset than the previous method with some improvement.

4.3. Experimental analysis

4.3.1. Effect of the number of LSTM neurons

The effect of the number of LSTM neurons on recognition accuracy is verified in this paper, and the relationship is shown in Figure 6. The experiments show that the number of neurons has little effect on the recognition rate after the number of neurons exceeds 200. Only a little attention is needed, the number of neurons should be roughly proportional to the dimensionality of the input features. For all data sets, the number of LSTM neurons is uniformly set to 512.

4.3.2. Influence of the number of features

In this section the effect of the number of features on the recognition accuracy is verified. Experiments were conducted on the UT Kinect dataset to test the recognition rates corresponding to different number of features, and the relationship is shown in Figure 7. From the figure, it can be observed that the recognition rate rises rapidly and slows down at a certain number; when the number of features exceeds 500, there is no significant improvement in the recognition rate. Therefore, it can be concluded that only some of the input features are valid.

5. Conclusion

To improve the recognition rate of human skeleton action recognition, this paper proposes a skeleton action recognition algorithm that combines geometric features with LSTM networks based on deep
learning. The method selects geometric features based on the distance between a joint and a selected line to describe the human skeleton, and inputs them into a time-selective LSTM network for training. Through experiments, the algorithm achieves 99.36% and 99.30% recognition rates on the SBU Interaction dataset and UT Kinect dataset, respectively, and the experimental results demonstrate the effectiveness of the algorithm.

References
[1] Shotton J, Sharp T, Kipman A, et al.(2013)Real-time human pose recognition in parts from single depth images. J. Communications of the ACM., 56(1):116-124.
[2] Chen C, Zhuang Y, Nie F, et al.(2011) Learning a 3D human pose distance metric from geometric pose descriptor. J. Transactions on Visualization & Computer Graphics., 17(11): 1676-1689.
[3] Vemulapalli R,Arrate F,Chellappa R. (2014) Human action recognition by representing 3D skeletons as points in a Lie group, J. Conference on Computer Vision and Pattern Recognition., 35(7):44-49.
[4] Shao Z,Li Y.(2020)Integral invariants for space motion trajectory matching and recognition. J. Pattern Recognit., 48(8),2418-2432.
[5] Chen Shengdi, Wei Wei, He Bingqian, et al. (2019) Action recognition base on improved deep convolutional neural network. J. Application Research of Computers., 36(4):1-7.
[6] Liang Yuqiang, Cheng Jinjie, Ye Qihan.(2017) Robot motion understanding based on AdaBoost and BP networks. J. Electronic Science and Technology., 30(8):63-66.
[7] Veeriah V, Zhuang N, Qi G J.(2015) Differential recurrent neural networks for action recognition. J. IEEE International Conference on Computer Vision., 336(35):627-638.
[8] Liu J,Shahroudy A, Xu D, et al.(2016) Spatio-temporal LSTM with trust gates for 3D human action recognition. J.14th European Conference on Computer Vision (ECCV)., 77(19):202-206.
[9] Song S,Lan C, Xing J, et al.(2017) An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In :31st AAAI Conference on Artificial Intelligence, San Francisco, pp.763-770.
[10] Zhang S, Yang Y, Xiao J,et al.Fusing .(2018) Geometric features for skeleton-based action recognition using multilayer LSTM networks. J.IEEE Transactions on Multimedia., 20(9):2330 - 2343.
[11] Liu J,Wang G, Hu P,et al.(2017).Global context-aware attention LSTM networks for 3D action recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Hawaii,pp-254-259.
[12] Yun K,Honorio J, Chattopadhyay D, et al. (2012) Two-person interaction detection using body-pose features and multiple instance learning. In :IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Rhode Island,pp.333-341.
[13] Ji Y, Ye G, Cheng H.Interactive. (2014) Body part contrast mining for human interaction recognition. In :IEEE International Conference on Multimedia and Expo Workshops, Chengdu,pp.301-305.
[14] Li W, Wen L, Chuah M C, et al.Category. (2015) Blind human action recognition: a practical recognition system. In: IEEE International Conference on Computer Vision, Santiago, pp. 99-105.
[15] Du Y, Wang W, Wang L.Hierarchical. (2015) Recurrent neural network for skeleton based action recognition. In : IEEE Conference on Computer Vision and Pattern Recognition, Boston, pp. 23-29.
[16] Zhu Y, Chen W, Guo G.Fusing.(2013) Spatiotemporal features and joints for 3D action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland. pp. 84-87.
[17] Anirudh R, Turaga P, Su J, et al. (2015) Elastic functional coding of human actions: from vector-fields to latent variables. In: IEEE Conference on Computer Vision and Pattern Recognition, Boston. pp. 156-169.