IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 19, 2020, accepted April 28, 2020, date of publication May 6, 2020, date of current version May 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2992740

Global Spatio-Temporal Attention for Action Recognition Based on 3D Human Skeleton Data

YUN HAN[1], SHENG-LUEN CHUNG[2], QIANG XIAO[3], WEI YOU LIN[2], AND SHUN-FENG SU[2], (Fellow, IEEE)

[1]School of Computer Science, Neijiang Normal University, Neijiang 641100, China
[2]Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei 10607, Taiwan
[3]School of Foreign Languages, Neijiang Normal University, Neijiang 641100, China

Corresponding author: Yun Han (han198010@163.com)

This work was supported by the Scientific Research Fund of Sichuan Provincial Education Department under Project 16ZA0315.

ABSTRACT The human skeleton joints captured by RGB-D camera are widely used in action recognition for its robust and comprehensive 3D information. Presently, most action recognition methods based on skeleton joints treat all skeletal joints with the same importance spatially and temporally. However, the contributions of skeletal joints vary significantly. Hence, a GL-LSTM+Diff model is proposed to improve the recognition of human actions. A global spatial attention (GSA) model is proposed to express the different weights for different skeletal joints to provide precise spatial information for human action recognition. The accumulative learning curve (ALC) model is introduced to highlight which frames contribute most to the final decision making by giving varying temporal weights to each intermediate accumulated learning results. By integrating the proposed GSA (for spatial information) and ALC (for temporal processing) models into the LSTM framework and taking the human skeletal joints as inputs, a global spatio-temporal action recognition framework (GL-LSTM) is constructed to recognize human actions. Diff is introduced as the preprocessing method to enhance the dynamic of the features, thus to get distinguishable features in deep learning. Rigorous experiments on the largest dataset NTU RGB+D and the common small dataset SBU show that the algorithm proposed in this paper outperforms other state-of-the-art methods.

INDEX TERMS Human action recognition, global attention model, accumulative learning curve, LSTM, spatio-temporal attention.

I. INTRODUCTION

Human action recognition has a wide range of applications [1], such as human-computer interaction, video surveillance, health care, entertainment, etc. Its application has become one of the research hotspots in the field of computer vision [2]. After several decades of development, research on human action recognition has made a series of important progress [3], among which two are most influential. One is the change of the type of information used, from the traditional RGB to the current and popular RGB-D. RGB-D not only contains RGB and depth information, but also extracts the 3D skeleton joints and expresses the movements of the human body more concisely and accurately. For example, Chen *et al.* [4] constructed the depth action maps by using depth camera Kinect, Evangelidis *et al.* [5] used skeletal quads feature for action recognition from Kinect, Ohn-Bar *et al.* [6] combined depth information and skeleton joints for human action recognition, and Saini *et al.* [7] used the skeleton joints to accomplish the interaction monitoring system between two persons. The other is the transformation from traditional learning to today's deep learning. Deep learning is goal-oriented and automatic, and the recognition effect is significantly better than traditional methods. For example, Zheng *et al.* [8] used deep learning to capture the long-term global motion dynamics in action sequences; Li *et al.* [9] adopted CNN and depth motion maps to accomplish real-time human action recognition. Zhang *et al.* [10] compared the result of different skeletal joints features based on LSTM. Zhu *et al.* [11] proposed the co-occurrence feature to enhance the result of action recognition. However, most of the current action recognition based on deep learning treats each

The associate editor coordinating the review of this manuscript and approving it for publication was K. C. Santosh.

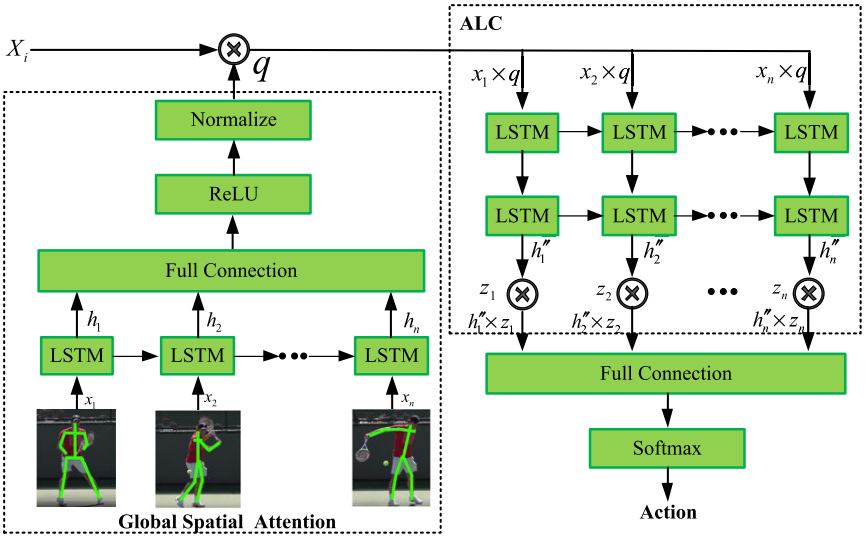

FIGURE 1. The proposed Global Spatio-Temporal Attention framework.

skeleton joint the same spatially, and gives each frame the same weight temporally. Intuitively, this is not in line with human cognition. In the sequence of actions, it may be completely different for the importance of each frame sequence to the recognition action; so does the effect of each joint on different actions. In response to this problem, the mainstream practice at present is to embed the attention model into deep learning. Attention is a concept proposed in cognitive neurology. It indicates that when the human brain receives external information, consciously or unconsciously, it selects a small portion of useful information to focus on and ignores the rest [12]. Sharma *et al.* [13] first proposed the attention model and achieved significant results. At present, in action recognition, an influential model is the spatio-temporal attention model STA-LSTM proposed by Song *et al.* [14]. It feeds the previous frame and the current frame to the LSTM framework to calculate the weight of each joint (spatial attention) and each action (temporal attention). The weight of each frame and each joint essentially reflects the significant change in the relationship between the two frames before and after, indicating the change of the local information of the action. This kind of attention is termed as ''local attention.''

The main disadvantage of the local attention model is that it only uses the local changes of the action sequence to obtain the attention weight. However, this is inconsistent with human cognition and accurate attention weight is not easy to obtain. Universally, only after reading the entire sequence of actions in a complete way, can it be reliable to determine which moments of action are more important and which joints weight greater in action recognition. Inspired by this observation, the present paper proposes a global spatio-temporal attention model as shown in Figure 1, which takes all frames of each action as inputs and obtains the weight of each joint for action recognition. Meanwhile, a weight parameter is given to each frame action, and the parameter value reflecting the weight of the frame is obtained by global learning. Spatially, a global attention model is used to determine the weight of each joint after observing the complete sequence of action. Temporally, the ALC model is used to determine the importance of each frame sequence. That is, data training is used to determine the weight of each frame action. Thus, this paper constructs global spatial attention with ALC action recognition model GL-LSTM.

It is also observed that in action recognition, the dynamic plays an important role. The traditional method based on LSTM mainly obtains the temporal relationship of actions, but insufficient attention is paid to the dynamic. Therefore, to improve the effect of action recognition, this paper introduces Diff [15] as the preprocessing method to strengthen the dynamic of features.

Compared with our previous work [16], there are two major improvements in this paper. First, it provides rich experimental results and detailed experimental analysis on NTU RGB+D dataset. In order to verify the adaptability of the algorithm, it is tested on the common small dataset SBU. Second, it introduces the Diff [15] feature in the preprocessing stage to enhance the dynamic of the features and get more distinctive features by deep learning for higher accuracy rate. The main contributions of this paper are:

1) Global spatial attention model: Different from the traditional local attention model, the global spatial attention model determines the importance of each skeleton joint from the perspective of the entire action sequence and therefore can express the importance of the skeletal joints more accurately.

2) Accumulative learning curve model: This model introduces a set of weight parameters that reflect the importance of each frame in the normal LSTM architecture. By using direct data training, the importance of each frame can be obtained effectively.

3) The spatio-temporal attention model and LSTM-based action recognition framework are integrated organically to

construct a simple and convenient training model, and the recognition effect is classical.

4) The method proposed in this paper combines human experience and deep learning to construct features, so that deep learning can focus on more distinctive features. According to human experience, dynamic features have a better effect on action recognition. Therefore, Diff is proposed as the basic feature of deep learning, which significantly improves the effect of action recognition.

II. RELATED WORK

The main focus of this paper is to construct an effective attention model to improve the effect of action recognition under the framework of RNN. Therefore, this section will introduce related works regarding the action recognition method based on RNN network and the application of attention mechanism in action recognition. At the same time, the method of using handcrafted features to enhance the dynamic characteristics is also introduced.

A. RNN FOR ACTION RECOGNITION

Recurrent Neural Network (RNN) is a popular model that uses the recurrent structure for sequential data modeling and feature extraction. It is mainly used to deal with timing characteristics of a sequence, such as speech recognition [17], machine translation [18] and motion recognition. The emergence of LSTM effectively solved the problem of poor memory of RNN, thus it gained wide attention and became one of the main methods to solve the sequence problem.

At present, the most frequently used model in action recognition is LSTM despite its' multifarious variants, such as Bidirectional LSTM [19], GRU [20]. At the same time, with the introduction of a depth camera, human skeleton joint is widely used in motion recognition for its simple structure and rich connotation. Therefore, human skeleton joints are often fed directly into the LSTM architecture to realize action recognition.

So far, the most frequently used structure in this regard is the ''LSTM+full connection layer+softmax layer.'' LSTM is used to obtain the temporal characteristics of actions, and full connection layer and softmax layer are used to classify actions. There are two typical LSTM concatenated architectures according to the choice of information location. One takes only the output of the last moment of the LSTM as the feature (referred to as the basic LSTM) and extracts the timing feature of the whole action sequence. This method has been widely used. However, for the action with a long-time sequence, some information about earlier time may be lost, which will affect its performance to some extent. The other uses the output information of each time as the final feature, which means the output of each time plays the same role in action recognition (hereinafter referred to as Equal Weight). This method is rarely used due to the fact that it is not particularly suitable to deal with such problems as action recognition. Therefore, most of the current LSTM based action recognition methods are based on the first structure. Zhang *et al.* [10] takes the joints of the human skeleton as input, and uses three-layer LSTM to realize the classification of actions, and obtains satisfactory results. In order to learn the effect of each segment of human joints in action recognition more precisely, Shahroudy *et al.* [21] proposed a P-LSTM network which divides human joints into five parts; each part is saved by a cell, and five cells are connected in series as the final cell. Zhu *et al.* [11] put forward the co-occurrence structure of skeleton joints on the LSTM framework, which related different actions to the corresponding joints and classifies the actions and joints together. In order to eliminate the influence of perspective change on action recognition, Zhang *et al.* [22] proposed view adaptive neural networks to achieve perspective alignment, and achieved good results. Sharma *et al.* [13] introduced the attention model on the basic LSTM to imitate human action recognition, which gained wide attention in the research of action recognition.

B. ATTENTION MECHANISM

Attention is a complex cognitive ability that human beings are born with. It is the ability to pay attention to some information while ignoring others [12]. The attention mechanism was first proposed by Itti *et al.* [23] in computer vision, which is mainly used to express the importance of different information. At present, attention model is widely used in various fields of deep learning, such as natural language processing [24], object recognition [25], [26], speech recognition [27], and so on. In the field of action recognition, Sharma *et al.* [13] has applied the attention model to extract the attention weight of each frame of image based on human joints, and then uses this parameter in the CNN framework based on RGB. Baradel *et al.* [28] proposed pose-conditioned spatio-temporal attention based on human joint data, which effectively improves the recognition effect. Li *et al.* [29] embedded soft attention model on LSTM. Yang *et al.* [30] introduced the attention model on the basis of the skeleton map. At present, an important attention model is STA-LSTM, which is proposed by Song *et al.* [14]. It has two parts: temporal attention and spatial attention, and both of them are calculated by the changing relationship between two adjacent frames. It is clear in structure and effective in the recognition effect.

The proposed GL-LSTM method in this paper differs from STA-LSTM as follows: first, STA-LSTM constructs a local attention model based on the relationship of two adjacent frames. The proposed GL-LSTM calculates attention weight based on all sequences of the whole action, that is, global attention model. Second, STA-LSTM is composed of three independent parts (Temporal attention, Spatio attention and LSTM main network) and it is difficult to achieve optimal results due to its complex training process. In contrast, GL-LSTM integrates the three parts organically, and constructs a simple network structure, which makes the training process relatively simple and re-usable.

C. HANDCRAFTED DYNAMIC FEATURE

In essence, action recognition mainly depends on two features: the static feature that reflects the appearance and shape of the human body, and the dynamic feature that reflects the changes of limbs. Among them, the dynamic characteristics play an important role in action recognition, especially in the movement with drastic changes of limbs. In RGB image sequence, the dynamic features are often achieved by optical flow technology [31] or trajectory technology [32], but this kind of method generally requires a large amount of computation. In 3D human joints, dynamic features are often constructed on the relationship between joints. Zhang et. al. [10] proposes geometric relational Joint-Line Distance features based on distances between joints and selected lines. Being applied to a cascade of three LSTM, these features and their variations attain better results than using raw skeleton data. JDM-CNN [33] codes pair-wise distances [34] between joints over a sequence of single or multiple person skeletons into color variations to capture temporal information. Based on human cognition, these basic features are constructed by hand, and then dynamic features are acquired by deep learning. It is better than modifying the deep network, and can be seen as a way of human-computer intelligence integration.

III. LSTM WITH GLOBAL SPATIAL ATTENTION AND ALC

The Global Spatio-Temporal Attention Model is proposed in order to express both spatial and temporal attentions as shown in Figure 1. The global spatial attention model will serve the purpose of spatial attention, while an accumulative learning curve (ALC) model for temporal attention. These two models are integrated into the LSTM framework for action recognition.

A. GLOBAL SPATIAL ATTENTION MODEL

The configuration of human joints expresses the spatial distribution of different joints of the human body, and is distinctive for the classification of action types. For different types of actions, the function of each joint in action recognition may be different. That is, the movement of one certain joint may be enough to determine some actions, while for other actions some other joints may be necessary. In order to express these ideas, the traditional method is to use the local spatial attention model as in (1).

$$X'_t = X_t a_t \tag{1}$$

where $X_t = (X_{t,1}, \ldots, X_{t,K})$ represents the input data at time t, K represents the number of skeleton joints in the human body, $a_t = (a_{t,1}, \ldots, a_{t,K})$ is weight at time t, and each $a_{t,i}$ represents the weight given to a skeleton joint k at time t. The local spatial attention weights are functions of the current frame and the previous frame, as expressed below:

$$a_t = f_s(x_{t,t-1}) \tag{2}$$

In this paper, it is believed that the importance of each joint in action recognition can be effectively determined only after the entire action sequence has been looked through. This spatial attention is called global spatial attention, which is expressed as:

$$X''_t = X_t q \tag{3}$$

Here, $q = (q_1, \ldots, q_K)$ is the attention weight of each skeleton joint, and K is the number of human joints. (3) is to determine the attention weight applicable to all frames of the current action after looking through the entire action sequence. That is, in the same action sequence, the same joint has the same weight in different frames, and different joints have different weights.

The weight of global spatial attention can be achieved by using the backpropagation algorithm to approximate the function by sending the LSTM outputs for all moments h_t together into a deep learning framework. $s = ReLU(\sum_{t=1}^{n} w_{hs} h_t + b_s)$ is used to approximate the joint weights needed. Further, normalization of the weights is necessary to prevent the fluctuation due to the data size variation. In other words, $q_i = 1 + \frac{\exp(s_i)}{\sum_{j=1}^{K} \exp(s_j)}$ represents the weight of each joint. As shown in Figure 2 in the dashed box, all the LSTM outputs h_t are first sent to the Full Connection, then to the ReLU to enhance the nonlinearity of the structure, finally to the Normalize to prevent data dispersion. Eventually, the joint points and weight of each frame q are multiplied to get the joint point data sequence with spatial attention X''_t. After that, it will be sent to the subsequent network structure to extract features and classification for action recognition.

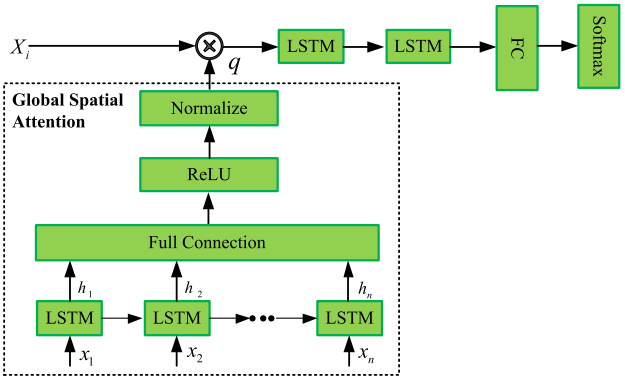

FIGURE 2. LSTM with Global Spatial Attention Model.

B. ACCUMULATIVE LEARNING CURVE (ALC) MODEL

At present, most action recognition methods attach the same importance to each frame in an action sequence. However, only a few key frames are important for human behavioral judgment. Less relevant frames are therefore ignored. This phenomenon is called temporal attention. The use of the additive learning curve model is proposed to express temporal attention. That is, when using the LSTM framework to identify actions, not only the output feature of each frame h''_t is considered, but also the different importance of each h''_t which

is represented by weight z_t of each h_t'', using $(z_1 h_1'', \ldots, z_t h_t'')$ to represent the features of the entire action video.

When inputting the actions in the video into the learning process of the LSTM network from the beginning, at output of LSTM, there will be a cumulative learning effect h_t'' from the initial time to the current time t at each moment. In other words, the temporal characteristics of the actions are extracted from the initial time to the current time t.

On the other hand, in order to consider the contribution of each cumulative learning effect, a separate network (see Figure 4) is designed to train its corresponding weights z_t. In general, not only the learning results h_t'' from the initial time to the current, but also the corresponding weights z_t are obtained. Therefore, the curve of resulting weights z_t (over time) is named as an accumulative learning curve, referred to as ALC, and Figure 3 is a typical ALC curve.

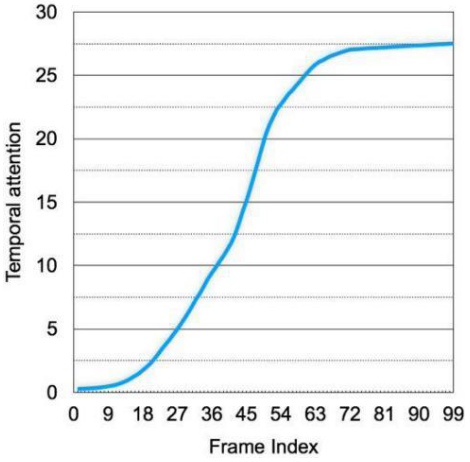

FIGURE 3. ALC curve.

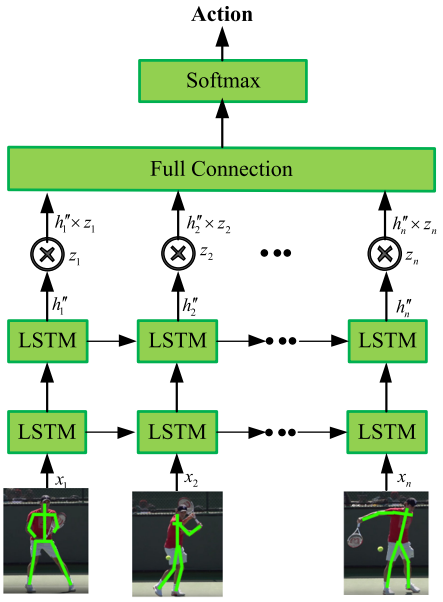

FIGURE 4. ALC model.

ALC, as shown in Figure 4, not only shows different effects of different h_t'' on action recognition, but also provides a more suitable way to use the Equal Weight framework. Let $o_t = z_t h_t''$, where $h_t'' = LSTM(LSTM(x_t))$ due to the use of a two-layer LSTM framework (c.f. Figure 4). Let $g = \sum_{t=1}^{n} w_{h\sim} o_t + b_\sim$ to represent the full connection layer, let $p(y = i|X) = \frac{\exp(g_i)}{\sum_{j=1}^{C} \exp(g_j)}$, $i = 1, \ldots C$ indicates the effect of the final Softmax classification. It shall be noted that the weights z_t are obtained directly by training, and thus all the video sequences need to be normalized to the same length. The desired feature can be obtained by multiplying z_t and h_t'' in the final test phase. Therefore, ALC z_t reflects the temporal weight distribution of most action sequences.

C. INTEGRATION OF SPATIAL MODEL AND TEMPORAL MODEL

To further improve the recognition effect, spatio-temporal attention framework as shown in Figure 1 is constructed by integrating the global spatial attention model and accumulative learning curve model. In this framework, not only the important skeleton joints in the spatial, but also the important action frames in the temporal domain are considered.

Compared with the spatial attention model and temporal attention model, this framework is more complex, thus making it difficult to train. Moreover, it is more likely to face the over-fitting issue. In order to alleviate the above problems, this paper adopts a second-order regularization strategy on the loss function which is defined as follows:

$$L = -\sum_{i=1}^{C} y_i \log \hat{y}_i + \lambda_1 ||w_{LSTM_{GSA}}||_2^2 + \lambda_2 ||w_{LSTM_{ALC}}||_2^2 \tag{4}$$

In (4), the first term represents the loss function by using cross-entropy, $y = (y_1, \ldots, y_C)^T$ represents the type of real action, $\hat{y}_i = p(C_i|X)$ represents the type of action calculated through this framework. The second term $||w_{LSTM_{GSA}}||_2^2$ represents the second-order regularity of the global spatial attention model parameter. The third term $||w_{LSTM_{ALC}}||_2^2$ represents the second-order regularization of the parameters of the additive learning curve model, λ_1 and λ_2 are the equilibrium factors.

IV. DIFF FOR TEMPORAL DYNAMIC FEATURE

To further improve the effect of action recognition, this paper proposes to build a Diff feature based on the motion track of human joints. It reflects the change of each joint between adjacent frames and enhances the dynamic of the features. Taking this feature as the preprocessing in deep learning network, the subsequent LSTM network can acquire more dynamic features, and its structure is shown in Figure 5. In other words, firstly, the 3D skeleton joint data of the human body is sent into Diff preprocessing to obtain more dynamic basic features; then, it is sent to the LSTM network for feature extraction and action recognition. The Diff feature

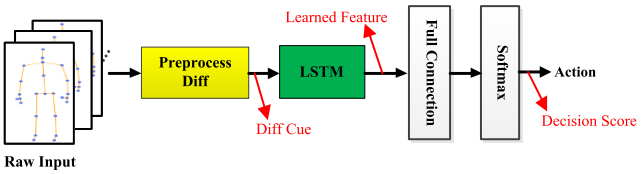

FIGURE 5. Action recognition by LSTM with Diff.

construction process is shown in Figure 6, and the detailed steps are as follows:

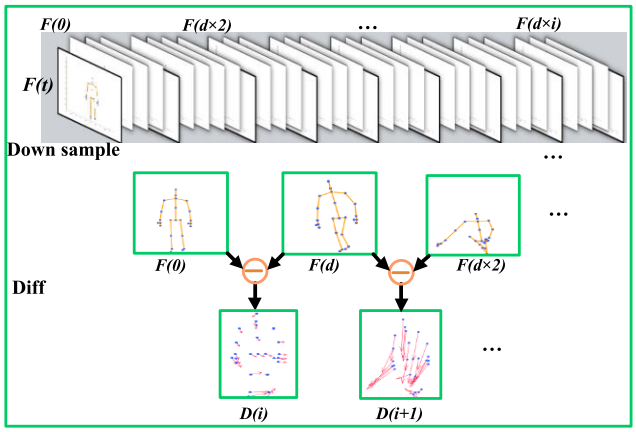

FIGURE 6. Extraction of temporal dynamic features by subtracting subsampled frames.

1) Downsample original human joint data. $F(i) = F(d \times i)$, index i is the index for the downsampled $F(i)$. That is, sample one frame of action every d time interval;

2) Subtract the sampling data of two adjacent frames to get displacement vector D_i. $d_k(i) = p_k(i+1) - p_k(i)$ where $p_k(i+1)$, $p_k(i)$ is the 3d coordinate of the k-th joint in $F(i+1)$, $F(i)$. Collectively, these 25 displacement vectors constitute the Diff cue, denoted by D_i.

From the above steps, it can be seen that Diff is easy to be calculated in the scene with 3D skeleton joints as input. As opposed to the trajectory method [32] where each displacement vector $d_k(i)$ is added back to the associated $p_k(i)$, the Diff here only keeps the displacement vectors. For recordings $F(t)$ of the same action performing instance by cameras of different locations and pose conditions, the respective Diff cues are expected to be different. However, the absolute amplitude changes along the whole process for all Diff cues are likely to be the same, serving a validate cue for discriminating different actions.

V. EXPERIMENTAL EVALUATION

In order to validate its effectiveness, the proposed algorithm is validated on two commonly used action recognition datasets with different scales, the largest NTU RGB+D dataset and the common small SBU dataset [35]. The following is an evaluation of the effect of the global spatial attention model and ALC model from the aspects of visualization, effect enhancement, comparison with classic attention model. The effect of Diff is evaluated from two aspects: recognition effect and distinguishing action types. Finally, the effectiveness of the proposed algorithm is evaluated more comprehensively by comparing with other state-of-the-art methods.

A. DATASET

1) NTU RGB+D DATASET

NTU RGB+D dataset [21] is currently the largest RGB-D action recognition dataset. NTU RGB+D dataset 60 is the version of NTU RGB+D dataset before 2019. As the largest RGB-D action data, it is highly praised and used by most action recognition methods based on deep learning. For the convenience of comparison, NTU RGB+D dataset 60 is selected as a test dataset in this paper. NTU RGB+D dataset 60 recorded 60 actions of 40 participants by three cameras in 80 viewing angles with a total of 56880 data, including 49 single actions, 10 health care actions and 11 actions by two individuals. It not only offers different forms of data, such as RGB, depth, skeleton joint points etc., but also provides two standard test methods: Cross subject (CS) and Cross view (CV). Cross subject mainly aims at predicting the action difference between different people, that is, the difference between different people when performing the same action. It takes 40320 data completed by 20 people numbered 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, 38 as the training set, selects randomly 5% as the validation set in the training set, and 16,560 sets of data completed by another 20 people as the test set. Cross view (CV) aims at the variability of the same kind of action from the different camera perspectives. It uses 37,920 data recorded by cameras numbered 2 and 3 as training set, extracts randomly 5% as a validation set in the training set, and 18,960 sets of data recorded by camera 1 as the test set.

2) SBU DATASET

At present, SBU is mainly used to test the performance of an action recognition algorithm based on deep learning in small-scale data. 282 action video sequences and 8 kinds of actions by 7 individuals are recorded by using Kinect V1. At the same time, it provides three kinds of information: RGB, depth and skeleton joints. The dataset is divided into 21 groups; each group contains 8 kinds of actions completed by 2 individuals. The dataset does not provide standard data test methods. At present, in this data set, most of the test methods used by other researchers are 50% cross-validation in the 21 groups of data, taking the average value as the final result. This paper also uses this method.

B. IMPLEMENTATION DETAILS

1) TEST PLATFORM AND SETTINGS

The test platform used in this paper consists of Intel (R) Core i7- 7700K @ 4.2GHz CPU, GEFORCE GTX 1080 TI GPU, with a Windows 10 operating system. Tensorflow is used as the development framework. The optimization method

is Adam, and the learning rate is set as 0.001. Meanwhile, both ALC and FC adopt Dropout to prevent over-fitting; the regularization parameters are set as 0.0001 and 0.00001 respectively. The initial weights for all parameters are set by using Glorot Initialization [36]. The initial value of all bias is set to 0.

2) DATA PREPROCESSING

In NTU RGB+D dataset 60, the participants may be one or two. For the convenience of programming, all actions are done by two individuals. That is, for the actions by only one person, joint point data of another individual will be added into data preprocessing (all joint position data of this person are 0). For the actions of two individuals, keep it as it is. In SBU dataset, each action is a two-individual action, and no additional processing is required.

3) TRAINING PROCESS

The mini-batch size is set to 128. And in mini-batch iteration, the loss is calculated by using Cross entropy. When Adam [37] adjusts the learning rate automatically, backpropagation through time (BPTT) [38] is used for reverse learning. According to the evaluation method of NTU RGB+D dataset, the validation set is verified every 100 iterations during the training; and only the model with the best recognition result is used for the final testing phase. Then, the selected test model is tested with the test set, and the test results are maintained; finally, the best test results are taken as the test results of the algorithm. In SBU dataset, due to its small scale of data, the parameter part changes Mini batch size to 16, and the other parts are the same as NTU RGB+D dataset.

C. VISUALIZATION

1) VISUALIZATION OF GLOBAL SPATIAL ATTENTION MODEL

The clapping action in NTU RGB+D dataset is randomly chosen to observe the visualization effect of spatial attention as shown in Figure 7. The red circle is used to represent the skeleton joints with large spatial attention weight, and the size of the circle represents the strength of the weight.

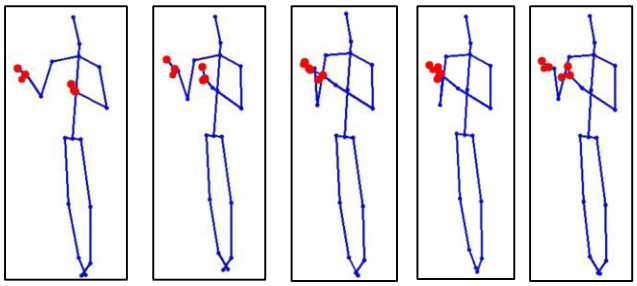

FIGURE 7. Visualization of global spatial attention for clapping action.

As can be seen from the figure, for the whole clapping sequence, the part with more weight is mainly on the hand. That is to say, the movement of the hand largely determines the type of action. The same conclusion is also true with the visualization of other actions. It is clear that the spatial attention model presented in this paper basically expresses the intuition of the proposed algorithm.

2) VISUALIZATION OF ALC MODEL

Similarly, in order to observe the effect of ALC in describing the expression of temporal attention weight, a trained test model ALC is randomly selected under the NTU-CS test. ALC visualization is shown in Figure 3. The result of the subtraction of two adjacent frames in the ALC curve is shown in Figure 8, which expresses clearly the weight change of each frame of the video in ALC. The curve in Figure 3 is the weight of each output arranged according to temporal sequence, which reflects the importance of the content contained in the video from the beginning to the end. It can be seen from the figure that the most important part of the action is mainly concentrated in the end part of the sequence. That is, the output comparatively later contains more information and plays a greater role in action recognition.

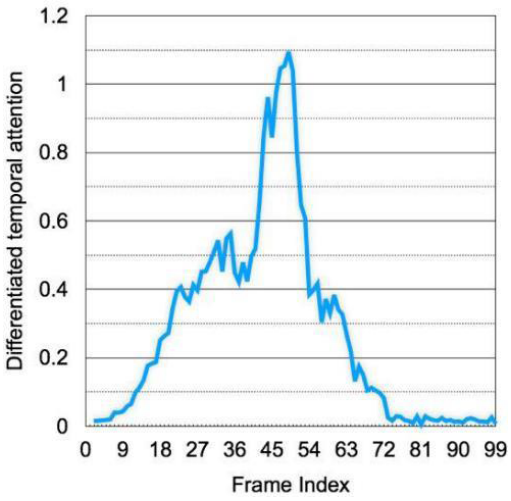

FIGURE 8. Differentiated temporal attention.

Figure 8 just makes up for this defect. It shows the importance of each frame of action in the whole recognition process. It can be seen from the figure that the most important part of the action is mainly concentrated in the middle. The actions at the beginning and end are less important. This is consistent with the message expressed in the video, and is also consistent with human cognition. It can be seen that the ALC model better represents the idea proposed by the algorithm.

D. ENHANCEMENT OF THE PROPOSED ATTENTION MODEL

1) GLOBAL SPATIAL ATTENTION MODEL

In order to further evaluate the spatial attention model proposed in this paper, recognition rate enhancement and efficiency will be analyzed. The recognition results of different methods are arranged into the histogram of recognition rate as shown in Figure 9. It can be seen from the figure that the

recognition rate by using pure LSTM is 66.8% under the NTU Cross Subject test, and that of using global spatial attention is 75.1%, with an 8.3% increase. Under the NTU Cross View test, the recognition rate by using pure LSTM is 77.5%, and that of using global spatial attention is 82.16%, with a 4.66% increase. Under the SBU dataset, the recognition rate by using global spatial attention is 7.34% higher than that of using pure LSTM. From the above experimental results, it can be seen that the recognition rate has been significantly improved by using the global spatial attention model.

Meanwhile, in order to further understand the effectiveness of the global spatial attention model, that is, what kind of action type it has better effect on, this paper examines the enhancement on NTU RGB+D dataset and sorts out top 10 actions that have better enhancement. Table 1 is the sequence of the 10 actions, namely clapping, putting on glasses, brushing teeth, drinking water, rubbing two hands, punch, taking off a hat/cap, writing, putting on a shoe, handshaking hands. On observing these actions it has been inferred that they have two common characteristics. Firstly, these actions are only closely related to a few limbs or parts of limbs, but hardly related to other parts of limbs. For example, clapping is mainly related to hands, not to other parts of limbs.

TABLE 1. ACCURACY (%) for Ram and GSA LSTM on NTU RGB+D dataset.

Action	Ram	GSA	Enhancement
Clapping	23.1	60.6	37.5
Put on glasses	37.5	71.6	34.1
Brush teeth	29.9	61.6	31.7
Drink water	38.4	59.0	20.5
Rub two hands	35.8	56.2	20.4
Punch	41.5	61.5	20.0
Take off a hat/cap	54.7	72.7	18.0
Writing	35.8	53.3	17.5
Put on a shoe	73.0	90.1	17.1
Handshaking hands	59.0	75.9	16.9

Secondly, recognition rate is generally lower on the pure LSTM framework that does not employ any attention models. The result is the same on SBU dataset. It can therefore be confirmed that the global spatial attention model effectively strengthens the importance of key limbs in action recognition and provides spatio-temporal information that significantly enhances the recognition accuracy.

2) ALC MODEL

Similar to the global spatial attention model, this part mainly evaluates the effectiveness of ALC model from the aspect of recognition enhancement. It mainly compares ALC with the two feature extraction methods basic LSTM (only the last output is taken as the feature) and Equal Weight (the output of each time is concatenated as the feature).

As shown in Figure 9, under the NTU Cross Subject test, the recognition rate by using pure LSTM (basic LSTM) is 66.8%, and that of Equal weight is 71.53%, and that of ALC is 77.4%. The recognition rate of ALC is better than the other two methods. Under NTU-Cross View test, the recognition rate of ALC is 83.79%, higher than 77.5% of pure LSTM and 77.09% of Equal Weight. In SBU dataset, similar results are presented. The recognition rate of ALC is 96.82%, while that of LSTM is only 86.7%. Therefore, ALC has an obvious effect on the recognition rate enhancement.

TABLE 2. ACCURACY (%) for Ram and ALC LSTM on NTU RGB+D dataset.

Action	Ram	ALC	Enhancement
Put on glasses	37.5	74.7	37.2
Brush teeth	29.9	66.8	36.9
Clapping	23.1	58.7	35.6
Rub two hands	35.8	62.3	26.5
Touch head	45.8	70.6	24.8
Drink water	38.4	61.6	23.2
Take off a hat/cap	54.7	77.5	22.8
Punch	41.5	63.0	21.5
Handshaking hands	59.0	80.1	21.1
Check time	56.6	74.3	17.7

Similarly, to further analyze the effect of ALC, 10 actions that have better enhancement are sorted out. They are putting on glasses, brushing teeth, clapping, rubbing two hands, touching head, drinking water, taking off a hat /cap, punching, handshaking hands, checking time in table 2. By looking at these 10 types of actions manually, there are two findings. On the one hand, the recognition rate is generally low if ALC is not used. On the other hand, the key frames of these actions are concentrated in the middle of time. In SBU dataset, a similar conclusion is presented. It can be seen that ALC mainly improves the effect of action recognition and highlights the importance of key frames by accurately modeling the weight distribution of action sequences.

E. COMPARISON WITH STA-LSTM MODEL

STA-LSTM is a classic spatio-temporal attention model in action recognition, and it is also the comparison target in this paper. The following is a comparison from three

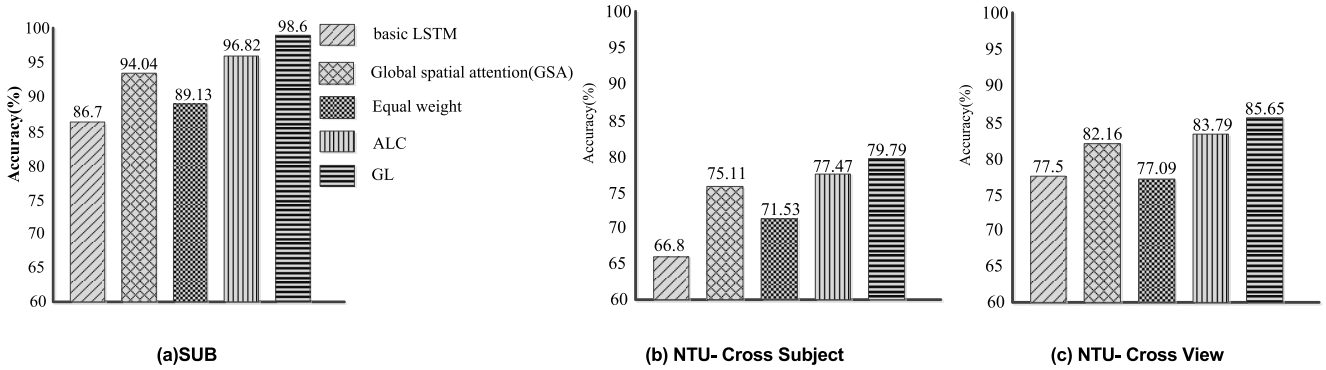

FIGURE 9. Accuracy on SBU and NTU RGB+D datasets.

aspects: spatial attention model, temporal attention model and spatio-temporal attention model.

1) SPATIAL ATTENTION MODEL

STA-LSTM uses local spatial attention, mainly considering the relationship between two adjacent frames to determine the weight of the skeleton joints. The global attention model proposed in this paper determines which joints are important and which joints are not important by all frames of the whole sequence. Global attention is more expressive than local attention because it contains more information.

TABLE 3. Accuracy (%) comparison on NTU RGB+D dataset.

Item	STA-LSTM		GL-LSTM		Enhancement	
	CS	CV	CS	CV	CS	CV
Spatial attention	71.9	80.4	75.1	82.16	3.2	1.76
Temporal attention	73.2	80.5	77.4	83.78	4.2	3.28
Spatio-temporal attention	73.4	81.2	**79.7**	**85.65**	6.3	4.45

In the CS test of NTU RGB+D dataset, as shown in table 3, the spatial attention recognition rates for STA-LSTM and GSA are 71.9% and 75.1% respectively; in the CV test, they are 80.4% and 82.16% respectively. In SBU datasets, as shown in Table 4, the spatial attention recognition rates for STA-LSTM and GSA are 88% and 94.04% respectively. The recognition rate of GSA is significantly higher than that of STA-LSTM. These experimental results show that the recognition of global spatial attention model outperforms that of local spatial attention model.

2) TEMPORAL ATTENTION MODEL

STA-LSTM uses local temporal attention, mainly considering the relationship of the before-after frames. In this paper, ALC is used to obtain the weight of each frame directly through data training, which reflects the weight distribution of the whole training data set in temporal dimension. Compared with local temporal attention, ALC considers the importance of the whole sequence globally. The workload of ALC is much smaller than STA-LSTM in both training and testing.

In terms of recognition rate, as shown in table 3, in NTU CV test, spatial attention accuracy for STA-LSTM and ALC are 73.2% and 77.4% respectively, while in NTU-CV test, they are 80.5% and 83.78% respectively. In SBU datasets, as shown in Table 4, the temporal attention accuracy for STA-LSTM and ALC are 89% and 96.82% respectively. It can be seen that the recognition rate of ALC proposed in this paper is significantly better than that of STA-LSTM in the two datasets, which is also consistent with the previous theoretical analysis.

TABLE 4. Accuracy (%) comparison ON SBU dataset.

Item	STA-LSTM	GL-LSTM	Enhancement
Spatial attention	88.0	94.04	6.04
Temporal attention	89.0	96.82	7.82
Spatio-temporal attention	91.5	**98.6**	7.1

3) SPATIO-TEMPORAL ATTENTION MODEL

STA-LSTM model consists of three parts, namely, spatial attention, temporal attention and main LSTM network. In terms of training mode, STA-LSTM is divided into four phases with 9 steps: 1) pre-train temporary attention model; 2) pre-train spatial attention model; 3) train the main LSTM network; 4) jointly train the whole network. On the whole, the training process is comparatively complex and it is difficult to get desirable results. The experimental results of STA-LSTM on NTU RGB+D dataset are the best example. Specifically, under NTU CS test, the recognition rate for spatial attention, temporal attention and spatio-temporal attention is 71.9%, 73.2%, and 73.4% respectively. The spatio-temporal attention rate is only 0.2% higher than that of temporal attention. Under NTU CV test, spatio-temporal attention is only 0.7% higher than that of temporal attention. Intuitively, the improvement of integrating spatio-temporal

attention model is not obvious, which is not consistent with human intuition.

In contrast, the framework proposed in this paper is relatively simple. It is only composed of global spatial attention and ALC, and the training method is also simple. The end to end method can be used directly without any additional steps. At the same time, the recognition rate of spatio-temporal attention is about 2% higher than that of temporal attention and spatial attention alone. Whether on NTU CS, NTU CV test, or SBU dataset, the overall recognition rate of GL-LSTM is higher than STA-LSTM. It can be seen that the spatio-temporal attention model proposed in this paper is better than STA-LSTM.

F. DIFF

To measure the effect of Diff on action recognition, tests are conducted on NTU RGB+D dataset and SBU dataset respectively. The test results are shown in table 5 and table 6.

TABLE 5. Recognition rate (%) with Diff on NTU RGB+D dataset.

Item	Raw-GL		Diff-GL		Enhancement	
	CS	CV	CS	CV	CS	CV
Spatial attention	75.1	82.16	81.3	87.7	6.2	5.54
Temporal attention	77.4	83.78	82.5	87.5	5.1	3.72
Spatio-temporal attention	79.7	85.65	84.4	90.2	4.7	4.55

TABLE 6. Recognition rate(%) with Diff on SBU dataset.

Item	Raw-GL	Diff-GL	Enhancement
Spatial attention	94.04	97.1	3.06
Temporal attention	96.82	98.3	1.48
Spatio-temporal attention	98.6	99.2	0.6

It can be seen from table 5 that the Diff method improved the recognition rate, whether it is NTU CS test or NTU CV test. To be exact, in spatial attention, the recognition rate is increased by 6.2% and 5.54% respectively; in temporal attention, increased by 5.1% and 3.72% respectively; in spatio-temporal attention, increased by 4.7% and 4.55% respectively. Similar results are presented on the SBU dataset shown in Table 6.

The reason: for action recognition, dynamic is one of the key characteristics to distinguish different actions, and plays a particularly important role. Diff improves the effect of action recognition by strengthening the dynamics of features to learn and distinguish features more effectively. In essence, Diff solves some problems encountered in action recognition (dynamic feature extraction), alleviates some pressure of deep learning, and enables deep learning to better play its own (learning feature) advantages and concentrate on feature extraction.

TABLE 7. ACCURACY (%) for Ram and Diff on NTU RGB+D dataset.

Action	LSTM +Ram	LSTM +Diff	Enhancement
Clapping	23.1	75.94	52.84
Rub two hands	35.8	73.1	37.3
Hopping	79.14	98.56	19.42
Punch	41.5	58.51	17.01
Hand waving	59.0	73.82	14.82

This paper further explored on what kind of actions Diff cue is effective. In order to ensure pure Diff cue effect, the original data and Diff processed data are sent to a three-layer general LSTM network. Table 7 lists the first five types of actions according to the enhancement rate. It can be seen from Table 7 that the five types of movements are all the types with obvious limb changes and large range of motion. That is to say, dynamic is the dominant feature in this kind of movements. Therefore, strengthening dynamic will help to improve the effect of action recognition. This coincides with the dynamics expressed by Diff. It can be seen that in the action sequence with intense and dynamic movement, Diff cue will notably improve the distinguishability of features.

G. COMPARISON WITH OTHER STATE-OF-THE-ART ALGORITHMS

In order to evaluate the effect of the proposed algorithm more accurately, the action recognition results of different algorithms on NTU RGB+D dataset 60 and SBU dataset are shown in table 8 and table 9. To ensure the reliability of the results, comparison of the algorithm with other state-of-the-art ones needs to meet four conditions at the same time: 1) use only 3D human skeleton as input and no other information; 2) use only original data without any data enhancement; 3) use only single stream information, and no multi-stream information integration; 4) use RNN-based approaches.

Table 8 shows the results of the proposed algorithm with other state-of-the-art algorithms on NTU RGB+D dataset. Lie group [39] and Dynamic skeletons [40] are the results of traditional machine learning; the rest is by deep learning. It can be seen from the table that the method based on deep learning is better than that based on traditional machine learning. Obviously, compared with traditional methods, the method based on deep learning has obvious advantages. Even though, there is a certain gap compared with integrated methods such as [47], [49], the GL-LSTM method proposed in this paper has obvious advantages compared with pure attention methods (such as STA-LSTM [14]). The GL-LSTM+Diff method has obvious advantages compared

TABLE 8. Recognition rate(%) on NTU RGB+D dataset.

Method	Modality	CS	CV
Lie group[39]	Skeleton	50.1	52.8
Dynamic skeletons[40]	Skeleton	60.2	65.2
ST-LSTM+Trust Gate [41]	Skeleton	69.2	77.7
A^2GNN[42]	Skeleton	72.74	82.8
STA-LSTM[14]	Skeleton	73.4	81.2
Ensemble TS-LSTM v2 [43]	Skeleton	74.6	81.2
URNN-2L-T[44]	Skeleton	74.6	83.2
Enhanced Skeleton visualization[45]	Skeleton	75.9	82.5
GCA-LSTM[46]	Skeleton	76.1	84
TSSI+SSAN[47]	Skeleton	80.9	86.1
ST-GCN[48]	Skeleton	81.5	88.3
GC-LSTM[49]	Skeleton	83.9	**92.3**
GL-LSTM(Ours)	Skeleton	79.7	85.65
GL-LSTM+Diff(Ours)	Skeleton	**84.4**	90.2

TABLE 9. Recognition rate (%) on SBU database.

Method	Modality	Accuracy
Joint Feature [10]	Skeleton	86.9
Co-occurrence LSTM [11]	Skeleton	90.4
STA-LSTM[14]	Skeleton	91.51
ST-LSTM+Trust Gate [41]	Skeleton	93.3
SkeletonNet[50]	Skeleton	93.47
ST-NBMIN[51]	Skeleton	93.3
TSSI+SSAN[47]	Skeleton	94
LSTM+FA+VF[52]	Skeleton	95
GL-LSTM(Ours)	Skeleton	98.6
GL-LSTM+Diff(Ours)	Skeleton	**99.2**

with other state-of-the-art methods, whether it is 84.4% under CS test or 90.2% under CV test. This proves that the method proposed in this paper has a good effect on large scale data sets.

Table 9 shows recognition rate of the algorithm proposed in this paper and other state-of-the-art algorithms on SBU dataset. Most algorithms adopt deep learning, except for joint feature in [10]. Compared with other deep learning methods, the recognition rate of the proposed GL-LSTM+Diff method reaches 99.2%, significantly higher than other state-of-the-art methods. This indicates that the method proposed in this paper has better adaptability on small scale datasets.

To sum up, the algorithm proposed in this paper has certain advantages in both large datasets and small datasets.

VI. CONCLUSION

To tackle the different importance of each joint and the different roles each frame plays in action recognition, this paper proposes the GL-LSTM model. By integrating the proposed GSA (for spatial information) and ALC (for temporal processing) models into the LSTM framework and taking the human skeleton joints as input, the global spatio-temporal action recognition framework is constructed to recognize human actions. Compared with other classic methods such as STA-LSTM, the algorithm proposed in this paper offers better performance, accuracy, least algorithmic complexity and training overheads. With the introduction of Diff cue, the recognition and dynamic of features are improved. The experimental results on the largest action dataset NTU RGB+D and the commonly used small SBU dataset show the effectiveness of the proposed GL-LSTM+Diff model over state-of-the-art models.

REFERENCES

[1] L. Wang, D. Q. Huynh, and P. Koniusz, "A comparative review of recent Kinect-based action recognition algorithms," *IEEE Trans. Image Process.*, vol. 29, no. 2, pp. 15–28, Jul. 2020.
[2] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "RGB-D-based human motion recognition with deep learning: A survey," *Comput. Vis. Image Understand.*, vol. 171, pp. 118–139, Jun. 2018.
[3] J. K. Aggarwal and L. Xia, "Human activity recognition from 3D data: A review," *Pattern Recognit. Lett.*, vol. 48, pp. 70–80, Oct. 2014.
[4] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3288–3297.
[5] G. Evangelidis, G. Singh, and R. Horaud, "Skeletal quads: Human action recognition using joint quadruples," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Stockholm, Sweden, Aug. 2014, pp. 4513–4518.
[6] E. Ohn-Bar and M. M. Trivedi, "Joint angles similarities and HOG_2 for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Portland, OR, USA, Jun. 2013, pp. 465–470.
[7] R. Saini, P. Kumar, B. Kaur, P. P. Roy, D. P. Dogra, and K. C. Santosh, "Kinect sensor-based interaction monitoring system using the BLSTM neural network in healthcare," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 9, pp. 2529–2540, Sep. 2019.
[8] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, and Z. Gong, "Unsupervised representation learning with long-term dynamics for skeleton based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, New Orleans, LR, USA, Feb. 2018, pp. 2644–2651.
[9] Y. Li, X. Ban, G. Yang, and J. Li, "Real-time human action recognition using depth motion maps and convolutional neural networks," *Int. J. High Perform. Comput. Netw.*, vol. 13, no. 3, pp. 312–320, 2019.
[10] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer LSTM networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Santa Rosa, CA, US Apr. 2017, pp. 148–157.

[11] R. Cui, A. Zhu, S. Zhang, and G. Hua, "Multi-source learning for skeleton -based action recognition using deep LSTM networks," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Phoenix, AZ, USA, Aug. 2018, pp. 3697–3703.

[12] J. H. R. Maunsell, "Neuronal mechanisms of visual attention," *Annu. Rev. Vis. Sci.*, vol. 1, no. 1, pp. 373–391, Nov. 2015.

[13] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Juan, Puerto Rico, Jul. 2016, pp. 321–327.

[14] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "Spatio-temporal attention-based LSTM networks for 3D action recognition and detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3459–3471, Jul. 2018.

[15] Y. Han, S.-L. Chung, S.-F. Chen, and S. F. Su, "Two-stream LSTM for action recognition with RGB-D-Based hand-crafted features and feature combination," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Miyazaki, Japan, Oct. 2018, pp. 3547–3552.

[16] Y. Han, S.-L. Chung, A. Ambikapathi, J.-S. Chan, W.-Y. Lin, and S.-F. Su, "Robust human action recognition using global spatial-temporal attention for human skeleton data," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Rio de Janeiro, Brazil, Jul. 2018, pp. 1–6.

[17] Y. Xie, R. Liang, Z. Liang, and L. Zhao, "Attention-based dense LSTM for speech emotion recognition," *IEICE Trans. Inf. Syst.*, vol. E102.D, no. 7, pp. 1426–1429, Jul. 2019.

[18] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. 28th Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, Dec. 2014, pp. 3104–3112.

[19] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, nos. 5–6, pp. 602–610, Jul. 2005.

[20] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in *Proc. IEEE 60th Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Boston, MA, USA, Aug. 2017, pp. 1597–1600.

[21] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1010–1019.

[22] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1963–1978, Aug. 2019.

[23] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[24] D. Kiela, C. Wang, and K. Cho, "Dynamic meta-embeddings for improved sentence representations," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, 2018, pp. 1466–1477.

[25] G. Sumbul, R. G. Cinbis, and S. Aksoy, "Multisource region attention network for fine-grained object recognition in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4929–4937, Jul. 2019.

[26] K. Xu, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc.32nd Int. Conf. Mach. Learn. (ICML)*, Lile, France, Jul. 2015, pp. 2048–2057.

[27] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 4960–4964.

[28] F. Baradel, C. Wolf, and J. Mille, "Pose-conditioned spatio-temporal attention for human action recognition," 2017, *arXiv:1703.10106*. [Online]. Available: http://arxiv.org/abs/1703.10106

[29] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. G. M. Snoek, "VideoLSTM convolves, attends and flows for action recognition," *Comput. Vis. Image Understand.*, vol. 166, pp. 41–50, Jan. 2018.

[30] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with visual attention on skeleton images," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Beijing, China, Aug. 2018, pp. 3309–3314.

[31] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *Int. J. Comput. Vis.*, vol. 12, no. 1, pp. 43–77, Feb. 1994.

[32] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 60–79, May 2013.

[33] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 624–628, May 2017.

[34] C. Li, Y. Hou, P. Wang, and W. Li, "Multiview-based 3-D action recognition using deep networks," *IEEE Trans. Human-Mach. Syst.*, vol. 49, no. 1, pp. 95–104, Feb. 2019.

[35] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Providence, RI, USA, Jun. 2012, pp. 28–35.

[36] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist. (AISTATS)*, Sardinia, Italy, May 2010, pp. 249–256.

[37] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015, pp. 512–517.

[38] P. J. Werbos, "Backpropagation through time: What it does and how to do it," *Proc. IEEE*, vol. 78, no. 10, pp. 1550–1560, Oct. 1990.

[39] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 588–595.

[40] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2186–2200, Nov. 2017.

[41] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal LSTM network with trust gates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3007–3021, Dec. 2018.

[42] C. Li, Z. Cui, W. Zheng, C. Xu, R. Ji, and J. Yang, "Action-attending graphic neural network," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3657–3670, Jul. 2018.

[43] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1012–1020.

[44] W. Li, L. Wen, M.-C. Chang, S. N. Lim, and S. Lyu, "Adaptive RNN tree for large-scale human action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1453–1461.

[45] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, Aug. 2017.

[46] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention LSTM networks," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1586–1599, Apr. 2018.

[47] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with spatio–temporal visual attention on skeleton image sequences," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2405–2415, Aug. 2019.

[48] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, Feb. 2018, pp. 7444–7452.

[49] H. Zhang, Y. Song, and Y. Zhang, "Graph convolutional LSTM model for skeleton-based action recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Shanghai, China, Jul. 2019, pp. 412–417.

[50] Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid, "SkeletonNet: Mining deep part features for 3-D action recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 731–735, Jun. 2017.

[51] J. Weng, C. Weng, J. Yuan, and Z. Liu, "Discriminative spatio-temporal pattern discovery for 3D action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 1077–1089, Apr. 2019.

[52] Z. Fan, X. Zhao, T. Lin, and H. Su, "Attention-based multiview re-observation fusion network for skeletal action recognition," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 363–374, Feb. 2019.

YUN HAN received the M.Sc. degree from Jiangsu University, Zhenjiang, China, in 2007, and the Ph.D. degree from Tongji University, Shanghai, China, in 2015. He is currently an Associate Professor with the School of Computer Science, Neijiang Normal University, China. His current research interests include deep learning, human motion analysis, computer vision, and vision surveillance.

SHENG-LUEN CHUNG received the B.S. degree from the Electronic Engineering Department, National Chiao-Tung University, Taiwan, in 1985, and the M.S.E. and Ph.D. degrees from the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, in 1990 and 1992, respectively.

Since 1992, he has been with the Electrical Engineering Department, National Taiwan University of Science and Technology, Taiwan, where he is currently a Professor. His research interests include control and automation, and deep learning.

QIANG XIAO received the B.A. degree in english from the Wuhan University of Science and Technology, Wuhan, China. He is currently an Associate Professor with the School of Foreign Languages, Neijiang Normal University. His research interests include artificial intelligence and computer aided translation, and Chinese-English translation: theory and practice.

WEI YOU LIN received the M.Sc. degree from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2018. He is currently an Engineer. His current research interests include deep learning, computer vision, and vision surveillance.

SHUN-FENG SU (Fellow, IEEE) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1983, and the M.S. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1989 and 1991, respectively. He is currently the Chair Professor with the Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei. He has published more than 160 refereed journals and conference papers in the areas of robotics, intelligent control, fuzzy systems, neural networks, and nonderivative optimization. His current research interests include computational intelligence, machine learning, virtual reality simulation, intelligent transportation systems, smart home, robotics, and intelligent control.

Dr. Su is a Chinese Automatic Control Society (CACS) Fellow. He is the President of the Taiwan Fuzzy System Association and a Vice President of the International Fuzzy Systems Association. He currently serves on the Board of Governors of the CACS, the Taiwan Society of Robotics, and the Taiwan Association of System Science and Engineering. He is currently an Associate Editor for the IEEE TRANSACTIONS ON CYBERNETICS and the IEEE TRANSACTIONS ON SYSTEMS, as well as an Area Editor for the *International Journal of Fuzzy Systems*.