SIN-CHN
SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2025, 22(4), 1546.
https://doi.org/10.62617/mcb1546

Article

Tennis serve recognition based on bidirectional long- and short-term memory neural networks

Jianhong Ni, Jing Wang*

Modern Education Technology Center, Hebei Institute of Physical Education, Shijiazhuang 050063, China
*** Corresponding author:** Jing Wang, 206793347@qq.com

CITATION

Ni J, Wang J. Tennis serve recognition based on bidirectional long- and short-term memory neural networks. Molecular & Cellular Biomechanics. 2025; 22(4): 1546. https://doi.org/10.62617/mcb1546

ARTICLE INFO

Received: 12 February 2025
Accepted: 27 February 2025
Available online: 17 March 2025

Abstract: The serve is a crucial technique in tennis, providing players the opportunity to master and organize their attacks during competitions. Current tennis serve training methods often lack sophisticated, data-driven tools, and existing recognition techniques rely on single-feature extraction methods focusing on isolated attributes like trajectory or speed. These methods do not fully utilize the comprehensive spatio-temporal information present in video data, resulting in limited accuracy and robustness in recognizing and analyzing different serving techniques. To address these limitations, this paper proposes a tennis serve recognition method using a bi-directional long short-term memory neural network (BiLSTM). Our approach first employs a modified convolutional neural network (CNN) to extract spatial features from images, enhanced by self-attentive weighting to improve feature extraction. It then uses BiLSTM to capture and represent important temporal features, thereby enhancing the model's ability to recognize and evaluate serving actions. Experimental results demonstrate that our method outperforms existing neural network models in server recognition tasks, effectively addressing the limitations of previous approaches.

Keywords: BiLSTM; CNN; spatial feature information; tennis serve; self-attentive weighting

1. Introduction

Machine learning methods have received more attention in areas such as sports, security prediction, and AI system security. For tennis players, serving and receiving training is also a key training content in daily training [1–5]. Through serving and receiving training, athletes' speed, reaction ability, physical fitness, etc., can be effectively cultivated, which has a good impact on their practical competition ability. Serving is a fundamental technique in tennis, providing athletes with a unique opportunity to initiate and control the point. The limitations of existing methods, especially the poor performance in dealing with complex serving movements, directly affect the improvement of athletes' skills. For example, a case study of professional tennis players showed that the traditional serving action recognition method has a low accuracy rate when dealing with complex movements, which makes it impossible for coaches to find the bottleneck of athletes' skills in time, thus affecting the effect of training. Improving serve performance is therefore a critical focus in tennis training and competition preparation. High-speed and accurate serving is an important means of scoring and winning. Because serving occupies an important position in tennis matches, tennis serving training based on artificial intelligence technology has received more and more attention. Correctly identifying tennis serving movements and timely standardizing incorrect movements will help athletes improve their serving level and increase their competitiveness [1].

There are primarily video analysis-based human motion recognition methods and sensor data-based motion recognition (MR) methods in the subject of MR. The human motion recognition approach based on video analysis has matured in its application and is the primary method used by sports scientists and professional coaches to investigate the biomechanics of various sporting motions. Research on video-based action detection has significant academic value and application potential; hence, this subject is rapidly becoming a research hotspot and challenging point in computer vision [6]. Existing research in motion recognition primarily utilizes video-based analysis, focusing on the processing of video frame sequences to extract spatiotemporal features that correspond to various sports actions, including tennis serves [7–9].

The early development of computer vision technology often faced challenges, leading researchers to rely heavily on smart wearable devices for human action recognition. These devices provide signals such as acceleration and angular velocity, which can be analyzed using traditional machine learning models like SVM and KNN [10–13]. However, these methods require manual feature extraction and heavily depend on the expert's sports background and research experience. In recent years, deep learning (DL) models have demonstrated impressive performance in abstracting data representations through self-learning capabilities and large-scale dataset training, making them increasingly applicable to video analysis for MR.

Traditional tennis serve recognition methods encounter significant obstacles, including sensitivity to lighting conditions and limitations in capturing comprehensive spatio-temporal information. To address these challenges, we propose an innovative approach that integrates both sensor-based and video-based modalities. Sensor-based methods provide precise motion measurements but lack detailed spatio-temporal context, while video-based approaches offer rich data yet struggle with long-term feature modeling and high computational demands. Our method uniquely combines convolutional neural networks (CNNs) for detailed spatial feature extraction with bi-directional long short-term memory (BiLSTM) networks for enhanced temporal feature representation, augmented by a novel self-attentive mechanism. This combination allows for more accurate and robust tennis serve recognition by effectively capturing both spatial and temporal aspects of the serve [14–16].

The combination of self-attention mechanism and BiLSTM actually provides a significant improvement for tennis serve action recognition, but its specific mechanism and advantages over other methods are still worth further exploration. First, the self-attention mechanism can help the model automatically focus on the key features in the input data without being disturbed by noise or irrelevant information. In tennis serve recognition, the spatial and temporal features of the action are highly correlated, and the self-attention mechanism can effectively identify the key action moments in the serve process by weighting the features of different time steps. For example, in the preparation stage, swing stage, and hitting stage of the serve action, different stages have different effects on the spatial and temporal features of the action. The self-attention mechanism dynamically adjusts the weights of the features, allowing the model to focus on these important moments more accurately in complex action recognition tasks. Compared with the traditional LSTM model, BiLSTM performs better in capturing long-term dependencies by introducing bidirectional time

series information. In tennis serve recognition, the serve action not only depends on the current state but also involves the correlation between the moments before and after the action sequence. BiLSTM can learn the information in the time series from both the front and back directions at the same time so as to more comprehensively capture the spatiotemporal characteristics of the serve process. Traditional LSTM only transmits information in one direction, which limits its ability to process long time series data. In particular, when recognizing the serve action, the connection between the previous and next stages is crucial. Combining the advantages of the self-attention mechanism and the BiLSTM network can further improve the performance of the model in tennis serve recognition. The self-attention mechanism is responsible for capturing the changes in local key features, while the BiLSTM effectively models the temporal dependencies of the entire action sequence. This combination enables the model to not only accurately extract the spatial features of the serve action but also handle temporal changes over a long time span, making the recognition process more accurate and robust.

In addition, one challenge facing tennis serve action recognition is that environmental factors (such as lighting, shooting angle, etc.) may affect the quality of video data. Sensor-based data provides accurate motion measurement but has limitations in modeling spatiotemporal backgrounds. Video data can provide rich contextual information, especially in capturing the details and dynamics of the action. However, traditional video-based action recognition methods are often limited by factors such as lighting and data noise, and it is difficult to ensure recognition results in unstable environments. Therefore, combining the self-attention mechanism and BiLSTM can not only improve the model's ability to extract spatiotemporal information but also enhance the model's robustness to environmental changes so that the server action recognition can still maintain a high accuracy in different actual scenarios.

This innovative method is not only of great significance in academic research but also shows great application potential in actual tennis training and match analysis. In training, coaches can use this method to analyze the details of the player's serve in real time, identify technical deficiencies in the serve, and provide personalized training plans for the player. In match analysis, it can help players and coaches analyze the opponent's serve strategy and provide data support for tactical decision-making.

In this research, we introduce a tennis serve recognition system that leverages CNN and BiLSTM to extract and synthesize spatial and temporal information from multidimensional time-series data. A key aspect of our contribution is the implementation of a self-attentive mechanism that adaptively assigns greater weight to significant features, thereby enhancing the model's sensitivity to critical elements. By employing BiLSTM instead of traditional LSTM, our approach captures long-term spatiotemporal dependencies more effectively. We validate the proposed method through extensive experiments on a publicly available dataset, demonstrating its superior performance compared to several existing models in tennis serve recognition, thereby highlighting its potential impact in the field of sports motion recognition. The contributions of this paper are as follows:

- We introduce a self-attention mechanism to improve feature extraction in tennis serve recognition.

- We enhance the existing LSTM model by integrating BiLSTM, which better captures long-term spatiotemporal dependencies.
- We develop an adaptive weight assignment strategy to increase the model's sensitivity to critical features.
- We demonstrate the superior performance of our approach through extensive experiments on a publicly available dataset.

2. Related work

Some scholars use CNN to train depth motion map data and extract static feature representation, using dense trajectory as dynamic feature information, and finally concatenating static and dynamic features as overall feature behavior representation to achieve good recognition results. However, this approach suffers from issues related to lighting conditions, expensive equipment, high computational costs, and limitations in perception range. In [17–19], based on 2D convolution, 3D convolution and 3D pooling are used throughout the network structure to extract the spatio-temporal features of the video, which preserves the temporal information between video frames and portrays the time-space correspondence. Some scholars used a dual-stream CNN model for MR [20–24]. This model extracts the static information of video with video frames as the input of the spatial streaming network and acquires the motion information between image sequences with optical streams as the input of the temporal streaming network, respectively. While effective for short time domain modeling, these methods fail to leverage long-term information from the entire video.

To address the long-distance temporal dependence problem, based on the dual-stream network, some scholars use a sparse sampling strategy to extract multiple short video clips and then construct a long-time temporal model on the temporal structure of multiple video clips. Mutegeki and Han, Zhao et al., and Heryadi and Warnars [25–27] connect an LSTM after a fully connected layer of CNN, which combines the predicted values of the video at each time node to classify the prediction of the whole video. Despite their success, these methods often overlook the significance of different regions within the scene, which play distinct roles in action representation [28–30].

Based on traditional feature extraction networks, some scholars introduced attention mechanisms and feature fusion strategies in the form of residual networks, allowing algorithmic models to obtain better recognition results [31]. Jain et al. [32] used a soft attention mechanism to focus on action-related regions and fed the weighted feature maps into a multilayer LSTM for the prediction of behavior categories. Deng et al. [33] introduced a pose attention mechanism to obtain robust human features by sharing attention parameters through human semantic-related nodes.

In summary, current human motion recognition methods exhibit several limitations: (1) 2-D and Image-Based Methods: Often struggle with lighting conditions and lack robustness against variations in video quality; (2) Feature Extraction Techniques: Typically focus on short-term dynamics and fail to utilize long-term temporal information effectively; (3) Sensor Data Approaches: Often require expensive equipment and high computational resources. To enhance the robustness and accuracy of motion recognition, our proposed method integrates sensor-based and video-based modalities, addressing these limitations by capturing comprehensive

spatio-temporal information and improving feature extraction through innovative mechanisms. This unique approach positions our work as a significant advancement in the field.

3. Background

This section introduces the concepts and formulas related to CNNs, LSTMs, and BiLSTMs, and we refer to the literature [18] for a formulaic description of models such as CNNs and LSTMs.

3.1. CNN

CNN is a class of feedforward neural networks containing convolutional operations and having a multilayer depth structure, and its network structure is shown in **Figure 1**.

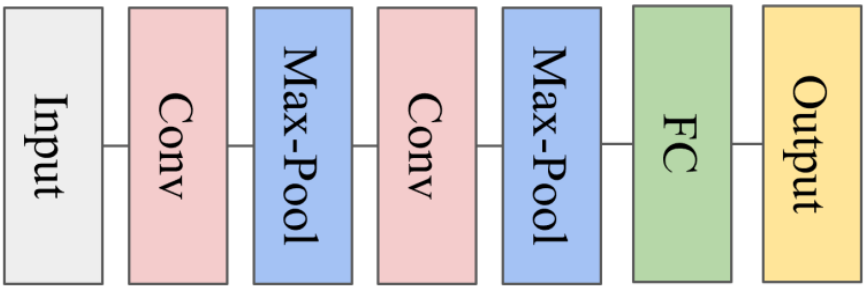

Figure 1. Diagram of CNN.

In **Figure 1**, Conv represents the convolutional layer, and the calculation principle of the 1-dimensional convolutional layer is as follows:

$$y_i^j = f_{\text{conv}}\left(w_j \otimes x_i + b_j\right) \tag{1}$$

where x_i is the input vector, w_j is the weight vector of the j-th convolution kernel, b_j is the bias term.

Pooling layers are used to aggregate feature information by reducing the data's resolution to control the number of parameters and overfitting, which is effectively a downsampling process, such as maximum pooling, mean pooling, etc. In reality, CNNs typically employ numerous convolutional and pooling layers in series and in alternation to extract more detailed features.

3.2. LSTM

LSTM is a special class of RNN, which can effectively improve the problems of gradient disappearance and long-period dependence faced by conventional RNNs, as shown in **Figure 2**.

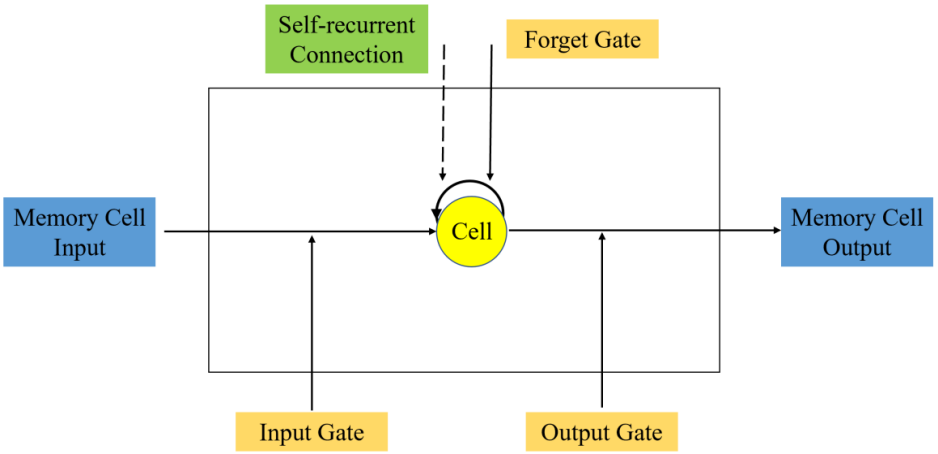

Figure 2. The structure of LSTM.

The input gate is calculated as:

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \tag{2}$$

where h_{t-1} is the output vector of $t-1$. The formulas for other gates are as follows:

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \tag{3}$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \tag{4}$$

$$\tilde{c}_t = \eta(w_c[h_{t-1}, x_t] + b_c) \tag{5}$$

$$c_t = f_t c_{t-1} + i_t \tilde{c}_t \tag{6}$$

$$h_t = o_t \eta(c_t) \tag{7}$$

where c_t is the value in the cell at time t, c_t is the update value in the cell at time t, η is the tanh function, σ is the sigmoid function. The structure of each neuron model is shown in **Figure 3**.

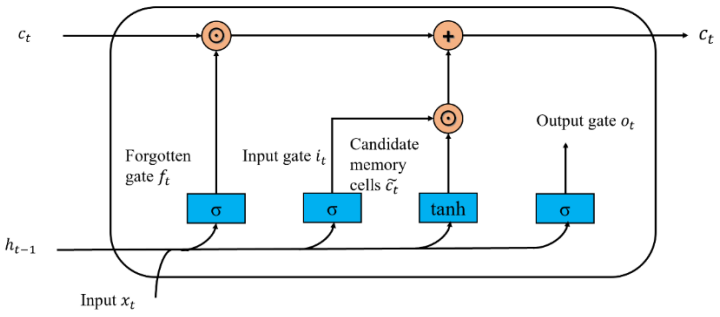

Figure 3. The structure of each neuron model.

3.3. BiLSTM

BiLSTM is an enhanced structure of LSTM consisting of two parts: Forward LSTM and backward LSTM, which can extract bi-directional characteristics of time series data from both forward and backward directions to provide superior outcomes. **Figure 4** depicts the architecture of BiLSTM.

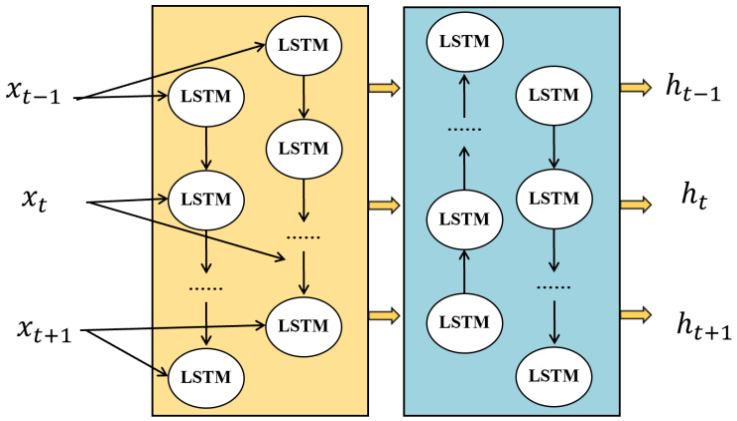

Figure 4. Structure of BiLSTM.

4. The proposed method

In this section, we propose a tennis serve recognition algorithm that integrates a self-attentive mechanism and BiLSTM (SAM-BiLSTM) to solve the problems that existing serve recognition methods have: Single feature extraction capability and the inability to fully utilize the spatio-temporal information contained in the data.

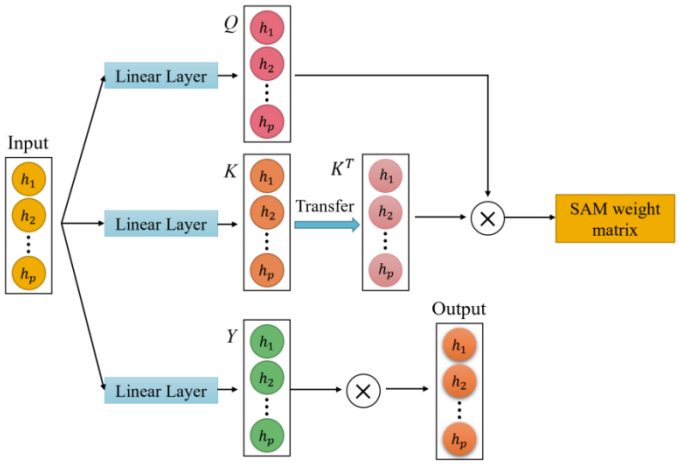

Figure 5. Framework of SAM.

The attention mechanism is a class of algorithms that replicates the human brain's attention mechanism and was initially developed for image classification and natural language processing applications [34,35]. The fundamental concept is to highlight the portions of the input data that are more pertinent to the output task in the form of weights, etc., based on the probability distribution of the data and the interrelationship of the variables, so as to extract more critical information to enhance the overall performance of the network model. There are various sorts of attention mechanisms, including location attention, input sequence attention, and self-attention. This study focuses on the Self-Attention Mechanism (SAM), which is more suited for practical engineering applications because it just pulls information from the input itself and does not need any additional information and has the benefits of fewer parameters and faster processing. **Figure 5** illustrates the construction of SAM.

As depicted in **Figure 5**, the input data are initially processed through three linear

layers to create the query matrix Q, the key matrix K, and the value matrix V. Softmax then calculates the self-attentive weight matrix A by multiplying the transpose of Q and K, dividing by the scaling factor, and calculating the product. The outcome of self-attentive weighting is obtained by multiplying the vector V by the weight matrix A. It is also possible to obtain the matrices Q and K using only two linear layers and to directly output the input after multiplying it with the weight matrix A.

$$SAM(h) = softmax\left(\frac{QK^T}{\sqrt{\alpha}}\right)V \tag{8}$$

where α is the scaling factor.

In addition, a convolutional attention module is introduced, in which channel attention and spatial attention are used in tandem, and the attention mechanism primarily relies on pooling operations, as the image's features are concentrated in a few local regions, and pooling helps to eliminate redundant information in order to focus more on the locally significant information. However, the features contained in the tennis serve video are widely dispersed, making it difficult to extract the remote dependencies among the features by pooling procedures for local regions, resulting in the loss of crucial information. In this research, we replace the pooling-based attention mechanism in the convolutional attention module with an original-based self-attentive mechanism. Because the self-attentive mechanism has a global perceptual field compared to pooling, the output after the self-attentive weighting takes into account the information of all the features and enables direct dot product fusion between features at different locations, irrespective of their distance, resulting in a more global feature extraction effect. The structure of SAM-BiLSTM is shown in **Figure 6**.

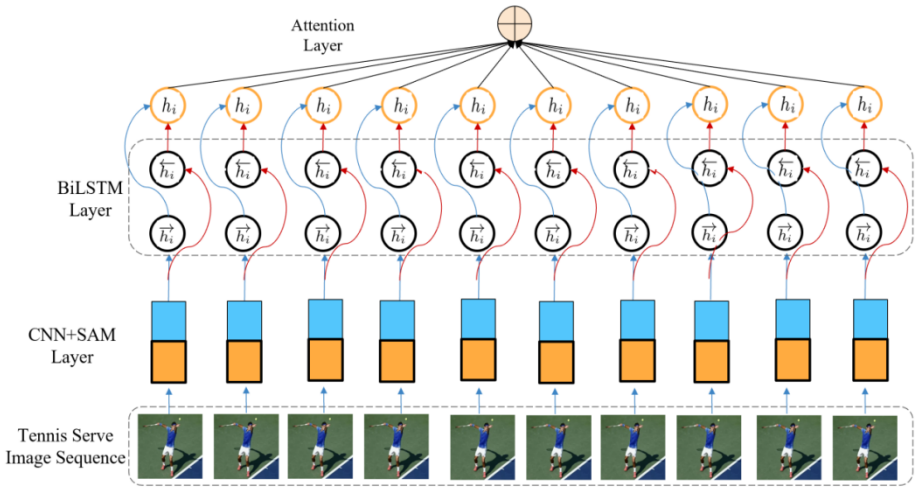

Figure 6. Framework of SAM-BiLSTM.

As illustrated in **Figures 5** and **6**, our proposed tennis serve recognition system consists of several key modules, each denoted by specific colors for clarity:

Input Module (Yellow): The input consists of tennis serve image sequences, generating feature vectors.

Linear Layers (Blue): Each input feature undergoes processing through linear layers, transforming the raw data into query (Q), key (K), and value (Y).

Attention Mechanism (Red and Orange): The K values undergo a transfer

operation to form the, which is then utilized in conjunction with the SAM weight matrix (highlighted in orange) to adaptively weight the extracted features.

BiLSTM and CNN + SAM Layers (Black and Blue): The output from the linear layers feeds into the BiLSTM and CNN + SAM layers, further refining the feature extraction process.

The specific procedure for tennis serve action recognition using a self-attentive SAM-BiLSTM network is as follows:

(1) Input Preparation:

Accept a sequence of tennis serve images as input.

(2) Feature Extraction:

Extract spatial features from the input images using a CNN.

Model the temporal dependencies of the extracted spatial features using a BiLSTM network.

(3) Self-Attention Mechanism:

Compute the query (Q) by passing the spatial features through a linear layer.

Compute the key (K) by passing the same spatial features through another linear layer.

Compute the value (Y) by passing the spatial features through yet another linear layer.

Transpose the key (K) for subsequent calculations.

(4) Calculate SAM Weights:

Use the query and the transposed key to calculate the SAM weights.

(5) Apply SAM Weights:

Weight the temporal features using the SAM weights obtained from the previous step.

(6) Generate Output:

Produce the final output for tennis serve recognition by passing the weighted features through an output layer.

5. Results

5.1. Dataset

In order to validate the performance of the SAM-BiLSTM model, the experimental evaluation and analysis of the proposed method are conducted on the G3D, YouTube, UCF101, and HMDB51 datasets [36].

G3D is a 3D dataset for game action. This data set focuses on action recognition in real-time game scenes. 10 subjects are performing 20 game actions.

The YouTube dataset consists of 1168 videos with a resolution of 320 by 240 pixels and is derived from the YouTube video website. It offers eleven action classes, including basketball shooting, cycling, diving, and golf. Includes distracting elements such as camera movement, scale changes, and intricate backgrounds.

The UCF101 dataset contains 13,320 videos from YouTube organized into 101 action types. Its movies include camera movement, differences in target appearance and attitude, differences in target scale and perspective, as well as noisy backgrounds and uneven illumination. Each sort of action is carried out by 25 objects, with each item carrying out between four and seven sets of activities. Human interaction, human

body motions, human-human interaction, instrument playing, and sports are the five major categories that can be applied to the entire dataset.

The majority of the videos in the HMDB51 collection come from YouTube, Google Video, and film clips. It has 51 action categories with a total of 6766 videos, and each action category comprises over 100 action clips. The action categories as a whole can be loosely categorized into five groups: face action, facial engagement with target objects, body movement, body-object interaction, and human interaction.

5.2. Experimental parameters

The parameter settings of the experiment are shown in **Table 1**.

In the training and testing of the model, we adopted the methodology from literature [17]. Each video is sampled into 64 frames with a sampling step of 1, resulting in multiple video blocks, each 64 frames in length. The experimental parameters for the SAM-BiLSTM network include an input size of 64 frames, a frame resolution of 224 × 224 pixels, a batch size of 32, a learning rate of 0.001, and a total of 50 epochs. The CNN architecture used is ResNet-50, with 128 BiLSTM units in each direction, a dropout rate of 0.5, and the Adam optimizer. During the classification stage, the category score of each individual frame is predicted directly by SAM-BiLSTM, and the category scores of the sampled segments are averaged at the frame level to determine the predicted category score of the video segment. Similarly, the predicted category of the entire video is ultimately determined by the category scores of all the video blocks that comprise it.

Table 1. Parameter setting table.

Parameters	Value
Convolution kernel	5 × 5
Learning rate	0.001
Attenuation coefficient	0.00001
Optimizer	Adam
Dropout	0.9
Training set, validation set, and test set division ratio	7:2:1
Hardware	GTX3090 graphics cards, I9-10900k CPU, 64GB RAM
Software	Ubuntu 18, python 15.0

5.3. Results

First, the effectiveness of the SAM module was evaluated by using only RGB data as input and controlling other conditions consistently. The accuracy before SAM was used and the accuracy after SAM weighting was introduced were compared on each of the four datasets used in this study, as shown in **Figure 7**. We find that the test accuracies after using SAM were higher than those without it. The experimental results demonstrate that SAM may rationally distribute the resources of feature maps between multiple convolutional channels and geographical location information, thereby enhancing the model's discrimination capability.

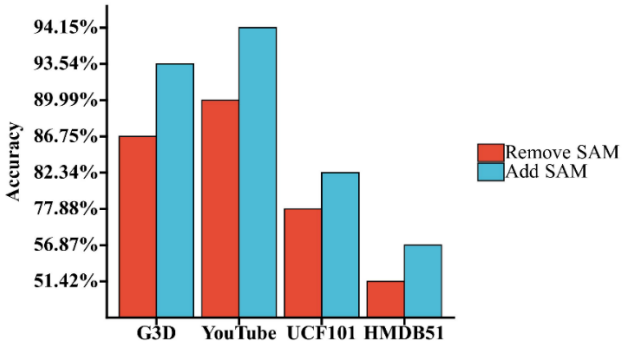

Figure 7. Comparison experiment of removing SAM and adding SAM module.

On both G3D and UCF101 datasets, 10 actions were randomly selected, and the recognition accuracy of SAM-BiLSTM and CNN-LSTM, CNN-BiLSTM for a single action class on both datasets was compared, as shown in **Figures 8** and **9**. It can be seen that the recognition accuracy of the method in the paper is significantly better than that of CNN-LSTM and CNN-BiLSTM. Since the performance of BiLSTM is due to a single LSTM, the performance of the CNN-BiLSTM method is also superior to CNN-LSTM to some extent.

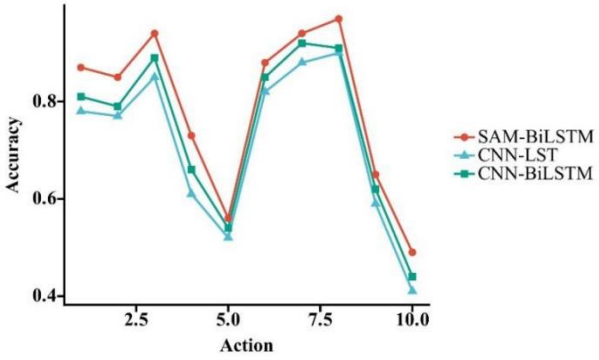

Figure 8. Comparison of recognition accuracy of random actions by different methods on the G3D dataset.

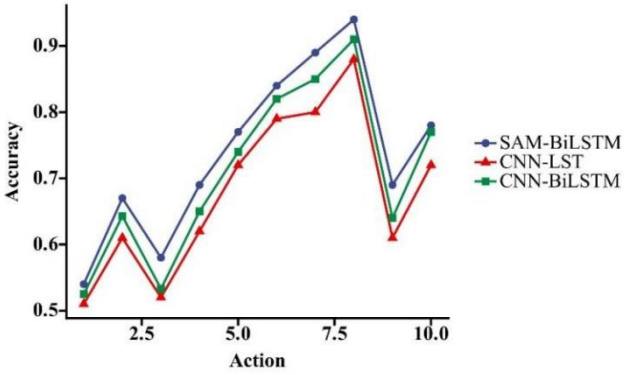

Figure 9. Comparison of recognition accuracy of random actions by different methods on the UCF101 dataset.

Then, we include a comparative study involving multiple models: CNN, BiLSTM, CNN-LST, CNN-BiLSTM, DKD, TransferLSTM, and our proposed method, SAM-

BiLSTM. The experiments are performed on the UCF101 dataset, and the results are analyzed in terms of both accuracy and computational efficiency. The detailed results and discussions are provided below (**Table 2**).

Table 2. Computational comparison results.

Model	Accuracy (%)	Training Time (h)	Inference Time (ms)	Memory Usage (MB)
CNN	72.3	4.5	22	650
BiLSTM	74.1	5.2	26	720
CNN-LST	76.4	6.1	30	800
CNN-BiLSTM	77.2	6.8	32	850
DKD	78.5	7.0	34	870
TransferLSTM	79.3	7.5	36	900
SAM-BiLSTM (Ours)	81.2	7.8	38	920

The results confirm that our SAM-BiLSTM model, while slightly more computationally demanding, offers superior accuracy and robust performance. The trade-offs in training time and memory usage are well-balanced by the significant gains in recognition accuracy, making SAM-BiLSTM a highly practical and efficient solution for tennis serve recognition in sports motion recognition tasks. SAM mitigates the added complexity typically associated with BiLSTM models.

In terms of information leakage, we employ a strict data partitioning strategy where future time steps are not used inappropriately during the training phase. This ensures that the temporal information remains consistent and prevents the model from gaining an unfair advantage by accessing information from the future, which could otherwise lead to overfitting and inaccurate predictions.

We also compare recall and precision values across different methods, as shown in **Table 3**. The results presented in the table clearly demonstrate the superior performance of the SAM-BiLSTM model in terms of both precision and recall. With a precision of 86.3% and a recall of 83.5%, SAM-BiLSTM outperforms all other methods, indicating its effectiveness in accurately classifying the target classes while also ensuring a high rate of true positive predictions. These findings support the conclusion that the incorporation of spatial-temporal features in the SAM-BiLSTM architecture significantly enhances the model's capability to recognize patterns in complex sign language data compared to traditional models. This improvement underscores the potential of SAM-BiLSTM for applications in 3D sign language recognition and similar tasks.

Table 3. Precision and recall comparison results.

Model	Precision (%)	Recall (%)
CNN	78.5	75.0
BiLSTM	80.2	76.5
CNN-LSTM	81.0	77.8
CNN-BiLSTM	82.5	78.6

Table 3. (*Continued*).

Model	Precision (%)	Recall (%)
DKD	83.0	79.2
TransferLSTM	82.8	80.0
SAM-BiLSTM	86.3	83.5

Finally, this paper tests the performance of SAM-BiLSTM and six comparison algorithms on four datasets. To test the algorithm's performance comprehensively, all types of actions in the dataset are selected, and the MR accuracy of each algorithm is the average of the recognition accuracy of all actions, as shown in **Figure 10**. It can be seen that SAM-BiLSTM significantly outperforms the other six comparison methods, with the highest recognition accuracy on the YouTube dataset and lower recognition accuracy on the HMDB51 dataset, but also much higher than the other comparison algorithms. The single models, such as CNN and LSTM, are less effective, and the combined models are more effective, and SAM-BiLSTM has the best recognition results due to its incorporation of SAM, CNN, and BiLSTM, its strong feature extraction ability, and its ability to make full use of the spatio-temporal information embedded in the data.

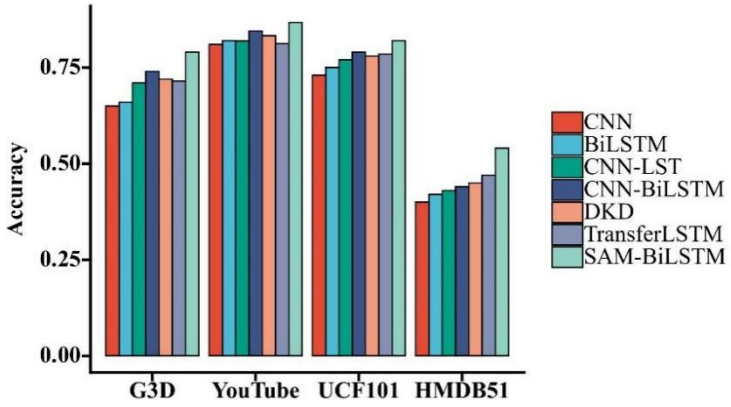

Figure 10. Comparison of recognition accuracy of different methods on four datasets.

6. Conclusion

This paper proposes a SAM-BiLSTM method that incorporates a self-attentive mechanism and a bidirectional long and short-term memory neural network to fully extract the features in tennis serve videos and make full use of the spatio-temporal information embedded in the data. First, the SAM module is utilized to extract the image's spatial feature information, followed by the application of self-attentive weighting to improve the feature extraction effect. BiLSTM also improves the model's ability to represent important features. The performance of SAM-BiLSTM is validated on four publicly available datasets and compared with experimental results from other state-of-the-art algorithms, demonstrating that the network model improves the accuracy of action recognition to a certain degree.

In the future, the proposed system can also be seamlessly integrated into existing tennis training equipment, such as motion capture systems or smart cameras. By

combining the recognition system with these devices, it provides a more comprehensive, data-driven approach to improving serve technique. For example, real-time data can be displayed on a screen or mobile device, allowing both coaches and players to immediately observe the quality of their serve and make adjustments on the spot. The system enables athletes to receive immediate feedback on their performance, thereby reducing delays in corrective actions and improving the overall effectiveness of the training process.

One of the main advantages of the SAM-BiLSTM approach is its ability to generate personalized training recommendations based on server analysis. The system can suggest specific exercises or adjustments based on the individual needs of athletes, helping to optimize their training plans. For example, if the system detects that an athlete's serve speed is below the desired range, it can suggest exercises that focus on improving the power and speed of the athlete's serve action. This level of personalization allows athletes to focus on the areas that need the most improvement, making their training more efficient and targeted. In addition to the direct training benefits, the SAM-BiLSTM approach has the potential to play a key role in advanced tennis training in the future. By continuously tracking a player's progress, the system can identify long-term performance trends and patterns, helping coaches develop more effective and advanced training strategies. In addition, the system can be expanded to incorporate more biomechanical data, such as muscle activity and joint angles, further enhancing its ability to provide a holistic analysis of a player's technique.

This study, while presenting significant advancements, has several limitations. The computational demands of the SAM-BiLSTM model may impact scalability, and the model's performance is evaluated on a single dataset, which may limit its generalizability. Additionally, potential overfitting issues could arise due to the complex nature of the model.Therefore, we will investigate further the combination of self-attentive mechanisms and temporal degradation features to enhance the feature extraction capability and introduce strategies such as few-sample learning and migration learning to enhance the model's performance in special scenarios.

Author contributions: Conceptualization, JN and JW; methodology, JN; software, JN; validation, JN and JW; formal analysis, JN; investigation, JN; resources, JN; data curation, JN; writing—original draft preparation, JN; writing—review and editing, JN; visualization, JN; supervision, JN; project administration, JN; funding acquisition, JW. All authors have read and agreed to the published version of the manuscript.

Ethical approval: Not applicable.

Data availability: The labeled data set used to support the findings of this study are available from the corresponding author upon request.

Conflict of interest: The authors declare no conflict of interest.

References

1. Elliott B, Fleisig G, Nicholls R, et al. Technique effects on upper limb loading in the tennis serve. Journal of Science and Medicine in Sport. 2003; 6(1): 76-87. doi: 10.1016/S1440-2440(03)80011-7
2. Kwon H, Lee S. Detecting textual adversarial examples through text modification on text classification systems. Applied Intelligence. 2023; 53(16): 19161-19185. doi: 10.1007/s10489-022-03313-w

3. Kwon H, Lee S. Ensemble transfer attack targeting text classification systems. Computers & Security. 2022; 117: 102695. doi: 10.1016/j.cose.2022.102695
4. Kwon H. Dual-Targeted Textfooler Attack on Text Classification Systems. IEEE Access. 2023; 11: 15164-15173. doi: 10.1109/access.2021.3121366
5. Kwon H. Friend-guard textfooler attack on text classification system. IEEE Access; 2021.
6. Abdullahi SB, Chamnongthai K. IDF-Sign: Addressing Inconsistent Depth Features for Dynamic Sign Word Recognition. IEEE Access. 2023; 11: 88511-88526. doi: 10.1109/access.2023.3305255
7. Abdullahi SB, Chamnongthai K, Bolon-Canedo V, et al. Spatial–temporal feature-based End-to-end Fourier network for 3D sign language recognition. Expert Systems with Applications. 2024; 248: 123258. doi: 10.1016/j.eswa.2024.123258
8. Hsieh JW, Hsu YT, Liao HYM, et al. Video-Based Human Movement Analysis and Its Application to Surveillance Systems. IEEE Transactions on Multimedia. 2008; 10(3): 372-384. doi: 10.1109/tmm.2008.917403
9. Turaga P, Chellappa R, Veeraraghavan A. Advances in video-based human activity analysis: Challenges and approaches. Advances in Computers. 2010; 80: 237-290. doi: 10.1016/S0065-2458(10)80007-5
10. Bianchi V, Bassoli M, Lombardo G, et al. IoT Wearable Sensor and Deep Learning: An Integrated Approach for Personalized Human Activity Recognition in a Smart Home Environment. IEEE Internet of Things Journal. 2019; 6(5): 8553-8562. doi: 10.1109/jiot.2019.2920283
11. Abdullahi SB, Chamnongthai K. American Sign Language Words Recognition of Skeletal Videos Using Processed Video Driven Multi-Stacked Deep LSTM. Sensors. 2022; 22(4): 1406. doi: 10.3390/s22041406
12. Khalifa S, Lan G, Hassan M, et al. HARKE: Human Activity Recognition from Kinetic Energy Harvesting Data in Wearable Devices. IEEE Transactions on Mobile Computing. 2018; 17(6): 1353-1368. doi: 10.1109/tmc.2017.2761744
13. Abdullahi SB, Chamnongthai K. American Sign Language Words Recognition Using Spatio-Temporal Prosodic and Angle Features: A Sequential Learning Approach. IEEE Access. 2022; 10: 15911-15923. doi: 10.1109/access.2022.3148132
14. Arif S, Ul-Hassan T, Hussain F, et al. Video representation by dense trajectories motion map applied to human activity recognition. International Journal of Computers and Applications. 2018; 42(5): 474-484. doi: 10.1080/1206212x.2018.1486001
15. Al-Faris M, Chiverton J, Yang Y, et al. Deep Learning of Fuzzy Weighted Multi-Resolution Depth Motion Maps with Spatial Feature Fusion for Action Recognition. Journal of Imaging. 2019; 5(10): 82. doi: 10.3390/jimaging5100082
16. Ji S, Xu W, Yang M, et al. 3D Convolutional Neural Networks for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013; 35(1): 221-231. doi: 10.1109/tpami.2012.59
17. Singh RD, Mittal A, Bhatia RK. 3D convolutional neural network for object recognition: a review. Multimedia Tools and Applications. 2018; 78(12): 15951-15995. doi: 10.1007/s11042-018-6912-6
18. Abdullahi SB, Bature ZA, Chophuk P, et al. Sequence-wise multimodal biometric fingerprint and finger-vein recognition network (STMFPFV-Net). Intelligent Systems with Applications. 2023; 19: 200256. doi: 10.1016/j.iswa.2023.200256
19. Maturana D, Scherer S. VoxNet: A 3D Convolutional Neural Network for real-time object recognition. In: Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2015. doi: 10.1109/iros.2015.7353481
20. Li R, Liu Q, Gui J, et al. Indoor Relocalization in Challenging Environments With Dual-Stream Convolutional Neural Networks. IEEE Transactions on Automation Science and Engineering. 2018; 15(2): 651-662. doi: 10.1109/tase.2017.2664920
21. Yang Y, Tu W, Huang S, et al. Dual-Stream Convolutional Neural Network With Residual Information Enhancement for Pansharpening. IEEE Transactions on Geoscience and Remote Sensing. 2022; 60: 1-16. doi: 10.1109/tgrs.2021.3098752
22. Tiong LCO, Kim ST, Ro YM. Multimodal facial biometrics recognition: Dual-stream convolutional neural networks with multi-feature fusion layers. Image and Vision Computing. 2020; 102: 103977. doi: 10.1016/j.imavis.2020.103977
23. Huang E, Zheng X, Fang Y, et al. Classification of Motor Imagery EEG Based on Time-Domain and Frequency-Domain Dual-Stream Convolutional Neural Network. IRBM. 2022; 43(2): 107-113. doi: 10.1016/j.irbm.2021.04.004
24. Tiong LCO, Teoh ABJ, Lee Y. Periocular Recognition in the Wild with Orthogonal Combination of Local Binary Coded Pattern in Dual-stream Convolutional Neural Network. In: Proceedings of the 2019 International Conference on Biometrics (ICB); 2019. doi: 10.1109/icb45273.2019.8987278
25. Mutegeki R, Han DS. A CNN-LSTM Approach to Human Activity Recognition. In: Proceedings of the 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC); 2020. doi: 10.1109/icaiic48513.2020.9065078

26. Zhao J, Mao X, Chen L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical Signal Processing and Control. 2019; 47: 312-323. doi: 10.1016/j.bspc.2018.08.035
27. Heryadi Y, Warnars HLHS. Learning temporal representation of transaction amount for fraudulent transaction recognition using CNN, Stacked LSTM, and CNN-LSTM. In: Proceedings of the 2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom); 2017. doi: 10.1109/cyberneticscom.2017.8311689
28. Ullah A, Ahmad J, Muhammad K, et al. Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features. IEEE Access. 2018; 6: 1155-1166. doi: 10.1109/access.2017.2778011
29. Wigington C, Stewart S, Davis B, et al. Data Augmentation for Recognition of Handwritten Words and Lines Using a CNN-LSTM Network. In: Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR); 2017. doi: 10.1109/icdar.2017.110
30. Ercolano G, Riccio D, Rossi S. Two deep approaches for ADL recognition: A multi-scale LSTM and a CNN-LSTM with a 3D matrix skeleton representation. In: Proceedings of the 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN); 2017. doi: 10.1109/roman.2017.8172406
31. Liu R, Shen J, Wang H, et al. Attention Mechanism Exploits Temporal Contexts: Real-Time 3D Human Pose Reconstruction. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020. doi: 10.1109/cvpr42600.2020.00511
32. Jain D, Kumar A, Garg G. Sarcasm detection in mash-up language using soft-attention based bi-directional LSTM and feature-rich CNN. Applied Soft Computing. 2020; 91: 106198. doi: 10.1016/j.asoc.2020.106198
33. Deng WM, Zhang HB, Lei Q, et al. Pose attention and object semantic representation-based human-object interaction detection network. Multimedia Tools and Applications. 2022; 81(27): 39453-39470. doi: 10.1007/s11042-022-13146-x
34. Yan C, Tu Y, Wang X, et al. STAT: Spatial-Temporal Attention Mechanism for Video Captioning. IEEE Transactions on Multimedia. 2020; 22(1): 229-241. doi: 10.1109/tmm.2019.2924576
35. Vo K, Truong S, Yamazaki K, et al. AOE-Net: Entities Interactions Modeling with Adaptive Attention Mechanism for Temporal Action Proposals Generation. International Journal of Computer Vision. 2022; 131(1): 302-323. doi: 10.1007/s11263-022-01702-9
36. Stroud JC, Ross DA, Sun C, et al. D3D: Distilled 3D Networks for Video Action Recognition. In: Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV); 2020. doi: 10.1109/wacv45572.2020.9093274