https://doi.org/10.53964/mit.2024003

https://www.innovationforever.com

Modern Intelligent Times

ISSN 2957-7942 (Online)

Open Access

Research Article

Human Pose Classification Based on Pose Long Short-Term Memory

Wanyi Li[1*], Jie Tan[2], Yingyin Fan[3]

[1]School of Computer Science, Guangdong University of Education, Guangzhou, Guangdong Province, China

[2]School of Mathematics, Guangdong University of Education, Guangzhou, Guangdong Province, China

[3]Library, Guangdong University of Education, Guangzhou, Guangdong Province, China

[*]**Correspondence to: Wanyi Li, PhD, Asssociate Professor** School of Computer Science, Guangdong University of Education, Guangzhou, 510303, Guangdong Province, China; E-mail: Chinaluther1212@163.com

Received: August 12, 2024 **Revised:** September 2, 2024 **Accepted:** September 20, 2024 **Published:** October 15, 2024

Abstract

Objective: At present, pose classification research is a hot topic in the field of computer vision, playing a crucial role in surveillance security, and motion data mining. The research serves as a crucial follow-up to three dimensional (3D) pose estimation and makes a significant contribution to the advancement of artificial intelligence. Pose classification is a complex problem involving a large amount of complicated data, making its digital modeling and further processing challenging.

Methods: This article proposes the Pose long short - term memory (LSTM) method to classify 3D poses reconstructed from 2D image sequences.

Result: Compared with traditional methods, the experimental tests show that the performance of the Pose LSTM is more superior in all aspects.

Conclusion: Pose LSTM is particularly effective for recognizing continuous action poses because they can capture temporal dependencies in sequential data. In the context of pose recognition, it can analyze the sequence of poses over time and understand how these poses transition from one to another, making it possible to accurately classify or predict actions that unfold over a series of frames.

Keywords: classification, pose, motion, data

Citation: Li W, Tan J, Fan Y. Human Pose Classification Based on Pose Long Short - Term Memory. *Mod Intell Times*, 2024; 2: 3. DOI: 10.53964/mit.2024003.

1 INTRODUCTION

In the realm of behavioral security monitoring, pose classification emerges as a key technology. This approach leverages advanced computer vision techniques to identify and categorize human postures and motions[1,2], enabling the detection and anticipation of potentially hazardous behaviors or unusual activities. Particularly critical in security-sensitive environments such as public transportation hubs, retail spaces, banking institutions, and other public venues, pose classification enhances

surveillance effectiveness and facilitates[3] timely responses to potential security threats.

The integration of pose classification into security systems combines real-time video analysis with machine learning algorithms that evaluate body positions and movement patterns. This capability is instrumental in identifying actions that deviate from normal behaviors, potentially indicating theft, vandalism, aggression, or other security concerns. For example, sudden movements or unusual postures near sensitive areas may trigger alerts that prompt security personnel to intervene.

Advancements in artificial intelligence have significantly enhanced the efficacy of pose classification technologies. Deep learning models[4], particularly those based on convolutional neural networks (CNNs) and their improved variants[5-9] excel at processing complex visual data from multiple camera angles, ensuring robust pose detection even in crowded or poorly lit environments. These models are trained on extensive datasets containing a wide range of human activities and can accurately differentiate between benign and potentially threatening behaviors.

Furthermore, the application of pose classification extends beyond simple anomaly detection. It provides valuable insights into crowd dynamics and behavior patterns, aiding in the design of safer public spaces. By analyzing how individuals move through and interact with these spaces, planners can implement design changes that enhance safety and reduce the likelihood of incidents.

Additionally, pose classification supports compliance with privacy regulations. Unlike facial recognition technologies, which involve more intrusive forms of personal identification, pose classification can ensure security without explicitly identifying individuals. This feature is particularly appealing in jurisdictions where privacy concerns are paramount.

Despite its many advantages, the deployment of pose classification in security systems is not without challenges. Issues such as data bias, the need for extensive training data, and the difficulty of handling ambiguous or overlapping poses can impact accuracy. Moreover, ethical concerns regarding surveillance and data management must be carefully addressed to maintain public trust and compliance with legal standards.

In the passing years, many research findings for three dimensional (3D) pose estimation have been proposed in the field of computer vision. These works supply the data 3D motion capture as illustrated in the Figure 1, such as two dimensional (2D) keypoints detection[10] and 3D motion reconstruction[11,12], which is providing a previous research foundation for posture classification. The 2D keypoints of the pose image are obtained through the image detection and annotation algorithms mentioned above. With these 2D annotated data, some advanced CNNs are used to reconstruct the 3D pose model. These 3D pose models are then used to train the proposed method (Pose long short - term memory (LSTM)) for 3D pose classification. The dataset is Humaneva dataset[13] in the Figure 1. The description of 3D poses is more accurate and effective than the forms in two-dimensional images. 3D poses are helpful for mathematical modeling and digital processing of various poses. Thus, a new method called pose long short-term memory (Pose LSTM) is proposed to achieve the human pose classification. Pose LSTM classifies 3D poses by directly processing and classifying the data from three-dimensional poses. Currently, many advanced algorithms extract features from two-dimensional image sequences and classify them based on the extracted image features. However, 2D image features often have ambiguity, thus their classification performance may be inferior compared to using three-dimensional image data for classification. The feature description of motion in 3D image will be more accurate than that in 2D image, because Pose LSTM establishes a mapping relationship between the 3D images model and their labels, thus the classification performance will be better. The contribution of this article is as follows: (1) Propose the pose long-term memory model for pose classification, which performs better than traditional methods; (2) Designed comprehensive experiment to test the proposed method (Pose LSTM).

2 RELATIVE WORK

For classification models, there are some traditional algorithms available. Compared to LSTM models, traditional machine learning methods like logistic regression[14] (LR), decision trees[15] (DT), random forest[16] (RF), support vector machine[17] (SVM), and basic neural networks[18] (such as feedforward neural network (FNN)) exhibit the following disadvantages when handling time-series data and complex pattern recognition tasks. There are the limits of these methods.

2.1 Logistic Regression

Lack of Capability to Handle Complex Nonlinear Patterns: These methods are based on linear assumptions and are inadequate for dealing with nonlinear relationships and complex data patterns, such as long-term dependencies in time-series data.

Inability to Capture Temporal Dynamics in Sequences: They do not account for the order in which data points appear, which is crucial for understanding temporal dynamics in time-series analysis.

2.2 Decision Trees and Random Forests

Limited Handling of Sequential Data: While powerful for classification and regression tasks within static

https://doi.org/10.53964/mit.2024003

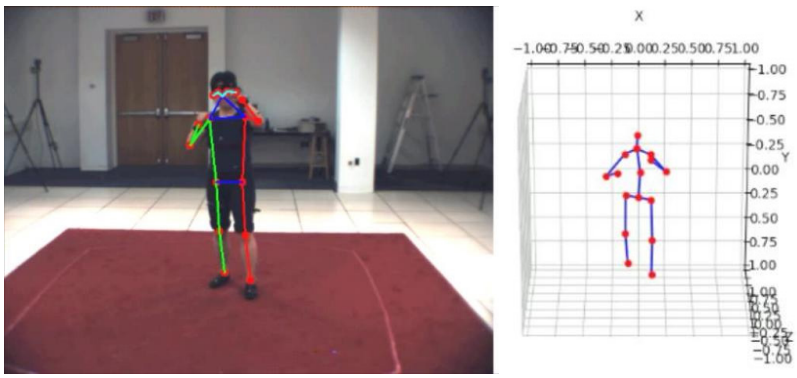

Figure 1. The previous work of the Pose Long short - term memory for Classification.

datasets, these models do not inherently process data in a sequence, making them less effective for tasks that require understanding of temporal relationships.

Susceptibility to Overfitting: Especially in the case of decision trees, there's a risk of overfitting to the noise in training data, which can degrade performance on unseen data.

2.3 SVMs

Challenges with Large-Scale Data: SVM can become computationally expensive as the size of the dataset increases, which is often the case with detailed time-series data.

Difficulty in Capturing Time-Series Dynamics: Without custom kernels designed to handle sequences, SVM struggle to model the time-dependent aspects inherent in time-series data.

2.4 Basic Neural Networks (FNN)

Lack of Memory for Past Inputs: These networks do not have a mechanism to remember previous inputs in a sequence, which is essential for tasks from which historical data influences current outcomes.

Inefficiency in Learning Temporal Patterns: Without recurrent structures, these networks cannot naturally learn dependencies over time, limiting their effectiveness in sequential prediction tasks.

In summary, while these traditional methods have their applications, their limitations become pronounced in scenarios requiring robust handling of temporal dynamics and complex, nonlinear relationships, areas where LSTM models excel due to their sophisticated design for sequence processing and long-term dependency management.

3 POSE LSTM FOR ACTION CLASSIFICATION

3.1 LSTM Model

Long Short-Term Memory (LSTM)[19-21] is a type of recurrent neural network (RNN)[22-24] architecture designed to model temporal sequences and their long-range dependencies more accurately than conventional RNNs. LSTMs are capable of learning long-term dependencies, which makes them suitable for various sequential data tasks such as language modeling, time series prediction, and more. There is the key Components of LSTM.

3.1.1 Forget Gate (f_t)

The forget gate decides what information should be discarded from the cell state. It takes the previous hidden state (h_{t-1}) and the current input (x_t) as inputs and applies a sigmoid activation function to generate a number between 0 and 1 for each cell state value.

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

3.1.2 Input Gate (i_t)

The input gate decides what new information should be stored in the cell state. It also takes the previous hidden state and the current input and applies a sigmoid activation function.

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

3.1.3 Candidate Cell State ($\hat{C}_t$)

A tanh layer creates a vector of new candidate values that could be added to the cell state. This is based on the previous hidden state and the current input.

$$\hat{C}_t = \tanh(W_c [h_{t-1}, x_t] + b_c) \quad (3)$$

3.1.4 Cell State (C_t) Update

The cell state is updated by combining the old cell state and the candidate cell state, modulated by the forget gate and the input gate.

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t \quad (4)$$

3.1.5 Output Gate (o_t)

The output gate decides what part of the cell state should be output. It applies a sigmoid activation function to the previous hidden state and the current input.

$$O_t = \sigma(W_o [h_{t-1}, x_t] + b_o) \quad (5)$$

https://doi.org/10.53964/mit.2024003

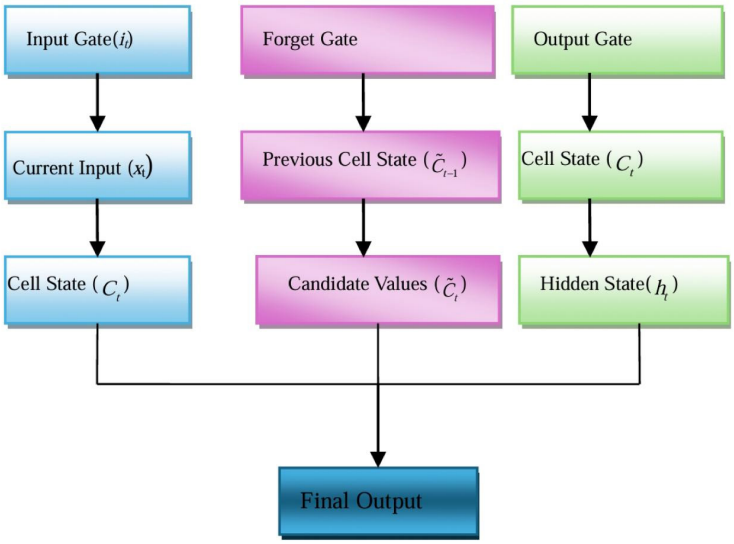

Figure 2. The working process of LSTM.

3.1.6 Hidden State Update (h_t)

The hidden state is updated by applying the output gate to the tanh of the cell state.

$$h_t = o_t * \tanh(C_t) \quad (6)$$

Notations from Equation (1) to Equation (6) are noted as follows. σ: Sigmoid activation function; tanh: Tanh activation function;

W_f, W_i, W_C, W_o: Weight matrices; b_f, b_i, b_C, b_o: Bias vectors; h_{t-1}: Previous hidden state; x_t: Current input; f_t: Forget gate output; i_t: Input gate output; $\hat{C}_t$: Candidate cell state; C_t: Cell state; o_t: Output gate output; h_t: Hidden state

3.2 Pose Classification Architecture Based on LSTM

The proposed Pose LSTM neural network is a deep learning model designed for pose classification tasks as shown in the Figure 3. This network leverages a Long Short-Term Memory (LSTM) layer to capture the temporal dependencies in the input data, followed by fully connected layers to classify the pose into one of the predefined classes. Compared to Bi-directional long short-term memory (BiLSTM) (one LSTM variant)[25], Attention-based LSTM (another LSTM variant)[26], and RNNs[22-24], Pose LSTM has several advantages and limitations. Pose LSTM is more computationally efficient compared to BiLSTM, as BiLSTM processes sequences in both forward and backward directions, leading to higher computational costs. However, BiLSTM can capture bidirectional dependencies in a sequence, which is useful for tasks requiring global context. If the task only relies on past information, Pose LSTM is more efficient. When compared to Attention-based LSTM, Pose LSTM has a simpler structure and does not use the complex attention mechanism, making it more efficient in terms of computation. However, if the pose data is long, and key points are not limited to the end of the sequence, Attention-based LSTM can improve classification accuracy. On the other hand, when the sequence is shorter or computational resources are limited, Pose LSTM may be more efficient. Compared to standard RNN, Pose LSTM solves the issue of vanishing gradients in long sequences, making it better for capturing long-term dependencies and handling complex sequences. While RNNs are simpler and faster, they are less stable in tasks with long time dependencies. In summary, Pose LSTM's simple structure and computational efficiency make it well-suited for medium-scale pose classification tasks, while BiLSTM and Attention-based LSTM may be more appropriate for tasks requiring the capture of complex or long-sequence information.

3.3 Architecture (Pose LSTM)

3.3.1 LSTM Layer

The network begins with an LSTM layer, which is configured with the following parameters: input_size: this is set to num_joints * num_dimensions, indicating that each time step's input is a flattened representation of the joint coordinates; hidden_size: this is set to 128, meaning the LSTM has 128 hidden units; num_layers: this is set to 2, indicating a two-layer stacked LSTM; batch_first: this parameter is set to true, indicating that the input tensor has the batch size as its first dimension.

3.3.2 Fully Connected Layers

The output from the LSTM is passed through two fully connected layers: fc1: This layer reduces the dimensionality from 128 to 64 and applies a ReLU activation function; fc2: This layer further reduces the dimensionality from 64 to num_classes, producing the final class scores.

3.3.3 Dropout Layer

A dropout layer with a dropout rate of 0.5 is applied between the two fully connected layers to prevent overfitting by randomly setting some of the activations to zero during training.

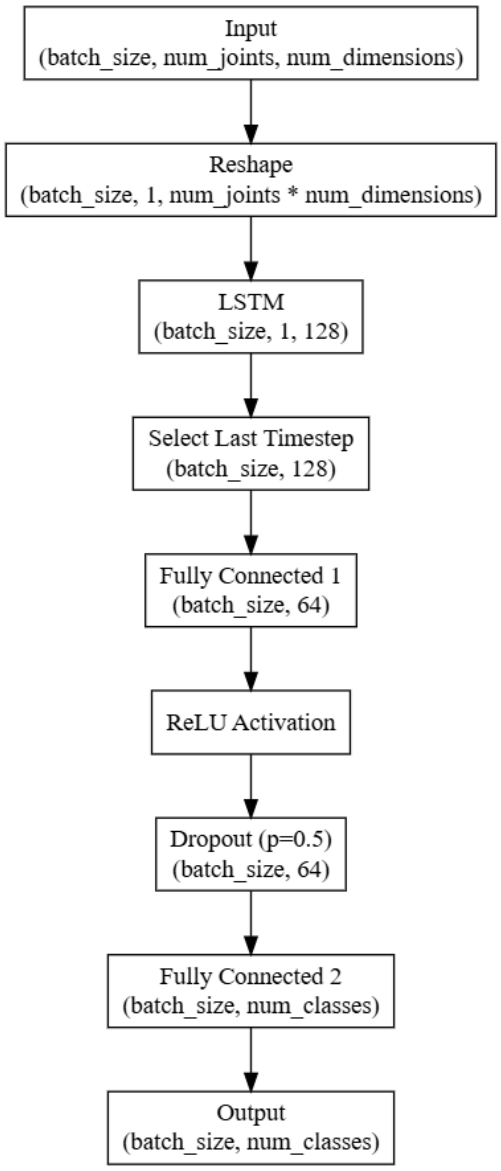

Figure 3. The network architecture of Pose LSTM.

3.4 Execution Steps

3.4.1 Reshaping Input

The input tensor x, which has the shape (batch_size, num_joints, num_dimensions), is reshaped to (batch_size, 1, num_joints * num_dimensions) to match the expected input shape for the LSTM layer.

3.4.2 LSTM Processing

The reshaped input is passed through the LSTM layer, and the output of the last time step is extracted. This output captures the temporal dependencies in the sequence.

3.4.3 Classification

The output from the LSTM is then passed through the first fully connected layer with ReLU activation, followed by a dropout layer. The cross entropy loss function, commonly used in classification tasks, is defined as.

$$\text{Loss} = -\sum_{c=1}^{c} y_c \log p_c \quad (7)$$

Where, C is the number of classes, y_c is the indicator for the true class (ground truth), $y_c = 1$ if the sample belongs to class c, and $y_c = 0$ otherwise. P_c is the predicted probability for class c.

Finally, the processed output is passed through the second fully connected layer to produce the final class scores.

4 EXPERIMENT AND EVALUATION

4.1 Pose LSTM Test

The actions of jogging, boxing smoking and walking are used to test the trained model (Pose LSTM). The data sets such as Humaneva[13] and Human 3.6M[27] can be used to train, valid and test the Pose LSTM. The accuracy of recognition is as the Table 1.

The accuracy of a classification model for each class is calculated as the ratio of the number of correctly predicted instances of that class to the total number of instances of that class. Mathematically, this can be expressed as follows.

$$A_i = \frac{TP_i}{N_i} \quad (8)$$

TP_i be the number of true positives (correct predictions) for class i, and N_i is the total number of instances for class i. The accuracy A_i for class i is given by Equation (8). The experiment total accuracy can be calculated by.

$$A_{total} = \frac{\sum_{i=1}^{N_c} TP_i}{\sum_{i=1}^{N_c} N_i} \quad (9)$$

The A_{total} is all the tested motion frame number, N_c is the number of classes for the experiment.

4.2 Comparison with Other Traditional Algorithms

The Pose LSTM is tested in this experiment, by comparison of several traditional algorithms including LR FNN, DT, RF, and SVM.The experimental results can be seen as the Table 2. Several traditional algorithms are compared with the Pose LSTM model. From the experiments in Table 2, it can be found that Pose LSTM is better than other traditional methods. The experiment randomly selected frames with relevant motion style. The total accuracy of LSTM can achieve 99.35%. During the experiments, the same frame sequences are selected to test all the methods. The evaluation metrics can be calculated as the following description. TP, FP, and FN are correspondingly denoting the true positives, false positives and false negatives.

Precision measures the proportion of true positive predictions out of all positive predictions made by the model. It indicates how many of the predicted positive samples are positives actually.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

https://doi.org/10.53964/mit.2024003

Table 1. The Accuracy of Recognition

Motion Style	Jog	Box	Smoke	Walk	Total
Correct prediction number	143	149	678	513	1483
Total frame number	144	149	721	526	1540
Accuracy	99.31%	100.00%	94.04%	97.53%	96.30%

Table 2. Experiments of Classification Among the Tested Method

Tested Method	Motion Style	Precision	Recall	F1-score	Support(Frames)	Total Accuracy (%)
Pose LSTM	jog	1.000	0.9984	0.9992	644	99.35%
	box	1.000	1.000	1.000	609	
	smoke	0.9922	0.9942	0.9932	3447	
	walk	0.9919	0.9895	0.9907	2485	
LR	jog	0.9780	0.8960	0.9352	644	88.96%
	box	1.000	1.000	1.000	609	
	smoke	0.8974	0.8932	0.8953	3447	
	walk	0.8325	0.8559	0.8440	2485	
FNN	jog	1.0000	0.9907	0.9953	644	97.63%
	box	1.0000	1.0000	1.0000	609	
	smoke	0.9910	0.9611	0.9758	3447	
	walk	0.9461	0.9879	0.9665	2485	
DT	jog	0.9798	0.9783	0.9790	644	94.60%
	box	0.9870	0.9967	0.9918	609	
	smoke	0.9493	0.9446	0.9469	3447	
	walk	0.9227	0.9272	0.9249	2485	
RF	Jog	1.0000	1.0000	1.0000	644	98.34%
	box	0.9951	1.0000	0.9975	609	
	smoke	0.9932	0.9730	0.9830	3447	
	walk	0.9636	0.9895	0.9764	2485	
SVM	jog	1.0000	1.0000	1.0000	644	91.96%
	box	1.0000	1.0000	1.0000	609	
	smoke	0.9415	0.8874	0.9137	3447	
	walk	0.8554	0.9235	0.8882	2485	

High precision means that when the model predicts a sample as positive, it is likely to be correct. A high precision means that when the model predicts a sample as positive, it is likely to be correct.

Recall measures the proportion of true positive predictions out of all actual positive samples. It indicates how many of the actual positive samples were correctly identified by the model.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (11)$$

A high recall means that the model identifies most of the positive samples, but it may include some false positives.

The F1-Score is the harmonic mean of Precision and Recall. It provides a single metric that balances both Precision and Recall, especially useful when the class distribution is imbalanced.

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

The F1-Score is useful when you need to balance Precision and Recall. It is particularly valuable in scenarios where false positives and false negatives have a similar impact.

Pose LSTM outperforms the traditional algorithms above in 3D data classification due to its unique ability to handle data characteristics and model structure. Pose LSTM is a type of RNN specifically designed for sequential data, allowing it to capture temporal dependencies and dynamic features, which are crucial for 3D pose classification since pose data typically unfolds over time. The memory units in Pose LSTM enable it to retain long-term contextual information and capture global patterns in pose variations.

https://doi.org/10.53964/mit.2024003

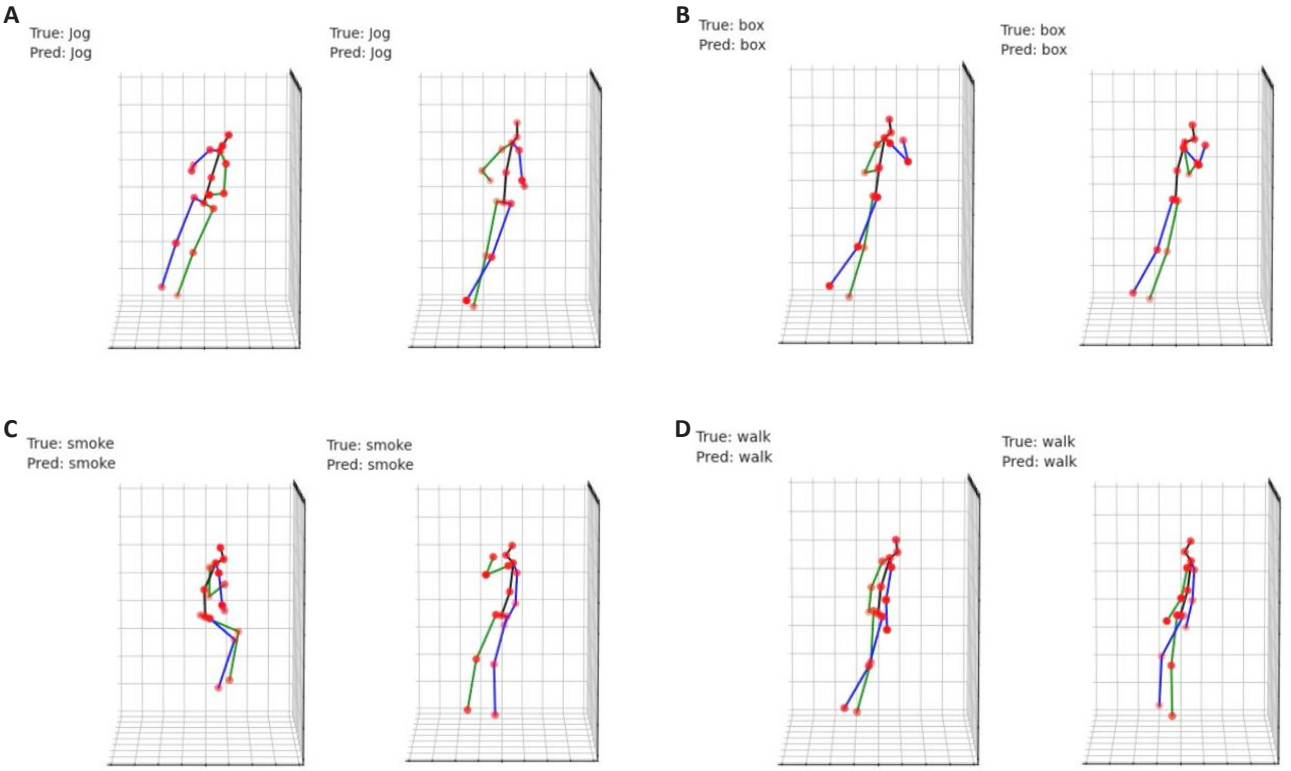

Figure 4. The tested frames for classification in the videos.

In contrast, traditional algorithms are more focused on static features and cannot effectively leverage the complex relationships in time series data. Moreover, while traditional models rely on manually extracted features, Pose LSTM can automatically learn features from raw data, reducing the complexity of feature selection. Structurally, its multi-layer recurrent design allows it to capture the nonlinear relationships within the data more effectively, whereas traditional algorithms often rely on limited decision rules for classification, making them less adaptable to complex 3D data. Therefore, LSTM can better establish the mapping relationships needed for accurate classification, improving its overall performance in 3D data tasks.

4.3 The Classification Test in the Videos

For testing the video classification performance of the Pose LSTM, a 3D pose frames are randomly selected for classification.

The video frames with 3D poses can be classified, and the results are shown in the following Figure 4. From the Figure 4, it can be found that the classification results are conforming to the category to which its posture belongs with good performance and accurate recognition. In the Figure 3, the true label is displayed by "Ture", and the predicted label is displayed by "Pred". The 3D skeleton model is used to describe the human posture, distinguishing left and right with different colors.

5 DISCUSSION

The primary advantage of pose LSTM stems from their ability to capture temporal dependencies and dynamics within sequential data. This capability is crucial for effectively modeling the sequence and duration of movements in pose classification.

Unlike models such as the traditional methods (SVM, DT, RF, FNN, and LR) which treat data points independently, Pose LSTM consider the entire sequence, allowing them to understand and predict based on the context established by previous data points in the sequence. This makes it particularly suited for tasks where the order and timing of events are predictive of the outcome, such as the sequences observed in pose movements.

Moreover, Pose LSTM can maintain information over longer sequences without the risk of losing important details to time, which is a common issue with simpler models that lack a mechanism to carry information across longer time spans. This attribute makes Pose LSTM more robust and accurate for complex pose classification scenarios where the precise ordering and timing of movements are critical.

Pose LSTM, Transformers (including the variants), and advanced CNNs each offer distinct advantages depending on the data characteristics they process. Pose LSTM excels at handling sequential data with temporal dependencies, making it highly effective for tasks that require capturing long-term relationships, such as language modeling and action recognition. Transformers and the variants, on the other hand, overcome LSTM's limitations by using attention mechanisms, allowing them to process entire

sequences simultaneously. This makes Transformers particularly adept at capturing global dependencies, especially in long-sequence tasks and natural language processing. Advanced CNNs, however, specialize in handling visual data like images and videos, as they can automatically extract complex spatial features. This makes CNNs ideal for tasks such as image classification and object detection. The choice of model depends on the task's requirements and the nature of the data, whether it emphasizes temporal dependencies or complex spatial patterns.

Overall, Pose LSTM, Transformers (including the variants), and advanced CNNs each have their strengths. Pose LSTM excels at handling sequential data, Transformers and the variants are strong in capturing long sequences and global dependencies, while advanced CNNs are effective at extracting spatial features from visual data. For pose classification tasks, Pose LSTM is particularly suitable, as it efficiently captures temporal dependencies and dynamic pose variations, which highlights its superior performance in this area.

6 CONCLUSION AND FUTURE WORK

Pose LSTM is particularly effective for recognizing continuous action poses because they can capture temporal dependencies in sequential data. In the context of pose recognition, it can analyze the sequence of poses over time and understand how these poses transition from one to another, making it possible to accurately classify or predict actions that unfold over a series of frames.

By remembering and relating past pose information to current data, Pose LSTM can handle variations in the speed and timing of actions, leading to more robust and accurate recognition of complex and continuous motion sequences. This capability is crucial when distinguishing between different types of actions or gestures that might look similar in individual frames but differ when viewed as part of a sequence.

There are some limitations in Pose LSTM, which will be primarily designed for temporal sequences and may struggle to effectively capture the spatial relationships between body parts critical in pose recognition. Pose LSTM process data sequentially, limiting the ability to parallelize operations, which can lead to slower performance compared to architectures like CNNs that handle multiple data points simultaneously. Pose LSTM might have difficulty with distinguishing between very similar actions that have subtle differences in motion patterns. For example, differentiating between walking and jogging might be challenging if the differences are minimal and highly context-dependent. The fine-grained action sensitivity needs to be improved.

To enhance the performance of models dealing with tasks such as pose recognition or sequential action identification, future improvements could significantly benefit from incorporating advanced techniques and exploring innovative architectures. One promising direction is the integration of attention mechanisms into Pose LSTM. Attention mechanisms can help the method focus on the most relevant parts of the input data, which is particularly useful in distinguishing subtle differences in actions or gestures. This can lead to more accurate and efficient processing, as the method allocates more resources to the crucial segments of the data.

Furthermore, exploring different variants of LSTM could yield improvements. For instance, variants like BiLSTM[25] process data in both forward and reverse directions, enhancing the context available to the model and potentially increasing its ability to recognize complex patterns. Gated Recurrent Units (GRUs) are another variant that simplifies the LSTM architecture by using fewer gates, which can reduce the model complexity and computational demands while maintaining similar performance.

Pose classification is an important research topic[28-30] in computer vision. There are some future important applications of pose classification. Firstly, pose classification technology can precisely capture a user's body movements, enabling augmented reality (AR) / virtual reality (VR) systems[31] to respond to user gestures in real-time, thereby enhancing the immersive experience. In virtual training and entertainment, recognizing user poses allows for more natural and interactive experiences. The Pose LSTM can be highly beneficial for pose classification in AR / VR systems because they excel at processing sequential data. In AR / VR systems, pose data is often collected as a time-series from sensors like Inertial Measurement Units or cameras, where each frame or data point is influenced by previous ones. The method can capture temporal dependencies and relationships within this sequence, allowing for more accurate classification of dynamic poses. By maintaining memory of previous frames, it can effectively handle complex, continuous movements, making them ideal for tasks like gesture recognition, body pose classification, and interaction tracking in AR / VR environments. Secondly, pose classification can be used to analyze an athlete's or fitness enthusiast's movements in real-time, providing corrective feedback, reducing the risk of injury, and optimizing training outcomes. This technology is beneficial for smart fitness devices and apps, helping users maintain correct exercise postures. Thirdly, in surveillance systems, pose classification can help identify suspicious behaviors or actions, improving the effectiveness of security monitoring. It can also be used to track and analyze crowd movements in public spaces, aiding in crowd management and safety measures.

Acknowledgements

This work was supported by the Special Funds for the

Cultivation of Guangdong College Students' Scientific and Technological Innovation. ("Climbing Program" Special Funds.) (No. pdjh2024b278), the Guangdong Provincial Special Program in Key Areas for Higher Education Institutions (New Generation Electronic Information (Semiconductors)) (No. 2024ZDZX1040), the Development of Guangzhou Philosophy and Social Science in 14th Five-Year Plan (No. 2023GZGJ171), the Educational Science Planning Project of Guangdong Province (No. 2022GXJK073, No. 2023GXJK125, No. 2024GXJK151), the Collaborative Project for the Development of Guangdong Province Philosophy and Social Science (No. 2023GD23XTY05), the National Undergraduate Innovation Training Project of China (No. 202414278015), the Special Support Program for Cultivating High-Level Talents of Guangdong University of Education (2022 Outstanding Young Teacher Cultivation Object: Wanyi Li). International Training Program for Excellent Young Scientific Research Talents from the Department of Education of Guangdong Province (Recipient of funding in 2023: Wanyi Li).

Conflicts of Interest

The authors declared no conflict of interest.

Author Contribution

Li W wrote the manuscript and designed the mechanism. Tan J performed the data analysis. Fan Y designed the experiment for testing the mechanism. All the authors contributed to writing the article, read and approved its submission.

Abbreviation List

2D, Two dimensional
3D, Three dimensional
AR, Augmented reality
BiLSTM, Bi-directional long short-term memory
CNNs, Convolutional neural networks
DT, Decision trees
FNN, Feedforward neural network
LR, Logistic regression
LSTM, Long short - term memory
RF, Random forests
RNN, Recurrent neural network
SVM, Support vector machine
VR, Virtual reality

References

[1] Khan A, Kim C, Kim JY et al. Sleep Posture Classification Using RGB and Thermal Cameras Based on Deep Learning Model. *Cmes-Comp Model Eng*, 2024; 140: 1729-1755.[DOI]

[2] Wang J, Deng H, Wang Y et al. Multi-sensor fusion federated learning method of human posture recognition for dual-arm nursing robots. *Inform Fusion*, 2024; 107: 102320.[DOI]

[3] Jain DK, Zhao X, Gan C et al. Fusion-driven deep feature network for enhanced object detection and tracking in video surveillance systems. *Inform Fusion*, 2024; 109: 102429.[DOI]

[4] Alzubaidi L, Bai J, Al-Sabaawi A et al. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *J Big Data-Ger*, 2023; 10: 46.[DOI]

[5] Salehi AW, Khan S, Gupta G et al. A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope. *Sustainability*, 2023; 15: 5930.[DOI]

[6] Zavala-Mondragon LA, Lamichhane B, Zhang L et al. CNN-SkelPose: a CNN-based skeleton estimation algorithm for clinical applications. *J Amb Intel Hum Comp*, 2020; 11: 2369-2380.[DOI]

[7] Jung M, Lee J, Kim J. A lightweight CNN-transformer model for learning traveling salesman problems. *Appl Intell*, 2024; 2: 1-17.[DOI]

[8] Linck I, Gómez AT, Alaghband G. SVG-CNN: A shallow CNN based on VGGNet applied to intra prediction partition block in HEVC. *Multimed Tools Appl*, 2024; 83: 73983-74001.[DOI]

[9] Mohammadpour L, Ling TC, Liew CS et al. A Survey of CNN-Based Network Intrusion Detection. *Appl Sci-Basel*, 2022; 12: 8162.[DOI]

[10] Waldmann U, Chan AHH, Naik H et al. 3D-MuPPET: 3D Multi-Pigeon Pose Estimation and Tracking. *Int J Comput Vision*, 2024; 3: 1-23.[DOI]

[11] Zhang XY, Zhou ZC, Han Y et al. Deep learning-based real-time 3D human pose estimation. *Eng Appl Artif Intel*, 2023; 119: 105813.[DOI]

[12] Ruescas-Nicolau AV, Medina-Ripoll EJ, Parrilla Bernabé E et al. Multimodal human motion dataset of 3D anatomical landmarks and pose keypoints. *Data Brief*, 2024; 53: 110157.[DOI]

[13] Sigal L, Balan AO, J Black M. Humaneva: synchronized video and motion capture dataset for evaluation of articulated human motion. *Int J Comput Vision*, 2006; 87: 3-27.[DOI]

[14] Mood C. Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. *Eur Sociol Rev*, 2010; 26: 67-82.[DOI]

[15] Grana C, Costantino R, Borghesani D. Optimized Block-Based Connected Components Labeling With Decision Trees. *IEEE Trans Image Process*, 2010; 19: 1596-1609.[DOI]

[16] Perdiguero-Alonso D, Montero FE, Kostadinova A et al. Random forests, a novel approach for discrimination of fish populations using parasites as biological tags. *Int J Parasitol*, 2008; 38: 1425-1434.[DOI]

[17] Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L et al. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 2020; 408: 189-215.[DOI]

[18] Scarselli F, Tsoi AC. Universal Approximation Using Feedforward Neural Networks: A Survey of Some Existing Methods, and Some New Results - ScienceDirect. *Neural Networks*, 1998; 11: 15-37.[DOI]

[19] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*, 1997; 9: 1735-1780.[DOI]

[20] Bilen B, Horasan F. LSTM Network Based Sentiment Analysis for Customer Reviews. *J Polytech*, 2022; 25: 959-966.[DOI]

[21] Geng YM. Design of English teaching speech recognition

system based on LSTM network and feature extraction. *Soft Comput*, 2023; 1: 1-11.[DOI]

[22] Moudgollya R, Sunaniya AK, Bhattacharjee RK. Efficient Multi-Object Tracking Using RNN and Siamese Re-Identification. *J Circuit Syst Comp*, 2024; 1: 2450298.[DOI]

[23] Golshanrad P, Faghih F. DeepCover: Advancing RNN test coverage and online error prediction using state machine extraction. *J Syst Software*, 2024; 211: 111987.[DOI]

[24] Talukdar K, Sarma SK. Deep Learning based Part-of-Speech tagging for Assamese using RNN and GRU. *Procedia Comput Sci*, 2024; 235: 1707-1712.[DOI]

[25] Liu G, Guo J. Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 2019; 337: 325-338.[DOI]

[26] Liao Y, Lin R, Zhang R et al. Attention-based LSTM (AttLSTM) neural network for Seismic Response Modeling of Bridges. *Comput Struct*, 2023; 275: 106915.[DOI]

[27] Ionescu C, Papava D, Olaru V et al. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *Ieee T Pattern Anal*, 2014; 36: 1325-1339.[DOI]

[28] Zheng C, Wu W, Chen C et al. Deep Learning-based Human Pose Estimation: A Survey. *Acm Comput Surv*, 2023; 56: 1-37.[DOI]

[29] Dubey S, Dixit M. A comprehensive survey on human pose estimation approaches. *Multimedia Syst*, 2023; 29: 167-195.[DOI]

[30] Garg S, Saxena A, Gupta R. Yoga pose classification: a CNN and MediaPipe inspired deep learning approach for real-world application. *J Amb Intel Hum Comp*, 2023; 14: 16551-16562.[DOI]

[31] Chang E, Lee Y, Billinghurst M et al. Efficient VR-AR communication method using virtual replicas in XR remote collaboration. *Int J Hum-Comput St*, 2024; 190: 103304.[DOI]