Basketball Motion Recognition Model Analysis Based on Perspective Invariant Geometric Features in Skeleton Data Extraction

Jiaojiao Lu
Sport Institute, Yantai Institute of Science and Technology, Yantai 265600, China

***Abstract*—The study proposes a recognition method based on skeleton data to address the basketball action recognition, especially those posed by viewpoint changes in videos. The key of this method is to extract geometric features of viewpoint invariance and combine them with spatio-temporal feature fusion techniques. In addition, the study constructs a dynamic topological map of the human skeleton based on long and short-term neural networks to improve the model performance. The experimental results showed that the research method had an average accuracy of 97.85% for Top-5 metrics on the Kinetics dataset and 97.82% for Top-5 metrics on the NTU RGB+D dataset. It is significantly better than the other three state-of-the-art methods. According to the experimental results, it achieves efficient and stable basketball action recognition, which is significantly superior to existing methods. This research not only provides a more efficient method for basketball motion recognition, but also provides valuable references for other sports action recognition fields.**

***Keywords*—Skeleton data; perspective invariance; geometric features; basketball recognition; spatio-temporal feature fusion**

I. Introduction

Basketball, as a global sport, not only attracts countless spectators and enthusiasts, but also becomes the focus of in-depth discussions in academic research and business. With the development of technology, analyzing the importance of basketball matches from a technical perspective is becoming increasingly prominent [1]. Among them, basketball action recognition technology plays an important role in player skill analysis and training assistance [2]. However, traditional visual basketball action recognition methods rely on the analysis of continuous video frames. They are limited by background noise, complex environments, and especially the change of camera viewpoints, which makes the action recognition face great challenges [3]. Skeletal data provide a more abstract and concise representation of movements than traditional methods, and are more effective in resisting the effects of perspective changes [4]. However, due to the poor fusion effect of spatio-temporal features in skeleton data, the accuracy of existing recognition models is not ideal [5]. To address this problem, the study proposes a novel basketball sports recognition model, which is based on viewpoint invariant geometric features and aims to improve the accuracy and robustness of skeleton data recognition. In addition, the study also explores the method of constructing a dynamic topological map of the human skeleton to further improve the model performance. This study consists of five parts. The first part is an overview of this study. The second part is a summary of relevant research. The third part has two sections. Firstly, a dynamic topology map of the human skeleton is constructed. The second section introduces a basketball motion recognition method based on perspective invariant geometric features. In the fourth section, the proposed method is tested on a dataset and in a real environment. The results are analyzed. The fifth section is a summary and outlook for this study. This study develops a new basketball action recognition system. It not only utilizes skeleton data more efficiently, but also shows significant improvement in both recognition accuracy and robustness. It is expected that this study can provide strong technical support for basketball game analysis, teaching assistance and other related applications.

II. Related Works

In the action recognition, feature extraction and fusion are the foundation of accurate recognition. Therefore, a large number of scholars have conducted in-depth research on this field [6]. Zhou W et al. proposed a novel feature fusion network to improve the performance of object detection under low light and uneven lighting conditions. This study applied a cross modal fusion module to fuse the corresponding size features of target detection and thermal modes. Then, the bidirectional reverse fusion module was used to achieve bidirectional fusion of foreground and background information. In the experimental results, the proposed feature fusion network was superior to other advanced methods [7]. Zhang X et al. proposed a network that combined multi-scale hierarchical feature fusion and mixed convolutional attention to solve the single image defogging in image recognition. By fusing multi-scale layered features, the haze level and image structure information were accurately estimated, resulting in the restored image containing less residual haze. According to the experimental result, it exceeded the most advanced defogging algorithm [8]. Choi H et al. proposed a new multimodal image feature fusion module to solve the transmission line inspection. The output of the multi branch feature extraction block was aggregated into an attention vector in the channel attention block. Each input feature was recalibrated. According to the experimental result, it was superior to the single mode input [9].

With the emergence of various optimization techniques, the improved action recognition models are gradually receiving attention. To solve the distinguishable feature extraction in skeleton action recognition model, Song Y F et al.

embedded an advanced separable convolutional layer into the Multiple Input Branches network. The research constructed an efficient Graph Convolutional Network (GCN) baseline. A composite extension strategy was designed to synchronously extend the width and depth. In the NTU 60 dataset testing, the accuracy reached 92.1% [10]. Li M et al. proposed a symbiotic graph neural network for 3D skeleton motion recognition. The network structure included a basic part, an action recognition part, and a motion prediction part. In the backbone network, a multi-scale GCN based on joints and parts was used to extract key features. Compared with the current methods, the method had better performance on all four datasets [11]. Li C et al. proposed a new temporal and spatial recalibration method to address the complex changes in skeletal joints in motion recognition. The research constructed a novel temporal attention mechanism based on residual learning to calibrate the frames of skeleton data. Compared with the most advanced methods, the research method significantly improved performance. It had the best results on six action recognition datasets [12].

In summary, there have been many studies in the action recognition, especially in feature extraction and fusion. However, most of these studies have focused on improving the performance of object detection under specific environmental conditions, specific problems in image processing, or specific applications. These approaches are often not applicable to complex sports scenarios, especially in basketball action recognition. They fail to adequately address the challenges posed by changes in perspective [13]. This research fills this gap by proposing a new basketball motion recognition model. It focuses on extracting viewpoint-invariant geometric features from skeleton data and incorporates spatio-temporal feature fusion techniques. Compared with existing studies, the innovation of the study is follows. It proposes an action recognition method for complex sports scenarios, which specifically addresses the viewpoint invariance in basketball. Further, the study effectively improves the accuracy and robustness of the model by constructing a dynamic topological map of the human skeleton based on long and short-term neural networks. In contrast to existing research in GCNs, symbiotic graph neural networks, and temporal re-space recalibration methods, the study focuses on combining action recognition with perspective invariance. It is an area rarely covered in existing research. As a result, the study not only makes significant improvements in the efficiency and accuracy of basketball action recognition, but also provides a new perspective to explore the action recognition problem. It provides a valuable reference for research in this area.

III. Construction of Basketball Motion Recognition Model Based on Perspective Invariant Geometric Features in Skeleton Data Extraction

To achieve more efficient and stable basketball motion recognition results, the study optimizes the model from two aspects. Firstly, based on short-term and short-term memory neural networks, a dynamic topology map of the human skeleton is constructed to provide more accurate data support for the recognition model. Then, a new spatio-temporal feature fusion method is proposed based on the perspective invariant geometric features extracted from skeleton data. On this basis, the research constructs a basketball motion recognition model.

A. Dynamic Topology Map Construction of Human Skeleton

Basketball action recognition based on human skeleton data aims to identify the types of actions represented by human skeleton time series. Traditional recognition methods have two categories. One is based on manual features, which mainly capture the dynamic relationships of joints through manual design features. The second is based on deep learning. It conducts end-to-end modeling with recurrent neural networks to capture joint information [14]. In practical scenarios, an action consists of multiple video frames. It is difficult to accurately display the dependency relationship between joints and bones by manually creating a topology map. The individual training and parameters for each video frame not only require a large amount of computation, but may also lead to catastrophic forgetting. Therefore, the research adopts a continuous learning method. The Long Short-Term Memory (LSTM) is utilized to dynamically construct the topology map of the human skeleton to improve the recognition performance of the basketball motion recognition model. The human skeleton sequence is composed of continuous human skeleton frames. Each frame is a set of joint coordinates and coordinates confidence. The definition of the human skeleton sequence is shown in Eq. (1).

$$VT = \{VT_1, VT_2, ..., VT_T\} \tag{1}$$

In Eq. (1), T represents the number of human skeleton frames. VT_i represents the i-th human skeleton frame in the human skeleton sequence. VT_i is composed of joint points containing spatio-temporal information, as shown in Eq. (2).

$$VT_i = \{V_1, V_2, ..., V_N\} \quad (1 \le i \le T) \tag{2}$$

In Eq. (2), N represents the joint points in the human skeleton frame. The i-th joint point V_i in the human skeleton frame is shown in Eq. (3).

$$V_i (1 \le i \le N) = (x_i, y_i, score_i) \tag{3}$$

In Eq. (3), (x_i, y_i) represents the position information of the joint point. $score_i$ represents the confidence information of the joint position. Therefore, the dimension of the human skeleton sequence is (T, N, C). C is the position vector dimension. The dynamic topology diagram of the human skeleton is shown in Fig. 1.

In data preprocessing, firstly, normalize the joint feature vectors. Then, convert skeleton sequences from different positions to the same position to promote convergence. A relationship graph is a many to many graph structure that exists between nodes. Convert multiple diagrams into a set of relationship triplets, as shown in Eq. (4) [15].

$$\{(u, r, v)\} \subseteq V \times E \times V \tag{4}$$

In Eq. (4), u and v are entities. r represents a relationship between entities. Therefore, encode multiple human bone sequence datasets into relational triplet sequence

datasets. Based on the multi relationship features of human skeleton data, Eq. (5) defines a sequence of triples.

$$RT = \left\{ (VT_i, r_{ij}, VT_j) \middle| 1 \le i \le T, 1 \le j \le T \right\} \tag{5}$$

After obtaining the relational triplet dataset, the features need to be decoupled. The feature decoupling module extracts entities and feature embedding vectors from a triplet dataset. Then, based on the association between independent joints in each action, the video frames of the triplet sequence are decomposed into multiple features to learn the embedding vectors of the entities. At the same time, encode the action types to learn the embedding vectors of relationship features. Entity feature decouples a single video frame into multiple features. Then the embedded feature vectors of each joint feature are learned separately. Therefore, the entity $VT_i \in V$ represented by each single video frame is transformed into a set of multiple independent joint points, as shown in Eq. (6).

$$VT_L = \left[V_1, V_2, ..., V_N \right] \tag{6}$$

In Eq. (6), $VT_L \in R^d$. d represents the vector dimension. After performing one-hot encoding on the action type, the study uses embedded feature vectors to represent action analogies, thereby constructing a dynamic topology map. The vectors obtained from the feature decoupling module are used to construct skeleton topology maps for different action categories. Firstly, the K-means is applied to cluster the relationship feature vectors representing action categories. Then the dataset is grouped based on the clustering center. Next, based on the action categories encoded by one-hot, combined with the attention mechanism and the partial update strategy, the study constructs a topology diagram for each type of action. Eq. (7) defines the i-th training set.

$$T_i = \{ (u_1^{T_i}, u_1^{T_i}, v_1^{T_i}), ..., (u_m^{T_i}, u_m^{T_i}, v_m^{T_i}) \} \tag{7}$$

In Eq. (7), m represents the instance number of T_i. For a skeletal relationship triplet, the research uses the attention mechanism to extract the relevant joint features of continuous video frames VT_i and their relationship r_{ij}. Triple (VT_i, r_{ij}, VT_j) is assigned N attention weights. Eq. (8) shows the importance α_r^i of the i-th joint point in the current relationship r.

$$\alpha_r^i = \frac{exp(\alpha_r^i)}{\sum_{j=1}^{N} exp(\alpha_r^j)} \tag{8}$$

In Eq. (8), α_r^j represents the importance of the j-th joint point in the current relationship r. Then, the research adopts the video frame VT_i with the highest attention weight and the first i joint points. Different features are used to construct a relationship topology diagram for this action category, as shown in Fig. 2.

In Fig. 2, a single skeleton frame contains 18 embedded feature vectors. The red leg nodes are the first six joint features most relevant to this action. The model experiences forgetting when updating parameters. Therefore, the study introduces a partial update strategy to dynamically adjust the topology map. When connecting a new skeleton relationship triplet, the model first identifies the parts related to the new data in the existing skeleton relationship graph. However, this only updates highly relevant features. Due to the complex connections between nodes, new data may affect multiple existing nodes. Therefore, activate the nodes directly or indirectly connected to it. Based on feature similarity, perform selective updates and further optimize computational efficiency. Finally, perform local updates on joint features that are highly similar to the new data.

B. Basketball Movement Recognition Based on Perspective Invariant Geometric Features

In modern basketball, numerous factors can easily affect target action recognition, such as local occlusion, rapid movement, and noise caused by inherent unstable factors in cameras [16]. These factors can cause the collected joint points to shake, resulting in a large amount of noise. Fig. 3 displays the schematic diagram of human joint shaking.

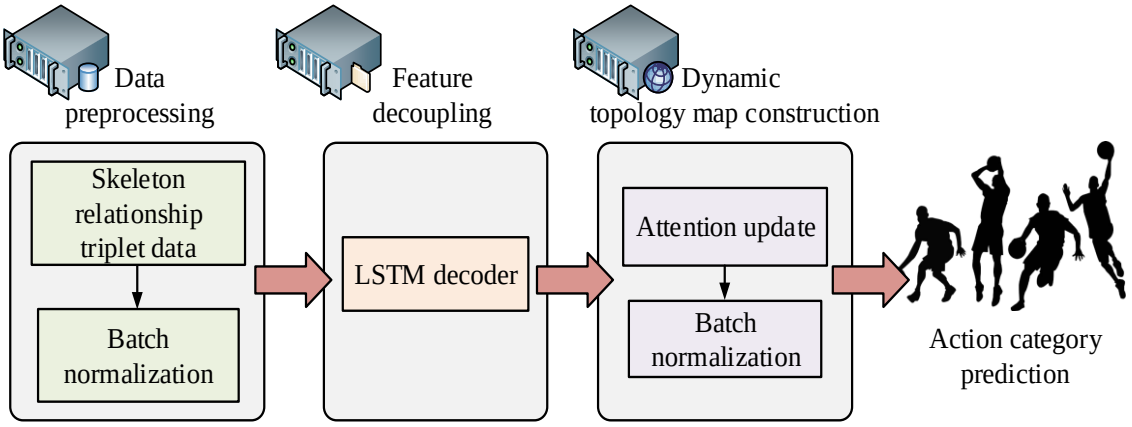


Fig. 1. The Process of human skeleton motion recognition based on dynamic topology graph.

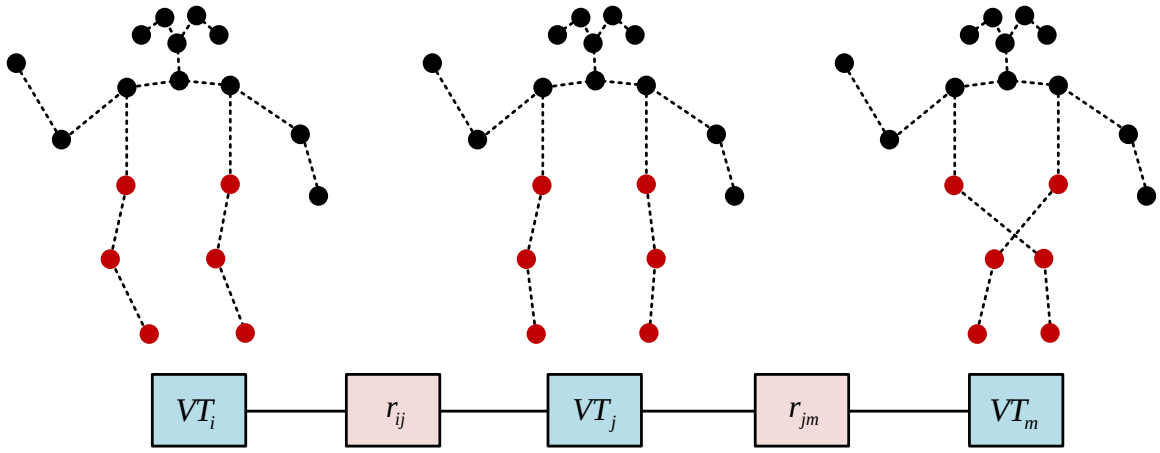


Fig. 2. Schematic diagram of feature encoding.

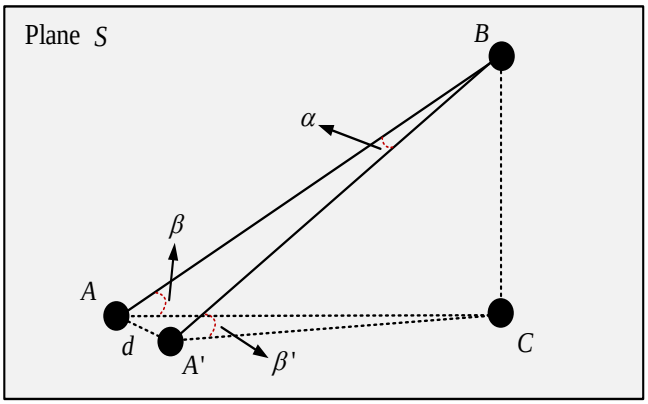


Fig. 3. Schematic diagram of human joint shaking.

In Fig. 3, when joint point A experiences shaking and moves to A', the shaking path is a periodic rotation around the joint BC. The rotation radius is AC, and the angle is α. The joint point A or conventional geometric characteristics used as network input may lead to system stability issues [17]. In addition, in cross perspective testing, apply two Kinect devices to calculate the three-dimensional coordinates of human bones. This calculation uses different projection centers, which may also introduce interference [18]. To address these challenges, a few feature representations that are not sensitive to three-dimensional spatial transformations of skeleton data are proposed. It is named the Planes of 3D Joint Motions Vector (P3DJMV). Cosine similarity is the primary distance comparison method for this research, primarily because it focuses on measuring directional similarity between vectors rather than the size. This property is particularly important for basketball motion recognition because the study focuses more on capturing and comparing features of the motion patterns rather than the absolute size of the action. In addition, cosine similarity typically performs well when dealing with high-dimensional data, which is particularly useful for research application scenarios that analyze complex motion data. Since it is based on directional similarity, cosine similarity naturally ignores differences in the size of the data. When comparing, it is possible to fairly handle different sizes of motion patterns. P3DJMV uses the cosine angle between vector $\underset{AB}{\rightarrow}$ and vector $\underset{AC}{\rightarrow}$ as the feature descriptor for skeleton data. In this way, even if point A shakes to point A', it will not follow the shaking due to point A shaking. In the plane S composed of A, B, and C, the angles β and β' are approximately equal. To explore the spatial variation of joint points, a feature representation based on node momentum (FRNM) is also designed, as shown in Fig. 4.

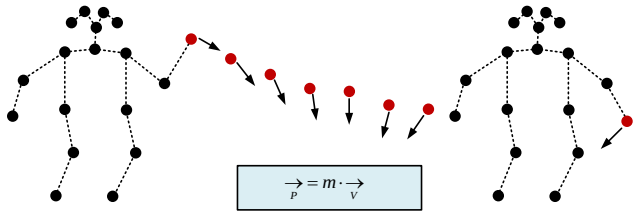


Fig. 4. Schematic diagram of feature FRNM.

Fig. 4 shows the FRNM visualization diagram of a basketball player's catch movement. The end node of the athlete's left hand is the target node, displaying its continuous changes in space. The arrow direction represents the motion direction of the target node. The length roughly reflects the motion speed. P3DJMV aims to simplify the three-dimensional spatial into a one-dimensional angular space, as displayed in Fig. 5.

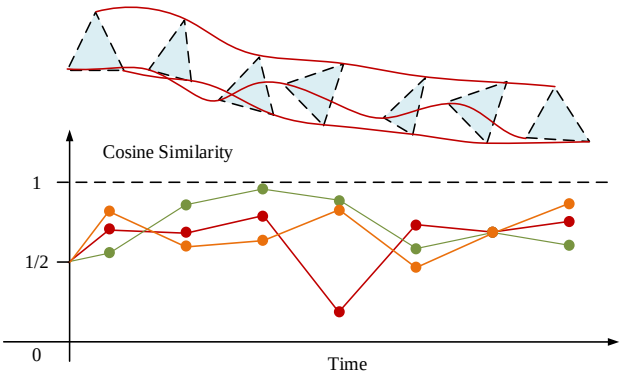


Fig. 5. P3DJMV feature visualization.

Unlike P3DJMV, FRNM essentially treats human joint points as a particle [19]. The particle mass is 1. Momentum represents the trajectory of particles. The human skeleton data contains 25 nodes. It is simplified to 20 nodes. Then, solve for the average and standard deviation of the entire dataset. The data meets the standard normal distribution through standardized processing. Afterwards, FRNM arranges groups from these 20 joints to extract possible planes, with a total of 1140 possibilities. Eq. (9) represents these planes expressions.

$$C_p = \{p_i, p_j, p_k\}, p \in N^+[1,1140]\ i, j, k \in N^+[1,20] \tag{9}$$

In Eq. (9), p_i, p_j, and p_k represent three joint points. p represents a plane. The three points p_i, p_j and p_k can obtain three vectors, as shown in Eq. (10).

$$\begin{cases} V_p(1) = (p_i^x - p_j^x, p_i^y - p_j^y, p_i^z - p_j^z) \\ V_p(2) = (p_i^x - p_k^x, p_i^y - p_k^y, p_i^z - p_k^z) \\ V_p(3) = (p_k^x - p_j^x, p_k^y - p_j^y, p_k^z - p_j^z) \end{cases} \tag{10}$$

In Eq. (10), the cosine value $V_p(1)$, $V_p(1)$, and $V_p(1)$ are shown in Eq. (11).

The P3DJMV is stacked in the tensor form of $F \times H \times W$. F represents the frames. H represents length. W represents the width. The node is abstractly represented as a physical particle. The mass is 1. The particle momentum is $\rho = mv$, and m is 1. If the trajectory of the physical particle is differentiated, then v can be obtained by differentiating the distance of the particle's motion per unit time, as shown in Eq. (12).

$$\begin{cases} A_p(1) = cos\alpha = \dfrac{V_p(1) \cdot V_p(2)}{V_p(1) * V_p(2)} \\ A_p(2) = cos\alpha = \dfrac{V_p(1) \cdot V_p(3)}{V_p(1) * V_p(3)} \\ A_p(3) = cos\alpha = \dfrac{V_p(3) \cdot V_p(2)}{V_p(3) * V_p(2)} \end{cases} \tag{11}$$

$$v = \lim_{\Delta t \to 0} \frac{y_{t+1} - y_{t-1}}{2\Delta t} \tag{12}$$

In Eq. (12), t represents the motion time. The FRNM is stacked as a geometric manifold $F \times H \times W$. F represents the frames. H represents length. W represents the width. Finally, P3DJMV and FRNM are fused and input into the prediction network. The research constructs the model based on 2D ResNet18. Its spatial down sampling remains unchanged. The first layer convolution is used to update the weights of each P3DJMV vector. In the spatio-temporal feature learning stage, each construction layer undergoes batch normalization and activation functions [20]. Finally, the two proposed features are fused after the fully connected layer. The research constructs a basketball motion recognition model based on perspective invariant geometric features extracted from skeleton data, as shown in Fig. 6.

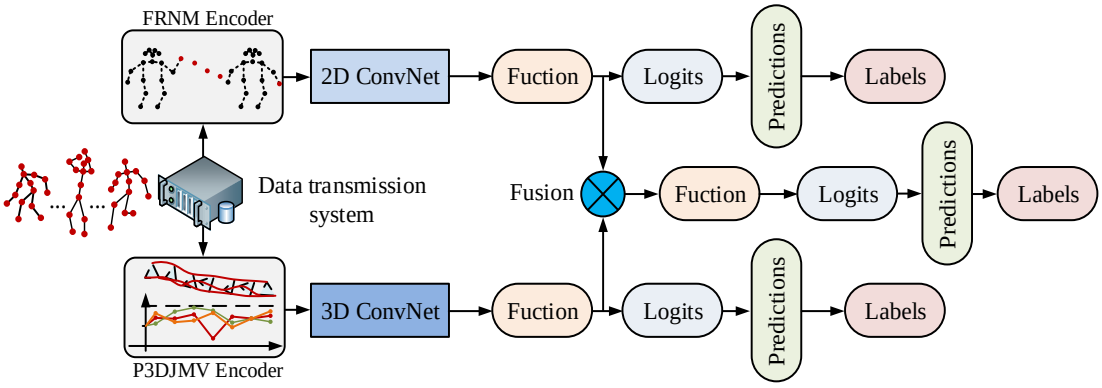


Fig. 6. Basketball movement recognition model based on skeleton data.

IV. Experimental Analysis

To improve the basketball motion recognition performance, the research proposed the basketball motion recognition model based on perspective invariant geometric features extracted from skeleton data. The dynamic topology map of the human skeleton was constructed to provide better output for recognition models. Then, a basketball motion recognition model based on perspective invariant geometric features was designed. To verify the effectiveness, test experiments were designed on datasets and real environments. The experimental results were analyzed.

A. Test Results in the Dataset

The dataset used in the experiment was two publicly available large-scale behavior recognition datasets, Kinetics and NTU RGB+D. Kinetic was a human behavior dataset that contained 300000 edited videos and 400 types of actions. The basketball movement segments were divided into training sets and validation sets. The second generation Kinetics depth sensor collection constituted the NTU RGB+D collection, with a total of 56000 edited action videos. To verify the validity of the research method, the experiments were conducted in the same environment. In the software architecture of the experimental system, this study

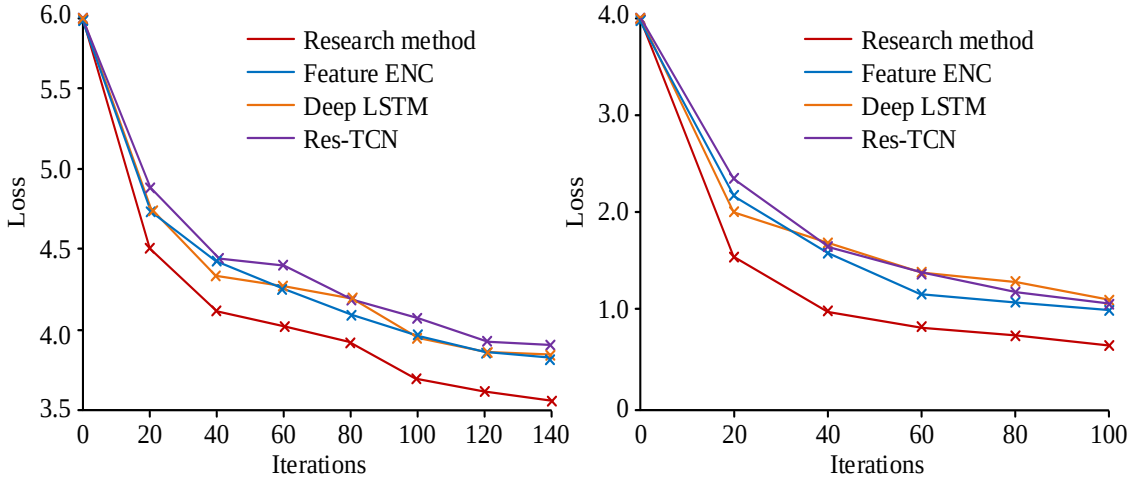
implemented five core modules, which were the Hikvision video capture module, the Noiton motion capture module, the Kinect somatosensory camera module, the Linux-based WiFi data acquisition module, and the synchronization control system used to coordinate these modules. Except for the Linux-based WiFi data acquisition module, all other modules ran on Windows 10 operating system. The research utilized the TCP/IP protocol to communicate with various collection subsystems, achieving data collection and storage operations. After precise testing, the system could ensure that the message synchronization error was within 100 milliseconds. In addition, the each acquisition module was flexible, which could work together on multiple hosts as well as operate independently. It enhanced the adaptability and reliability of the system. The experimental hardware environment was shown in Table I.

TABLE I. INTRODUCTION TO EXPERIMENTAL ENVIRONMENT

Name	Configuration Introduction
CPU	Intel Core i9-10920X CPU@3.50 GHz
GPU	NVIDIA1 GeForce RTX 3080Ti
Running memory	32GB
Operating system	Windows 10
Programming Language	Python
Development environment	Anaconda 3+python 3.6+pytorch 1.9

Top-1 and Top-5 classification accuracy were applied to evaluate recognition performance. The Top-1 indicator represented the probability that the first ranked category in the predicted category score vector was a true category. The Top-5 indicator represented the probability that the top five categories in the predicted category score vector contained the correct category. The proposed methods were compared with advanced methods, including Feature Encoding (Feature ENC) [21], Deep LSTM based on Recurrent Neural Network (RNN) [22], and Residual Temporal Convolutional Network (Res-TCN) based on Convolutional Neural Network (CNN) [23]. Among them, Feature ENC could effectively extract and encode key features, which improved the model's ability to recognize complex action patterns. The method demonstrated excellent performance in dealing with different types of action data, which was crucial for identifying diverse and complex basketball motions. Deep LSTM was suitable for dealing with time-series data. It combined the spatial feature extraction ability of CNN with the advantages of time series data processing. The residual network structure could avoid the gradient vanishing problem in deep network training, which represented the most advanced technology currently used to handle complex action sequences. The performance of different algorithms during training iterations was shown in Fig. 7.

Fig. 7 (a) showed the iterative loss curve on the Kinetics dataset. From the graph, the method proposed in the research showed significant differences from the other three methods after 20 iterations. The proposed method achieved lower loss values with faster iteration speed. When the iterations were 140, the loss of the research method was 3.52. The loss values for Feature ENC, Deep LSTM, and Res-TCN were3.85, 3.89, and 3.98, respectively. Fig. 7 (b) showed the iterative loss curve on the NTU RGB+D dataset. From the graph, the method demonstrated better performance after 20 iterations. When the iterations were 100, the loss value was 0.82. The other three methods were 1.18, 1.32, and 1.26, respectively. The results on the test set demonstrated the effectiveness of this method. The Top-1 indicator results of different algorithms were shown in Fig. 8.



(a) Kinetics dataset (b) NTU RGB+D dataset

Fig. 7. Curve of loss value with number of iterations.

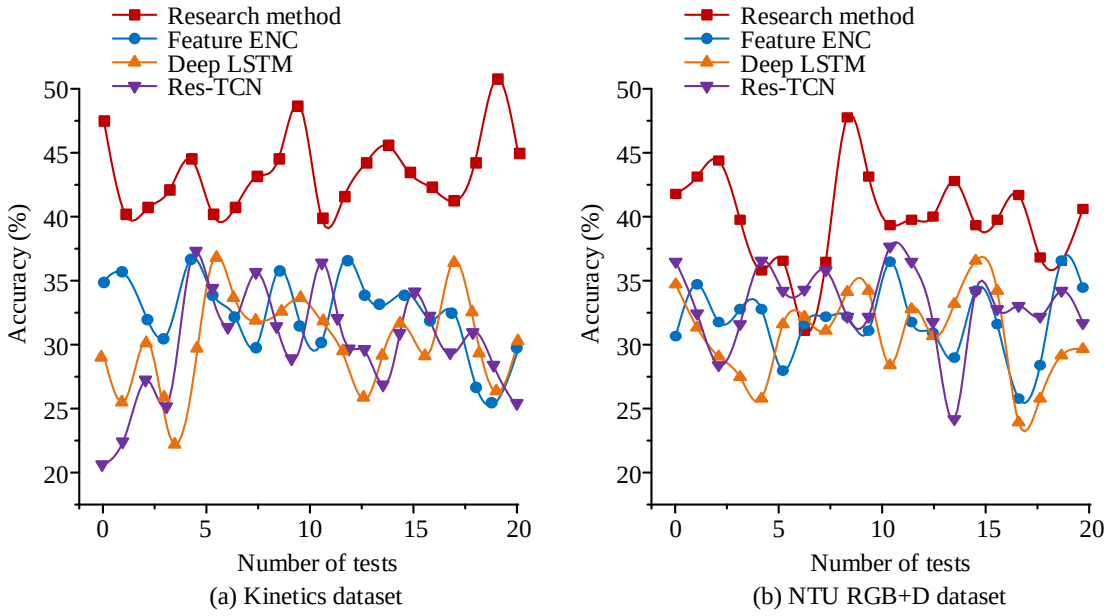


Fig. 8. Comparison results of Top-1 indicators.

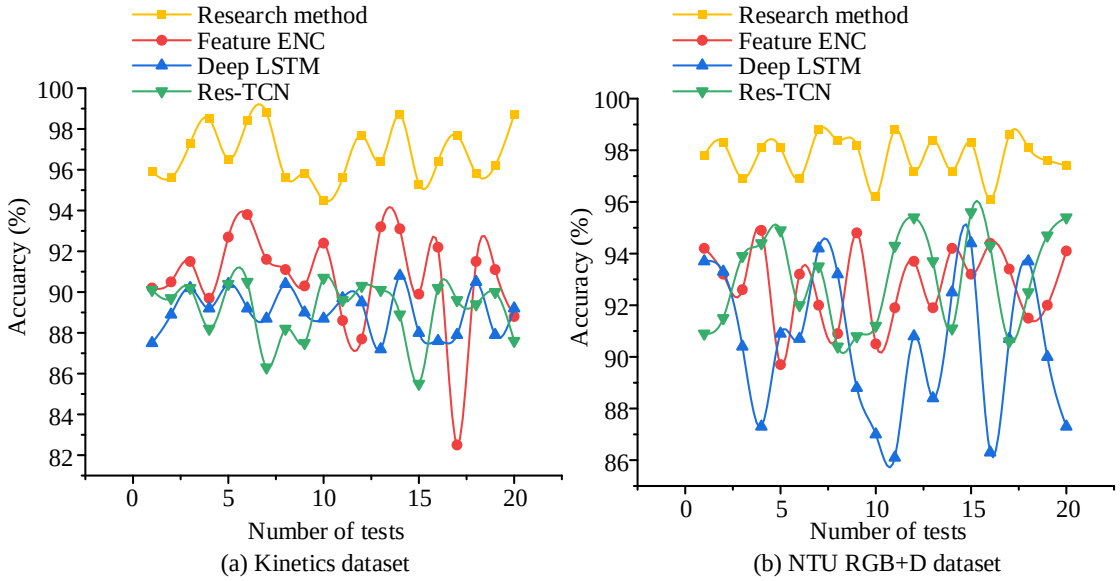


Fig. 9. Comparison results of Top-5 indicators.

Fig. 8 (a) showed the Top 1 standard accuracy. From the graph, the accuracy exceeded the other three methods. The average accuracy of the method in 20 tests was 43.52%, while the other three methods were 30.85%, 29.65%, and 27.34%, respectively. Fig. 8 (b) showed the Top 1 standard accuracy on the NTU RGB+D dataset. From the graph, the accuracy was higher than the other three methods in most tests. The average accuracy of the method in 20 tests was 40.52%. The other three methods were 34.12%, 32.07%, and 31.67%, respectively. From the experimental results, the research method had higher accuracy. The performance on the Kinetics dataset was superior to that on the NTU RGB+D dataset. The Top-5 indicator results of different algorithms were shown in Fig. 9.

Fig. 9 (a) showed the Top-5 standard accuracy on the Kinetics dataset. From the graph, the accuracy exceeded the other three methods. The average accuracy in 20 tests was 97.85%. The other three methods were 90.81%, 89.95%, and 89.27%, respectively. Fig. 9 (b) showed the Top-5 standard accuracy on the NTU RGB+D dataset. The accuracy also exceeded the other three methods. The average accuracy in 20 tests was 97.82%. The other three methods were 92.08%, 90.11%, and 91.73%, respectively. From the experimental results, compared to the current advanced three methods, the proposed method significantly improved the accuracy of the Top-5 evaluation index, verifying the effectiveness of this method.

B. Real Environment Application Analysis

To evaluate the performance of the basketball motion recognition model based on perspective invariant geometric features extracted from skeleton data in real environments, it performed motion recognition experiments on 80 volunteers. This study first used a confusion matrix to evaluate the performance of the basketball motion recognition model. The confusion matrix displayed the correspondence between the recognition results of each category and the actual results. It helped to understand the performance of basketball action recognition models in different categories, such as which categories were accurately identified and which categories were easily confused. By analyzing the confusion matrix, the researcher could identify and improve the weaknesses of the model, such as increasing the recognition rate or reducing misclassification. The research divided basketball motions into six categories, including shooting, dribbling, layups, passing, and grabbing the board. It performed 280 recognition tests for each action category. Fig. 10 displayed the test results.

Fig. 10 (a) showed the recognition results of the research method. The average correct recognition quantity for each category was 269.6. The expected correct recognition quantity was 280. Therefore, the average accuracy of the research method was 96.3%. Fig. 10 (b) displayed the recognition results of Feature ENC. The average correct recognition quantity for each category of Feature ENC was 247.8. The expected correct recognition quantity was 280. Therefore, the average accuracy of Feature ENC was 88.5%. Fig. 10 (c) showed the recognition results of Deep LSTM. The average correct recognition number for each category in Deep LSTM was 249.2. The expected correct recognition number was 280. The average accuracy of Deep LSTM was 89.0%. Fig. 10 (d) showed the recognition results of Res-TCN. The average correct recognition number for each category in Res-TCN was 240.4. The expected correct recognition number was 280. The average accuracy of the research method was 85.86%. From Fig. 10, the correctly identified color bands in research methods had darker colors, while the incorrectly identified color bands had lighter colors. This indicated that it still outperformed the other three advanced methods in real-world testing. To further validate the performance of the proposed model, the study increased the number of experiments to 20 and used recall rate as an indicator. The recall rate was the proportion of positive category samples (e.g., a specific basketball action) correctly recognized by the model to all actual positive category samples. This metric was particularly important for evaluating the recognition ability. The results were shown in Fig. 11.

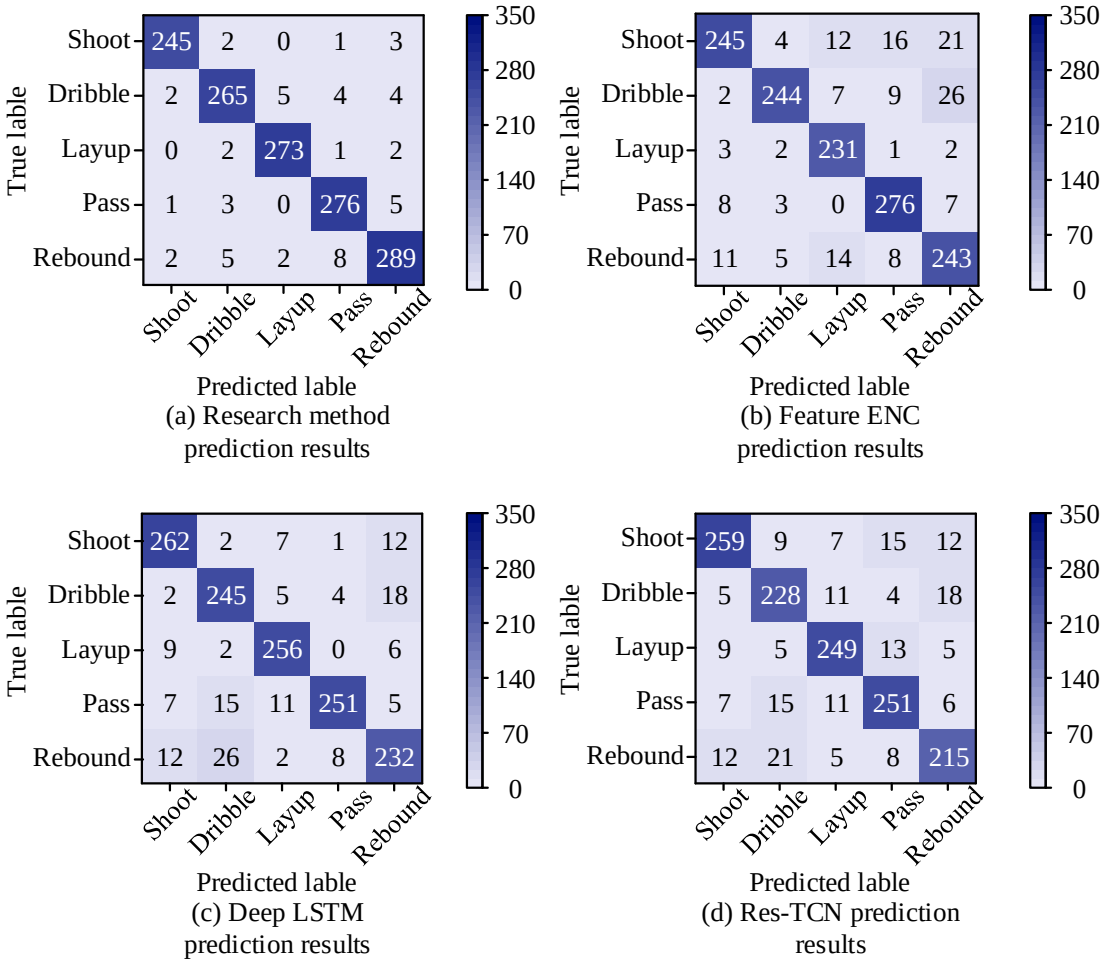


Fig. 10. Results of basketball motion recognition using different methods.

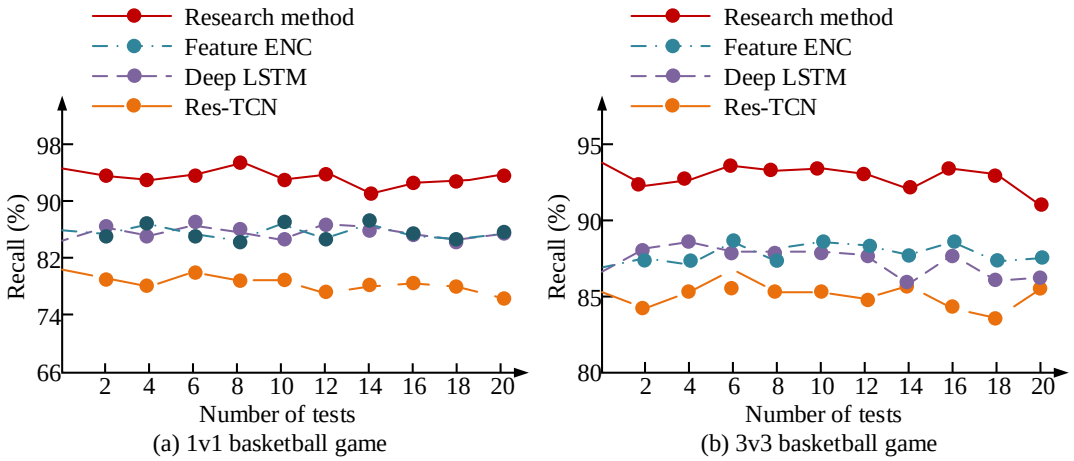


Fig. 11. Comparison of recall of different methods.

Fig. 11(a) showed the action recognition results in 1-to-1 basketball game. The recall of the research method was significantly higher than that of the other three methods. In 20 tests, the average recall rate was 96.35%, while the average recall rate of the other three methods was below 90%. Fig. 11(b) showed the action recognition results in 3-to-3 basketball game. The average recall of the research method reached 94.02% in 20 tests. The average recall of the other three methods was below 90%. From the recall experiments, the recognition performance of the constructed model was significantly better than other methods in different environments. Finally, to verify the computational complexity, the research compared the running times of the four recognition models. The results were shown in Fig. 12.

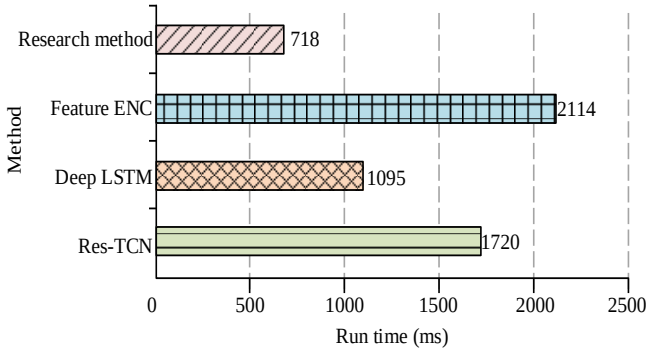


Fig. 12. Comparison of runtime of different recognition methods.

In Fig. 12, the running time of the research method was 718ms. The Feature ENC was 2114ms. TheDeep LSTM was 1095ms. The Res-TCN was 1720ms. Compared to the other methods, the running time of the proposed method was decreased by 66.03%, 34.43%, and 58.26%. According to the results, the basketball motion recognition model based on perspective invariant geometric features extracted from skeleton data had faster recognition speed.

V. Conclusion

To improve the utilization of human bone data and build a more accurate and efficient basketball motion recognition model, the study first constructs a dynamic topology map of human skeleton. Then it serves as input to the recognition model, thereby removing interference from environmental information. Then, based on the perspective invariance in skeleton data extraction, the study proposes two feature representation methods, namely the 3D joint motion vector plane and the feature representation method based on node momentum. By fusing two feature representations, the spatio-temporal feature fusion of skeleton data is achieved, thus constructing a new basketball motion recognition model. In real-world testing, the proposed method had an average accuracy of 96.3% for each category. The average accuracy of Feature ENC was 88.5%. The average accuracy of Deep LSTM was 89.0%. The average accuracy of Res-TCN was 85.86%. The results validated the effectiveness of this research. It demonstrated excellent performance in experiments. However, this research experiment uses single-mode data. Therefore, the recognition effect of multimodal data has not been verified yet. In the future, this method will be further optimized. Combined with other advanced machine learning technologies, this method will be improved to achieve better performance in a wider range of application scenarios.

Reference

[1] ZHAO J, SHE Q, MENG M, CHEN Y. Skeleton Action Recognition Based on Multi-Stream Spatial Attention Graph Convolutional SRU Network. ACTA ELECTONICA SINICA, 2022, 50(7): 1579-1585.

[2] Qin X, Li H, Liu Y, Yu J, He C, Zhang X. Multi-stage part-aware graph convolutional network for skeleton-based action recognition. IET Image Processing, 2022, 16(8): 2063-2074.

[3] Zhang P, Zhang J, Elsabbagh A. Lower limb motion intention recognition based on sEMG fusion features. IEEE Sensors Journal, 2022, 22(7): 7005-7014.

[4] Shu X, Yang J, Yan R, Song Y. Expansion-squeeze-excitation fusion network for elderly activity recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(8): 5281-5292.

[5] Gao S, Yun J, Zhao Y, Liu L. Gait-D: Skeleton-based gait feature decomposition for gait recognition. IET Computer Vision, 2022, 16(2): 111-125.

[6] Song Z, Zhang Y, Liu Y, Yang K, Sun M. MSFYOLO: Feature fusion-based detection for small objects. IEEE Latin America Transactions, 2022, 20(5): 823-830.

[7] Zhou W, Guo Q, Lei J, Yu L, Hwang J N. ECFFNet: Effective and consistent feature fusion network for RGB-T salient object detection.

IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(3): 1224-1235.

[8] Zhang X, Wang J, Wang T, Jiang R. Hierarchical feature fusion with mixed convolution attention for single image dehazing. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(2): 510-522.

[9] Choi H, Yun J P, Kim B J, Jang H, Kin S W. Attention-based multimodal image feature fusion module for transmission line detection. IEEE Transactions on Industrial Informatics, 2022, 18(11): 7686-7695.

[10] Song Y F, Zhang Z, Shan C, Wang L. Constructing stronger and faster baselines for skeleton-based action recognition. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(2): 1474-1488.

[11] Li M, Chen S, Chen X, Zhang Y, Wang Y, Tian Q. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(6): 3316-3333.

[12] Li C, Xie C, Zhang B, Han J, Zhen X, Chen J. Memory attention networks for skeleton-based action recognition. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(9): 4800-4814.

[13] Zhang C, Liang J, Li X, Xia Y, Di L, Hou Z, Huan Z. Human action recognition based on enhanced data guidance and key node spatial temporal graph convolution. Multimedia Tools and Applications, 2022, 81(6): 8349-8366.

[14] Zhang Z, Wang S, Liu C, Xie R, Hu W, Zhou P. All-in-one two-dimensional retinomorphic hardware device for motion detection and recognition. Nature Nanotechnology, 2022, 17(1): 27-32.

[15] Oslund S, Washington C, So A, Chen T, Ji H. Multiview Robust Adversarial Stickers for Arbitrary Objects in the Physical World. Journal of Computational and Cognitive Engineering, 2022, 1(4): 152-158.

[16] Choudhuri S, Adeniye S, Sen A. Distribution Alignment Using Complement Entropy Objective and Adaptive Consensus-Based Label Refinement for Partial Domain Adaptation. Artificial Intelligence and Applications. 2023, 1(1): 43-51.

[17] Suneetha M, Prasad M V D, Kishore P V V. Sharable and unshareable within class multi view deep metric latent feature learning for video-based sign language recognition. Multimedia Tools and Applications, 2022, 81(19): 27247-27273.

[18] Zhang K, Li Y, Wang J, Cambria E, Li X. Real-time video emotion recognition based on reinforcement learning and domain knowledge. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(3): 1034-1047.

[19] Giannakeris P, Petrantonakis P C, Avgerinakis K, Vrochidis S, Kompatsiaris I. First-person activity recognition from micro-action representations using convolutional neural networks and object flow histograms. Multimedia Tools and Applications, 2021, 80(15): 22487-22507.

[20] Van Amsterdam B, Funke I, Edwards E, Speidel S, Collins J, Sridhar A, Stoyanov D. Gesture recognition in robotic surgery with multimodal attention. IEEE Transactions on Medical Imaging, 2022, 41(7): 1677-1687.

[21] Zhou Z, Dong X, Li Z, Yu K, Ding C Yang Y. Spatio-temporal feature encoding for traffic accident detection in VANET environment. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(10): 19772-19781.

[22] Torres J F, Martínez-Álvarez F, Troncoso A. A deep LSTM network for the Spanish electricity consumption forecasting. Neural Computing and Applications, 2022, 34(13): 10533-10545.

[23] Shang Z, Liu H, Zhang B, Feng Z, Li W. Multi-view feature fusion fault diagnosis method based on an improved temporal convolutional network. Insight-Non-Destructive Testing and Condition Monitoring, 2023, 65(10): 559-569.