IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Keyframe Selection via Deep Reinforcement Learning for Skeleton-based Gesture Recognition

Minggang Gan, Jinting Liu, Yuxuan He, Aobo Chen, and Qianzhao Ma

***Abstract*— Skeleton-based gesture recognition has attracted extensive attention and has made great progress. However, mainstream methods generally treat all frames as equally important, which may limit performance, especially when dealing with high inter-class variance in gesture. To tackle this issue, we propose an approach that models a Markov decision process to identify keyframes while discarding irrelevant ones. This paper proposes a deep reinforcement learning double-feature double-motion network comprising two main components: a baseline gesture recognition model and a frame selection network. These two components mutually influence each other, resulting in enhanced overall performance. Following the evaluation of the SHREC-17 and F-PHAB datasets, our proposed method demonstrates superior performance.**

I. Introduction

With the widespread application of photoelectric motion capture, Leap Motion, and depth sensors, it becomes easier to obtain 3D joint locations. Skeleton-based sequences, in contrast to 2D RGB-based action recognition methods, contain compact 3D locations of crucial hand joints that are robust to changes in lighting, camera angles, and other background [1]. Skeleton data provide a substantial reduction in computing cost compared to video data. Additionally, the hand serves as one of the most effective interaction tools for humans. Consequently, skeleton-based gesture recognition has gained significant research [2] attention across multiple fields, including human-computer interaction [3], [4], Human-robot collaboration [5], [6], and human behavior understanding [7].

However, despite these advantages, developing a precise recognition system remained challenging due to the high intra-class variance arising from the various subjects that could perform the same gesture. Moreover, human observers can discern discriminative information from several still frames. Therefore, we can discard a large number of non-keyframes in the skeleton data without affecting the recognition accuracy.

Although most hand action recognition methods have made great progress, these methods based on deep learning generally regard all frames as equally important. One strategy is that all the frames participate in the training of the deep learning model [8], [9], which leads to computational redundancy. The other strategy is to uniformly or continuously sample [10]–[12], which cannot guarantee the best recognition result. Not all skeleton frames contain useful information in the recognition process based on the deep learning method, and only a few keyframes within the skeleton sequence significantly contribute to the recognition results. Redundant and ambiguous frames in the hand skeleton data can lead to incorrect recognition results. Absolutely, discarding irrelevant frames is a crucial step in reducing noise and optimizing the gesture recognition process. Irrelevant frames may not contribute significantly to the gesture's distinctive characteristics and can introduce unnecessary variability. By selectively focusing on keyframes and eliminating irrelevant ones, the gesture recognition model can achieve higher accuracy and efficiency, leading to more reliable results.

We propose a deep reinforcement learning double-feature double-motion Network (DRL-DDNet) to select keyframes in a sequence. The DRL-DDNet consists of two main components: a double-feature double-motion Network (DD-Net) and a frame selection network (FS-Net). The FS-Net employs reinforcement learning to find the optimal strategy in a Markov decision process (MDP), aiming to identify distinguished frames and discard irrelevant ones in the sequence. The DD-Net and FS-Net directly impact each other's performance. The DRL-DDNet uses reinforcement learning to select keyframes, which requires a lot of computation and resource consumption. We use the smaller and faster advantages of the DD-Net to improve the processing speed of the algorithm. During training, DD-Net rewards FS-Net, enabling FS-Net to select keyframes that represent the most informative frames for recognition. Subsequently, we input the selected keyframes into the DD-Net for skeleton-based gesture recognition. The symbiotic relationship between DD-Net and FS-Net is a crucial aspect of the DRL-DDNet. The performance of the DD-Net benefits from the frame selection process, which ensures that only key frames can be recognized, avoiding recognition errors caused by ambiguous frames. On the other hand, the training of the FS-Net is affected by the recognition accuracy of the DD-Net, which provides better rewards for faster and more accurate keyframe selection. The step-by-step interaction of DD-Net and FS-Net forms an integrated and efficient system. The training process of DRL-DDNet ensures that the model learns to focus on the keyframes, leading to enhanced skeleton-based gesture recognition accuracy.

To sum up, the contributions of this paper can be summarized:

1) Introducing a Markov process to discard ambiguous and redundant frames in the hand skeleton training set, leading to improved training quality and enhanced recognition accuracy of the network.

This work was supported by the National Key Research and Development Program of China under Grant 2020YFB1708500. (Corresponding author: Minggang Gan.)

The authors are with the Key Laboratory of Complex System Intelligent Control and Decision, School of Automation, Beijing Institute of Technology, Beijing 100811, China

2) Proposing a novel deep reinforcement learning double-feature double-motion network for skeleton-based gesture recognition, consisting of interconnected components that continuously enhance recognition performance through interaction.
3) Demonstrating that the proposed algorithm achieves the most advanced performance on two datasets, namely SHREC-17 and F-PHAB.

The remainder of this paper is organized as follows. Related work on hand recognition methods and reinforcement learning are briefly reviewed in Section II. In Section III, we provide details information about the proposed algorithm. Section IV details the experimental setup and results. and the conclusions are presented in Section V.

II. Related Work

A. Skeleton-based action-gesture recognition

The rapid development of deep neural networks [13] has attracted the attention of an increasing number of researchers exploring deep learning methods for action and gesture recognition. Various deep neural network architectures, such as Convolutional Neural Networks (CNN) [14], [15], Recurrent Neural Networks (RNN) [16], [17], and Long Short-Term Memory (LSTM) [18] have been developed to effectively capture the temporal and spatial features of hand or body data for accurate recognition. For example, Nunez et al. [19] proposed an architecture that combines a CNN with an LSTM recurrent network for skeleton-based action-gesture recognition. The CNN focuses on extracting spatial features, while the LSTM recurrent network captures patterns related to time evolution. Hou et al. [20] proposed the Spatial-Temporal Attention Residual Temporal Convolutional Network (STA-Res-TCN) to learn different levels of attention for skeleton-based dynamic gesture recognition. Alberto Sabater et al. [21] proposed a skeleton-based hand motion representation model with excellent generalization capabilities across various action domains and camera perspectives.

However, many existing methods [22], [23] suffer from large models and slow execution speed. By studying skeleton sequence properties, Yang et al. [10] proposed a more lightweight DD-Net for hand and body action recognition. In the recognition process, while all joints are considered, it is found that only a handful of important joints are essential for the recognition. Ferda Ofli et al. [24] represent the actions as a sequence of the most informative joints for action recognition, which can avoid taking into account non-information that often brings noise and degrades performance.

Most of the aforementioned methods overlook the varying importance of frames and treat each frame equally. However, not all skeleton frames contain useful information throughout the entire recognition process based on deep learning methods. Redundant and ambiguous frames in the hand skeleton data can lead to incorrect recognition results. To overcome this limitation, it is essential to identify and utilize the keyframes for recognition.

B. Deep reinforcement learning for action recognition

Reinforcement learning (RL) has shown efficient representation of extracted sequences and strong generalization ability. Deep Reinforcement Learning (DRL) has been increasingly applied to address the activity recognition problem, serving various purposes such as finding most information in sequence and optimizing network structures [25]. Human activity recognition via DRL can be expressed as a search for optimal solutions. In this context, the RL agent constantly interacts with the environment to try and error, to identify the framework set that contains the most information and achieve accurate identification.

To select the most informative frames for action recognition, Tang et al. [26] proposed a deep progressive reinforcement learning method for skeleton-based action recognition. Similarly, to reduce the computational burden of the learning model in untrimmed video analysis, it is important to sample frames effectively. Wu et al. [27] solved the frame sampling problem in untrimmed video analysis through multi-agent reinforcement learning. Dong et al. [28] proposed an attention-aware sampling method for action recognition, which preserves the most discriminative frames and discards the irrelevant frames. Xu et at. [29] proposed a feature selection network (FSN) that selects the most representative features to improve recognition performance. For group activity recognition, Hu et al. [30] proposed a method based on DRL to progressively refine the low-level and high-level features relations of group activities.

However, there is a lack of investigation into modeling Markov decision processes for skeleton keyframes of gestures, which exhibit completely different characteristics from skeleton-based hand models. Therefore, we emphasize the improvement of hand action recognition baselines, and our model employs a modeling Markov process to select keyframes more effectively.

III. Methodology

A. Model Architecture

The overview of the proposed DRL-DDNet framework is illustrated in **Fig. 1**. The input to a DRL-DDNet is the hand skeleton sequence of three-dimensional coordinates. We use the hand skeleton as input to train the DD-Net. Next, we integrate the trained model into the FS-Net as a reward for reinforcement learning. We feed the hand skeleton data into the FS-Net as a state. We use the prediction results of the DD-Net as a reward, adjust the sequence through actions, and use a continuous trial and error mechanism to help us discard irrelevant frames and generate keyframes. Subsequently, we input the obtained keyframes into the DD-Net again for further training, resulting in an improved the DD-Net, which we then use as a reward for the FS-Net. This alternating training process leads to the final classification result.

B. Baseline model

In this study, we use the DD-Net [10] as the baseline model, which is smaller in size and faster in processing speed

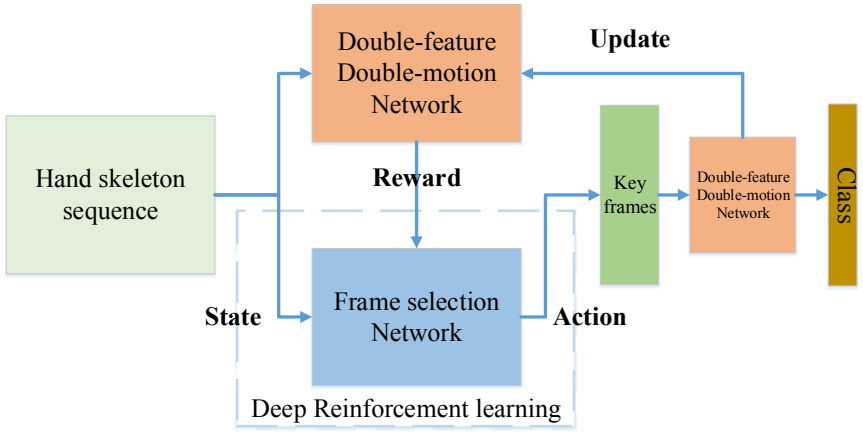


Fig. 1. The pipeline of our proposed method.

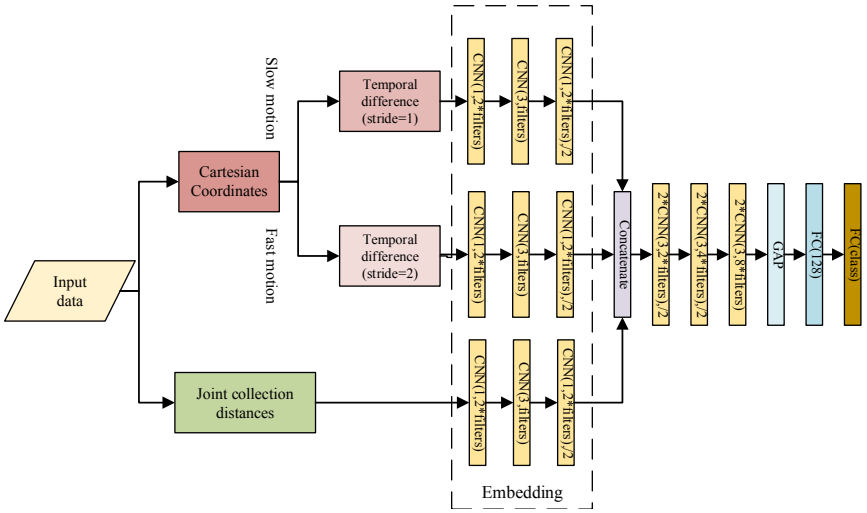


Fig. 2. The network architecture of the DD-Net.

compared to traditional network architectures, and has excellent Action recognition accuracy and generalization ability are very necessary as rewards in reinforcement learning. To train the baseline model, we use the standard cross-entropy loss, ensuring a robust and efficient training procedure. The network architecture of the DD-Net is shown in **Fig. 2**. We supplement the more details of the figure in IV-A.2. This network consists of joint collection distances (JCD) features and two-scale motion features, which are used to model location-viewpoint invariant features and global scale-invariant motions.

1) Joint Collection Distances: The DD-Net introduces the JCD feature to tackle the limitations of the two primary existing feature representations in skeleton-based action recognition. Cartesian coordinate features can undergo significant changes when dealing with skeleton rotation or displacement. Geometric coordinates as input features necessitate extensive feature redesign when transitioning from one dataset to another. Euclidean distances between collective joints are calculated to form a symmetric matrix, and JCD features are obtained by selecting the lower triangular matrix without the diagonal to reduce redundancy.

To be more specific, the total number of joints in a hand skeleton is represented as N. The 3D coordinates of the i-th joint in the k-th frame are denoted as $J_i^k = (x, y, z)$. The hand skeleton in the k-th frame, comprising N joints, is described as $E^k = \{J_1^k, J_2^k, ..., J_N^k\}$. The JCD feature of E^k can be written as:

$$JCD^k = \begin{bmatrix} \left\|\overline{J_2^k J_1^k}\right\| & & & \\ \vdots & \ddots & & \\ \vdots & & \ddots & \\ \left\|\overline{J_N^k J_1^k}\right\| & \dots & \dots & \left\|\overline{J_N^k J_{N-1}^k}\right\| \end{bmatrix}, \quad (1)$$

where, $\|\overline{J_i^k J_j^k}\| (i \neq j)$ denotes the Euclidean distance between J_i^k and J_j^k.

2) Two-scale Motion Feature: Both fast and slow motions should be taken into account while learning a robust global motion feature. To achieve this, we incorporate two distinct global motion features, one capturing fast motion and the

other representing slow motion. These two-scale global motion features are integrated into the DD-Net for enhanced motion feature representation.

To ensure the robustness of our model in recognizing both fast and slow motions, the DD-Net introduces a two-scale global motion feature. This feature consists of a fast global motion and a slow global motion.

The following equation can be used to generate the two-scale motions:

$$\begin{aligned} M_{slow}^{k} &= E^{k+1} - E^{k}, k \in \{1, 2, 3, \cdots, K-1\}, \\ M_{fast}^{k} &= E^{k+2} - E^{k}, k \in \{1, 3, \cdots, K-2\}, \end{aligned} \tag{2}$$

where, M_{slow}^{k} and M_{fast}^{k} denote the slow motion and the fast motion at the frame k, respectively. Additionally, E^{k+1} and E^{k+2} refer to the motion information of one frame and two frames behind E^{k}, respectively.

Then, the dynamic change of joint correlation in different actions is solved by embedding the JCD feature and the two-scale motion feature into the latent vector at each frame. Finally, the network learns the temporal information by employing a 1D ConvNet layer, leading to classification results via Global Average Pooling (GAP) and Fully Connected (FC) layers.

C. Frame Selection Network

Skeleton-based gesture recognition involves discriminative actions that may occur sparsely in several frames. Consequently, not every frame in the sequence is equally informative for the recognition task. To avoid the negative effects of irrelevant frames, it is vital to discard them.

We formulate the process of selecting keyframes as a MDP [26], as illustrated in **Fig. 3**. The input to the MDP is the entire sequence, treated as the state S. The actions A of the MDP represent the adjustments made by the agent's choice in the next time step, and the reward R is the prediction result of the DD-Net. The agent interacts with the environment, receiving rewards and updating its state. The objective of the agent is to learn a policy that maximizes the total expected reward, leading to the selection of the keyframes, denoted by m. For a more comprehensive understanding of the select frames process, **Fig. 4** provides a detailed overview. The network also includes structures for extracting features from selected frames, including convolutional and pooling. Following this, each frame undergoes adjustments after the full connection and softmax, resulting in a probability score denoting its likelihood of being selected as a keyframe. The final output is the adjusted action. The primary objective of this network architecture is to optimize the frame selection process, ensuring the preservation of keyframes while discarding irrelevant ones.

State. The state S is composed of two components, S_a and S_b. Specifically, S_a consists of two tensors, G and M. The tensor G represents all frames of the hand skeleton sequence for training, with a shape of $g \times N \times 3$, where g, N, and 3 denote the number of frames, joints, and dimensions, respectively. The tensor M contains selected keyframes and has a shape of $m \times N \times 3$, where m is the number of selected

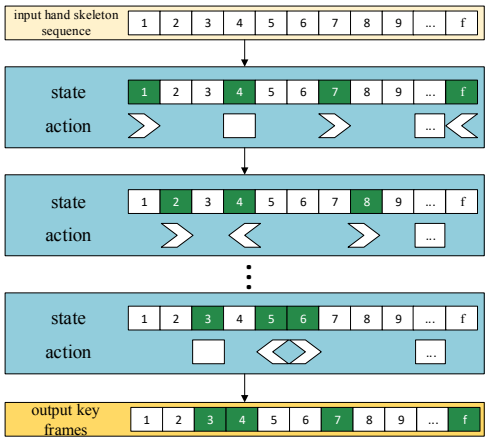


Fig. 3. Process of selecting keyframes in hand skeleton sequence.

frames. On the other hand, S_b represents the index of the selected frame.

Action. The action in the output of the FS-Net represents the adjustment direction for each selected frame. Actions represent three types of action shift strategy $j \in (0, 1, 2)$ as ?shift left? (action 0), ?retain? (action 1) and ?shift right? (action 2), and shift step is set to be 1 frame. The FS-Net emits a vector $A \in \mathbb{R}^{m \times 3}$, where $A_{ij} \in [0, 1]$ denotes the probability of choosing the shift strategy j for the ith selected frame.

Reward. The reward function $r(S, A)$ reflects the effectiveness or goodness of the action taken by the agentwith respect to the current state S. We use the prediction results of the DD-Net as rewards. Initially, in the first iteration, we set the reward r to 1 if the prediction is correct, and -1 otherwise. For subsequent iterations $(n > 1)$, we define the r_0 reward as follows:

$$r_0 = sgn\left(p_c^n - p_c^{n-1}\right), \tag{3}$$

where, c is the ground truth label of the skeleton sequence, and p_c^n represents the probability of predicting the skeleton sequence as class c at the nth iteration. The function sgn stands for the sign function, which returns -1 for a negative input, and 1 for a positive input. The value of r_0 is in the range $(-1, 1)$. If the predicted class changes from incorrect to correct after one iteration, a strong reward of Ω is given. Otherwise, a strong punishment of $-\Omega$ is enforced. The overall reward function can be expressed as follows:

$$r = \begin{cases} \Omega & , \text{ if stimulation} \\ -\Omega & , \text{ if punishment} \\ r_0 & , \text{ otherwise.} \end{cases} \tag{4}$$

Training with policy gradient. We train the FS-Net to obtain a policy function π_θ by maximizing the following reward [26]:

$$R = \sum_{t=0}^{\tau} \gamma^t r_t, \tag{5}$$

where, γ^t is the discount factor. r_t is computed by Equation 4. τ represents the maximum iteration step. In the process

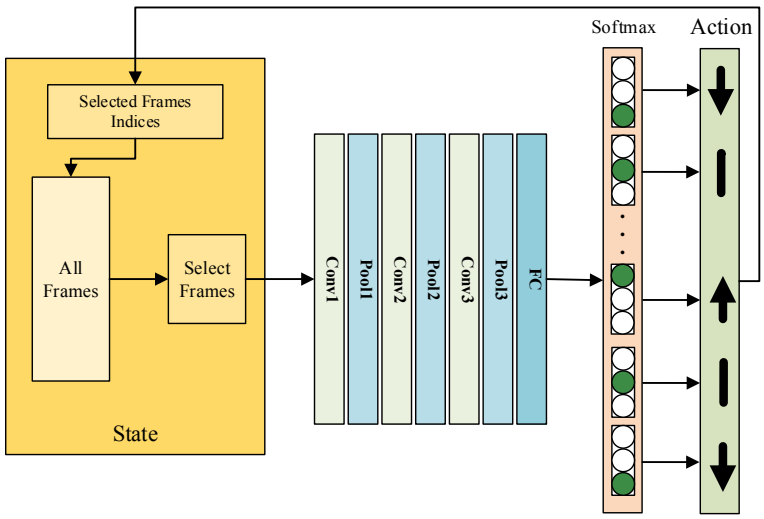


Fig. 4. The frame selection Net architecture for selecting the keyframes in hand skeleton sequence.

of selecting keyframes, as the number of selected frames increases, the complexity of the network will increase exponentially by 3^M, making the application of deep Q-learning [31] computationally infeasible. To address this challenge, we choose the policy gradient method (reinforcement learning) as an alternative approach.

The cross-entropy loss is calculated as follows to maximize the discounted reward:

$$l(\theta) = -\frac{1}{m}\sum_{t=1}^{m}\log\left(\pi_\theta\left(S_t, A_t\right)\right). \tag{6}$$

More details for training the proposed FS-Net are summarized in **Algorithm 1**.

Algorithm 1: keyframe selection training

Input: training set V_{tr}, baseline model DD-Net
Output: parameters θ of the frame select network
Initialize θ of the frame selection network
for $epoch \leftarrow 1, 2, \cdots, E$ **do**
  for V_t *in* V_{tr} **do**
    random sample m frames and its index from the training sequence V_t;
    for $i \leftarrow 1, 2, \cdots, m$ **do**
      choice action $\{a_n\}_{n=1}^{m}$ from policy π_θ;
      update the select frames by action a_n;
      update the state to sample sequence by action shift strategy $j \in (0, 1, 2)$;
      get the new prediction by the DD-Net;
      compute reward using r;
    end
    compute the loss by $l(\theta)$;
    update θ;
  end
end
return θ

IV. Experiments

In this section, we have provided a detailed description of the dataset, baseline method, and implementation details in our experiments. Subsequently, we evaluated our proposed method's performance on two datasets, comparing it against state-of-the-art methods to assess its effectiveness and superiority.

A. Datasets and experimental setup

1) Experimental Datasets: We evaluate the DRL-DDNet on two gesture recognition datasets: SHREC-17 [32] and F-PHAB [33].

SHREC-17 The SHREC-17 dataset serves as a valuable resource for researchers in the human-computer interaction domain, which is collected by RealSense depth cameras. Captured from a frontal third-person view, the dataset offers a diverse range of operations categorized into 14 and 28 classes of granularity. The dataset includes 1960 sequences for training and 840 sequences for validation. Notably, the actions are performed by 28 different users, introducing inter-user variance that closely reflects real-world scenarios.

F-PHAB The F-PHAB dataset contains motion sequences from various real-life scenarios, recorded from a first-person viewpoint. This dataset involves hand movements performed by six subjects and includes 45 action categories. We split the 1175 action sequences into three sets for training and validation, using different training-to-validation ratios. Specifically, we have employed 1:3, 1:1, and 3:1 splits for the action videos, facilitating a comprehensive evaluation of different training scenarios. Additionally, we adopted the leave-one-person-out cross-person protocol, which involves training on five subjects and testing on the remaining one.

2) Baseline Methods: In our experimental design, each hand sequence is represented as a tensor with dimensions $T \times N \times 3$. T denotes the number of sampling frames. We conduct a comparative study with two scenarios: without sampling, where no processing for the sampling frames,

and sampling, where 90% of the entire frame is randomly selected for temporal enhancement [10]. We determine the optimal number of sampling frames for subsequent experiments and set T to 32. We set the hand sampling point N to 22 for the SHREC-17 dataset, and we set the hand sampling point N to 20 for the F-PHAB dataset. 3 denotes the 3D coordinates.

The network architecture of the DD-Net is shown in **Fig.2**. In the figure, "2*CNN(3,2*filters), /2" denotes two 1D ConvNet layers (with kernel of 3, filters is the channels) and a Maxpooling (strides = 2). Similar ConvNet layers are defined in the same manner.

3) Implementation Details: The proposed method was implemented using the Tensorflow platform, and the network architecture was built on two Nvidia GTX 2080Ti GPUs for efficient computation. We used our proposed algorithm, which comprises two main components: DD-Net and FS-Net, both trained from scratch. We set the parameters for training the DD-Net based on the reference [10]. During training, we employed the Adam optimizer with an initial learning rate that decayed from 10^{-3} to 10^{-5}, conducting 600 training iterations. The value of filters in **Fig. 2** was set to 64. Regarding the FS-Net, we trained it using a dropout rate of 0.5, ReLU activation functions, and the Adam optimizer with a learning rate of 10^{-5}. We set the parameter Ω (in Eq. (4)) to 5. Additionally, we empirically set the discount factor γ^t (in Eq. (5)) to 0.7. To reduce the computational cost of selecting keyframes, we set the number of interactions between the two networks to 5.

B. Design evaluation

1) Results on the SHREC-17 Dataset: **Comparison with state-of-the-art results.**

We expect the original sequence to get a useful frame sequence after the Markov decision, which can improve the action recognition results. To provide convincing arguments for verification purposes, we used the above network to verify the SHREC-17 dataset.

The results of our proposed method are compared with several recent state-of-the-art methods, which include CNN+LSTM [19], STA-Res-TCN [20], MFA-Net [34], DD-Net [10], TCN+SUM [21], MMEGRN [35]. **Table I** shows the recognition accuracy of our proposed method and SOTA methods on the SHREC datasets. Our method achieves remarkable results, with an accuracy of 96.5% for the 14 gestures and 93.8 for the more fine-grained 28 gestures. Importantly, our model exhibits a substantial accuracy improvement of 1.9% and 1.9% for the 14 and 28 gestures, respectively, over the baseline model, showing the superiority of our proposed approach. Additionally, compared to the CNN+LSTM method [19], our model outperforms with a remarkable improvement 6.7% and 7.5% for the 14 and 28 gestures, respectively. Furthermore, The DRL-DDNet achieves competitive performance with the state-of-the-art algorithms [35]. Notably, for the 28 gestures, our method demonstrates even greater superiority. These results highlight the significant advancements and competitiveness of our proposed model in skeleton-based gesture recognition tasks.

TABLE I

COMPARISONS WITH SOTA METHODS ON THE SHREC-17 DATASET IN ACCURACY (%).

Methods	14 gestures	28 gestures
CNN+LSTM [19]	89.8	86.3
STA-Res-TCN [20]	93.6	90.7
MFA-Net [34]	91.3	86.6
DD-Net [10]	94.6	91.9
TCN+SUM [21]	93.6	91.4
MMEGR [35]	96.4	93.2
DRL-DDNet	**96.5**	**93.8**

TABLE II

COMPARISON OF ACCURACY (%) OF DIFFERENT SAMPLING METHODS FOR DIFFERENT NUMBERS ON THE SHREC-17 DATASET.

Methods	14 gestures			28 gestures		
	N=10	N=20	N=32	N=10	N=20	N=32
DD-Net (without sampling)	94.34	94.11	94.05	89.93	90.93	91.78
DD-Net (sampling)	**94.57**	94.22	94.74	89.67	91.59	91.89
DRL-DDNet	94.22	**96.41**	**96.53**	**91.05**	**91.89**	**93.81**

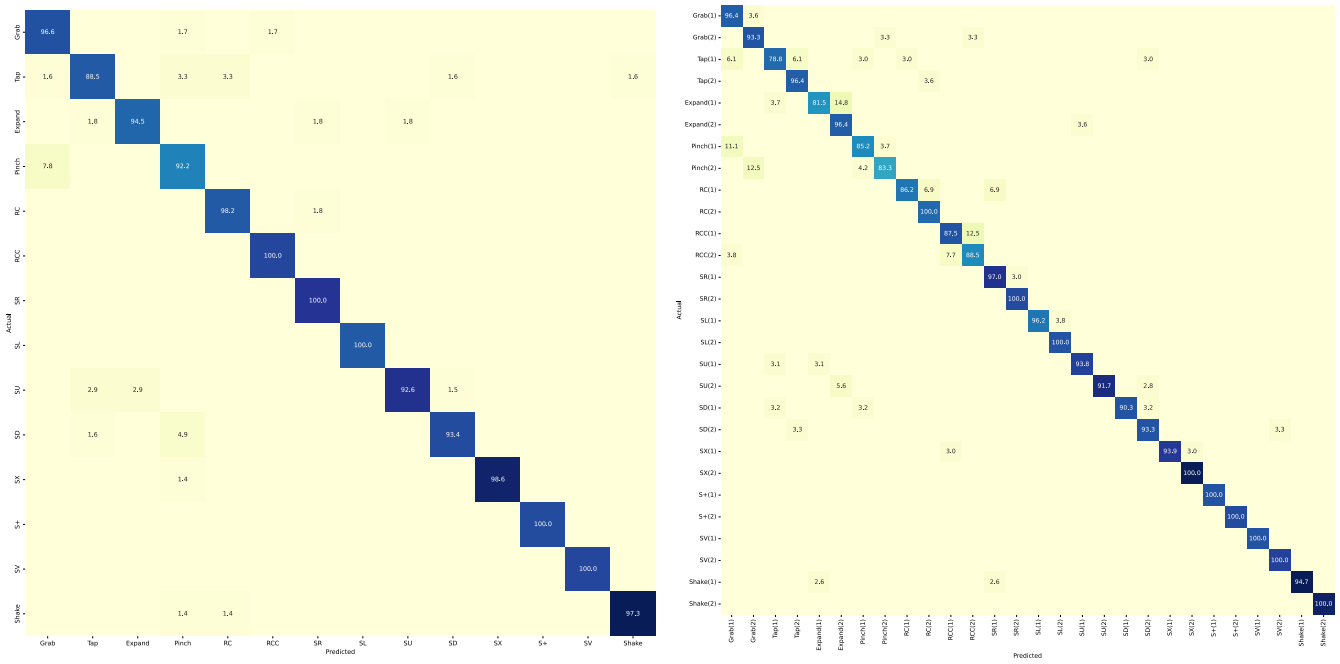
In **Fig. 5a** and **Fig. 5b**, we show the confusion matrices for the 14 and 28 gestures, respectively, allowing a detailed analysis of the recognition results for each type of action. Remarkably, our proposed DRL-DDNet achieves a recognition rate higher than 90.0% for the 14 gestures, with an outstanding 100% recognition accuracy achieved in five categories of actions. These results demonstrate exceptional performance and effectiveness of our proposed model in accurately recognizing a diverse set of gestures. During the collection process of sequence actions, there were instances of unclear expression of actions, which led to the inclusion of more redundant or ambiguous frames in the training data. In such cases, our proposed algorithm demonstrated improved performance compared to the DD-Net.

Performance evaluation with different numbers of frames selected.

We evaluate the impact of selecting different numbers of frames with the DRL-DDNet model and baseline methods. To this end, we choose to vary the number of selected frames to 10, 20, and 32. We also compare these results with scenarios involving both without sampling and sampling. As shown in **Table II**, the performance of our algorithm steadily improves as the number of sampled frames increases. This improvement is intrinsically related to the Markov decision process, which significantly enhances the recognition performance by acquiring more keyframes.

2) Results on the F-PHAB Dataset: **Comparison with state-of-the-art results.**

In our experiments on the F-PHAB dataset, we conducted a thorough comparison between our method and state-of-the-art skeleton-based gesture recognition methods using two



(a) (b)

Fig. 5. (a) Confusion matrix of SHREC dataset with 14 gestures. (b) Confusion matrix of SHREC dataset with 28 gestures.

TABLE III
COMPARISONS WITH SOTA METHODS ON THE F-PHAB DATASET IN ACCURACY(%).

Methods	Protocol			
	1:3	1:1	3:1	cross-person
DD-Net(without sampling)	91.95	93.91	95.41	80.06
DD-Net(sampling)	91.95	94.96	95.96	79.99
TCN+SUM [21]	**92.90**	95.93	**96.76**	**88.70**
DRL-DDNet	92.59	**96.00**	96.43	81.80

evaluation protocols. The results can be found in **Table III**.

The first protocol involved exploring various training-to-validation ratios. When the ratio is set to 1:3, our proposed method not only outperforms the baseline models but also achieves par performance with the TCN+SUM method [21]. Moving to the 1:1 setting ratio, our method achieves a substantial 2.09% and 1.04% improvement over the DD-Net (without sampling) and the DD-Net (sampling), respectively, and exhibits superior performance compared to the TCN+SUM. As the training dataset size increased with a 3:1 ratio, the DRL-DDNet consistently maintains an advantage over the baseline model. Experimental results show that modeling the Markov model can effectively select keyframes in the dataset to improve the performance of the model.

The second protocol, leave-one-person-out cross-person evaluation, demonstrate that the DRL-DDNet method led to an improve the DD-Net(without sampling) recognition accuracy by 1.74%, and it further enhanced the DD-Net (sampling) classification accuracy by 1.81%. Nevertheless, there remained a performance gap of approximately 6.9% compared to the TCN+SUM method. The baseline model's inherent limitation lies in its inability to generate view-point agnostic motion representations [21]. This deficiency substantially constrains its capacity to effectively accommodate the diverse hand movement styles observed among different subjects. Consequently, the model experiences a discernible degradation in its cross-person generalization ability.

V. CONCLUSION

In this paper, we proposed a deep reinforcement learning double-feature double-motion Network (DRL-DDNet) to improve the performance of skeleton-based gesture recognition by discarding irrelevant frames. The Markov decision process, implemented as the Frame Selection Network, assists the Networks in adaptively focusing on keyframes while excluding irrelevant ones that often introduced unnecessary noise. The keyframes are then used to train the DD-Net again, continuously improving the recognition performance through alternating training. The experimental results confirmed the effectiveness of our proposed DRL-DDNet, demonstrating significant accuracy enhancements compared to other state-of-the-art methods. During the keyframe selection process in reinforcement learning, we observed that the number of sampled keyframes was limited to the length of the gesture sequence. Previous studies have employed padding or truncation operations on the skeleton sequence to fix the length. However, this approach may hinder the ability to dynamically model the entire sequence. In the future, we plan to explore an adaptive sampling strategy to achieve better temporal modeling of gesture sequences,

ultimately leading to more efficient skeleton-based hand action recognition performance.

REFERENCES

[1] Fei Han, Brian Reily, William Hoff, and Hao Zhang. Space-time representation of people based on 3d skeletal data: A review. *COMPUTER VISION AND IMAGE UNDERSTANDING*, 158:85–105, MAY 2017.

[2] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[3] Hai Liu, Hanwen Nie, Zhaoli Zhang, and You-Fu Li. Anisotropic angle distribution learning for head pose estimation and attention understanding in human-computer interaction. *Neurocomputing*, 433:310–322, 2021.

[4] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13264–13273, 2021.

[5] Debasmita Mukherjee, Kashish Gupta, Li Hsin Chang, and Homayoun Najjaran. A survey of robot learning strategies for human-robot collaboration in industrial settings. *Robotics and Computer-Integrated Manufacturing*, 73:102231, 2022.

[6] Shufei Li, Ruobing Wang, Pai Zheng, and Lihui Wang. Towards proactive human–robot collaboration: A foreseeable cognitive manufacturing paradigm. *Journal of Manufacturing Systems*, 60:547–552, 2021.

[7] Inwoong Lee, Doyoung Kim, and Sanghoon Lee. 3-d human behavior understanding using generalized ts-lstm networks. *IEEE Transactions on Multimedia*, 23:415–428, 2020.

[8] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018.

[9] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016.

[10] Fan Yang, Yang Wu, Sakriani Sakti, and Satoshi Nakamura. Make skeleton-based action recognition model smaller, faster and better. In *Proceedings of the ACM multimedia asia*, pages 1–6. 2019.

[11] Dongliang He, Zhichao Zhou, Chuang Gan, Fu Li, Xiao Liu, Yandong Li, Limin Wang, and Shilei Wen. Stnet: Local and global spatial-temporal modeling for action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8401–8408, 2019.

[12] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017.

[15] Pichao Wang, Wanqing Li, Chuankun Li, and Yonghong Hou. Action recognition based on joint trajectory maps with convolutional neural networks. *Knowledge-Based Systems*, 158:43–53, 2018.

[16] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015.

[17] Hongsong Wang and Liang Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 499–508, 2017.

[18] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot. Global context-aware attention lstm networks for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1647–1656, 2017.

[19] Juan C Nunez, Raul Cabido, Juan J Pantrigo, Antonio S Montemayor, and Jose F Velez. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. *Pattern Recognition*, 76:80–94, 2018.

[20] Jingxuan Hou, Guijin Wang, Xinghao Chen, Jing-Hao Xue, Rui Zhu, and Huazhong Yang. Spatial-temporal attention res-tcn for skeleton-based dynamic hand gesture recognition. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[21] Alberto Sabater, Iñigo Alonso, Luis Montesano, and Ana C Murillo. Domain and view-point agnostic hand action recognition. *IEEE Robotics and Automation Letters*, 6(4):7823–7830, 2021.

[22] Mohammadreza Zolfaghari, Gabriel L Oliveira, Nima Sedaghat, and Thomas Brox. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2904–2913, 2017.

[23] Quentin De Smedt, Hazem Wannous, and Jean-Philippe Vandeborre. Skeleton-based dynamic hand gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2016.

[24] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation*, 25(1):24–38, 2014.

[25] Bahareh Nikpour, Dimitrios Sinodinos, and Narges Armanfard. Deep reinforcement learning in human activity recognition: A survey. 2022.

[26] Yansong Tang, Yi Tian, Jiwen Lu, Peiyang Li, and Jie Zhou. Deep progressive reinforcement learning for skeleton-based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5323–5332, 2018.

[27] Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, and Shilei Wen. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6222–6231, 2019.

[28] Wenkai Dong, Zhaoxiang Zhang, and Tieniu Tan. Attention-aware sampling via deep reinforcement learning for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8247–8254, 2019.

[29] Zheyuan Xu, Yingfu Wang, Jiaqin Jiang, Jian Yao, and Liang Li. Adaptive feature selection with reinforcement learning for skeleton-based action recognition. *IEEE Access*, 8:213038–213051, 2020.

[30] Guyue Hu, Bo Cui, Yuan He, and Shan Yu. Progressive relation learning for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 980–989, 2020.

[31] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[32] Quentin De Smedt, Hazem Wannous, Jean-Philippe Vandeborre, Joris Guerry, Bertrand Le Saux, and David Filliat. Shrec'17 track: 3d hand gesture recognition using a depth and skeletal dataset. In *3DOR-10th Eurographics Workshop on 3D Object Retrieval*, pages 1–6, 2017.

[33] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018.

[34] Xinghao Chen, Guijin Wang, Hengkai Guo, Cairong Zhang, Hang Wang, and Li Zhang. Mfa-net: Motion feature augmented network for dynamic hand gesture recognition from skeletal data. *Sensors*, 19(2):239, 2019.

[35] Adam AQ Mohammed, Jiancheng Lv, Md Islam, Yongsheng Sang, et al. Multi-model ensemble gesture recognition network for high-accuracy dynamic hand gesture recognition. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–14, 2022.