Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 12, No. 3, 2025

Lightweight Dynamic Gesture Recognition based on shufflenetv2-Mamba Hybrid Architecture

Jiaxuan Chai, Mingge Sun *, Dongxuan Huang, Sen Ye

School of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin, Jilin 132022, China
*** Corresponding author:** Mingge Sun

Abstract: Dynamic gesture recognition has important application value in human-computer interaction of mobile terminal, but the existing methods generally face the problems of high computational complexity and insufficient time sequence modeling ability. Therefore, this paper proposes a lightweight dynamic gesture recognition model based on shufflenetv2 Mamba (Shuma) hybrid architecture. In this model, Mamba's state space sequence modeling module is embedded into the shufflenetv2 backbone network to achieve efficient spatio-temporal feature fusion. First, part of the convolution operation is replaced in the downsampling bottleneck layer of shufflenetv2, and Mamba's linear complexity is used to capture the long-range dependence between video frames; Secondly, a multi-scale feature dynamic fusion mechanism is designed, which combines channel shuffle and cross layer feature stitching to enhance the collaborative representation ability of local details and global motion patterns of continuous gestures. In order to further optimize the deployment efficiency, layered quantization and structured pruning technology are introduced to compress the model parameters to 2.1MB. Experiments on a specific dynamic gesture data set including first person and home monitoring show that the accuracy of gesture classification is 89.7%, which reduces the computational overhead by about 43.6% compared with the traditional 3d-cnn and cnn-lstm models. This study provides an efficient solution for real-time dynamic gesture interaction in resource constrained scenes, and verifies the effectiveness of the fusion of lightweight convolution and sequential state space model.

Keywords: Dynamic Gesture Recognition; Lightweight Model; ShufflenetV2; Mamba; Spatio Temporal Feature Fusion.

1. Introduction

In recent years, with the rapid development of augmented reality (AR), intelligent wearable devices and human-computer interaction technology, dynamic gesture recognition, as one of the core technologies of natural interaction, has gradually become a research hotspot in the field of computer vision[1]. By recognizing the trajectory and posture changes of the hand, the system can analyze the user's intention in real time, and plays an important role in virtual reality control, smart home control, barrier free communication and other scenes. However, dynamic gesture recognition faces two core challenges: first, gesture movements are highly time-space dependent[2], and it is necessary to capture local details (such as finger bending angle) and global temporal correlation (such as continuous frame motion trend)[3]. Traditional 3D convolutional network (3D-CNN) and recurrent neural network (LSTM/GRU)[4] are difficult to achieve efficient reasoning at the mobile terminal due to high computational complexity or limited long-range modeling ability; Second, the actual application scenario has strict requirements on the lightweight and deployment efficiency of the model. Although the existing lightweight networks [5] (MobileNet and Shufflenetv2) perform well in still image classification, they lack targeted optimization of temporal dynamic features, resulting in insufficient accuracy of complex gesture recognition.

To solve the above problems, researchers have explored a variety of improvement directions in recent years: on the one hand, the efficiency of video feature modeling is optimized through spatio-temporal separation convolution (TSM) [5] or temporal attention mechanism (TAM) [6]; On the other hand, with the help of transformer's global self-attention, the ability of long sequence modeling is improved, but the secondary computational complexity still limits its application in resource constrained scenarios. At the same time, the proposal of state space models (SSMS), especially Mamba architecture[8], provides a new idea for sequential data modeling - based on selective scanning mechanism and hardware aware design, it realizes long-range dependent capture with linear complexity, showing significant advantages in the fields of language, audio and so on [9]. However, how to combine Mamba's efficient sequence modeling ability with lightweight visual network to build a dynamic gesture recognition model[10] that takes both accuracy and efficiency into account is still a subject to be explored.

This paper proposes a lightweight dynamic gesture recognition model based on shufflenetv2 Mamba hybrid architecture. The core is to propose a modular heterogeneous fusion design: embed Mamba blocks in the shufflenetv2 backbone network, use channel shuffle and multi-scale feature fusion mechanism to jointly extract local spatial features and global time dependence, break through the modeling bottleneck of traditional single-mode architecture, and design grouped state space (SSM) and hierarchical quantization strategy[11] according to the characteristics of video frame sequence, while reducing the memory occupation of Mamba modules, maintain sensitivity to complex gesture actions.

The experimental part is based on the fine dynamic gesture recognition data set and the self built multi scene gesture video. The results show that: compared with the mainstream dynamic gesture models (CNN-LSTM, Dyhand), the recognition accuracy of this method is improved by 1.8% with 43.6% parameter reduction (2.1MB), which provides a reliable solution for real-time gesture interaction in resource

constrained environments.

2. Related Work

This paper divides the related work into three main parts, including the advantages of ShuffleNetV2-Mamba structure; The comparison between the existing mainstream model and the model proposed in this paper, and the impact of each part of the model on the model.

2.1. CNN-LSTM

CNN-LSTM model combines convolutional neural network (CNN) and long-term and short-term memory network (LSTM), extracts the spatial features (such as hand shape and position) in the gesture video frame through CNN, and then uses LSTM to capture the dynamic changes of the timing of gesture action. This architecture can effectively integrate static image features and context information of continuous actions, and has excellent performance in dynamic gesture recognition. It can achieve high accuracy and support real-time interaction, especially when processing short-term gesture sequences (such as clicking and sliding). It is suitable for VR/AR and other scenes requiring low delay response.

The Spatio-temporal modeling ability of the model can not only recognize the local details of gestures, but also capture the coherence of actions, and has strong scalability in multimodal input (such as combining skeleton data). However, the model has high computational complexity, especially in the long sequence processing, it is prone to efficiency bottlenecks; At the same time, the model is highly dependent on the amount of training data and annotation quality, and its robustness may decline in complex background or occluded scenes. It needs to rely on data enhancement or additional sensors (such as depth camera) to assist optimization.

2.2. Transformer.

The transformer model was proposed by Vaswani et al. In 2017. The model uses a codec architecture. It was first applied in the NLP field, and then widely used in CV and timing analysis. Its core is to capture the sequence relationship through multi head attention and positional encoding, and can still establish long-distance correlation without convolution or circulation[12].

In dynamic gesture recognition, transformer can process the sequence of video frames and analyze the continuity and local details of gesture actions through attention mechanism. However, due to the limitations of transformer's structure, the model relies heavily on high computing resources and data sets, which limits the computational efficiency of the model and affects the ability to handle complex sequential tasks.

2.3. Mamba

Mamba is a new type of deep learning architecture based on the selective state space model, which aims to solve the efficiency bottleneck problem caused by the complexity of secondary calculation when the transformer processes sequential tasks. The selectivity mechanism and hardware aware algorithm enable it to show the same performance as transformer in audio and video tasks.

The model uses the scan mechanism to replace the traditional convolution or attention calculation, and combines the GPU memory hierarchy (SRAM and HBM) to reduce the read and write overhead of the video memory, and removes the attention module and MLP in the transformer, which is only realized by the Mamba block (SSM+MLP) stack, with a more compact structure and higher parameter utilization. The computational complexity also shows linear scaling effect when the sequence length is extended, which ensures the excellent performance of the model in long sequence tasks.

3. Network Design

This paper uses shufflenetv2 as the backbone network, its core is the channel split and channel shuffle mechanism, and integrates Mamba block to innovatively propose an efficient gesture recognition method. Its core highlights are the introduction of Mamba block and the low parameter of the network.

SMF embeds Mamba blocks in the bottleneck block of the traditional shufflenetv2 network to replace part of the convolution operation. Its key design is through the hidden state in the state space model (SSM):

$$h_t = Ah_{t-1} + Bx_t \tag{1}$$

Where h_t is the hidden state at time t, x_t is the t vector of the input sequence, a and B are the state transition matrix and input projection matrix, and their outputs: where h_t is the hidden state at time t, x_t is the t vector of the input sequence, A and B are the state transition matrix and input projection matrix, and their outputs are:

$$y_t = Ch_t + Dx_t \tag{2}$$

C is the output projection matrix and D is the connection matrix.

This improvement enhances the ability of the model to deal with complex feature sequences, and the design of selective scanning mechanism in SMF enables it to dynamically adjust SSM parameters[13], which is that the model can adaptively focus on key temporal features according to the gesture trajectory. The SSM discretization process is to convert the state transition matrix and input projection matrix into discrete parameters through zero order hold (ZOH):

$$\begin{aligned} A &= e^{\Delta A} \\ B &= (\Delta A)^{-1}(e^{\Delta A} - I)\Delta B \end{aligned} \tag{3}$$

Δ is a learnable time step parameter.

3.1. Network Model Structure

The data flow dependency of this study is shown in Figure 1. The model provides dynamic scalability while taking into account the efficiency performance balance, and can dynamically adjust the channel division ratio to adapt to the constraints of different hardware resources.

Local convolution flow ensures the real-time performance of high-resolution feature extraction.

Through the shallow feature extraction module, the forward flow of the model realizes the collaboration of local perception and global sequence modeling through the Shuffle-Mamba block stacked in the multi-stage feature fusion, uses attention pooling to compress the space-time dimension, and outputs the gesture category probability through the full connection layer and SoftMax layer after retaining the discriminative features.

In the module level dependency, the input characteristic graph is divided into two channels and sent to the local and global information processing flows respectively. The local convolution flow adopts the mechanism of grouping convolution and channel shuffling to extract spatial details and reduce the computational complexity. The global sequence flow paves the feature map into a sequence, and

uses SSM to model the long-distance Spatio-temporal dependence to capture the dynamic continuity of gestures. Finally, the learnable weight is used to fuse the two stream features to balance the local details and global features.

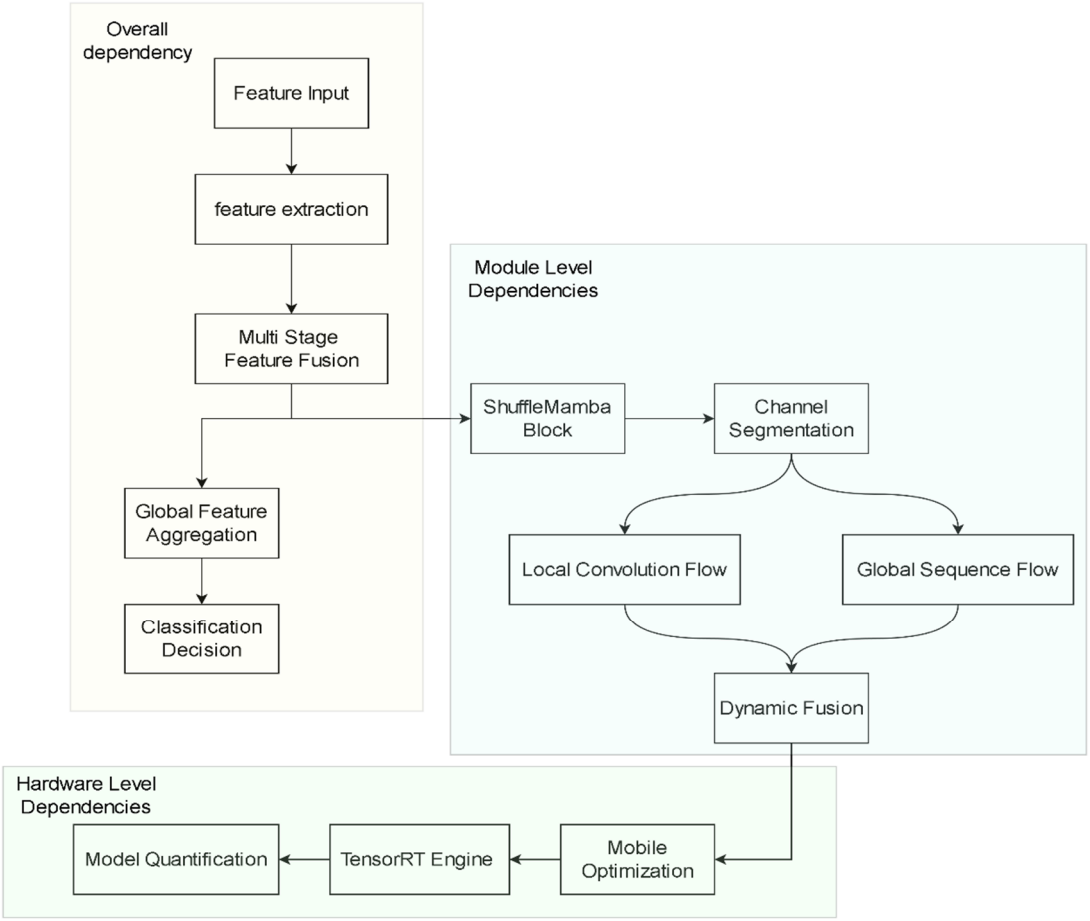

Fig 1. overall network process

At the hardware level, the Pytoch model is converted to ONNX format to adapt to the mainstream device architecture, and the model is quantified to reduce the model volume. Combined with gradient based unstructured pruning, the calculation is further compressed to enable it to run on edge devices.

3.2. Backbone Network Shufflenetv2

3.2.1. Core Design Principles

The design goal of shufflenetv2 is to optimize the reasoning efficiency of mobile terminals and embedded edge devices. Its network design focuses not only on the theoretical amount of computation (Flops), but also on the actual running speed and memory access cost (MAC). According to the network structure, when the input and output channels are equal, the memory access cost (MAC) is the minimum. In order to ensure the memory exchange efficiency, the backbone network needs to balance the channels. In the optimized computational efficiency, deep convolution:

$$\mathrm{FLOPs}_{\mathrm{depthwise}} = k^2 \cdot C_{\mathrm{in}} \cdot H \cdot \quad (4)$$

Pointwise convolution:

$$\mathrm{FLOPs}_{\mathrm{pointwise}} = 1^2 \cdot C_{\mathrm{in}} \cdot C_{\mathrm{out}} \cdot H \cdot W \quad (5)$$

The total amount of calculation is the sum of the above two, compared with the standard convolution calculation $k^2 C_{\mathrm{in}} C_{\mathrm{out}} HW$,can reduce the amount of calculation by $1/C_{\mathrm{in}} + 1/k^2$ times.

The two main technical features of backbone network: packet convolution and channel shuffling will also affect the overall reasoning speed. Although packet convolution will reduce flops, it will increase Mac. According to the experiment, the larger the packet is, the slower the reasoning speed is. When the number of packets increases from 1 to 8, the GPU reasoning speed decreases to 1/4. In dealing with network fragmentation, the concept structure will reduce the parallelism and increase the synchronization overhead, so the network structure adopts a simple single branch structure to improve the efficiency of parallel devices.

3.2.2. Structural Features

The input channel of the common unit in the shufflenetv2 structure is divided into two parts. One part is processed by deep separable convolution and 1x1 convolution. The other part is direct identity mapping. Finally, the stitching is through channel shuffle to enhance the interaction of information. At the same time, the channel segmentation is canceled, and the spatial down sampling and channel expansion are directly realized through the convolution with step size of 2 to maintain low Mac.

The channel shuffling mechanism can improve the expression ability of the model and reduce the computational fragmentation. Usually, the shuffling operation is performed after splicing, combined with convolution operation to achieve efficient feature fusion. The deep separable convolution is adopted to replace the standard convolution, which reduces the parameters and the amount of calculation, and has different network widths to adapt to the equipment with different resource constraints.

3.2.3. Shufflenetv2 Structure Integrated with Mamba Block

The purpose of fusing the shufflenetv2 structure of Mamba block is to solve the balance problem between computational efficiency and temporal modeling ability in dynamic gesture recognition by combining the efficient spatial feature extraction ability of lightweight convolutional network and Mamba's long sequence modeling advantage. Its core structure is shown in Figure 2.

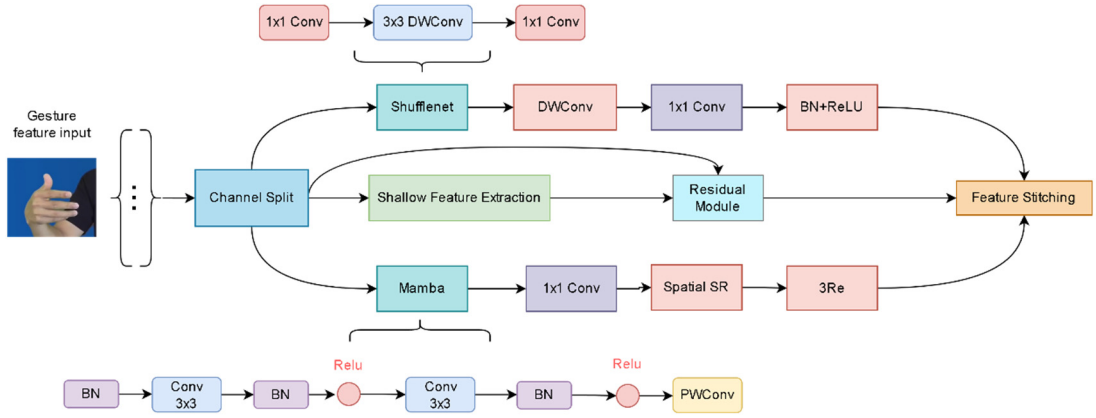

Fig 2. Fusion model structure

After inputting the dynamic gesture video sequence, it is divided into three parts through channel segmentation. The first part extracts the low-order local spatial details of finger posture through a series of calculations such as the initial convolution layer of shufflenet module, the depth separable convolution layer and point-by-point convolution. In the second part, the feature map is rearranged into sequences and input into Mamba's state space model (SSM) to capture the motion dependence between frames, such as the waving trajectory, and then the spatial dimension is restored by reverse rearrangement. The last part is the feature fusion of the residual layer and the two layers after the simple shallow feature extraction, and the three parts are output and spliced to perform the channel shuffling operation. After the final feature is pooled by the global average, the gesture category probability is output through the classification layer.

4. Experiment

In this paper, shufflenetv2-Mamba is compared with the existing popular model to evaluate the performance of the model. The evaluation results on the above data sets show that the overall performance of the model proposed in this paper is excellent.

4.1. Experimental Data and Evaluation Criteria

The data set used in this experiment contains the specific dynamic gestures in the first person and home monitoring. In these dynamic data sets, including seven kinds of dynamic gestures such as sliding up and sliding down under different lighting conditions and different viewing angles. In these different types of data sets, the number of single categories of all gestures is the same. Due to the influence of illumination, the model needs to have strong detail extraction ability and noise tolerance[14]. On the data set of partial gesture inching, the model must be able to capture small details and aggregate local information. Through the classification task on these data sets, it is proved that the model can face the gesture control of diverse scenes such as smart home. The Table 1 for details of datasets.

Table 1. Dataset details

Dataset	Task	Number of samples	Training/Validation/Testing
Fine dynamic	Multi classification	350	280/35/35
First person	Multi classification	700	560/70/70
Oblique top	Multi classification	700	560/70/70
Opposite	Multi classification	420	336/42/42

On the above data sets, ACC and macro average are used to evaluate the model, where ACC represents the proportion of all samples correctly classified.

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_i = \hat{y}_i) \tag{6}$$

Macro average is to directly average the indicators of all categories, focusing on the performance of small categories.

$$\text{Macro-Precision} = \frac{1}{C}\sum_{i=1}^{C} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{7}$$

4.2. Experimental Details

This model is trained and reasoned on the above data sets, and compared with the mainstream models (MobileNet, ViM), etc. The experiment was carried out on two 24GB NVIDIA Geforce RTX 3090ti using python3.8 and pytorch2.4.0. After initial parameter adjustment, the training superparameter is selected as epoch 200 and batch size 16. For all data sets, the optimizer uses Adam, and the learning rate is dynamically adjusted from 1e-3 to 1e-5.

4.3. Result Analysis

The results of the comparative experiment are shown in Table 2, and the data adopts the percentage system. The results show that in some scenarios, the model proposed in this paper is superior to the existing mainstream models. Compared with the same mobile terminal model MobileNet, Shuma's ACC is improved by 3% on average. In the data set of dynamic gesture smart home application scenarios above the slope, the ACC is improved by 4.8% compared with 3D-ResNet. Similarly, Shuma also performs well in other data sets.

Table 2. performance comparison between Shuma and other models

Model	Fine dynamic		First person		Oblique top		Opposite	
	ACC	MP	ACC	MP	ACC	MP	ACC	MP
MobileNetXt+CA[15]	73.2	61.7	81.1	73.2	64.4	51.3	85.6	72.6
ViM[16]	84.1	**64.2**	86.5	74.6	70.2	53.1	89.6	74.1
3D-LSTM[17]	81.9	63.0	83.3	73.5	69.3	51.7	87.4	74.8
3D-ResNet[18]	80.4	61.1	84.2	71.5	65.2	51.4	84.8	73.1
PoseNet[19]	**84.2**	63.3	86.2	75.1	**70.9**	**54.0**	89.1	75.3
Inception-LSTM[20]	80.2	61.1	81.6	73.3	69.5	53.0	86.3	73.4
Shuma	81.7	62.9	**87.7**	**76.2**	70.0	53.7	**89.7**	**75.4**

In order to verify the generalization ability of the model in this paper, data enhancement is carried out on the proposed dataset[22] to verify the ability of the model to process noisy data. The accuracy of all data sets is averaged, and five experiments are carried out on each model to eliminate the interference of special case data on the experimental results. The experimental results take the noise adding ratio of 0.2, as shown in Figure 3.

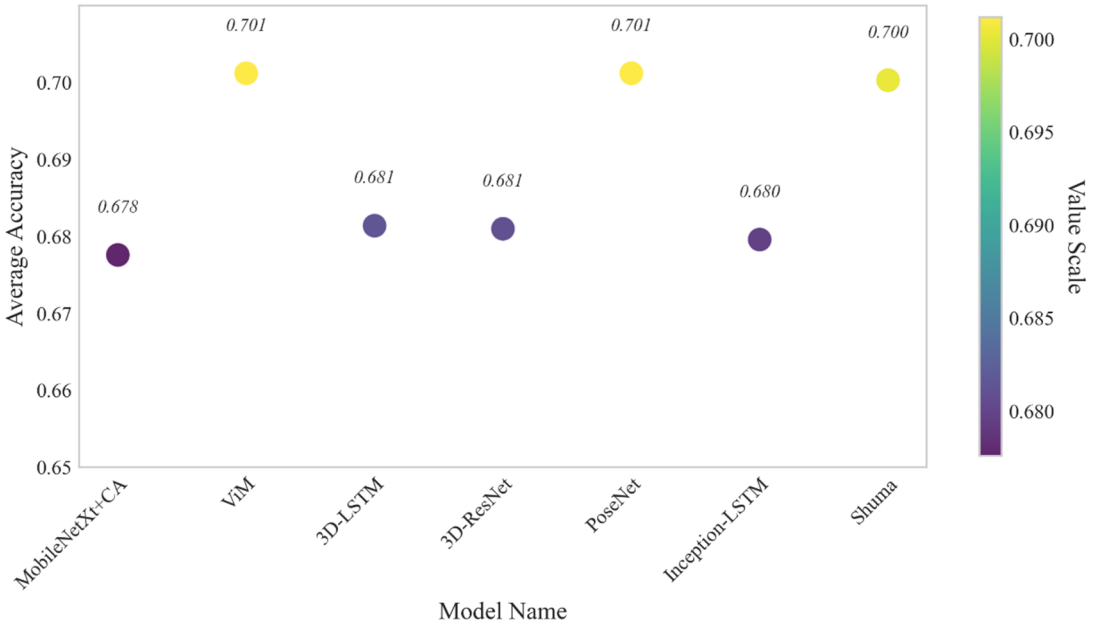

Fig 3. Comparison of average accuracy of model denoising

The results show that the average accuracy of Shuma is also better than most models when the data set is added with noise[21], which further proves its excellent performance in dynamic gesture recognition.

5. Conclusion

The Shuma model proposed in this paper is a dynamic gesture recognition model based on improved shufflenetv2 and Mamba. Combining the lightweight architecture of shufflenetv2 model with the long time-series dependence of Mamba results, the model has unique advantages in the deployment and recognition accuracy of marginalized devices. Combining with the different widths of the backbone network, it can take into account different computing devices, and significantly improves the scalability of the model. In the case of noise interference in the scene, Shuma model can still maintain the performance superior to other models, highlighting its ability to effectively deal with gesture recognition in complex scenes. In the future, we will continue to explore the application of Shuma model and other lightweight models in gesture control tasks, and further optimize the structure and parameters of the model to achieve better performance and be competent for more complex scenes.

References

[1] Hu J ,Liu S ,Liu M , et al.ST-CGNet: A spatiotemporal gesture recognition network with triplet attention and dual feature fusion[J]. Pattern Recognition,2025,167111767-111767.

[2] Shaopeng C, Xueyu H .LM-Net: a dynamic gesture recognition network with long-term aggregation and motion excitation[J]. International Journal of Machine Learning and Cybernetics, 2023, 15(4):1633-1645.

[3] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. -C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018.

[4] X. Zhang, X. Zhou, M. Lin and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018.

[5] TSM: Temporal Shift Module for Efficient Video Understanding. [J].IEEE transactions on pattern analysis and machine intelligence,2020, PP.

[6] Z. Liu, L. Wang, W. Wu, C. Qian and T. Lu, "TAM: Temporal Adaptive Module for Video Recognition," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp.

[7] Gunawardane H S D P, MacNeil R R ,Zhao L , et al.A Fusion Algorithm Based on a Constant Velocity Model for Improving the Measurement of Saccade Parameters with Electrooculography [J].Sensors,2024,24(2).

[8] Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces[J]. arXiv preprint arXiv:2312.00752, 2023.

[9] Tu C J, Chuang L Y, Chang J Y, et al. Feature selection using PSO-SVM [J]. IAENG International journal of computer science, 2007, 33(1).

[10] Gao Q, Chen Y, Ju Z, et al. Dynamic hand gesture recognition based on 3D hand pose estimation for human–robot interaction[J]. IEEE Sensors Journal, 2021, 22(18): 17421-17430.

[11] Zhang W, Wang J, Lan F. Dynamic hand gesture recognition based on short-term sampling neural networks[J]. IEEE/CAA Journal of Automatica Sinica, 2020, 8(1): 110-120.

[12] De Smedt Q, Wannous H, Vandeborre J P. Skeleton-based dynamic hand gesture recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2016: 1-9.

[13] Prakash K S, Kunju N. An optimized electrode configuration for wrist wearable EMG-based hand gesture recognition using machine learning[J]. Expert Systems with Applications, 2025, 274: 127040.

[14] Ma Q, Gu Z, Gao X, et al. Intelligent Hand-Gesture Recognition Based on Programmable Topological Metasurfaces [J]. Advanced Functional Materials, 2025, 35(1): 2411667.

[15] Pintelas E, Livieris I E, Tampakas V, et al. MobileNet-HeX: Heterogeneous Ensemble of MobileNet eXperts for Efficient and Scalable Vision Model Optimization[J]. Big Data and Cognitive Computing, 2025, 9(1): 2.

[16] Wu R, Liu Y, Liang P, et al. H-vmunet: High-order vision mamba unet for medical image segmentation[J]. Neurocomputing, 2025: 129447.

[17] De Jesus N M, Festijo E D, Apolinario G F D G, et al. Multi-Location and Multi-Feature LMP Forecasting: A 2D Spatiotemporal LSTM-CNN Approach[C]//2025 15th International Conference on Power, Energy, and Electrical Engineering (CPEEE). IEEE, 2025: 207-214.

[18] Boitel E, Mohasseb A, Haig E. MIST: Multimodal emotion recognition using DeBERTa for text, Semi-CNN for speech, ResNet-50 for facial, and 3D-CNN for motion analysis[J]. Expert Systems with Applications, 2025, 270: 126236.

[19] Shahid M A, Raza M, Sharif M, et al. Pedestrian POSE estimation using multi-branched deep learning pose net[J]. PloS one, 2025, 20(1): e0312177.

[20] Zhang W. Dynamic pose recognition based on deep learning: Developing a CNN model for choral conductor pose recognition [J]. Journal of Computational Methods in Sciences and Engineering, 2025: 14727978251323068.

[21] Huang S, Zhang H, Li X. Enhance vision-language alignment with noise [C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(16): 17449-17457.

[22] Falisse A, Uhlrich S D, Chaudhari A S, et al. Marker data enhancement for markerless motion capture[J]. IEEE Transactions on Biomedical Engineering, 2025.