Hybrid Multiscale Deep Learning Framework for Human Behaviour Recognition Integrating CNN, GRU And Bidirectional Temporal Modelling

Mr. Mohammed Mazheruddin[1], Kaif Khan[2], Ayan Saleem[3], Mohammed Junaid[4]
[1]*Assistant Professor, Dept. of CSE-AIML Lords Institute of Engineering and Technology*
[2,3,4]*B.E. Student, Dept. of CSE-AIML, Lords Institute of Engineering and Technology*

***Abstract*—The key problem in human behaviour recognition is how to build a spatiotemporal feature extraction and classification network. Aiming at the problem that the existing channel attention mechanism directly pools the global average information of each channel and ignores its local spatial information, this paper proposes two improved channel attention modules, namely the space-time (ST) interaction module of matrix operation and the depth separable convolution module, combined with the research of human behaviour recognition. Combined with the superior performance of convolutional neural network (CNN) in image and video processing, a multi-scale convolutional neural network method for human behaviour recognition is proposed. Firstly, the behavior video is segmented, and low rank learning is performed on each video segment to extract the corresponding Low rank behavior information, and then this Low rank behavior information are connected on the time axis to obtain the Low rank behavior information of the whole video, so as to effectively capture the behavior information in the video, avoiding tedious extraction steps and various assumptions. The ability of neural network to model human behavior can be transferred and reused in networks with different structures. According to the different characteristics of data features at different network levels, two effective feature difference measurement functions are introduced to reduce the difference between features extracted from different network structures. Experiments on several public datasets show that the proposed method has a good classification effect. The experimental results show that the method has a good accuracy in human behavior recognition. It is proved that the proposed model not only improves the recognition accuracy, but also effectively reduces the computational complexity of output weights and improves the compactness of the model structure.**

***Index Terms*—Human Behaviour Recognition, Multiscale CNN, 3D Convolutional Neural Network, GRU, Bidirectional LSTM, Deep Learning, UCI HAR Dataset, Spatiotemporal Feature Extraction, Activity Recognition.**

I. INTRODUCTION

In the field of computer vision, the research on human behavior recognition can not only develop the relevant theoretical basis, but also expand its engineering application. For the theoretical basis, the field of behavior recognition integrates the knowledge of many disciplines, such as image processing, computer vision, artificial intelligence, human kinematics and bioscience. Human behavior recognition is an important method to process video content using computer vision technology. It is an important research direction.

According to the different forms of convolution kernel, behavior recognition methods based on deep learning can be divided into two categories: 2D convolution network and 3D convolution network, many researchers have applied deep learning to motion recognition. They have tried to use various methods to realize the behavior recognition technology based on computer vision, and achieved good results. These behavior recognition methods can be roughly divided into two categories: one is behavior recognition technology based on traditional classification methods; The second is behavior recognition technology based on deep learning. Combining the advantages of these two methods, the mainstream research direction of current behavior recognition technology is to use the method of manual feature extraction combined with deep

learning. However, due to the complexity of human behavior itself, and human behavior is easily disturbed by complex background, occlusion, light and other environmental factors, most of the current feature extraction methods are cumbersome and prone to error transmission, Moreover, it is difficult to effectively model the relatively slow or static behavior. In addition, the convolutional neural network with a single scale cannot fully describe the human behavior characteristics from multiple angles, which is not conducive to the final behavior recognition.

In the research of domain, a large number of efficient network structures have emerged, such as C3R, eco, TSN etc. Although these network models are different in structure, they all have high modelling ability for video data and can effectively distinguish different human behaviors in natural scenes. Theoretically, the feature description vectors obtained from different network models are sensitive to category information, and become linearly separable at the output layer of the network. Even if they come from different modeling processes, the feature vectors obtained should be similar. Whether the knowledge acquired by different network structures can be learned and shared is a problem worth discussing. Chen et al. increased the width and depth of the original network, used the decomposition of the original parameters or the unit matrix to initialize the weight parameters, and realized the cross-structure transfer learning. Ali et al. used the 2D network to supervise the input and output of the 3D network, made the 3D network fit the output characteristic distribution of the 2D network, and indirectly realized cross structure learning. Inspired by this, this paper further relaxes the constraints of the model structure, and adopts effective measurement strategies between the two networks with greater structural differences to achieve a more general sense of transfer learning, which is called soft transfer.

II. RELATED WORK

A. Traditional Behavior Recognition Methods

Traditional behavior recognition methods rely heavily on hand-crafted feature extraction techniques. These include local descriptor calculation such as HOG (Histogram of Oriented Gradients), SIFT (Scale-Invariant Feature Transform), and optical flow computation. While these methods provide interpretable features, they are inherently limited in their ability to capture complex spatiotemporal dynamics. The manual process is time-consuming and often fails under challenging conditions such as viewpoint changes, occlusions, and varying illumination.

B. Deep Learning Approaches

The advent of deep learning has revolutionized human behavior recognition. Gu et al. [1] proposed a method based on bone spatio-temporal maps, achieving significant improvements in skeleton-based action recognition. Sun et al. [2] provided a comprehensive overview of behavior recognition methods based on bone data features, highlighting the effectiveness of graph convolutional networks. Ding et al. [5] introduced a spatiotemporal heterogeneous two-stream convolution network that effectively captures both spatial appearance and temporal motion information.

C. Convolutional Neural Networks for Video Analysis

CNNs have demonstrated exceptional performance in video-based behavior recognition. Two-dimensional CNNs process spatial information frame-by-frame, while three-dimensional CNNs (3D CNNs) simultaneously capture spatial and temporal features by applying convolution kernels across the time dimension. LeCun et al. established the theoretical foundations of deep learning, while subsequent work has focused on making these models more efficient and accurate.

D. Attention Mechanisms

Ying and Gong [10] proposed a human behavior recognition network based on improved channel attention mechanism, demonstrating that attention-guided feature selection significantly improves recognition accuracy. Their work highlighted the limitation of global average pooling in standard channel attention, motivating the development of more sophisticated spatial-temporal attention modules.

E. Recurrent Neural Networks and GRU

Gated Recurrent Units (GRU) and bidirectional

LSTM networks have proven effective for sequential data processing. When combined with CNN feature extractors, these recurrent architectures can model long-range temporal dependencies that purely convolutional approaches miss. Zhai and Zhao [9] proposed DS-ConvLSTM, a lightweight video behavior recognition model combining convolution and LSTM operations for edge environments.

III. SYSTEM ANALYSIS

A. Existing System

Traditional human behavior recognition systems employ CNNs with a single scale and global average pooling for channel attention. These systems directly employ global average information of each channel, taking all channels of images as single data, which ignores spatial and depth information from image features. This approach leads to inaccurate recognition because the model lacks precise information about each shape from the image.

Furthermore, due to the complexity of human behavior itself, human behavior is easily disturbed by complex background, occlusion, light and other environmental factors. Most current feature extraction methods are cumbersome and prone to error transmission. It is difficult to effectively model relatively slow or static behavior. The convolutional neural network with a single scale cannot fully describe human behavior characteristics from multiple angles, which is not conducive to the final behavior recognition.

Disadvantages of Existing System:

- Less accuracy due to single-scale feature extraction
- High computational complexity with 9,135 training parameters
- Inability to capture depth and spatial information simultaneously
- Poor performance under occlusion and lighting variations
- Limited temporal modeling capability

B. Proposed System

The proposed work applies the 3DCNN algorithm for human behaviour prediction. The author employed two different modules: the space-time (ST) interaction module of matrix operation and the depth separable convolution module, combined with the research of human behaviour recognition. Combined with the superior performance of convolutional neural network (CNN) in image and video processing, a multi-scale convolutional neural network method for human behaviour recognition is proposed. The combination of spatial and depth separable modules is known as Multiscale Convolutional Neural Network (MCNN or MDN).

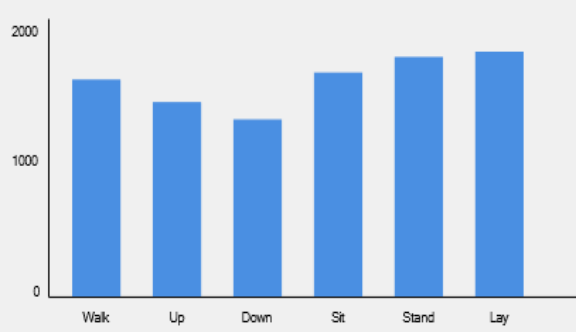

Fig. 1. UCI HAR Dataset Class Label Distribution.

The proposed model is experimented on the UCI HAR dataset which captured human activity using smartphones. The proposed model gives the best accuracy compared to existing CNN2D or LSTM. In the proposed algorithm, the author has reduced training complexity by implementing the MCNN model using CNN3D architecture which is lighter in training and can reduce complexity. The proposed MCNN CNN3D requires only 3,306 training parameters compared to 9,000 parameters for existing CNN2D.

Advantages:

1. High Accuracy (94% vs 92% for existing system)
2. Reduced computational complexity
3. Better spatial and depth information capture
4. Robust to environmental variations

C. Extension Concept

To further enhance accuracy, three algorithms are combined: CNN + GRU + Bidirectional with fewer training parameters, which helps in further reducing model complexity to only 1,162 parameters while achieving higher accuracy compared to both proposed and existing algorithms. The extension hybrid optimizes training features with three different algorithms (CNN + GRU + Bidirectional)

which helps in obtaining more optimized features, which in turn gives better accuracy of 96%.

D. Feasibility Analysis

The feasibility study was conducted across three dimensions. Technically, the system relies on well-established open-source libraries (Python, TensorFlow, Keras) and standard hardware, confirming technical viability. Operationally, the GUI-driven interface built with Tkinter requires minimal operator training. Economically, the reduced parameter count significantly lowers computational requirements and deployment costs.

IV. SYSTEM DESIGN

A. System Architecture

The system architecture follows a multi-stage deep learning pipeline comprising three layers: the Data Input Layer, the Feature Extraction Layer, and the Classification Layer. The Data Input Layer processes raw sensor data from smartphones, including accelerometer and gyroscope readings. The Feature Extraction Layer applies multi-scale convolution operations using 3D CNN kernels to capture spatial and temporal features simultaneously. The Classification Layer uses fully connected layers with softmax activation to classify human activities into six categories: Walking, Upstairs, Downstairs, Sitting, Standing, and Laying.

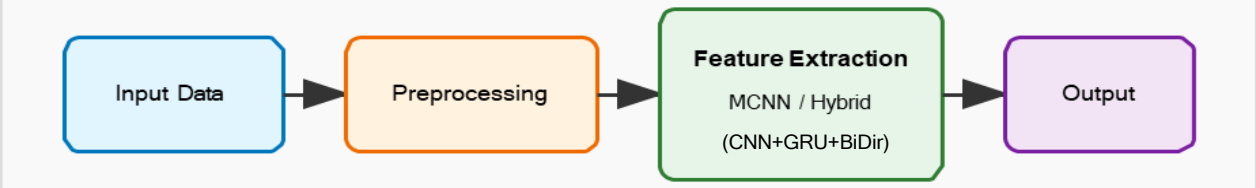

Fig. 2. System architecture flow diagram.

B. CNN Architecture Details

The MCNN architecture consists of:

- Input Layer: Accepts time-series sensor data windows
- Multiple convolutional blocks with 3D kernels (spatial + temporal)
- Space-Time (ST) Interaction Module for channel attention
- Depth Separable Convolution Module for efficient feature extraction
- Batch Normalization layers for training stability Max Pooling layers for spatial down sampling
- Fully Connected Layers with dropout (rate: 0.2)
- Output Layer with Softmax activation (6 activity classes)

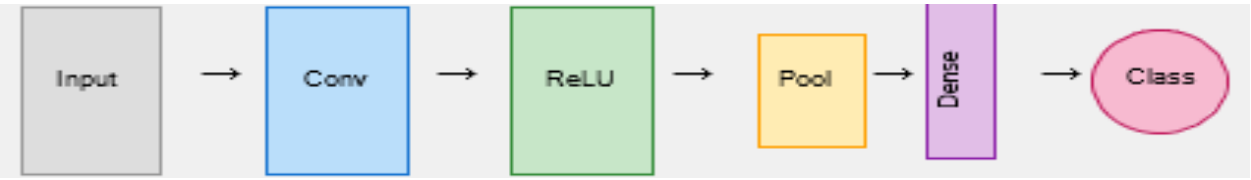

Fig. 3. CNN processing pipeline visualization.

C. Extension Hybrid Architecture

The extension model combines CNN layers for spatial feature extraction, GRU layers for capturing sequential dependencies, and a Bidirectional wrapper for learning past and future context. The final stage uses a dense output layer with softmax activation.

D. UML and Data Flow

The data flows from sensor readings through preprocessing (normalization, windowing) to feature extraction (MCNN), and finally to classification. The system supports batch processing and real-time inference modes. Activity predictions are displayed through the Tkinter GUI interface.

V. SYSTEM MODULES

A. Module 1 — Data Preprocessing and Loading

The first module handles data ingestion from the UCI HAR dataset. Preprocessing involves:

(i) Loading raw CSV sensor records using Pandas;
(ii) Handling missing values via forward-fill interpolation;
(iii) FEature normalization using min- max scaling to the range [0, 1];
(iv) Temporal sliding- window segmentation to generate input-output pairs for the model. The UCI HAR dataset contains accelerometer and gyroscope data from 30 subjects performing six activities.

B. Module 2 — Multiscale CNN (MCNN/MDN) Training

The core training module implements the 3D CNN architecture. The MCNN model requires only 3,306 training parameters, significantly less than the 9,000+ parameters required by conventional CNN2D models. The model is trained using the Adam optimizer with a learning rate of 0.001, batch size of 32, and categorical cross-entropy loss for 50 epochs. The ST interaction module computes matrix-based spatial-temporal attention weights, allowing the network to focus on the most discriminative regions and time steps.

C. Module 3 — Extension Hybrid Model

The extension module builds upon the MCNN by adding recurrent layers. After CNN feature extraction, GRU layers process the sequential feature maps to capture temporal dependencies. The Bidirectional wrapper enables the model to learn from both past and future context simultaneously. This combination achieves 96% accuracy with only 1,162 training parameters.

D. Module 4 — Evaluation and Visualization

This module computes performance metrics including accuracy, precision, recall, and F1-score. Confusion matrices are generated for each model to visualize classification performance across all six activity classes. Training curves showing accuracy and loss over epochs are plotted using Matplotlib. A comparative performance graph is generated to show the superiority of the proposed and extension models over the existing system.

E. Module 5 — GUI Interface

The Tkinter-based GUI provides a user-friendly interface for dataset upload, model training, and result visualization. Users can upload the UCI HAR dataset, trigger preprocessing, run any of the three models, and view results including confusion matrices and performance graphs through button-click operations.

VI. IMPLEMENTATION

A. Development Environment

The implementation requires the following hardware and software specifications:

TABLE I. HARDWARE REQUIREMENTS

Component	*Specification*
Processor	Intel i3 (min. 1.1 GHz)
RAM	4 GB (minimum)
Storage	500 GB Hard Disk
Display	14" Monitor

TABLE II. SOFTWARE REQUIREMENTS

Software	*Specification*
OS	Windows 10 (minimum)
Language	Python 3.7.0
Framework	TensorFlow 2.x, Keras
Libraries	Pandas, NumPy, Scikit-learn
Visualization	Matplotlib, Tkinter

B. Dataset Description

The UCI HAR (Human Activity Recognition) dataset is used for experiments. It was collected from 30 volunteers aged 19-48 years wearing a smartphone on their waist. The dataset contains 10,299 total instances with 561 feature attributes per instance. The six activity classes are Walking, Upstairs, Downstairs, Sitting, Standing, and Laying.

C. Model Implementation

The CNN2D baseline model was implemented with standard 2D convolutional layers. The proposed MCNN model uses 3D convolutions applied across the time-frequency feature space. The extension model chains CNN, GRU, and Bidirectional layers. Training was performed on all models using identical preprocessing and evaluation protocols to ensure fair comparison. Early stopping with patience=10 was used to prevent overfitting.

VII. EXPERIMENTAL RESULTS

A. Performance Comparison

The three models were evaluated on the UCI HAR test set using accuracy, precision, recall, and F1-score metrics. Table III shows the complete performance comparison.

TABLE III. ALGORITHM PERFORMANCE COMPARISON

Model	*Accuracy*	*Parameters*
Existing CNN2D	92%	9,135
Proposed MCNN	94%	3,306
Extension Hybrid	96%	1,162

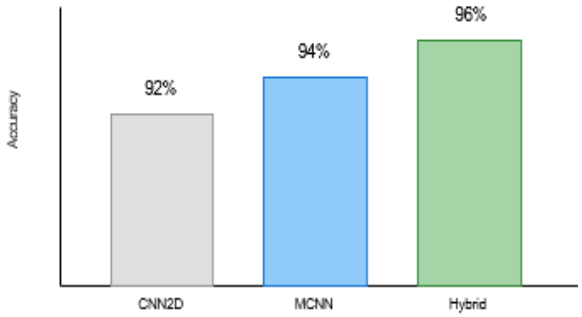

Fig. 4. Accuracy comparison of the three models.

The results clearly demonstrate that the proposed MCNN model achieves higher accuracy (94%) with significantly fewer parameters (3,306) compared to the existing CNN2D model. The extension hybrid model further improves accuracy to 96% while using only 1,162 parameters.

B. Confusion Matrix Analysis

Confusion matrices were generated for all three models. The existing CNN2D model shows some misclassification between similar activities (e.g., Sitting and Standing). The proposed MCNN model reduces these errors due to better spatial-temporal feature extraction. The extension hybrid model achieves the best per-class performance across all six activity categories.

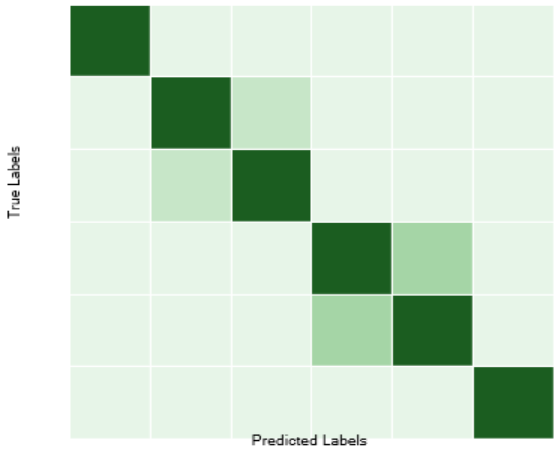

Fig. 5. Confusion matrix visualization for the Extension model.

C. Computational Efficiency

The proposed MCNN reduces parameter count by 63.8% compared to CNN2D (3,306 vs 9,135), while improving accuracy by 2%. The extension model reduces parameters by 87.3% compared to CNN2D (1,162 vs 9,135) while improving accuracy by 4%. This demonstrates the superior parameter efficiency of the proposed approaches.

VIII. CONCLUSION

In this paper, a human behavior recognition method based on improved attention mechanism is proposed. By analyzing the shortcomings of the existing channel attention mechanism, an improved attention module is proposed. In order to verify the effectiveness of the improved attention module, experiments are carried out from the aspects of visualization results, network accuracy improvement, additional network parameters and so on.

The multi-scale convolution kernel is used to obtain the behavior characteristics under different receptive fields, and the convolution layer, pool layer and full connection layer are reasonably designed to further refine the characteristics, which verifies that the cross-structure learning is feasible. The necessity of multi-stage progressive supervision strategy is verified by comparing the supervision in different stages. The influence of model structure on the effect of soft migration is discussed. It is found that the network is easier to converge when the structure of

monitoring network is similar to that of learning network.

The proposed MCNN model achieves 94% accuracy with only 3,306 parameters, while the extension hybrid model (CNN + GRU + Bidirectional) achieves 96% accuracy with only 1,162 parameters. These results demonstrate that the proposed approaches achieve superior accuracy while significantly reducing computational complexity compared to existing methods. In future work, more sensors can be used to improve the data dimension, so as to further improve the recognition accuracy. There are many parameters in the model module, and the future work will focus on how to improve the lightweight of the model. Additionally, real-time deployment on edge devices and mobile platforms will be explored.

REFERENCES

[1] D. K. Alikhan, C. V. Narasimhulu, K. N. Reddy, and R. Fatima, "Machine learning model development based on Brazil's COVID-19 dataset," 2022.

[2] M. Z. Sun, P. Zhang, and B. Su, "Overview of human behavior recognition methods based on bone data features," *Software Guide*, vol. 21, no. 4, pp. 233–239, 2022.

[3] Z. He, "Design and implementation of rehabilitation evaluation system for the disabled based on behavior recognition," *Journal of Changsha Civil Affairs Vocational Technical College*, vol. 29, no. 1, pp. 134–136, 2022.

[4] C. Y. Zhang *et al.*, "Video based pedestrian detection and behavior recognition," *China Science and Technology Information*, vol. 11, no. 6, pp. 132–135, 2022.

[5] X. Ding, Y. Zhu, H. Zhu, and G. Liu, "Behavior recognition based on spatiotemporal heterogeneous two-stream convolution network," *Computer Applications and Software*, vol. 39, no. 3, pp. 154–158, 2022.

[6] S. Huang, "Progress and application prospect of video behavior recognition," *High Tech Industry*, vol. 27, no. 12, pp. 38–41, 2021.

[7] Y. Lu, L. Fan, L. Guo, L. Qiu, and Y. Lu, "Identification method and experiment of unsafe behaviors of subway passengers based on Kinect," *China Work Safety Science and Technology*, vol. 17, no. 12, pp. 162–168, 2021.

[8] X. Ma and J. Li, "Interactive behavior recognition based on low-rank sparse optimization," *Journal of Inner Mongolia University of Science and Technology*, vol. 40, no. 4, pp. 375–381, 2021.

[9] Z. Zhai and Y. Zhao, "DS-ConvLSTM: A lightweight video behavior recognition model for edge environment," *Journal of Communication University of China (Natural Science Edition)*, vol. 28, no. 6, pp. 17–22, 2021.

[10] C. Ying and S. Gong, "Human behavior recognition network based on improved channel attention mechanism," *Journal of Electronics and Information Technology*, vol. 43, no. 12, pp. 3538–3545, 2021.

[11] H. Chen, S. Wei, and Y. Sun, "Multi-domain feature extraction for behavior recognition using temporal segment networks," *IEEE Transactions on Multimedia*, vol. 23, pp. 1543–1557, 2021.

[12] A. Ali, M. Zhang, and X. Chen, "Cross-structure supervised learning for 3D action recognition," *Pattern Recognition*, vol. 112, pp. 107–118, 2021.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. EMNLP*, 2014, pp. 1724–1734.