Journal of Computer Science and Cybernetics, V.41, N.3 (2025), 245–264
DOI no. 10.15625/1813-9663/22770

MAMBA-MHAR: AN EFFICIENT MULTIMODAL FRAMEWORK FOR HUMAN ACTION RECOGNITION

TRUNG-HIEU LE[1,2], THAI-KHANH NGUYEN[1,2], TUAN-ANH LE[2],
MATHIEU DELALANDRE[3], TRUNG-KIEN TRAN[4], THANH-HAI TRAN[1], CUONG PHAM[5,*]

[1]*School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, 01 Dai Co Viet Street, Bach Mai Ward, Ha Noi, Viet Nam*
[2]*Dai Nam University, 01 Pho Xom, Phu Luong Ward, Ha Noi, Viet Nam*
[3]*Polytechnic University of Tours, France*
[4]*Institute of Information Technology, AMST, 17 Hoang Sam, Nghia Do Ward, Ha Noi, Vietnam*
[5]*Posts and Telecommunications Institute of Technology, Nguyen Trai Street, Mo Lao Ward, Ha Noi, Viet Nam*

Abstract. Human Action Recognition (HAR) has emerged as an active research domain in recent years with wide-ranging applications in healthcare monitoring, smart home systems, and human–robot interaction. This paper introduces a method, namely Mamba-MHAR (Mamba based Multimodal Human Action Recognition), a lightweight multimodal architecture aimed at improving HAR performance by effectively integrating data from inertial sensors and egocentric videos. Mamba-MHAR consists of double Mamba-based branches, one for visual feature extraction - VideoMamba, and the other for motion feature extraction - MAMC. Both branches are built upon recently introduced Selective State Space Models (SSMs) to optimize the computational cost, and they are lately fused for final human activity classification. Mamba-MHAR achieves significant efficiency gains in terms of GPU usage, making it highly suitable for real-time deployment on edge and mobile devices. Extensive experiments were conducted on two challenging multimodal datasets UESTC-MMEA-CL and MuWiGes, which contain synchronized IMU and video data recorded in natural settings. The proposed Mamba-MHAR achieves 98.00% accuracy on UESTC-MMEA-CL and 98.58% on MuWiGes, surpassing state-of-the-art baselines. These results demonstrate that a simple yet efficient fusion of multimodal lightweight Mamba-based models provides a promising solution for scalable and low-power applications in pervasive computing environments.

Keywords. Mamba, selective state space model, selection mechanism, HAR, multimodal fusion, visual sensor, inertial sensor.

*Corresponding author.
E-mail addresses: hieult@dainam.edu.vn (T.H.Le); khanhnt@dainam.edu.vn (T.K.Nguyen); anhlt@dainam.edu.vn (T.A.Le); mathieu.delalandre@univ-tours.fr (M.Delalandre); t2kien@gmail.com (T.K.Tran); hai.tranthithanh1@hust.edu.vn (T.T.H.Tran); cuongpv@ptit.edu.vn (C. Pham).

1. INTRODUCTION

Human Activity Recognition (HAR) is a research field that focuses on the automatic identification of human actions using data collected from various sensors [1]. Applications of HAR span across healthcare monitoring, fitness tracking, smart homes, and human-computer interaction [2–5]. With the growing availability of wearable and ambient sensing technologies, multimodal HAR (multiHAR) which combines multiple data sources such as accelerometer, gyroscope, audio, and video has emerged as a promising approach to improve recognition accuracy and robustness in real-world environments [6]. However, leveraging multiple modalities often leads to increased computational complexity and memory usage, making it challenging to deploy these models on edge devices like smartphones, smartwatches, or embedded IoT systems, which have limited processing power and storage. As a result, there is a growing need to develop lightweight and efficient multimodal HAR models that balance performance with the constraints of low-resource hardware platforms.

Several approaches have been proposed for tackling HAR by using video and motion data, which are typically extracted using deep learning models such as CNNs [7, 8], LSTMs [9], or Transformers [10, 11]. For fusion, features from different modalities are often fused either at early stages (early fusion) or at later stages (late fusion) of the network. CNNs are commonly used for spatial feature extraction from video frames or sensor signals, but they tend to be computationally expensive due to convolution operations. Transformers, on the other hand, are powerful in capturing long-range dependencies but require large-scale data to generalize effectively [12]. When combining video and motion streams, especially using cross-attention mechanisms or complex fusion strategies, these models can become heavy and resource-intensive, making them difficult to deploy on low-power hardware platforms such as mobile devices or embedded systems.

Recently, State Space Models (SSMs) have been introduced as a promising alternative for sequence modeling, offering a compelling balance between performance and efficiency [13, 14]. The core idea behind SSMs is to model sequences using linear state dynamics combined with learned input and output projections, which allows them to capture long-range dependencies with significantly lower computational cost and memory usage compared to traditional architectures like CNNs or Transformers. SSMs have demonstrated strong performance in various sequence-based tasks such as language modeling [15], time-series forecasting [16], and audio processing [17]. In the context of HAR, a few studies have begun to explore the use of SSMs, but mostly in single-modality settings (e.g., using only inertial data [18] or video data [19]), where they have shown competitive or even superior results compared to conventional models. However, their potential in multimodal HAR remains underexplored. This motivates our work to investigate the integration of SSMs into a lightweight yet effective architecture for multiHAR, aiming to take advantage of their efficiency while leveraging complementary information from multiple sensing modalities.

This paper introduces a novel multimodal recognition framework named Mamba-MHAR, which exploits the efficiency of SSMs for both RGB and IMU data streams. The proposed architecture is designed to effectively integrate features extracted from multiple sensor modalities, including visual data (video frames) and inertial signals (accelerometer and gyroscope). Mamba-MHAR consists of two Mamba blocks, one dedicated to processing visual inputs and the other tailored for handling sensor data. Each block independently learns modality-specific representations, thereafter, these learned features are integrated using a late fusion

strategy, allowing the model to preserve modality-specific advantages while leveraging complementary information across channels. In summary, the contributions of this paper are threefold.

- First, we propose to investigate two Mamba-based architectures MAMC [17] and VideoMamba [19] for the task of Human Action Recognition (HAR). While VideoMamba has been applied to video recognition, MAMC was originally designed and evaluated on radio signals. Its performance has not yet been explored for HAR tasks.
- Second, we introduce a new framework, Mamba-MHAR, that adopts a late fusion strategy to effectively combine the high-level features extracted from the two Mamba blocks, enabling the model to leverage complementary information from different modalities. To the best of our knowledge, this is the first work to investigate the use of Mamba models for multimodal human activity recognition.
- Finally, we evaluate the proposed framework on two publicly available benchmark datasets: UESTC-MMEA-CL [20] and MuWiGes [21]. Experimental results demonstrate that our method outperforms both unimodal baselines and several existing multimodal approaches in terms of recognition accuracy. The proposed Mamba-MHAR is lightweight with memory and computational requirements optimized for RAM and GPU efficiency, making it highly suitable for deployment on edge devices.

The rest of the paper is structured as follows: Present the relevant works in Section 2. We introduce our framework in Section 3. Experiments and conclusions are presented in Sections 4 and 5, respectively.

2. RELATED WORKS

This section provides a brief overview of related works in Human Activity Recognition (HAR) using multimodal sensor data, with a focus on combining motion sensors and visual sensors, as well as the use of State Space Models (SSMs). Previous studies have shown that integrating data from IMU sensors and images can improve HAR performance by capturing both temporal and spatial information. At the same time, SSM-based approaches, especially Mamba and its recent variants, have become promising solutions for modeling long sequences efficiently, while being lightweight enough for deployment on mobile and edge devices.

2.1. Multimodal fusion for HAR

Early multimodal systems in human action recognition (HAR) have shown marked improvements by integrating heterogeneous sensor modalities, such as vision, depth, and inertial data. These systems leverage the complementary nature of spatial and motion signals to significantly enhance recognition accuracy and robustness in real-world applications. For instance, Radu et al. [22] proposed deep neural network architectures with modality-specific branches and feature concatenation, employing CNN and DNN backbones to jointly infer human activity and environmental context. Yen et al. [23] demonstrated the effectiveness of combining depth and inertial modalities, showing that spatial structure from depth sensors and motion dynamics from inertial data yield richer, more discriminative features. Expanding on these insights, Imran et al. [24] developed a multi-stream deep learning framework for

fusing RGB-D and inertial sensor data, where spatial features were extracted using CNNs and temporal dependencies modeled through LSTM layers. Their results confirmed that multimodal integration provides superior performance, particularly in complex or occluded scenes. Xin Chao et al. [25] proposed a late fusion approach in which separate classifiers CNN for depth data and MLPs for each inertial sensor were trained independently, and final decisions were obtained through weighted majority voting. In a more recently, Yadav et al. [26] introduced a two-stream decision-level fusion model, where CNNs processed spatial cues from video frames while LSTMs captured temporal motion from inertial sensors. This architecture proved highly effective in dynamic and unconstrained settings. Likewise, Wei et al. [27] presented a CNN-based fusion system that combined video and inertial data at the feature level, demonstrating resilience to variations in human motion and environment.

In continuous action recognition, Dawar and Kehtarnavaz [28] applied a deep fusion strategy to jointly learn from visual and inertial streams, outperforming unimodal baselines especially in the presence of overlapping or subtle movements. Most recently, Tang et al. [29] provided a comprehensive review of wearable multimodal HAR systems, highlighting the growing adoption of hierarchical and stacked CNN-LSTM fusion architectures, optimized for both real-time inference and computational efficiency.

2.2. State space models

2.2.1. Background

State Space Models (SSMs) have recently become prominent in sequence modeling due to their scalability and efficiency, especially when handling long-range dependencies in sequential data [13]. Among these, Mamba stands out as a selective SSM that integrates time-varying parameters and a hardware-aware design, enabling efficient training and inference [30]. By leveraging selective scan operations and a recomputation mechanism, Mamba avoids storing intermediate activations in the forward pass and reconstructs them during backpropagation, leading to reduced GPU memory usage and improved computational performance.

State Space Models (SSMs) are a class of mathematical models that employ first-order differential equations to describe the temporal evolution of hidden (latent) states in dynamic systems. These models also incorporate a secondary equation to connect the hidden states with the observable outputs. Given a known input sequence $x(t) \in \mathbb{R}^D$, the system transitions through a latent state $h(t) \in \mathbb{R}^N$, from which the output sequence $y(t) \in \mathbb{R}^N$ can be derived as follows

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \tag{1}$$
$$y(t) = \mathbf{C}h(t), \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ and $\mathbf{B}, \mathbf{C} \in \mathbb{R}^{N\times D}$ are the system matrices to be learned. To model discrete sequences, the above functions can be discretized using a time step Δ

$$h(n) = \bar{\mathbf{A}}h(n-1) + \bar{\mathbf{B}}x(n), \tag{3}$$
$$y(n) = \mathbf{C}h(n), \tag{4}$$

where $\bar{\mathbf{A}} = \exp(\Delta\mathbf{A})$ and $\bar{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - I) \cdot \Delta\mathbf{B}$. Obtained transforming the continuous form $(\Delta, \mathbf{A}, \mathbf{B}, \mathbf{C})$ into the discrete form $(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \mathbf{C})$. This transformation allows the system to operate in a linear recurrent fashion, which notably improves computational efficiency during both training and inference phases [13].

To improve the capacity of modeling long-range dependencies embedded in the hidden state h with respect to the input signal x, the HiPPO matrix is utilized to augment the discrete transition matrix $\bar{\mathbf{A}}$ by the Structured SSM (S4) [14]. Concurrently, by employing approximate diagonalization techniques, the matrix $\bar{\mathbf{A}}$ can be simplified to a Normal Plus Low-Rank (NPLR) format, consisting of a limited number of normal and low-rank terms.

$$\mathbf{A}_{nk} = -\begin{cases} \sqrt{(2n+1)(2k+1)} & \text{if } n > k \\ n+1 & \text{if } n = k \\ 0 & \text{if } n < k \end{cases} \tag{5}$$

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^* - \mathbf{P}\mathbf{Q}^\top = \mathbf{V}\left(\mathbf{\Lambda} - (\mathbf{V}^*\mathbf{P})(\mathbf{V}^*\mathbf{Q})^*\right)\mathbf{V}^* \tag{6}$$

where $\mathbf{V} \in \mathbb{R}^{N\times N}$ is unitary, $\mathbf{\Lambda}$ is diagonal, and $\mathbf{P} \in \mathbb{R}^{N\times 1}$, $\mathbf{Q} \in \mathbb{R}^{N\times 1}$ are low-rank matrices.

This transformation significantly reduces the computational burden of the recurrent architecture, lowering the complexity from $O(L^2)$ to $O(L)$.

2.2.2. Selective SSM

Recent innovations in State Space Models (SSMs) have led to the development of Mamba, a selective and hardware-efficient architecture that significantly enhances both computational and memory performance for long sequence modeling tasks [30]. Built upon the foundation of structured SSMs (SSSMs), Mamba integrates three key components: (i) a gating mechanism akin to the Gated Linear Unit (GLU), (ii) residual connections, and (iii) normalization layers such as LayerNorm or RMSNorm. The gating mechanism enables input-conditioned modulation, while the residual and normalization components facilitate the training of deeper networks by ensuring stable gradient propagation.

Unlike conventional time-invariant SSMs that rely on fixed parameters $(\mathbf{A}, \mathbf{B}, \mathbf{C})$, Mamba introduces input-dependent dynamics by computing the time step Δ, and matrices $\mathbf{B}$ and $\mathbf{C}$ directly from the input. These dynamic components are then used to generate a discretized transition matrix defined as $\bar{\mathbf{A}} = \exp(\Delta\mathbf{A})$, allowing the model to adjust the influence of past versus recent inputs based on context. A large Δ emphasizes current observations, while a small Δ retains long-term dependencies.

To improve runtime and memory efficiency, Mamba adopts a hardware-aware strategy. During each forward pass, parameters Δ, $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are fetched from high-bandwidth memory (HBM) and transferred to static RAM (SRAM), where all computation is performed in-place. This design eliminates the need to store intermediate hidden states, as they can be recomputed during the backward pass. As a result, the total memory access complexity per forward step is reduced to $\mathcal{O}(BLN)$, where B is the batch size, L is the sequence length, and N is the hidden state dimension. Compared to Transformer-based architectures, which require $\mathcal{O}(BLN)$ due to attention operations, Mamba achieves a memory reduction factor of $\mathcal{O}(N)$. This efficiency makes Mamba particularly suitable for real-time inference and deployment on edge devices with constrained computational and memory resources.

2.2.3. Mamba architecture

Mamba [30] introduces input-dependent parameters including the step size Δ and transition matrices $\mathbf{B}$ and $\mathbf{C}$. These are computed as functions of the input x through learnable linear projections

$$\Delta, \mathbf{B}, \mathbf{C} = f(x).$$

The recurrence dynamics is governed by a discretized version of the SSM:

$$\bar{\mathbf{A}} = \exp(\Delta \mathbf{A}), \quad \bar{\mathbf{B}} = \text{discretize}(\Delta, \mathbf{A}, \mathbf{B}),$$

where $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ are the input-conditioned state transition and input matrices, respectively.

During the forward pass, Mamba avoids storing intermediate states by recomputing them during the backward pass. It also transfers key matrices to static RAM (SRAM) for in-place computation, significantly reducing the volume of memory transactions between SRAM and high-bandwidth memory (HBM). The total memory complexity is reduced to $\mathcal{O}(BLN)$, where B is the batch size, L is the sequence length, and N is the hidden state size.

2.2.4. Mamba in downstream tasks

Thanks to the efficient design, Mamba has inspired various domain-specific adaptations. For vision tasks, LightM-UNet combines Mamba with a lightweight UNet structure for real-time medical image segmentation [31], while 2D-SSM [32] and DenseMamba introduce variations focused on spatial modeling and shallow state fusion for better feature retention [33]. ConvSSM merges SSMs with convolutional recurrence (ConvLSTM) to allow parallel scan operations and faster training [34]. In time-series domains, TimeMachine applies Mamba to multivariate forecasting tasks with long-term temporal dependencies [35], and MotionMamba introduces hierarchical spatio-temporal blocks for human motion synthesis [36]. More recently, lightweight variants of Mamba have been proposed for human activity recognition (HAR) using visual sensor and IMU sensor data. VideoMamba extends Mamba's selective scanning along spatial and temporal axes to model video dynamics efficiently [19]. In parallel, MAMC (Mamba for Multimodal Cross Attention) introduces a low-latency pipeline for fusing 1D motion signals from IMU sensors, using a soft-thresholding denoising block and Residual Selective SSM for robust real-time processing [17]. In this study, we adopt VideoMamba [19] and MAMC [17] as the visual and motion backbones in our proposed multimodal HAR framework. Their lightweight design and complementary strengths allow accurate and efficient recognition, even on edge devices.

3. PROPOSED METHOD

3.1. Proposed framework for multimodal human action recognition

Our proposed framework that combines multimodality for recognition of action / gesture is illustrated in Fig 1. The architecture comprises three stages:

1) **Pre-processing**: We pre-process raw data before inputting it into the Mamba models.

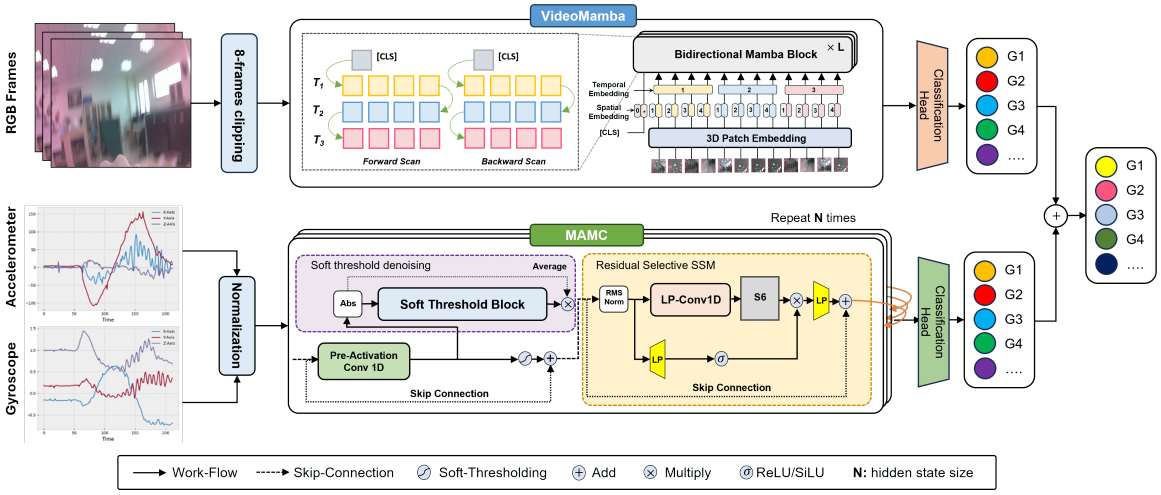

Figure 1: Our proposed lightweight multimodal framework Mamba-MHAR for human action recognition.

- For the video stream, we resize each frame to $W \times H$ and apply a random sampling technique to extract a T-frame clip for each video sample. The input video is represented as a tensor $\mathbf{X}_v \in \mathbb{R}^{3\times T\times H\times W}$.
- For the IMU stream, we smooth the signal using an average filter, resample the signal, and compute the magnitude for the accelerometer and gyroscope signals. The motion data is represented as a tensor $\mathbf{X}_m \in \mathbb{R}^{8\times P}$ where P is the length of each motion signal.

2) **Single model learning**: The pre-processed data are input separately into each Mamba model to extract features and output the corresponding classification scores, denoted as $\mathbf{s}_v$ and $\mathbf{s}_m$. VideoMamba is deployed for the video stream, while MAMC is used for the IMU stream.

3) **Multimodal fusion**: Finally, we apply late fusion of the two score vectors using a simple addition rule to generate the final score vector $\mathbf{s}$ for multimodal classification.

In the following, we will detail the Mamba models for processing each data stream and the multimodal fusion.

3.2. Visual modality learning

Numerous models leveraging 3D CNN architectures such as TSM [37], SlowFast [38], and Transformer-based methods [39] have been widely adopted for HAR using RGB data. While some models focus on improving recognition accuracy, others reduce computational overhead for deployment on resource-constrained devices. Given the demand for real-time inference on compact hardware, our work emphasizes lightweight models that maintain high performance. VideoMamba [19] emerges as a compelling solution due to its efficient memory usage, neural architecture optimization, and strong temporal modeling via ensemble techniques. As a result, we employ VideoMamba as the video modality learning in our framework.

VideoMamba is a selective State Space Model (SSM) tailored for video data. It adopts the Vision Transformer (ViT) design [40], but replaces the conventional self-attention blocks with Bidirectional Mamba (B-Mamba) blocks to retain linear computational complexity while capturing long-range dependencies effectively. Given an input video $\mathbf{X}_v$ of dimensions $3 \times T \times H \times W$, each frame is first divided into non-overlapping spatial patches of size 16×16. A 3D convolution with kernel size $1 \times 16 \times 16$ is applied to embed these patches into a sequence of tokens. The total number of patches is computed as $N_{\text{patch}} = T \times \frac{H}{16} \times \frac{W}{16}$. An additional [CLS] token, $\mathbf{X}_{\text{cls}}$, is prepended to this sequence to aggregate global information for downstream classification. To retain positional information that is not inherently captured by the SSM architecture, both spatial and temporal positional embeddings are added. The final input to the VideoMamba encoder is formulated as

$$\bar{\mathbf{X}}_{\mathbf{v}} = [\mathbf{X}_{\text{cls}}, \mathbf{X}_{\mathbf{v}}] + \mathbf{p}_s + \mathbf{p}_t, \tag{5}$$

where $\mathbf{p}_s \in \mathbb{R}^{(hw+1)\times C}$ denotes the learnable spatial position embeddings and $\mathbf{p}_t \in \mathbb{R}^{T\times C}$ denotes the temporal position embeddings. This embedding formulation ensures that the model can encode both spatial layout and temporal dynamics. The resulting token sequence is then passed through L stacked B-Mamba blocks. Each block performs bidirectional scanning along the temporal axis, both forward and backward, allowing the model to integrate contextual information from both preceding and succeeding frames. This approach is conceptually similar to bi-directional RNNs but maintains the efficiency and scalability of state space models.

We choose the VideoMamba-Tiny [19] variant due to its compact size (7M parameters) and its strong capability in temporal modeling. The Spatial-First scanning strategy is adopted, processing patches at identical spatial locations across time first, which is empirically shown to effectively capture motion continuity. Training is conducted on 8-frame video clips, each resized to a spatial resolution of 224×224 pixels, with a learning rate of 0.001, batch size of 32, and 35 epochs. Data augmentation includes random cropping, horizontal flipping, and color jittering. Dropout with a probability of 0.5 is used to prevent overfitting. The depth multiplier is set to 1.0, and training utilizes the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The loss function is categorical cross-entropy, suitable for multi-class classification. Upon completion of training, the VideoMamba-Tiny model outputs a feature vector $\mathbf{F}_{\mathbf{v}} = f(\bar{\mathbf{X}}_v, \Phi_v) \in \mathbb{R}^{192}$ with Φ_v representing the VideoMamba model parameters. This vector encapsulates the spatiotemporal characteristics of the input video, which consists of frames resized to 224×224 pixels. Through linear layers, we obtain the score vector $\mathbf{s}_v = Linear(\mathbf{F}_{\mathbf{v}})$ to classify the action.

3.3. Motion modality learning

Motion feature extraction aims to process an 8-channel motion signal, comprising tri-axial acceleration, tri-axial gyroscope, and their respective magnitudes. While traditional models such as CNNs [23, 41] and Transformers [42, 43] have been widely applied to analyze such time series data, they face notable limitations in capturing long-term dependencies inherent in sequential motion signals. CNNs are often constrained by their local receptive fields, whereas Transformers, despite their global attention mechanism, can be computationally expensive and memory-intensive when applied to long 1D sequences. Considering the requirements for real-time and resource-efficient deployment, we adopt the MAMC (Mamba-based Automatic

Modulation Classification Architecture) model as a motion feature extractor, owing to its lightweight design and effective modeling capabilities over long-range temporal structures. MAMC is a selective State Space Model (SSM) architecture tailored to extract meaningful features from long 1D radio signals while maintaining computational efficiency [17]. We adapt the model to the sensor time-series data domain, we use 8 input data channels, instead of 2 channels like the original version. The model comprises two main components: a soft-thresholding denoising block, and a residual selective SSM block.

The denoising module in the MAMC architecture refers to the Deep Residual Shrinkage Network that begins by applying a 1D convolutional layer to the multi-channel IMU time-series data $\mathbf{X}_m \in \mathbb{R}^{8\times P}$. This convolution captures short-range temporal dependencies and local patterns. The output is then passed through a batch normalization to extract localized signal features from the input time-series data. To generate a dynamic noise-suppressing threshold, the processed signal undergoes absolute value computation and global average pooling (GAP) to produce a summary representation. This summary is passed through two linear projection (LP) layers and a sigmoid activation to yield an adaptive soft-threshold τ. The soft-thresholding operation is then applied element-wise to the original input feature $\mathbf{X}_m$ to suppress low-amplitude noise while preserving informative signal components. The soft-thresholding function is formally defined as

$$y = \begin{cases} x + \tau & \text{if } x < -\tau \\ 0 & \text{if } -\tau \leq x \leq \tau \\ x - \tau & \text{if } x > \tau, \end{cases} \tag{6}$$

where $x \in \mathbf{X}_m$ is a sample in the motion signal $\mathbf{X}_m$, y denotes the corresponding output feature and τ is the learned threshold parameter. This function zeros out values near zero, effectively reducing noise, while maintaining higher-magnitude signal components. To ensure stable training, its derivative is piecewise constant, preventing gradient vanishing and explosion

$$\frac{\partial y}{\partial x} = \begin{cases} 1 & \text{if } x < -\tau \text{ or } x > \tau \\ 0 & \text{if } -\tau \leq x \leq \tau. \end{cases} \tag{7}$$

The denoised output y is subsequently passed into a residual selective SSM block, which consists of Root Mean Square (RMS) normalization, a 1D convolutional layer (to mix or expand channels), a selective SSM for modeling long-term temporal dependencies, and a final linear projection layer. Skip connections are employed to facilitate efficient gradient flow and mitigate information loss. By stacking N such MAMC blocks, the architecture forms a deep sequential model that resembles Transformers but replaces self-attention with efficient SSMs to maintain linear complexity.

Given the original motion signal $\mathbf{X}_m \in \mathbb{R}^{8\times P}$, the output after the final MAMC block $\mathbf{F_i} = f(\mathbf{X}_i, \Phi_i) \in \mathbb{R}^G$, where Φ_i is MAMC model's parameters. Through the classification head, it produces the score $\mathbf{s}_i = \mathit{Linear}(\mathbf{F}_i)$. For training, we use a batch size of 16 and a learning rate of 0.001. These hyperparameters were selected to ensure a balance between learning stability and computational efficiency during model convergence. In our implementation, to better adapt to the characteristics of motion data, we modified MAMC [17] to

create a smaller version, referred to as MAMC-s. MAMC-s differs from the original MAMC in the following architectural aspects (as shown in Table 1).

Table 1: Parameter comparison between MAMC and MAMC-s

Model	#SSM Layers	#FC layers	Input Length (samples per activity)	Parameter (M)
MAMC [17]	4	2	512–1024	16.8
MAMC-s (Ours)	1	1	256	0.125

3.4. Fusion of visual and motion features

To effectively leverage complementary information from multiple modalities, we adopt a late fusion strategy in our framework. In this approach, each modality is first processed independently by its respective model to produce a modality-specific classification score vector: $\mathbf{s}_v$ for the video stream and $\mathbf{s}_m$ for the IMU stream. These score vectors represent the confidence levels across target classes as predicted by each modality. Rather than combining features at an earlier stage (early fusion), we defer the integration to the decision level, allowing each model to specialize in learning from its own data domain. Finally, we apply weighted fusion of scores as defined below.

$$\mathbf{s} = \alpha \mathbf{s}_m + (1 - \alpha)\mathbf{s}_v, \tag{8}$$

where α is a hyperparameter. A higher value of α indicates that decision-making relies more on motion data, whereas a lower value shifts the reliance toward the video stream.

Here, $\mathbf{s} \in \mathbb{R}^N$ is an N-dimensional vector, where N is the total number of classes. Each element s_i in the vector $\mathbf{s}$ represents the aggregated confidence score for class i. The final predicted class label $\hat{y}$ is obtained by selecting the class with the highest score using the arg max operation

$$\hat{y} = \arg \max_{i \in \{1,\ldots,N\}} s_i \tag{9}$$

This decision rule enables the system to combine evidence from both modalities in a straightforward yet effective manner, enhancing robustness, especially in scenarios where one modality may be unreliable or noisy.

4. EXPERIMENTS

4.1. Datasets

In this work, we evaluate our proposed method using two publicly available benchmarks: UESTC-MMEA-CL [20] and MuWiGes [21]. Both datasets are designed to evaluate multimodal human action recognition algorithms by providing synchronized multimodal data streams. Each sample includes time-aligned RGB video frames along with inertial measurements from accelerometers and gyroscopes. Figure 2 illustrates an example from each dataset. On the left, a sequence from the UESTC-MMEA-CL dataset demonstrates the action of "cooking". On the right, a gesture sample from the MuWiGes dataset illustrates the execution of the gesture "*G*5", captured via a wrist-mounted camera.

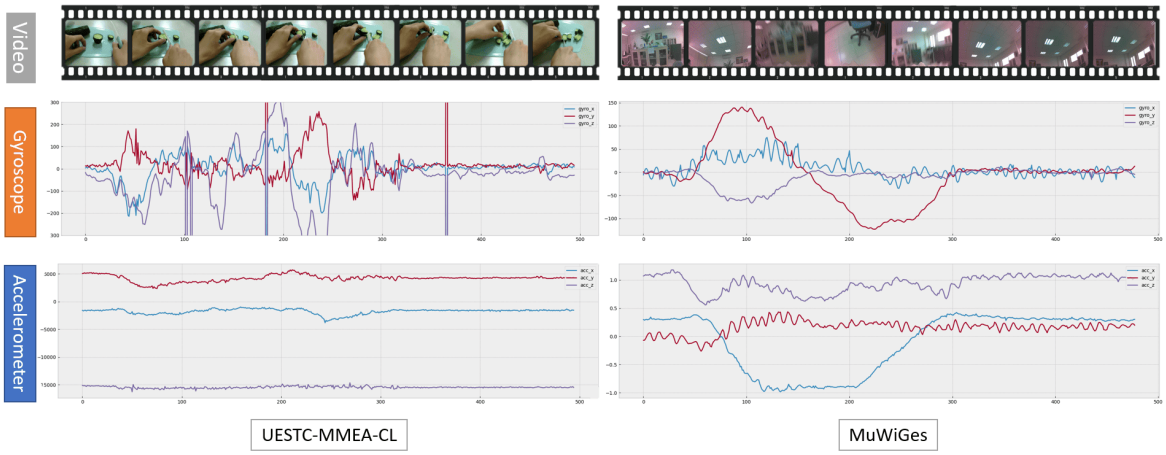

Figure 2: Representation of various modalities captured by sensors in two datasets. The first row illustrates the raw RGB frames from the datasets UESTC-MMEA-CL [20], and MuWiGes [21]. The second and third rows present the corresponding accelerometer and gyroscope signals.

- **UESTC-MMEA-CL dataset** serves as a multimodal benchmark designed for the continuous recognition of egocentric human activities. It contains recordings of 32 distinct daily actions (such as climbing stairs, drinking, shopping, playing cards, etc.) collected across various environments, including both indoor and natural settings, from 10 different participants. Data acquisition was conducted using smart glasses equipped with a front-facing camera and built-in inertial measurement unit (IMU) sensors. The visual modality was captured at a resolution of 640×480 pixels with a frame rate of 25 FPS, while the IMU recorded accelerometer and gyroscope signals at 25 Hz. Each activity class is represented by approximately 200 synchronized sequences, comprising first-person RGB video and corresponding motion sensor data.

- **MuWiGes dataset** was collected using a wrist-worn device composed of a camera and IMU sensors. The camera captures images at a resolution of 1280×720 pixels with a frame rate of 30 FPS, while the inertial module, including accelerometers and gyroscopes, samples motion data at 50 Hz. The dataset are collected by 50 subjects (33 male and 17 female), aging from 10 to 65 years, each performing 12 predefined hand gestures across diverse indoor settings such as homes and office environments. Every participant executed each gesture in a natural manner, between 2 and 12 repetitions. The collected data were temporally segmented using automatic annotation of gesture start and end points, resulting in a total of 5,408 multimodal samples combining RGB video and motion signals. For evaluation purposes, two standard protocols are supported: cross-subject and cross-scene settings. In this work, we adopt the cross-subject evaluation protocol to investigate the generalization ability of the proposed model across different individuals.

Table 2 summarizes the key characteristics of the two multimodal datasets used in this study. Although both datasets provide synchronized video and inertial sensor data, they

Table 2: Summary of two datasets used in experiments

Datasets	UESTC-MMEA-CL [20]	MuWiGes [21]
Activity type	Daily activity	Hand gesture
Camera mounting	Head	Wrist
IMU mounting	Head	Wrist
Scenario	Natural	Indoor (Home, Office)
Data modalities	RGB + Acc + Gyro	RGB + Acc + Gyro
Total subjects	10	50
Number of classes	32	12
Total samples	6522	5048
Train/test splitting	4553 / 1316	3276 / 1772

differ in the types of actions performed, sensor placement, and environmental settings. The UESTC-MMEA-CL dataset employs a head-mounted camera and an IMU sensor, enabling the capture of egocentric video sequences that include both hands and the frontal scene, offering a comprehensive field of view for daily activity recognition. In contrast, the MuWiGes dataset features a wrist-mounted device that captures both visual and inertial signals. Although this configuration is well-suited for capturing fine-grained hand gestures, it inherently limits the camera's field of view to a narrow region around the hand. This constraint introduces additional complexity to the gesture recognition task due to frequent occlusions and limited contextual information.

4.2. Experimental results

To ensure a fair comparison with prior studies, we adhere to the same data-splitting protocols previously adopted for each dataset. Table 3 and Table 4 provide a summary of performance results across the two benchmark datasets, highlighting the effectiveness of each method in utilizing visual cues, inertial data, and their late fusion. In these tables, entries marked with an asterisk (*) indicate results obtained through our own re-implementation and evaluation on the respective datasets. Other values are taken from the original papers. All training and evaluation procedures were conducted on an ASUS E900 G4 workstation equipped with a dual Intel Xeon 4210R CPU, 64GB RAM, and an NVIDIA A30 GPU with 24GB memory.

In the following experiments, we set $\alpha = 0.5$, which represents the best combination of video and IMU streams. An ablation study analyzing the impact of α on performance is presented in Subsection 4.2.3. In the following experiments, we also report results using the original MAMC for the inertial stream, as well as its combination with VideoMamba-Tiny, referred to as **Mamba-MHAR-l**.

4.2.1. Results on UESTC-MMEA-CL

Table 3 presents a comprehensive comparison of human action recognition performance across various models using different modalities namely inertial, RGB, and their fusion on the UESTC-MMEA-CL dataset. For inertial modality, MAMC-s* produces lower accuracy (71.24%) compared to GAFormer (79%). However, it offers significant computational advantages, requiring only 0.125M parameters, 0.9 GFLOPs, and 0.0021 seconds of inference

Table 3: Experimental results on the UESTC-MMEA-CL dataset. Bold values represent the best results.

Method	Modality	Accuracy (%)	Parameter (M)	Inference Time (s)	FLOPS (G)
Xu et al. (2023) [20]	Inertial	59.70	-	-	-
GAFormer (2023) [44]	Inertial	79.00	13.49	0.017	2.1
MAMC* (2024)[17]	Inertial	74.77	16.8	0.246	0.7
MAMC-s*	Inertial	71.24	0.125	0.0021	0.9
Xu et al. (2023) [20]	RGB	92.60	-	-	-
SNRO (2024) [45]	RGB	85.81	-	-	-
MoViNet (2023) [21]	RGB	96.32	8	0.033	9.8
VideoMamba-Tiny* (2024) [19]	RGB	**97.13**	**7**	**0.009**	**1.1**
Xu et al. (2023) [20]	Fusion	95.60	-	-	-
CMR-MFN (2023) [46]	Fusion	95.29	-	-	-
Finetune + PR (2023) [20]	Fusion	92.20	-	-	-
Mamba-MHAR-l (our)	Fusion	98.10	23.8	0.255	1.8
Mamba-MHAR (our best)	**Fusion**	**98.00**	**7.1**	**0.011**	**2.0**

time. In the case of the RGB modality, all evaluated models generally exhibit higher computational complexity in terms of both parameter count and FLOPs compared to their inertial-only counterpart. Among these, VideoMamba-Tiny* not only reaches the best accuracy (97.13%) but also maintains efficiency with just 7M parameters, 1.1 GFLOPs and 0.009s inference time.

Regarding the fusion of inertial and RGB data, all evaluated methods exhibit improved accuracy, highlighting the complementary strengths of the two modalities. Notably, the proposed Mamba-MHAR, which is a fusion of MAMC-s and VideoMamba-Tiny, achieves the highest overall performance, with an accuracy of 98.00%. This result surpasses traditional fusion baselines, including Xu et al. (95.6%) [20] and CMR-MFN [46] (95.29%). In addition to its superior accuracy, Mamba-MHAR maintains a favorable balance between performance and efficiency, with a modest model size of 7.1M parameters, computational cost of 2.0 GFLOPs, and a fast inference time of 0.011 seconds. It is also noticed that all activities on the UESTC-MMEA-CL dataset are well recognized by the proposed method as shown in Figure 3, confirming its robustness and effectiveness in multimodal human action recognition tasks.

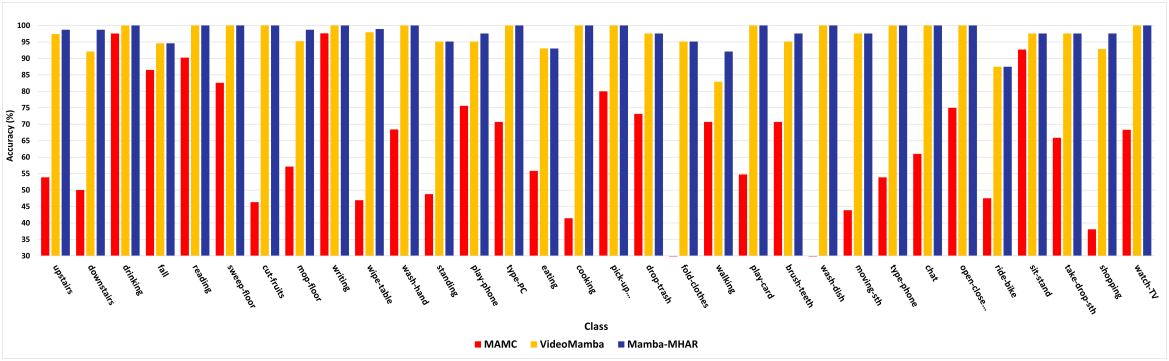

Figure 3: Comparison of the performance for each gesture on the UESTC-MMEA-CL dataset.

Table 4: Experimental results on the MuWiGes dataset. Bold values represent the best results.

Method	Modality	Accuracy (%)	Parameter (M)	Inference Time (s)	FLOPS (G)
Nguyen et al. (2023) [21]	Inertial	95.60	3.51	0.033	9.8
GAFormer* (2023) [44]	Inertial	98.33	13.49	0.017	2.1
MAMC* (2024) [17]	Inertial	96.2	16.8	0.246	0.7
MAMC-s*	Inertial	97.3	**0.125**	**0.0021**	0.9
EfficientNet3D-b0 (2023) [21]	RGB	52.94	4.72	-	0.06
MobileNet3D_v2_1.0x (2023) [21]	RGB	67.42	2.4	-	1.1
C3D (2023) [21]	RGB	70.88	63.37	-	77.3
R3D-50 (2023) [21]	RGB	88.54	46.2	-	80.1
MoViNet (2023) [21]	RGB	91.55	8	0.033	9.8
VideoMamba-Tiny* (2024) [19]	RGB	97.36	7	0.009	1.1
Mamba-MHAR-l (ours)	Fusion	98.31	23.8	0.255	1.8
Mamba-MHAR (our best)	**Fusion**	**98.58**	**7.1**	**0.011**	**2.0**

4.2.2. Results on MuWiGes

Table 4 reports the performance results on the MuWiGes dataset, which exhibit a trend consistent with those observed on the UESTC-MMEA-CL dataset. For the inertial modality, the MAMC* model provides a compelling trade-off, achieving 97.3% accuracy, which is slightly below GAFormer, while significantly reducing complexity, with only 0.125M parameters, 0.9 GFLOPs, and a minimal inference time of 0.0021 seconds. For the RGB modality, VideoMamba-Tiny* yields the highest accuracy in its group, reaching 97.36%, while preserving a lightweight architecture (7M parameters) and rapid inference (0.009 seconds). This result highlights the model's strength in balancing accuracy and computational efficiency.

Again, when fusing both modalities, the proposed Mamba-MHAR stands out as the best-performing model with an accuracy of 98.58%. Notably, Mamba-MHAR maintains an exceptional trade-off between accuracy and computational efficiency, as its performance metrics, including the number of parameters, GFLOPS, and inference time, remain consistent across different datasets. This robustness underscores the model's generalizability and deployment potential in real-world multimodal scenarios.

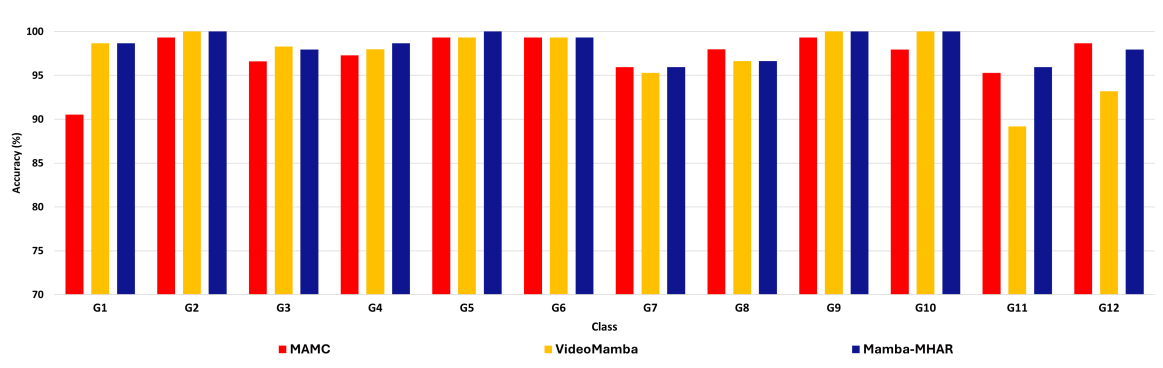

Figure 4: Comparison of the performance for each gesture on the MuWiGes dataset

Figure 4 presents the classification accuracy of the proposed Mamba-MHAR model in comparison with its unimodal counterparts, MAMC and VideoMamba, across twelve hand

gesture classes in the MuWiGes dataset. Overall, Mamba-MHAR consistently outperforms the individual models in most gesture categories (with the exception of G3, G8, and G12), highlighting the effectiveness of RGB and IMU fusion within the Mamba-based architecture.

Figure 5 presents the confusion matrices illustrating the classification performance of the proposed models on both the UESTC-MMEA-CL and MuWiGes datasets. The top row shows the confusion matrices for the 12 gesture classes in the MuWiGes dataset, while the bottom row depicts results for the 32 activity classes in UESTC-MMEA-CL. Each matrix highlights the per-class accuracy and misclassification patterns, enabling visual assessment of model strengths and weaknesses across different gesture types.

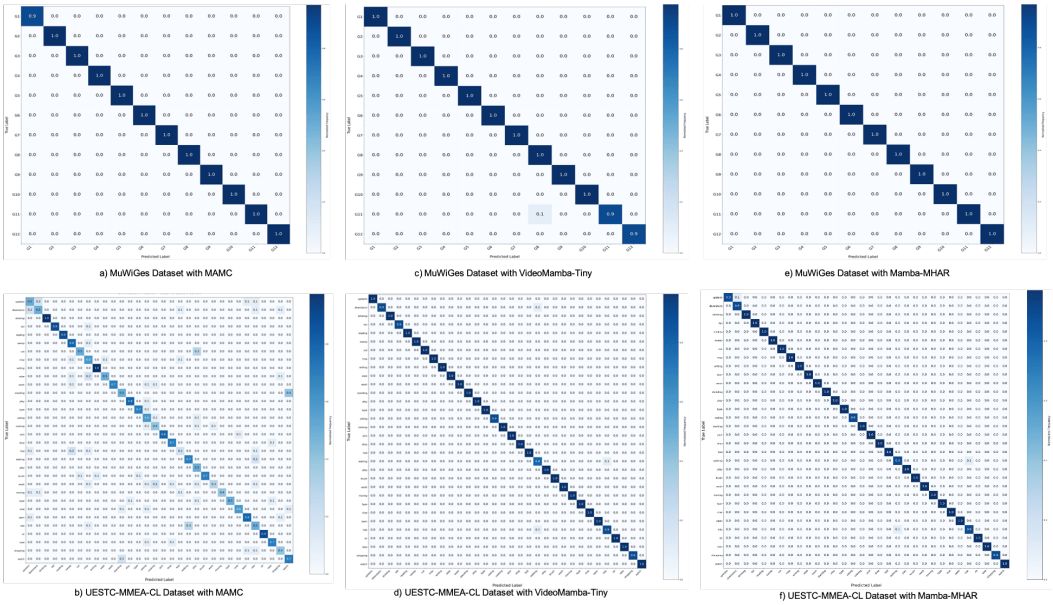

Figure 5: The confusion matrix of human action recognition on MuWiGes (first row) and UESTC-MMEA-CL (second row) datasets. Columns represent the confusion matrix by MAMC (first column), VideoMamba-Tiny (second column) and our Mamba-MHAR (last column).

4.2.3. Ablation study

In this subsection, we try to answer three questions: i) Is multimodal recognition better than unimodal recognition ?; ii) Which modality impacts more on recognition performance ? iii) Is MAMC-s is better than the original MAMC ?

Is multimodal recognition better than unimodal recognition ? In both experiments, we observed that Mamba-MHAR outperformed Video-Mamba-Tiny with an slight improvement of 1.22% on the MuWiGes dataset and 0.87% on the UESTC-MMEA-CL dataset. However, a detailed analysis of per-gesture accuracy on the UESTC-MMEA-CL dataset revealed that incorporating inertial data particularly improved the recognition of some activities such as “downstairs,” “shopping” and “walking” by 6.61%, 4.7% and 9.19% respectively. On the MuWiGes dataset, we observed an accuracy increase of 6.76% and

4.76% for gestures G11 and G12, respectively. These results suggest that many gestures or actions are better recognized when video data is combined with motion sensor input. Additionally, the motion model MAMC-s is highly lightweight, making the fusion strategy both effective and efficient in many cases.

Which modality impacts more on recognition performance ? To further assess the contribution of each modality in our fusion framework, we vary the weight α in the fusion function (Eq. (8)) from 0.1 to 0.9 in increments of 0.1. Figure 6 presents the accuracy of the proposed model on two datasets across different values of α. On the MuWiGes dataset, the accuracy remains relatively stable, with a slightly higher value at $\alpha = 0.7$, suggesting that motion data has a greater impact on performance. This is consistent with the nature of the dataset: the RGB video streams are captured from a wrist-worn camera, which primarily records background scenes and rarely captures hand movements. As a result, the visual data is less reliable for recognizing certain activities. In contrast, on the UESTC-MMEA-CL dataset, the accuracy decreases as α increases, indicating that placing greater emphasis on motion data reduces performance. Overall, $\alpha = 0.5$ yields high accuracy on both datasets and was therefore used in the experiments discussed in the previous section.

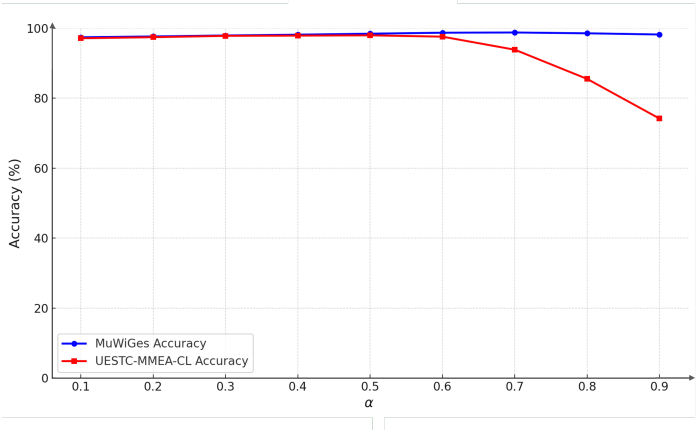

Figure 6: Evaluation of Mamba-MHAR performance with different α values

Is MAMC-s is better than the original MAMC ? As mentioned earlier, in this work, instead of using the original MAMC [17] for the inertial stream, we modified it to create a smaller version, referred to as MAMC-s. Tables 3 and 4 show that MAMC-s achieves higher accuracy with fewer parameters on the MuWiGes dataset. On the UESTC-MMEA-CL dataset, MAMC-s yields a 3.50% drop in accuracy compared to the original MAMC. However, when combining MAMC-s with VideoMamba-Tiny to form Mamba-MHAR, the model achieves competitive accuracy compared to Mamba-MHAR-l (which combines the original MAMC and VideoMamba-Tiny). Notably, Mamba-MHAR is approximately three times lighter in terms of parameter size and has significantly faster inference time than Mamba-MHAR-l. This makes it an optimal trade-off between computational cost, memory usage, and performance.

5. CONCLUSIONS

This paper presented a novel model, Mamba-MHAR, for multimodal human action recognition. Our proposed method utilized multimodal data, which passes through two independent feature extractors VideoMamba for the video stream and MAMC for the motion stream to obtain high-level representations from each. The scores outputed from each stream were then combined in late fusion. We evaluated our proposed model on two benchmark datasets (UESTC-MMEA-CL and MuWiGes), demonstrating that Mamba-MHAR achieved very competitive accuracy. With UESTC and MuWiGes datasets show superior accuracies of 98.00%, and 98.58% compared to single modality models and some existing fusion models. In addition to its strong recognition performance, Mamba-MHAR maintains a high computational efficiency, requiring only 7.1M parameters and 1.1 GFLOPs, with a low inference time of 0.011 seconds. In future work, we aim to expand this framework by exploring alternative feature extractors, adapting it for continuous activity streams, and testing its robustness across a wider range of users and diverse real-world environments.

ACKNOWLEDGMENT

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA2386-23-1-4064.

REFERENCES

[1] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019.

[2] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[3] Y. Guan and T. Plötz, "Ensembles of deep lstm learners for activity recognition using wearables," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous technologies*, vol. 1, no. 2, pp. 1–28, 2017.

[4] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What makes multi-modal learning better than single (provably)," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 944–10 956, 2021.

[5] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu, "Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 54, no. 4, pp. 1–40, 2021.

[6] F. Shafizadegan, A. R. Naghsh-Nilchi, and E. Shabaninia, "Multimodal vision-based human action recognition using deep learning: A review," *Artificial Intelligence Review*, vol. 57, no. 7, p. 178, 2024.

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[8] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE Tnternational Conference on Computer Vision*, 2015, pp. 4489–4497.

[9] H. Li, A. Shrestha, H. Heidari, J. Le Kernec, and F. Fioranelli, "Bi-lstm network for multimodal continuous human activity recognition and fall detection," *IEEE Sensors Journal*, vol. 20, no. 3, pp. 1191–1201, 2019.

[10] X. Zhou, J. Yuan, L. Fan, X. Niu, K. Zha, and X. Liu, "Msmft: Multi-stream multimodal factorised transformer for human activity recognition," *IEEE Sensors Journal*, vol. 25, no. 6, pp. 10 402–10 416, 2025.

[11] T. Agrawal, M. Balazia, P. Müller, and F. Brémond, "Multimodal vision transformers with forced attention for behavior analysis," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3392–3402.

[12] R. Pramanik, R. Sikdar, and R. Sarkar, "Transformer-based deep reverse attention network for multi-sensory human activity recognition," *Engineering Applications of Artificial Intelligence*, vol. 122, p. 106150, 2023.

[13] A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré, "Combining recurrent, convolutional, and continuous-time models with linear state space layers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 572–585, 2021.

[14] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[15] D. Y. Fu, T. Dao, K. K. Saab, A. W. Thomas, A. Rudra, and C. Ré, "Hungry hungry hippos: Towards language modeling with state space models," *arXiv preprint arXiv:2212.14052*, 2022.

[16] J. Hu, D. Lan, Z. Zhou, Q. Wen, and Y. Liang, "Time-ssm: Simplifying and unifying state space models for time series forecasting," *arXiv preprint arXiv:2405.16312*, 2024.

[17] Y. Zhang, Z. Zhou, Y. Cao, G. Li, and X. Li, "Mamc-optimal on accuracy and efficiency for automatic modulation classification with extended signal length," *IEEE Communications Letters*, vol. 28, no. 12, pp. 2864–2868, 2024.

[18] S. Li, T. Zhu, F. Duan, L. Chen, H. Ning, C. Nugent, and Y. Wan, "Harmamba: Efficient and lightweight wearable sensor human activity recognition based on bidirectional mamba," *IEEE Internet of Things Journal*, vol. 12, no. 3, pp. 2373–2384, 2024.

[19] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, and Y. Qiao, "Videomamba: State space model for efficient video understanding," in *European Conference on Computer Vision*. Springer, 2024, pp. 237–255.

[20] L. Xu, Q. Wu, L. Pan, F. Meng, H. Li, C. He, H. Wang, S. Cheng, and Y. Dai, "Towards continual egocentric activity recognition: A multi-modal egocentric activity dataset for continual learning," *IEEE Transactions on Multimedia*, vol. 26, pp. 2430–2443, 2023.

[21] H.-Q. Nguyen, T.-H. Le, T.-K. Tran, H.-N. Tran, T.-H. Tran, T.-L. Le, H. Vu, C. Pham, T. P. Nguyen, and H. T. Nguyen, "Hand gesture recognition from wrist-worn camera for human–machine interaction," *IEEE Access*, vol. 11, pp. 53 262–53 274, 2023.

[22] V. Radu, C. Tong, S. Bhattacharya, N. D. Lane, C. Mascolo, M. K. Marina, and F. Kawsar, "Multimodal deep learning for activity and context recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, pp. 1–27, 2018.

[23] C.-T. Yen, J.-X. Liao, and Y.-K. Huang, "Human daily activity recognition performed using wearable inertial sensors combined with deep learning algorithms," *IEEE Access*, vol. 8, pp. 174 105–174 114, 2020.

[24] J. Imran and B. Raman, "Evaluating fusion of rgb-d and inertial sensors for multimodal human action recognition," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 1, pp. 189–208, 2020.

[25] X. Chao, Z. Hou, and Y. Mo, "Czu-mhad: a multimodal dataset for human action recognition utilizing a depth camera and 10 wearable inertial sensors," *IEEE Sensors Journal*, vol. 22, no. 7, pp. 7034–7042, 2022.

[26] S. K. Yadav, M. Rafiqi, E. P. Gummana, K. Tiwari, H. M. Pandey, and S. A. Akbara, "A novel two stream decision level fusion of vision and inertial sensors data for automatic multimodal human activity recognition system," *arXiv preprint arXiv:2306.15765*, 2023.

[27] H. Wei, R. Jafari, and N. Kehtarnavaz, "Fusion of video and inertial sensing for deep learning–based human action recognition," *Sensors*, vol. 19, no. 17, p. 3680, 2019.

[28] N. Dawar and N. Kehtarnavaz, "Action detection and recognition in continuous action streams by deep learning-based sensing fusion," *IEEE Sensors Journal*, vol. 18, no. 23, pp. 9660–9668, 2018.

[29] J. Ni, H. Tang, S. T. Haque, Y. Yan, and A. H. Ngu, "A survey on multimodal wearable sensor-based human action recognition," *arXiv preprint arXiv:2404.15349*, 2024.

[30] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[31] W. Liao, Y. Zhu, X. Wang, C. Pan, Y. Wang, and L. Ma, "Lightm-unet: Mamba assists in lightweight unet for medical image segmentation," *arXiv preprint arXiv:2403.05246*, 2024.

[32] E. Baron, I. Zimerman, and L. Wolf, "2-D SSM: A general spatial layer for visual transformers," *arXiv preprint arXiv:2306.06635*, 2023.

[33] W. He, K. Han, Y. Tang, C. Wang, Y. Yang, T. Guo, and Y. Wang, "Densemamba: State space models with dense hidden connection for efficient large language models," *arXiv preprint arXiv:2403.00818*, 2024.

[34] J. Smith, S. De Mello, J. Kautz, S. Linderman, and W. Byeon, "Convolutional state space models for long-range spatiotemporal modeling," *Advances in Neural Information Processing Systems*, vol. 36, pp. 80 690–80 729, 2023.

[35] M. A. Ahamed and Q. Cheng, "Timemachine: A time series is worth 4 mambas for long-term forecasting," in *ECAI 2024: 27th European Conference on Artificial Intelligence*, vol. 392, 2024, p. 1688.

[36] Z. Zhang, A. Liu, I. Reid, R. Hartley, B. Zhuang, and H. Tang, "Motion mamba: Efficient and long sequence motion generation," in *ECCV*. Springer, 2024, pp. 265–282.

[37] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7083–7093.

[38] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF ICCV*, 2019, pp. 6202–6211.

[39] J. Yang, X. Dong, L. Liu, C. Zhang, J. Shen, and D. Yu, "Recurring the transformer for video action recognition," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2022, pp. 14 063–14 073.

[40] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF ICCV*, 2021, pp. 6836–6846.

[41] M. H. Rahmani, R. Berkvens, and M. Weyn, "Chest-worn inertial sensors: A survey of applications and methods," *Sensors*, vol. 21, no. 8, p. 2875, 2021.

[42] T.-H. Le, T.-H. Tran, and C. Pham, "Human action recognition from inertial sensors with transformer," in *2022 International Conference on MAPR*. IEEE, 2022, pp. 1–6.

[43] Y. Shavit and I. Klein, "Boosting inertial-based human activity recognition with transformers," *IEEE Access*, vol. 9, pp. 53 540–53 547, 2021.

[44] T.-H. Le, T.-K. Nguyen, T.-K. Tran, T.-H. Tran, and C. Pham, "Gaformer: Wearable imu-based human activity recognition with gramian angular field and transformer," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023, pp. 297–303.

[45] J. Jiao, Y. Dai, H. Mei, H. Qiu, C. Gong, S. Tang, X. Hao, and H. Li, "Slightly shift new classes to remember old classes for video class-incremental learning," *arXiv preprint arXiv:2404.00901*, 2024.

[46] H. Wang, S. Zhou, Q. Wu, H. Li, F. Meng, L. Xu, and H. Qiu, "Confusion mixup regularized multimodal fusion network for continual egocentric activity recognition," in *Proceedings of the IEEE/CVF ICCV*, 2023, pp. 3560–3569.

Received on April 23, 2025
Accepted on July 01, 2025