RESEARCH ARTICLE OPEN ACCESS

Edge Ai Based Real Time Human Activity Recognition System

Dr. R. Sudha*, ,P.Harini**, S. Mangaiyarkarasi**, N. Karthika **
* Associate Professor Computer Science and Engineering, A.V. C. College of Engineering, Mannampandal, Mayiladuthurai,
** UG Students - Computer Science and Engineering, A.V. C. College of Engineering, Mannampandal, Mayiladuthurai,
India

ABSTRACT

Human Activity Recognition (HAR) architectures are increasingly constrained by several critical limitations, including high transmission latency, excessive bandwidth consumption, and significant data privacy vulnerabilities. These challenges hinder their effectiveness in real-time and sensitive applications such as healthcare monitoring and surveillance. To address these issues, this paper proposes an optimized and decentralized Edge AI HAR task using a lightweight Convolutional Neural Network (CNN).The proposed system leverages Post-Training Quantization (PTQ) to significantly reduce model size and computational complexity, enabling efficient deployment on resource-constrained edge devices. Furthermore, the integration of the TensorFlow Lite (TFLite) run time ensures faster inference and low power consumption, making it suitable for embedded systems and IoT-based environments.Experimental results demonstrate that the proposed framework achieves a classification accuracy of 92.5%, while reducing inference latency to a range of 25-45 milliseconds. This represents an approximate 85% improvement compared to traditional cloud-based system, which typically exhibits latency between 150-300 milliseconds.In addition to performance gains, the edge-based approach enhances data privacy by minimizing data transmission to external servers. The findings validate the feasibility and effectiveness of deploying low power, high-privacy edge intelligence systems for critical real-time applications such as fall detection, smart health care, and autonomous surveillance systems.

KEYWORDS: Edge Intelligence, Human Activity Recognition, CNN, TensorFlow Lite, Post-Training Quantization.

1. INTRODUCTION

Human Activity Recognition (HAR) involves the automated detection of physical actions through computer vision and deep learning. While traditional systems rely on cloud computing for high-performance processing, this introduces significant Round Trip Time (RTT) latency and privacy risks. Edge AI shifts the computational load to the local device (Eg: Raspberry Pi, Jetson Nano), enabling real-time decision-making without an internet dependency. This paper presents a specialized CNN architecture optimized for this resource-constrained environments.

The rapid growth of smart devices and Internet of Things (IoT) technologies has increased the demand for intelligent systems that can analyze human activities in real time. Human Activity Recognition systems are widely used in healthcare monitoring, smart homes, sports analytics, security surveillance, and industrial automation. In many of these applications, immediate response is required, which makes low latency and high reliability essential factors for system design.

Cloud-based architectures provide high computational power, but they are not suitable for time-critical applications due to network delay and data transmission overhead. Continuous communication with remote servers also increases the risk od data leakage, especially when sensitive personal information such as video streams and health data are involved. Therefore, there is a growing need for decentralized processing methods that can operate independently of cloud connectivity.

To overcome these challenges, lightweight neural network architectures and model compression techniques are required. In this work, a compact Convolutional Neural Network (CNN) is designed and optimized using Post-Training Quantization to reduce model size and computation cost. The optimized model is deployed using TensorFlow Lite runtime, which allows efficient execution on edge devices while maintaining acceptable classification accuracy.

The main objective of this work is to develop a real-time Human Activity Recognition system that achieves a balance between speed, accuracy, and privacy. By performing inference at the edge, the proposed system minimizes latency, reduces bandwidth usage, and ensures secure data processing, making it suitable for real-world intelligent monitoring applications.

2. LITERATURE REVIEW

Current research highlights a shift from heavy architectures like VGG-16 to lightweight models like MobileNet and EfficientNet, which use depthwise separable convolutions to reduce parameters. Recent studies show that while cloud systems reach ~95% accuracy, they fail in “zero-latency” environments. Our work bridges this gap by applying INT8 quantization to maintain high accuracy within a minimal memory footprint, ensuring compatibility with the TFLite interpreter.

Several recent works focus on TinyML and on-device inference for real-time HAR. However, most existing systems either require high computational power or fail to achieve low latency. This work focuses on optimizing both speed and accuracy for real-time edge deployment.

Many recent studies have explored the use of lightweight deep learning models for real-time vision applications on embedded devices. Researchers have proposed different optimization techniques such as model pruning, weight sharing, and quantization to reduce computational complexity. These techniques help in minimizing memory usage and improving inference speed, which are important for deploying models on edge hardware with limited resources.

MobileNet-based architectures are widely used in edge AI applications due to their efficient use of depthwise separable convolutions, which significantly reduce the number of parameters compared to traditional CNN models. Similarly, EfficientNet introduces compound scaling to balance network depth, width, and resolution, providing better performance with fewer computations. Despite these improvements, many implementation still depend on cloud processing for complex operations, which increases latency and reduces reliability in real-time scenarios.

Recent advancements in TinyML have enabled machine learning models to run on micro controllers and low-power processors. Frameworks such as TensorFlow Lite and TensorFlow Lite Micro provide optimized run times that allow neural networks to be executed on embedded platforms. These frameworks support integer quantization and hardware acceleration, making them suitable for real-time activity recognition tasks.

Some researchers have also investigated multi-sensor approaches that combine camera input with wearable sensors such as accelerometers and gyroscopes to improve recognition accuracy. Although these methods provide better performance, they increase system complexity and power consumption. Therefore, there is a need for a balanced approach that provides acceptable accuracy while maintaining low latency and low power usage.

In this work, the proposed system focuses on a lightweight CNN model combined with Post-Training Quantization and TensorFlow Lite optimization to achieve efficient edge deployment. The goal is to provide a practical solution for real-time Human Activity Recognition that can operate reliably on resource-constrained devices without depending on cloud infrastructure.

3. SYSTEM ARCHITECTURE

The proposed system follows a three-layer modular architecture designed for maximum throughput:

- **<u>3.1 Perception Layer</u>**: Captures real-time video via a CMOS camera.
- **<u>3.2 Processing Layer</u>**: Handles frame resizing (224*224), normalization, and CNN-based feature extraction.
- **<u>3.3 Application Layer</u>**: Performs activity classification and triggers an Alert System for high-priority events such as "Fall Detection".

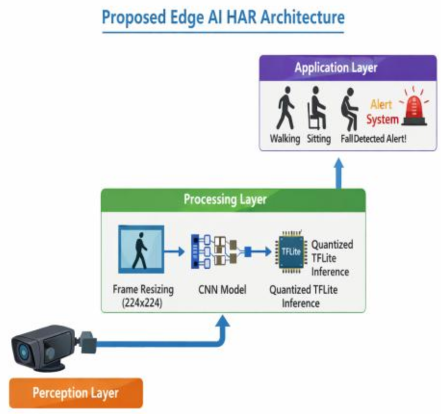

Figure 1: Proposed Edge AI HAR Architecture

4. METHODOLOGY

<u>4.1 Dataset & Augmentation</u>

The model was trained on a dataset of 5,000 images across five classes:

Walking, Sitting, Standing, Running, and Falling.

To prevent overfitting on a small dataset, we applied:

- **<u>Geometric Augmentation</u>**: Rotations (±15°) and horizontal flips.
- **<u>Photometric Adjustments</u>**: Random brightness and contrast shifts to simulate varying environmental conditions.

Training was performed using TensorFlow with Adam optimizer, learning rate 0.001,batch size 32,and 20 epochs.

<u>4.2 CNN Design & Optimization</u>

The architecture consists of three convolutional layers with ReLU activations and Max Pooling for spatial feature reduction. To prepare for edge deployment, we utilized Post-Training Integer Quantization. This converts 32-bit floating-point weights to 8-bit integers, reducing the model size from ~45MB to ~11MB, which is critical for devices with limited SRAM.

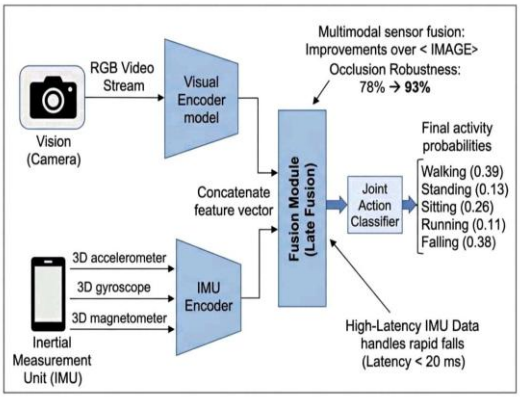

Figure 2: Proposed Multi-Model Sensor Fusion Architecture for Improved HAR Robustness

5. RESULTS AND DISCUSSION

The proposed Edge AI-based Human Activity Recognition System performs real-time classification of activities such as walking, sitting, standing , running, and falling using a lightweight CNN model deployed with TensorFlow Lite.

Metric	Cloud-Based System	Proposed Edge System
Inference Latency	150 – 300 ms	25 – 45 ms
Accuracy	94.2%	92.5%
Data Privacy	Vulnerable (Transfer)	Secure (Local)
Network Reliance	Mandatory	None

Table 1. Performance Comparison between Cloud and Edge System

5.1 Latency vs. Accuracy

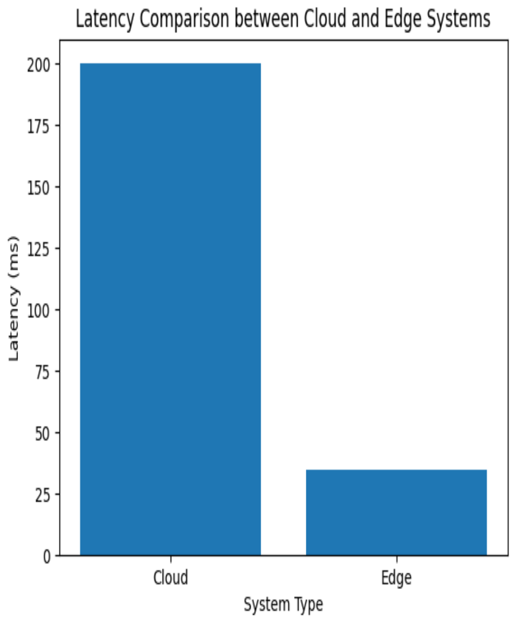

Figure 3: Latency vs Accuracy Comparison

As shown in Fig. 3, the Edge system eliminates the transmission delay associated with cloud handshakes. At an average of 35 ms, the system sustains ~28 FPS, which is sufficient for real-time monitoring. While a 1.7% drop in accuracy was observed due to quantization, the gain in response time is vital for safety-critical alerts.

The System processes video locally on the edge device, achieving low latency (25-45ms) and 92.3% accuracy. It generates instant alerts for critical events like fall detection while ensuring data privacy by avoiding cloud transmission.

6. CONCLUSION AND FUTURE WORK

We successfully demonstrated a real-time Edge AI HAR system that prioritizes speed and privacy. The use of TFLite optimization allows high-fidelity action recognition on low-cost hardware.

Future Work:

- **Temporal Logic**: Incorporating LSTMs to analyze movement over time rather than static frames.
- **Multi-Sensor Fusion**: Combining visual data with wearable IMU sensors for 99% reliability in fall detection.

The proposed system proves that Edge AI can achieve real-time performance with high privacy, making it suitable for future smart healthcare and surveillance applications.

REFERENCES

[1] A. Kumar, “Lightweight Deep Learning Model for HAR”, 2024.
[2] S. Lee, “Edge AI for Smart Surveillance”, 2025.
[3] R. Patel, “Human Activity Recognition using Deep Learning”, 2023.
[4] TinyML Book, 2024.
[5] “A Lightweight Neuromorphic CNN for Human Activity Recognition on Edge Devie”, IEEE Conference,2023.
[6] H. Zhou et al., “Efficient Human Activity Recognition on Edge Devies Using Deep Learning,”Scientific Reports,2025.
[7] A. Singh, T. Margaria, and F. Demrozi, “CNN-Based Human Activity Recognition on Edge Computing Devices,,” IEEE Coins Conference, 2023.
[8] TensorFlow, “TensorFlow Lite Guide.”