Embedded AI-Driven Multimodal Human Activity Recognition Using RGB–D Sensor Data and Cross-Modality Attention

Journal:	*IEEE Sensors Journal*
Manuscript ID	Sensors-92002-2025
Manuscript Type:	Regular Paper
Date Submitted by the Author:	02-Jun-2025
Complete List of Authors:	Biswas, Sougatamoy; National Institute of Technology Rourkela Department of Computer Science and Engineering Biswas, Sourav; National Institute of Technology Rourkela Department of Computer Science and Engineering Biswas, Anupriya; National Institute of Technology Rourkela Department of Computer Science and Engineering Nandy, Anup; National Institute of Technology Rourkela Department of Computer Science and Engineering Naskar, Asim Kumar; National Institute of Technology Rourkela
EDICS:	SYST

Embedded AI-Driven Multimodal Human Activity Recognition Using RGB–D Sensor Data and Cross-Modality Attention

Sougatamoy Biswas, Sourav Biswas, Anupriya Biswas, Anup Nandy, *Senior Member, IEEE,* Asim Kumar Naskar, *Member, IEEE,*

Abstract—**Human Activity Recognition (HAR) plays an important role in developing intelligent systems for applications such as assisted living, surveillance, and human-robot interaction. Traditional HAR approaches often rely on a single data modality, such as 2D RGB images or 3D point clouds. While 2D images provide detailed visual and motion cues, they lack depth information. In contrast, 3D data offer rich spatial and depth information but are often low in resolution and lack detailed texture information. These limitations make it challenging to achieve robust activity recognition using a single data modality. To address these challenges, we propose a sensor-driven multimodal HAR framework that integrates RGB and depth data from both 2D and 3D image streams. The system employs dual neural branches for modality-specific feature extraction and introduces a novel Cross-Modality Scaled Dot-Product Attention (CMSDPA) module to adaptively fuse visual and spatial features from 2D and 3D images. The Intel RealSense Multimodal Activity Dataset (IRMAD) used in this study is acquired using the Intel RealSense Depth Camera D435 sensor, which captures synchronized RGB and depth data. The framework is deployed on the Jetson Orin Nano edge AI device and achieves a recognition accuracy of 98.15%. Experimental results demonstrate that our model effectively leverages multimodal sensor data, enabling accurate, robust, and deployable HAR in real-world environments.**

Index Terms—**Human Activity Recognition, Multimodal Sensing, Edge AI Computing, Human-robot interaction, Convolutional Neural Networks**

I. Introduction

Human Activity Recognition (HAR) has become increasingly important in applications such as smart surveillance, healthcare monitoring, and human–robot interaction [1]. These systems require accurate and robust interpretation of human motion across diverse conditions, including variable lighting, partial occlusion, and dynamic backgrounds. Traditionally, HAR methods rely on 2D RGB image sequences and convolutional neural networks (CNNs) to learn motion and appearance features. However, the absence of depth information in RGB data limits their generalization in complex real-world environments. Specifically, RGB-based systems often struggle under poor illumination or when visual appearance alone is insufficient for reliable activity discrimination. Consequently, there is a growing need for lightweight HAR models that efficiently integrate multimodal data while being suitable for deployment on resource-constrained edge AI platforms.

Recent studies have incorporated 3D data such as point clouds [2] and depth maps, which provide structural and spatial context to complement 2D visual information. While the fusion of RGB and depth modalities has shown promise, many existing multimodal HAR frameworks employ conventional fusion techniques like direct concatenation or static weighting [3], [4]. These approaches fail to capture the dynamic contextual dependencies between modalities and frequently encounter poor alignment between texture-rich RGB features and low-density geometric depth data. Furthermore, many state-of-the-art methods rely on graph-based models or multi-branch networks that are computationally expensive [5], making them impractical for real-time deployment on embedded systems.

To address the limitations of existing methods, we propose a multimodal HAR framework that integrates RGB and depth information from both 2D and 3D streams using a novel Cross-Modality Scaled Dot-Product Attention (CMSDPA) mechanism. The proposed architecture, illustrated in Fig. 1, comprises two parallel branches: a 2D CNN that processes RGB and depth frames to extract fine-grained appearance features, and a 3D CNN that operates on voxelized representations to learn spatiotemporal patterns. The proposed CMSDPA mechanism enables effective cross-modal fusion by allowing 2D features to selectively emphasize and incorporate the most informative 3D feature representations. Additionally, a trainable scaling parameter dynamically modulates the strength of interaction, allowing the model to adapt fusion behavior based on activity context. This approach enhances recognition accuracy while maintaining a low computational overhead, enabling efficient deployment on edge devices.

Sensor selection plays a critical role in designing efficient and deployable HAR systems [6]. In this work, we utilize the Intel RealSense D435, an RGB–depth sensor capable of capturing synchronized color and depth streams in real time [7]. Its active stereo design and onboard depth processing ensure consistent performance across diverse lighting conditions and complex environments. The sensor's capability to produce both high-resolution RGB images and reliable 3D point clouds from a single device eliminates the need for multi-sensor calibration. To support real-time inference and portability, the system is integrated with the NVIDIA Jetson Orin Nano, an edge AI

Sougatamoy Biswas, Sourav Biswas, Anupriya Biswas and Anup Nandy are with the Department of Computer Science and Engineering, National Institute of Technology Rourkela, Odisha, India (e-mail: 521cs6015@nitrkl.ac.in, 224cs1016@nitrkl.ac.in, 224cs1001@nitrkl.ac.in, nandya@nitrkl.ac.in).

Asim Kumar Naskar is with the Department of Electrical Engineering, National Institute of Technology Rourkela, Odisha, India (e-mail: naskara@nitrkl.ac.in).