IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 February 2024, accepted 9 March 2024, date of publication 18 March 2024, date of current version 26 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3378515

RESEARCH ARTICLE

HADE: Exploiting Human Action Recognition Through Fine-Tuned Deep Learning Methods

MISHA KARIM[1], SHAH KHALID[1], ALIYA ALERYANI[2], NASSER TAIRAN[2], ZAFAR ALI[3], AND FARMAN ALI[4]

[1]School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad 44000, Pakistan
[2]Department of Computer Science, College of Computer Science, King Khalid University, Abha 62529, Saudi Arabia
[3]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[4]Department of Applied AI, School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Seoul 03063, South Korea

Corresponding author: Shah Khalid (shah.khalid@seecs.edu.pk)

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through Small Groups RGP.1/369/44.

ABSTRACT Human Action Recognition (HAR) is a vital area of computer vision with diverse applications in security, healthcare, and human-computer interaction. Addressing the challenges of HAR, particularly in dynamic and complex environments, is essential to advancing this field. The strength of the HADE framework is its carefully curated dataset, which was primarily derived from smartphone camera footage. This dataset encompasses a wide range of human actions captured in various settings, providing a robust foundation for training our novel HAR models, HADE I and HADE II. These models have been specifically designed and optimized for parallel processing on GPUs, showing significant improvements in the efficiency of both training and inference processes. Through a comprehensive evaluation, the HADE framework demonstrated a remarkable improvement in HAR accuracy, achieving 83.57% accuracy on our custom dataset. This marks a considerable enhancement over existing methodologies and underscores the efficacy of the HADE approach in accurately interpreting complex human actions. The framework's potential applicability in healthcare in the domain of neurological patient care is particularly noteworthy, where it can aid in early detection and facilitate personalized treatment plans. Future research should focus on expanding the range of actions covered by HAR and exploring avenues for real-time processing. The introduction of the HADE framework not only makes a substantial contribution to the field of computer vision but also paves the way for its practical application across various sectors.

INDEX TERMS Human action recognition, computer vision, machine learning, SlowFast, I3D ResNet50, real-time action recognition.

I. INTRODUCTION

The distinction between human action and activity recognition is fundamental in various domains, such as video surveillance and sports analysis, involving the categorization and identification of actions in videos and images [1]. Although the terms 'action recognition' and 'activity recognition' might be used interchangeably in some contexts, the difference in terminology does not significantly alter the underlying principle [2]. Action recognition systems are designed to identify discrete actions, such as 'sitting,' whereas activity recognition systems encompass broader sequences of actions, like 'sitting down,' enhancing the understanding of complex behaviors. This differentiation is pivotal for a myriad of applications. Human Action Recognition (HAR), a critical component of Computer Vision (CV) [3] and Machine Learning (ML) finds its utility in areas like security monitoring, healthcare, human-computer

The associate editor coordinating the review of this manuscript and approving it for publication was Turgay Celik.

interaction, and sports analysis. Benchmark datasets, including the widely referenced KTH dataset [4] serve as a standardized metric for evaluating HAR algorithms. It's noteworthy that the delineation between actions and activities can blur, contingent on the context; for instance, clapping during a sports event or walking with a purpose like commuting to work or leisurely strolling in a park can be considered activities [5], [6].

Efficient action recognition is crucial for improving efficiency, security, and decision-making across various fields such as image processing [7], sign language recognition [8], artificial intelligence [9], and human-computer interaction [10], facilitating intelligent systems in making informed decisions and reacting aptly. Nonetheless, in specific areas like in-home nursing, elder care, and anomaly detection [11], [12], the complex nature of human motion introduces significant challenges. These obstacles encompass dealing with occlusions, managing variations in noise levels during data collection, and fulfilling the demands for real-time processing. Such challenges underscore the pressing need for innovative HAR strategies capable of adeptly handling these complexities while ensuring high levels of accuracy and efficiency.

Real-time HAR technologies have begun to play a pivotal role in adaptive sectors such as security and healthcare, facilitating the capability for anticipatory decision-making and swift responses [13], [14]. These real-time systems are adept at identifying actions within various contexts, offering critical insights to decision-makers [15]. Their proficiency in navigating diverse situations enhances the practical utility of HAR systems in real-world applications. This is particularly valuable in the security and healthcare fields [16], where the demand for immediate action is high, such as in secure monitoring systems aimed at thwarting potential security threats [17], [18], [19], [20], [21].

Recent studies in HAR have increasingly focused on ML and deep learning techniques, including well-regarded models such as two-stream networks and Convolutional Neural Network (CNN)s [22], [23], [24]. Applying CNNs at the frame level has demonstrated higher accuracy over traditional manual feature extraction methods. Moreover, processing sets of frames simultaneously has further enhanced this effectiveness. These technological advancements have notably increased the precision of HAR models, marking significant progress in the field.

In recent years, the domain of image and video analysis in the Internet of Things (IoT) has witnessed significant advancements, particularly with the integration of neuroheuristic approaches. Such methodologies have been effectively utilized in diverse applications, from smart home energy-management systems [25] to automated guided vehicle navigation [26]. Our study, centered on the Human Action in Diverse Environments (HADE) framework, contributes to this evolving landscape by enhancing Convolutional Neural Networks (CNNs) for HAR, a field in which neuro-heuristic methods have shown promising results, especially in surveillance video analysis [25], [27]. The innovative approach of the HADE framework is positioned to address some of the critical challenges faced by contemporary video analysis systems.

Recent advancements in the fields of deep learning and artificial intelligence (AI) have led to notable breakthroughs in HAR. These advancements are characterized by the integration of diverse techniques and the adoption of specific design strategies, marking significant progress in the domain. For example, cascaded design strategies [28], hybrid approaches [29], and the integration of HAR into smart living support [30] showcase developments have shown enhancements in action recognition systems, resulting in better accuracy and understanding of context. Recent work in this area focuses on explorations in complex hierarchical feature reduction models, as earlier noted in a study by Serpush et al. [31]. A thorough examination of vision-based HAR by Jegham et al. [32] provides a comprehensive analysis of vision-based HAR. In addition, the progress of advanced three-dimensional CNN-based models has led to a shift towards neural architectures with increased resilience and adaptability [33]. Our objective is to broaden the practical application of HAR systems and to improve their capacity for accurately detecting human behavior. We are committed to refining existing technologies through the introduction of an innovative framework designed to address the core challenges within HAR. This effort highlights the use of cutting-edge technologies to develop a system that is both more precise and adaptable, capable of functioning efficiently in a variety of settings.

Our approach considers diversity and consistency using state-of-the-art feature extraction techniques and machine learning algorithms. It further enhanced our robustness in results, as it reduced the blurring effect caused by occlusion and changes in lighting. In this research, we introduce a highly efficient unsupervised learning strategy for sequences, employing advanced feature extraction techniques and machine learning algorithms. This breakthrough significantly enhances the accuracy of detected human action systems, effectively addressing the difficulties posed by subject occlusion and fluctuating light conditions in varied settings. The HADE I and HADE II convolutional neural network (CNN) models accurately categorize activities using motion data from the HADE dataset. Based on our modeling results, the SlowFast model demonstrated greater performance compared to the HADE II model. This improvement greatly enhances the effectiveness of human action recognition systems in numerous applications and improves the accuracy of action recognition in HAR systems. The HADE dataset is crucial as it supplies the essential motion data for the HADE I and HADE II CNN algorithms to precisely detect activities. Moreover, the study revealed that the SlowFast model surpassed the HADE II model, leading to a significant enhancement in the performance of action detection in

HAR systems. These discoveries significantly improve the accuracy and dependability of HAR systems. Our goal is to expand the limits of human action recognition by utilizing breakthroughs in the field and offering new viewpoints. The objective of our investigation was to give crucial highlights that will contribute to the advancement of more precise and dependable systems.

- We present the novel HADE dataset, capturing essential motion information and enabling precise action categorization using the HADE I and HADE II CNN models.
- Our approach employs State-of-the-Art (SOA) feature extraction techniques and advanced machine learning algorithms to effectively mitigate challenges arising from occlusion, shifting lighting conditions, and background noise.
- We also provide a comprehensive benchmark evaluation that demonstrates the performance of our model, consistently achieving SOA outcomes.

The remainder of this paper is organized as follows. Section II presents a thorough review of SOA HAR and emphasizes the limitations of their contributions. Section III introduces the curated and comprehensive HADE dataset and explains its configuration in the HADE I and HADE II models to capture spatial and temporal information. Section IV provides details of the dataset, GPU-based computing resources, training configuration, and the evaluation metrics used in the experiments. In Section V, we present a comprehensive account of the experimental setup including the hardware and software configurations employed, which are crucial for evaluating the results of this study. Finally, Section VI discusses our research and proposes directions for future research.

Considering the significant advancements in HAR, this research seeks to address the following research questions:

1) What are the differences between the HADE I and HADE II models in terms of accuracy, precision, and computational efficiency when applied to various HAR scenarios?
2) How do advanced preprocessing approaches and GPU parallel processing affect the speed and scalability of HAR systems when processing various action datasets?
3) How can the HADE technique be enhanced for specific real-world applications, especially in the healthcare sector while maintaining a balance between accuracy and real-time processing capabilities?

II. RELATED WORK

HAR encompasses various learning methods, including supervised and unsupervised approaches, to achieve the accurate recognition and classification of human movements. In this section, we provide an overview of the relevant methodologies employed in HAR, focusing on the models and their information factors, and highlight the crucial role of benchmarking techniques in optimizing system performance.

Recent advancements in HAR have been characterized by diverse and innovative approaches. The work presented by Chen et al. [34] demonstrates a hybrid approach that combines multiple techniques to enhance recognition accuracy. Similarly, [28] explored a cascaded design strategy that offered a nuanced method for action recognition. Within the realm of smart living, Diraco et al. [30] offered insights into the utilization of HAR in smart services and applications, emphasizing the incorporation of HAR in everyday technologies. A hierarchical feature reduction model is introduced by Serpush et al. [31] to identify complex human behaviors. A comprehensive study of vision-based hyperparameters for human activity recognition (HAR) and their evolution was recently published by Jegham et al. [32]. Wang et al. [33]'s study on 3D CNN-based models emphasizes how important it is to use cutting-edge neural network designs to boost the effectiveness of Human Activity Recognition (HAR) systems. The works highlight the dynamic and complex character of the most recent HAR research, illustrating the field's continuous progress toward advanced, accurate, and flexible action recognition systems.

Using labeled data during the training phase has shown promising outcomes in Human Activity Recognition with supervised learning methods. For example, in a study by Liu et al. [35], they implemented a supervised learning-based APSR framework and achieved an accuracy of 86.5% on a specific dataset. Creating a framework that includes constructing a network to generate features for each body part and performing weighted pooling based on relevance scores in the word-embedding space. The utilization of labeled data during training guides the learning process, and the pooling result is employed for one-shot recognition. The UT-Kinect dataset [36] consists of RGB and depth videos, and it has been extensively utilized within HAR research. Supervised learning techniques have delivered an accuracy of 94.8% in the recognition of ADL, through the UT-Kinect dataset. These results suggest that large labeled training sets and supervised learning can significantly improve the performance of HAR applications.

During the past decade, HAR has traditionally been performed using supervised learning techniques that work through labeled activity data to learn patterns of interest (e.g., statistics, structures, and features). However, a growing body of recent work focuses on unsupervised learning, including various approaches that autonomously discover insights and patterns through activity data using methods like clustering, dimensionality reduction, and generative modeling. Such methods enable HAR systems to discover patterns and insights without guidance (e.g., labels) and thereby enable their operation across diverse dynamic and open-ended real-world scenarios. For instance, the demonstrated recognition system is easily adapted to new sensor platforms and activities and was never trained nor tested with activities in the presence of obstacles or taken from a person skiing. Consequently, this recent trend in unsupervised HAR research holds great promise for the development of general and adaptive context-aware recognition systems. Notably, PCANet-TOP [37], SCAR [38], and TSN [39]

have demonstrated promising results on benchmark datasets, attaining accuracies of up to 83.9%, 92.9%, and 82.8%, respectively. These findings contribute to the development of more robust HAR systems. However, some researchers have argued that unsupervised learning approaches may be inherently less accurate than supervised methods that rely on labeled data [40]. Although unsupervised techniques may not achieve the same level of accuracy, they provide valuable tools for exploratory analysis, data comprehension, and feature extraction. Despite their potentially lower accuracy, unsupervised learning approaches complement a broader range of techniques in HAR research, thereby enhancing the overall understanding and versatility of the field.

Recent advancements in deep learning models have revolutionized HAR, thereby surpassing the limitations of traditional approaches. Deep learning models, such as SlowFast [22], I3D [41], and their variants, exhibit exceptional capabilities in the autonomous learning of intricate patterns and representations, leading to enhanced performance and potential in HAR systems. These models are highly accurate and robust, particularly for handling variations in the lighting, pose, and background. To address computational costs, smaller variants of models, such as C3D, R-C3D, TSN, and TSM, are recommended, and multigrid training techniques enhance the generalization capabilities.

Researchers can also use the 3DPW dataset [42], [43] for 3D human pose estimations and feature extractions. Using a simple deep-learning model, researchers can extract features that are instrumental in training a classifier for action label prediction. This technology enables the capture of detailed 3D human pose and motion data, aligning them to advance HAR by leveraging sensor data.

Recent developments in HAR underscore the field's rapid evolution propelled by deep learning and augmented reality. Studies such as Tangina et al. [44] offer a taxonomy-based analysis of HAR techniques, illuminating the strides made in both deep learning and traditional approaches. This comprehensive review not only showcases the current state of HAR but also identifies gaps in understanding the full complexity of human activities. The introduction of the SepCNN model by Chunzhao et al. [27] resulted in a significant improvement in AR-P300 recognition accuracy. This advancement highlights the potential of AR to enhance HAR, particularly in complicated environments through fine-grained visual recognition and multimodal systems. In actual life, human actions are more unpredictable, making it difficult to apply these advanced methods.

Adaptive HAR systems such as LAPNet-HAR demonstrate a field shift towards accommodating dynamic and diverse data qualities. The LAPNet-HAR framework proposed by [45] represents adaptive learning in processing sensor-based data streams, and addresses the critical issue of catastrophic forgetting in HAR. Nonetheless, as argued by [46], modeling complex human behavior through machine learning is inherently limited by the incompressibility and unpredictability of social systems. This critique highlights a significant gap in HAR research: the need for models that capture the nuances of human actions and address the societal and ethical implications of these technologies. The challenge lies in developing HAR systems that are not only technically proficient but also socially responsible, recognizing and mitigating the potential for historical and societal bias.

In addition, the clustering-based DeepTransHAR model with GRU layers [47] demonstrated superior performance, achieving an accuracy of 86.89% with reduced training time. Multi-view actions such as the Northwestern-UCLA dataset [48] can be a valuable resource for training and evaluating models for action recognition tasks that involve the soft weighting of body parts, temporal modeling, and classification based on RGB-D videos focused on human body components. This contributes to a better generalization, robustness, and overall performance of the model in real-world scenarios, where actions may be observed from various viewpoints and performed by different individuals. Some other har system works might acquire complex spatiotemporal patterns and adapt to varied real-world settings, thus spiraling up in challenges and reaching state-of-the-art performance.

The integration of communication and sensor technologies has spearheaded research on HAR is providing novel insights for human behavior analysis and has enhanced the performance significantly. For example, interior human behavior analysis has been greatly facilitated by the exploitation of Channel State Information (CSI) and the Fresnel zone model [29]. Furthermore, deep neural network algorithms introduce innovative approaches for integrating image information across video sequences in HAR [49], resulting in significant performance improvements compared with previous methods. These advancements have contributed to ongoing progress in HAR research, highlighting the potential for further enhancing the accuracy and effectiveness of action recognition systems.

The techniques utilized by the developers to ensure that it offers real-time processing in the HAR are quite amazing and inventive. For example, the use of I3D-ResNet50 [50], in the method of modeling through pruning techniques gives remarkable effectiveness in terms of the performance of the system. The redundancy and computational efficiency problems in the HAR systems have been handled elegantly. For example, state-of-the-art accuracy in the Human Activity Pruning with Deep Neural Network (HAP-DNN) architecture has led to computational efficiency problems as obtained by the authors in [51]. Depth cameras offer a significant advantage in terms of spatial information compared to other types of cameras. This is due to their ability to directly capture 3D space information. Furthermore, it has been noted that the algorithms used to compute this information have been enhanced, resulting in superior feature representation. Therefore, the findings of our research suggest that the implementation of segmentation techniques can

yield highly favorable outcomes in real-world scenarios. The authors of the study also present the advancements made in segmentation algorithms and the incorporation of more detailed spatial information through the adoption of depth cameras [52].

Skeleton-based graph designs and Graph Convolutional Networks (GCNs) have emerged as promising approaches to action recognition in HAR. A SkeletonPose method proposed by [53] improves 3D human pose estimation by integrating human skeleton constraints. Unlike exclusive deep-learning methods, they combine data-driven and calculation techniques. When tested with the Human Eva dataset [54], it utilizes a skeleton length before reducing the predicted skeleton errors and enhancing accuracy. This aligns with the trends in skeleton-based approaches and Graph Convolutional Networks for action recognition in HAR research. Jiagang Zhu [55] utilizes the H3D-MHAD dataset and achieved an average accuracy of 93.5% to enhance action recognition. Their two-stage model initially extracts features from 3D skeletal data, effectively transforming them into a graph-based representation, which is the cornerstone of skeleton-based graph design. These methods utilize graphs to represent actions, with nodes representing body joints, and edges intuitively capturing the movements and structure of the human body [56]. GCNs have demonstrated impressive results on benchmark datasets, further highlighting their effectiveness in action recognition [57], [58], [59]. Another innovative approach, known as Spatio-Temporal Attention-based Transformer (STAR), effectively represents the cross-modal features for action recognition [60]. These advancements have contributed to the overall progress of HAR research, paving the way for more accurate and effective action-recognition systems in various application domains.

Recent literature on image and video analysis highlights a trend toward neuroheuristic methods, particularly within IoT applications. These approaches, for instance, have been effective in enhancing different tasks, including but not limited to energy optimization in intelligent residence tasks [25] and tasks of enhancing precision and reliability of automated guided vehicles [26]. In this context, neuroheuristic analysis has been applied and emerged as an efficient approach in the contexts of surveillance for the detection related to real-time behaviors [61] Our work with the HADE framework goes well with this background in the research area focused on the improvement of HAR, especially by exploring advanced deep learning models. We position our work in the present research and, therefore, in a way, also in further developing the HADE framework.

The HAR study has conducted more research on the capability of smartphone sensors to capture and analyze human actions on a larger scope. The use of accelerometers and wearable sensors has yielded encouraging outcomes in HAR, providing a handy and inconspicuous method for monitoring human actions [62], [63], [64]. These investigations add to the progress in monitoring human activity and enhance our comprehension of human behavior in different situations.

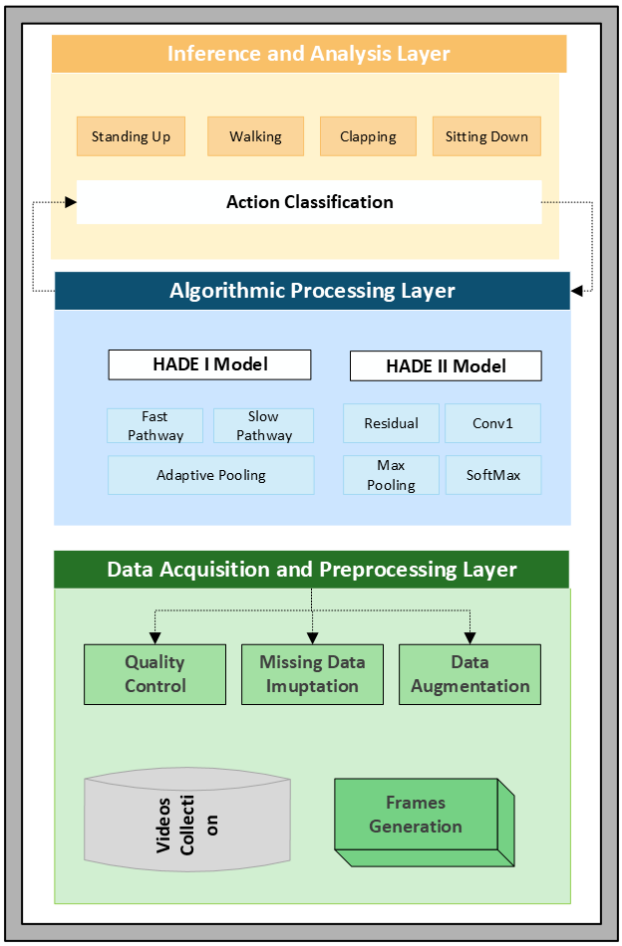

FIGURE 1. The HADE Architecture for Action Recognition.

We review methodology and advances in HAR and related studies. Current research contributes to the field, but its limits must be addressed. Using tagged data for training limits the scalability and generality of supervised learning systems. Deep learning methods are effective yet computationally expensive. These expenses can be reduced by using smaller variations and multigrid training. Smartphone sensor technologies are promising but need improvement. In the next section, we will propose a framework to overcome these restrictions and address the related issues. We use supervised learning and powerful machine-learning algorithms to improve action recognition accuracy, efficiency, and robustness.

III. HADE: PROPOSED ARCHITECTURE

We present the HADE model for 3D action recognition. HADE combines the strengths of HADE I and HADE II models to recognize and categorize human actions in 3D space. These models possess a keen ability to capture the complex spatial and temporal aspects of various scenes and are equipped to handle a range of lighting variations,

TABLE 1. Overview of research studies: models, methods, datasets, results, contributions, and limitations.

Reference	Model	Method	Dataset	Results	Contributions	Limitations
[22]	SlowFast	Deep Learning	Kinetics-400, Something-Something V2	79.8% accuracy	Handles variations in lighting, pose, background	Requires additional parameters, high computational cost
[35]	ASPR framework	Supervised Learning	UT-Kinect	86.5% accuracy	Utilizes labeled data for training, one-shot recognition	Limited scalability due to reliance on labeled data
[37]	PCANet-TOP	Sparse Coding (Unsupervised Learning)	UCF-101, HMDB51	Up to 83.99% accuracy	Adaptability to real-world scenarios	Potentially lower accuracy compared to supervised methods
[38]	SCAR	Two-stream Networks	COCO, YouTube-VOS	74.2% accuracy on COCO, 80.3% on YouTube-VOS	Practical technique for object segmentation in videos	Routing map learning unclear, affects robustness, generalization
[41]	I3D ResNet50	Resolution Convolutional Networks	UCF-101, HMDB51	90.5% on UCF-101, 87.6% on HMDB51	Segments-based 3D ConvNet captures spatial, temporal features	Limited exploration of generalization ability in other domains
[47]	DeepTransHAR	Deep Learning (Clustering)	Northwestern-UCLA	86.89% accuracy	Reduced training time, robustness	Best performance with specific datasets
[29]	HAP-DNN	Communication, Sensor Technology	UCI-HAR, OPPORTUNITY	99.38% accuracy on WISDM, 99.23% on OPPORTUNITY	Hybrid attention-based DNN for multi-sensor pruning	Complex network, requires extensive training data
[50]	Optimization (Pruning)	3D Convolutional and Two-stream Models	Kinetics, Hand Gesture Dataset	30.7% on Kinetics, 94.0% on Hand Gesture Dataset	Efficient real-time action recognition on mobile devices	Lack of detailed implementation, statistical analysis
[57]	STAR or ST-GCN	Skeleton-based Graph Convolutional Networks	Kinetics, NTU RGBD	52.8% on Kinetics, 88.3% on NTU RGBD	Motion energy/entropy-based segment selection strategy	Does not consider graph structure variations among actions, subjects
[58]	CNN based LSTM	Raw Accelerometer Signaling	WISDM	96.7% accuracy	Automatic learning of complex features from raw signals	Limited analysis of different performance segments
[60]	MPA based CNN	Optimized CNN based Graph Convolutional Networks	UCI-HAR, WISDM	98.2% on UCI-HAR, 97.8% on WISDM	Captures spatial and temporal connections among human joints	Uneven datasets handling unclear, could impact efficiency
[25]	Heuristic-Based Algorithm	IoT-Enabled Smart Homes	Custom	Optimized Trade-offs	Integrates demand response, renewable sources	Limited to specific home environments
[26]	Neuro-Heuristic Model	Pallet Detection for Vehicles	Custom	High Accuracy	Combines image processing with neural networks	Specific to vehicle navigation contexts
[61]	Neuro-Heuristic Framework	Surveillance Video Analysis	Centralized IoT	Real-Time Detection	Enhances behavior analysis using CNNs	Requires high computational power

diverse poses, and complex backgrounds. HADE uses these capabilities to improve 3D action recognition accuracy and reliability. This is supported by a carefully curated dataset that provides a thorough evaluation and forms the basis of precise action-recognition systems.

Figure.1 outlines the proposed architecture, which comprises three layers. These three layers are the Data Acquisition and Preprocessing Layer, the Algorithmic Processing Layer, and the insight generation and visualization layer. In the next three sections, we describe these layers in further details.

A. DATA ACQUISITION AND PREPROCESSING LAYER

The human action recognition (HAR) system depends on the efficient management of the HADE dataset's gathering and organization. This layer includes a variety of pre-processing

methods to ensure that the data is properly organized for subsequent processing. A smartphone camera was used to collect the data, and the action scenes were saved as.mp4 video clips. Each video file was between 4 to 5 seconds and during the recording, they were at a frame rate of 30 frames per second (fps). The frames were then organized in separate folders according to subject and action name.

1) PROPOSED DATASET - HADE

The HADE dataset comprises 699 samples of four fundamental actions, as listed in Table 2, which lists the number of samples available for each action.

TABLE 2. Action sampling in video collection.

Action #	Action Name	# of Samples
1	Clapping	178
2	Sitting down	171
3	Standing up	171
4	Walking	178
Total		699

The HADE dataset stands out from other datasets for 3D action recognition because of its breadth of action coverage and environmental diversity. This variety is crucial to the generalization of trained models, effectively preparing them to handle a wide range of real-world scenarios.

Figure 2 presents a diverse selection of clapping action samples, highlighting the variability and richness of the actions within the dataset.

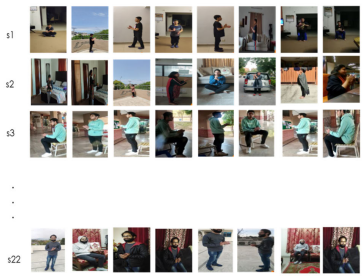

FIGURE 2. Selected Samples of Clapping Action from 4 Randomly Chosen Subjects out of 22.

2) COMPARISON WITH OTHER DATASETS

A comparison of the 3D action recognition datasets in Table 3 highlights the diversity of the datasets and identifies their unique features of the proposed dataset (HADE). This provides a comprehensive overview of the datasets that are currently available and can help researchers choose an appropriate dataset for their specific research objectives.

These datasets were compared to HADE based on the sampling size and diversity, which influenced our preprocessing decisions for informed analysis.

a: UTKINECT-ACTION3D DATASET

UTKinect-Action3D captures ten actions from ten subjects using RGB, depth, and skeletal joint data at 30 fps [62]. Challenges include viewpoint variations, self-occlusions, and intraclass differences. Researchers have employed alignment, normalization, and augmentation for this enhancement.

b: CMU MOCAP3D DATASET

The CMU Mocap3D dataset, which contains 2600+ motion trials by 144 actors in diverse settings, yielded nearly 600 samples. Captured using a Vicon optical system with 12 cameras at 120 Hz [65], it provides the 3D marker positions, skeleton joint angles, and BVH animation files. Challenges encompass viewpoint variations and self-occlusions, which are addressed via meticulous data alignment and normalization during preprocessing, thereby enhancing the usability of the dataset.

c: HUMANEVA DATASET

The HumanEva dataset contains 87,000 frames of synchronized motion capture and video data recorded by seven cameras at 60 Hz [54]. Illustrating six typical actions carried out by four people in real-world scenarios comes with its own set of difficulties, such as changes in lighting and obstacles obstructing the view. Addressing these issues requires careful data alignment and synchronization during the preprocessing phase.

d: NORTHWESTERN-UCLA DATASET

Each of the ten action categories in the Northwestern-UCLA dataset is completed twice by ten subjects. The Kinect cameras captured the movements, encompassing data related to the skeleton, depth, and RGB colours. Obstructions and differences in vision are two challenges. To rectify these deficiencies and enhance the dataset's suitability for analysis, it underwent preparation through the implementation of normalization, data augmentation, and temporal alignment techniques.

e: BMHAD DATASET

The BMHAD dataset includes 11 action categories performed five times by 12 individuals in various scenarios, resulting in a total of 660 action samples. Acquired data consists of RGB, depth, and inertial sensor information, along with ground-truth labels and bounding boxes [67]. The tasks involved extracting depth frames, removing background noise, extracting features, and using machine learning algorithms for accurate action classification, leading to a comprehensive dataset analysis.

f: MSR ACTIONPAIRS DATASET

The MSR ActionPairs dataset represents an invaluable resource within the field of computer vision, specifically tailored for the examination of 3D action recognition through depth sequences. This dataset encapsulates a collection of 12 distinct actions, each executed by ten different subjects, thereby offering a comprehensive framework for the analysis and interpretation of human motion in a three-dimensional context. The diversity of subjects and

TABLE 3. Comparing 3D action recognition datasets.

Dataset	Preprocessing	Actions	Samples	Diversity
UTKinect-Action3D [62]	Data alignment, normalization, and data augmentation	3	90	Limited
CMU Mocap3D [65]	Data alignment and normalization	6	600	Limited
HumanEva [54]	Data alignment and synchronization	15	1080	Limited
Northwestern-UCLA [66]	Temporal alignment of data, normalization, and data augmentation	10	100	Limited
BMHAD [67]	Depth frame extraction, background noise reduction, feature extraction	10	100	Limited
MSR ActionPairs [68]	Depth frame extraction, noise reduction, feature extraction using HON4D	15	150	High
3DPW [42]	Data acquisition, conversion, alignment, normalization, and data augmentation	15	450	High
ViBe-3D [69]	Data alignment, model parameter prediction, adversarial learning, and incorporation of temporal information	10	400	High
H3D-MHAD [70]	Data import, data extraction, quality control, encoding, partitioning, and data scaling	10	500	High
HADE	Data collection, data cleaning, data augmentation, and temporal consistency check	4	900	High

actions within the dataset provides a robust basis for the development and evaluation of computational models aimed at understanding and classifying human movements with a high degree of accuracy. Unique to this dataset, it pairs actions exhibiting similar motion patterns and shapes but different relationships [71], [72]. The preprocessing pipeline includes depth-frame extraction, noise reduction, and feature extraction using a HON4D histogram [73].

g: 3DPW DATASET

The 3DPW dataset provides real-world 3D human pose estimations using synchronized video and motion capture data. It features 60 outdoor video sequences, detailed 2D and specialized 3D pose information, camera poses, 3D body scans and human models [42]. Preprocessing encompasses various stages such as data acquisition, conversion, alignment, normalization, and data augmentation, ensuring that the dataset is primed for accurate and comprehensive analysis.

h: VIBE-3D DATASET

With 22,000 different video clips, VIBE-3D provides a specialised dataset and framework for 3D human pose and shape estimation [69]. Preprocessing entails careful procedures including data alignment, model prediction, adversarial learning, and temporal information integration to handle issues like temporal learning and view variations.

i: H3D-MHAD DATASET

3D LiDAR and inertial sensors were used to enable 3D human action detection with the H3D-MHAD dataset. With annotations providing 3D point clouds, skeletal joint locations, and sensor readings, it covers 12 activity categories [70]. Data extraction, encoding, partitioning, scaling, quality assurance, and encoding are all included in pre-processing.

Figure 3 visually compares the actions and sizes listed in Table 3.

3) PRE-PROCESSING HADE DATASET

A methodical approach was used in the production of the video data to ensure accuracy and consistency. To ensure a consistent analysis, the videos were segmented into 16-frame chunks to standardize the input size. Techniques for choosing frames that capture the subtleties of the material include uniform, random, and dense sampling. To maintain data consistency, the frames were scaled to 256 pixels by 256 pixels, a standard resolution. It is necessary to record the motion dynamics for video analysis. Temporal information is added via optical flow computation, which calculates the motion between frames. Gaussian blur, color jittering, and random flips were used to enhance the dataset's variety and strengthen the model's robustness. Large video files, however, call for efficient data handling. This is made possible by temporal jittering, which guarantees a variety of contexts, and temporal sampling, which chooses clips to cut down on processing expenses. To stabilize the training, the pixel values were normalized to have a mean of 0 and a standard deviation of 1. To capture the temporal context, the frames were stacked in a 16. The data were organized into batches of size $\times$ num_frames $\times$ height $\times$ width $\times$

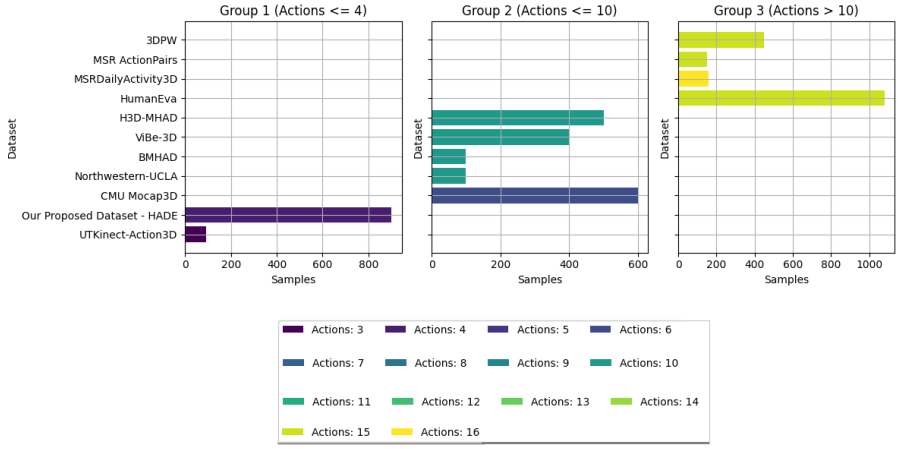

FIGURE 3. Comparing total actions based on the number of total samples of each dataset in Table 3.

channels for efficient computation. Action labels are encoded in numerical formats such as one-hot encoding or integer labels for precise model interpretation.

Pre-processing techniques were used during the dataset curation process to guarantee the quality and usability of the dataset for training advanced 3D action recognition models. Among these methods are the following ones.

- Data Augmentation (DA): Rotation, rotating, and cropping are used to enhance dataset size using data augmentation. These adjustments increased dataset size and variety. So, models get more training cases, to assist them. These methods increase knowledge and help identify data trends, which improves model performance. This improves model accuracy and generalizability across many conditions.
- Missing Data Imputation (MDI): Incomplete or missing samples in datasets are addressed by MDI data preparation approaches. These methods include interpolation, which estimates missing data points by using adjacent data points; extrapolation, which extends data analysis beyond the original observation range to infer missing values; and mean imputation, which replaces missing values with the mean of the available data. These methods help researchers assure dataset integrity and completeness, improving analysis dependability.
- Quality Control (QC): To ensure dataset quality and dependability, QC techniques are used. This process involves detailed error checking, identifying gaps, and fixing issues like corrupted files and incorrect labels. These methods ensure the dataset's stability and quality, making it useful for analyses and model training. A rigorous dataset validation reduced biases and errors that could affect training.

The dataset pre-processing process can be represented as follows:

$$X \xrightarrow{\text{DA}} X_a \xrightarrow{\text{MDI}} \hat{X} \xrightarrow{\text{QC}} \tilde{X} \tag{1}$$

By applying these preprocessing techniques, the dataset became more suitable for subsequent analysis and model training, resulting in improved reliability and accuracy of the obtained results.

Algorithm 1 Data Preprocessing and Augmentation for HADE Dataset

Require: Raw video dataset X
Ensure: Pre-processed and augmented dataset $\tilde{X}$
1: Divide video into 16-frame clips
2: Apply frame selection techniques (uniform, random, dense sampling)
3: Resize frames to 256×256 pixels
4: Compute optical flow for temporal information
5: Apply data augmentation (random flips, color jittering, Gaussian blur)
6: Apply temporal sampling and jittering
7: Normalize pixel values (mean = 0, std dev = 1)
8: Stack frames in a 16-frame format
9: Organize data into batches
10: Encode action labels
11: Process dataset through DA to obtain
12: Process dataset through MDI to obtain
13: Process dataset through QC to obtain final dataset
14: **Return** Pre-processed and augmented dataset

B. ALGORITHMIC PROCESSING LAYER

The Algorithmic Processing Layer focuses on using supervised learning to improve HAR performance. This includes training and fine-tuning models like HADE I and HADE II with labeled data from the HADE dataset.

The training process involved the following steps:

1) FINE-TUNING THE PARAMETERS

To update the model parameters during training, a gradient-based optimization algorithm was employed. The parameters were updated by multiplying the gradient of the objective

function by the learning rate (α), as shown in Eq. (2).

$$\Delta\theta = \alpha \cdot \nabla\theta \tag{2}$$

The key parameters involved in the fine-tuning process and their specific roles in the training process are outlined below:

- **Learning Rate (α)**: Set at 0.0001, this learning rate was optimized to ensure a balance between efficient learning and the risk of overshooting the minimum during gradient descent. This value was experimentally determined to achieve a steady convergence.
- **Weight Decay (wd)**: Set at 0.0001 for L2 regularization, this parameter helps in preventing overfitting by penalizing large weight values, ensuring the model's generalizability.
- **Momentum (mom)**: Setting the momentum value at 0.9 significantly enhances the optimization process, facilitating quicker convergence towards the optimal solution and effectively navigating the challenges posed by local minima. This momentum parameter acts as a catalyst, propelling the optimization algorithm forward, thereby optimizing the learning pace and improving the overall efficiency of the model's training phase.

$$\text{vel} = \text{mom} \times \text{vel} - \alpha \times \text{grad} \tag{3}$$

  In Eq. (3), the term vel represents the integral of velocity in the parameter update process, highlighting the role of momentum in how parameters are adjusted over time.
- **Learning Rate Decay**: A decay rate of 0.1 gradually lowers the learning rate, enabling finer tuning of weights during training and reducing fluctuations around the best solution.
- **Batch Size**: A batch size of 5 balances computational effort and generalization, aiding in effective gradient calculation.
- **Total Epochs**: The model underwent training for 15 epochs, a period chosen from practical experience, showing adequate learning without excessive fitting to training data.

The combination of these parameter values and the optimization process was critical for achieving high accuracy in our models. By carefully fine-tuning these parameters, we ensured that our models adapted and improved their performance throughout the training process, ultimately contributing to accurate results.

Parameter	Value
Learning Rate (α)	0.0001
Weight Decay (wd)	0.0001
Momentum (mom)	0.9
Learning Rate Decay	0.1
Batch Size	5
Total Epochs	15

2) HADE I MODEL

The HADE I model, which captures spatial and temporal information by separating the slow-moving and fast-moving parts of a video into distinct pathways, is introduced. The architecture is explained, including the slow pathway, fast pathway, and the fusion of their outputs. Eqs.s were used to describe the operations involved in the HADE-I model. A network flow diagram was included to visualize the processing of the video streams through the pathways and the final output in the following steps.

Step 1. *Determining the standard architecture features to capture spatiotemporal connections.*

The SlowPath in Eq. 4 represents the slow pathway.

$$I_s = \text{SlowPath}(I) \tag{4}$$

and FastPath functions represent the fast pathways in Eq. 5.

$$I_f = \text{FastPath}(I) \tag{5}$$

where I represents the input video and I_s and I_f are the outputs of the slow and fast pathways, respectively. Similarly, the Fuse function in Eq. 6 combines the outputs of defined slow and fast pathways.

$$y = \text{Fuse}(I_s, I_f) \tag{6}$$

where y denotes the final prediction of the action in the video.

Figure 4 depicts the HADE I architecture, illustrating the processing of video streams through the slow and fast pathways. The intermediate feature maps generated by these pathways were fused, and the combined feature map underwent spatial average pooling, Softmax, and additional processing to produce the final output.

Step 2. *Fine-Tuning the Parameters.* The same method was applied to tune the parameters used in Eq. (2).

Step 3. *Training the HADE I model.*

The HADE approach with the HADE I model, outlined in Algorithm 2, was employed for action recognition. The algorithm begins by preprocessing the dataset of the video clips, including tasks such as resizing, normalization, and augmentation. The dataset was divided into training and validation datasets. Subsequently, training parameters, such as the learning rate, batch size, and optimization algorithm settings, were initialized. The HADE I model M is initialized using Eqs.4, 5, and 6. The algorithm iteratively samples batches of data from the training set and passes them through the HADE I model to obtain predicted outputs. The trained model was evaluated using a validation set, and its performance was calculated. If the validation performance satisfied a predefined convergence criterion, the training was continued; otherwise, the training was stopped. Ultimately, the trained HADE I model was returned as the output of the algorithm.

The integration of the HADE I model into our system, along with iterative training and validation monitoring, enabled effective action recognition by leveraging the model's capabilities in capturing temporal and spatial features accurately.

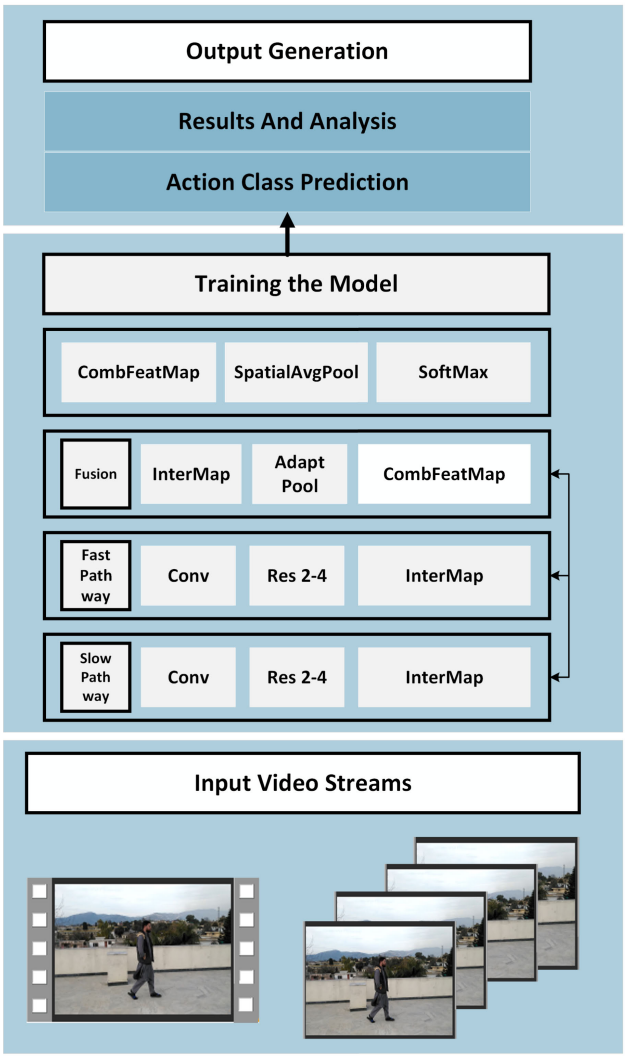

FIGURE 4. HADE I Network Flow Diagram.

3) HADE II MODEL

The HADE II model is built for video classification, using 3D convolutions and connections that help it pick up on patterns over time and space. It's set up to handle video data by breaking it down into features that can be analyzed. The model is explained with equations that outline its key parts and how they work together. A network flow diagram was included to illustrate the connectivity and output of the HADE II model.

Step 1. *Determining the standard architecture features to capture spatiotemporal and residual connections.*

$$\mathbf{W}_{\text{Conv}} = \text{Conv1}(7 \times 7, \text{stride} : 2) \tag{7}$$

In Eq. 7, a 7×7 Convolutional layer 1 (Conv1) layer with a stride of 2, capturing low-level features from the input video frames.

$$\mathbf{W}_{\text{MaxPool}} = \text{MaxPool}(3 \times 3, \text{stride} : 1) \tag{8}$$

Algorithm 2 Training HADE I Model

Require: Dataset D video clips
Ensure: Trained Dataset D on HADE I model M

```
Preprocess dataset D
Split dataset D into training and validation sets
Initialize training parameters
Initialize HADE II model M using layers defined in eq.(4), (5), and (6).
Initialize the initial performance as prev_val_perf
Sample a batch of data from the training set
Perform a forward pass through the model M to obtain predicted outputs
Compute the loss between the predicted outputs and the corresponding ground truth labels
Compute the gradients of the loss for the model parameters using backpropagation
Update the model parameters based on the gradients using gradient descent:
    Compute the parameter update by eq.(2)
Evaluate the trained model M on the validation set
Calculate the validation performance val_perf if val_perf ≥ prev_val_perf then
    Update model M based on training data;
    Set prev_val_perf as val_perf;
    else
    Break the training loop;
return Trained HADE I model M
```

In Eq. 8, a 3×3 Max pooling layer (MaxPool) layer with a stride of 1 reduces the spatial dimensions.

$$\mathbf{W}_{\text{Res}} = \text{Res}(2, 3, 4) \tag{9}$$

In Eq. 9, $\mathbf{W}_{\text{Res}i}$ represents the 16 residual blocks with varying block and filter sizes and i is the block number.

$$\mathbf{W}_{\text{AvgPool}} = \text{AvgPool}(7 \times 7, \text{stride} : 1) \tag{10}$$

Eq. 10 signifies a 7×7 average pooling layer, employed with a stride of one, aiming to diminish the spatial dimensions within the model.

The HADE II model's architecture is designed for gradual downsampling and alteration of features, as illustrated in Figure 5. This diagram visually represents the model's flow, highlighting how input video streams are processed through the Inflated 3D (I3D) configuration settings, culminating in the model's output.

Step 2. *Training the HADE II model.*

The HADE approach, specifically through the HADE II model as outlined in Algorithm 3, is applied for action recognition. This process begins with preprocessing the video clip dataset, including resizing, normalization, and augmentation. The dataset is then split into training and validation sets. Initial training parameters such as learning rate, batch size, and optimization algorithm settings are set up. The HADE II model, denoted as M, is initialized

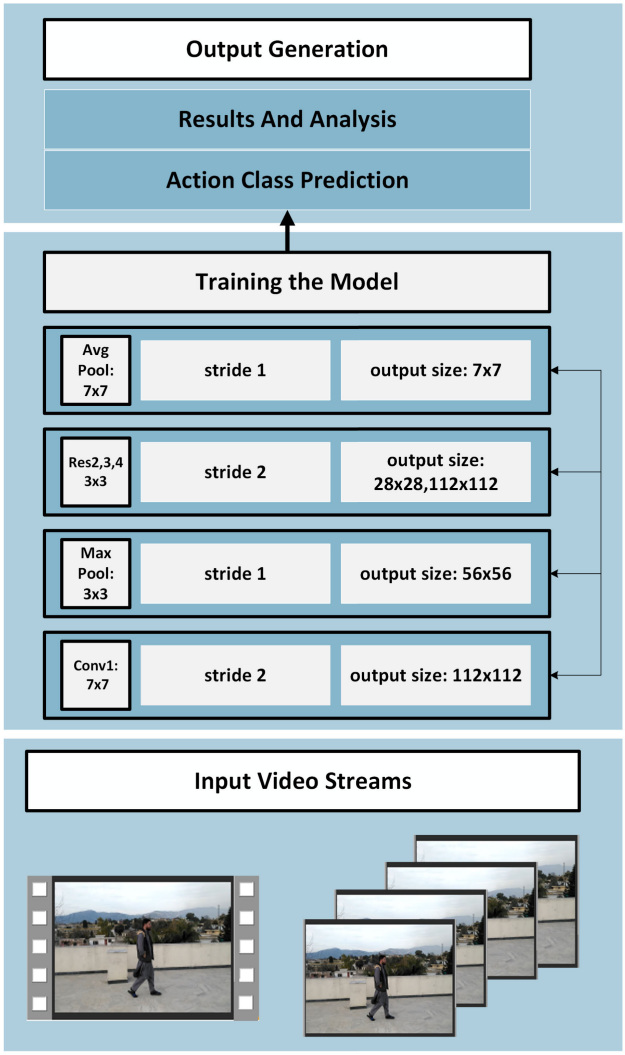

FIGURE 5. HADE II Mode Network Flow Diagram.

according to equations 7, 8, and 9. Through iterative sampling of data batches from the training set for processing by the HADE II model, predicted outputs are obtained. The model's performance is assessed using the validation set, and based on whether it meets a specific convergence criterion, training is either continued or halted. The outcome of this algorithm is the fully trained HADE II model, ready for action recognition tasks.

The incorporation of the HADE II model into our framework significantly bolsters the reliability and efficiency of human action recognition tasks. The advanced architecture of HADE II and the iterative procedures used in its training are credited with this improvement. The model's contribution to enhancing the overall system performance in identifying complicated human movements is highlighted by this strategic integration, which makes it easier to identify human actions in a more nuanced and precise manner.

Algorithm 3 Training HADE II Model

Require: Dataset D video clips
Ensure: Trained Dataset D on HADE II model M
1: Preprocess dataset D
2: Split dataset D into training and validation sets
3: Initialize training parameters
4: Initialize HADE II model M using layers defined in eq.(7), (8), and (9)
5: Initialize the initial performance as prev_val_perf
6: Sample a batch of data from the training set
7: Perform a forward pass through the model M to obtain predicted outputs
8: Compute the loss between the predicted outputs and the corresponding ground truth labels
9: Compute the gradients of the loss for the model parameters using backpropagation
10: Update the model parameters based on the gradients using gradient descent:
11: Compute the parameter update by eq.(2)
12: Evaluate the trained model M on the validation set
13: Calculate the validation performance val_perf **if** $val\_perf \geq prev\_val\_perf$ **then**
Update model M based on the training data;
Set $prev\_val\_perf$ as val_perf;
else
Break the training loop;
14: **return** Trained HADE II model M

Algorithm 4 HADE Model

Require: Trained HADE I and HADE II models
Ensure: Trained HADE Action Recognition Model
1: Initialize Trained Models:
2: i3d_model ← Algorithm 3
3: slowfast_model ← Algorithm 2
4: evaluate_model(i3d_model, slowfast_model, val_set)
5: $trained\_model \leftarrow (val\_acc\_i3d \geq val\_acc\_slowfast)?i3d\_model : slowfast\_model$
6: evaluation_results ← evaluate_model(trained_model)
7: **return** trained_model

4) UTILIZING TRAINED MODELS FOR INSIGHTS GENERATION AND VISUALIZATION

This section presents a detailed comparison of HADE I and HADE II models, which are built for action recognition. After evaluating both models using a validation dataset, Algorithm 4 was applied to select the more suitable model for leading action recognition efforts. The chosen model was then crucial in analyzing new data, leveraging its understanding of spatial and temporal dynamics for accurate action identification in videos. This analysis extends beyond just accuracy, including metrics like precision, recall, and F1-score, offering a holistic view of the models' action recognition capabilities and effectiveness.

Section V compares HADE II and HADE I models on the validation set, focusing on performance metrics like accuracy. It uses visualizations to pinpoint misclassifications, dissect decision-making processes, and explore the spatial and temporal cues critical for action recognition, offering a deep dive into each model's effectiveness and insights for improvement.

C. FEATURE EXTRACTION

We identify key video parts using three steps: SlowPath, FastPath, and Fusion, detailed in equations 4, 5, and 6. This approach helps us capture and combine crucial details efficiently. However, it presents challenges such as the complexity of combining features, the need for precise adjustment of settings, missing broader context, and high computational requirements, all critical for enhancing system performance and reliability.

$$y = X \xrightarrow{\text{SlowPath}} X \xrightarrow{\text{FastPath}} X_{\text{Fuse}} \tag{11}$$

where:

- SlowPath applies operations to extract relevant features.
- FastPath performs operations to extract complementary features.
- Fuse combines the extracted features from the SlowPath and FastPath to obtain the fused feature (y).

By utilizing both SlowPath and FastPath, the feature-extraction process enhances the capability of the model to capture and represent diverse levels of information. This improvement leads to enhanced feature representations that can be effectively utilized in various downstream tasks.

1) LIMITATIONS OF THE FEATURE EXTRACTION PROCESS

Despite its benefits, the feature-extraction process has several limitations that must be considered when designing and implementing HAR systems. These limitations include the following.

1) **Complexity of Feature Fusion:** Integrating features from SlowPath and FastPath adds complexity, requiring careful crafting and tuning of fusion methods to maximize data utility and avoid loss of valuable insights.
2) **Sensitivity to Parameter Settings:** Parameter settings in data analysis, including kernel size, stride, and pooling, are crucial for extracting meaningful features necessary for clear differentiation.
3) **Limited Contextual Information:** Feature extraction from videos focuses on local movements' specifics and timing but may overlook broader contextual actions necessary for comprehensive understanding.
4) **Computational Complexity:** Deep learning-based feature extraction in videos demands significant computational resources and time, challenging scalability, and real-time processing.

Acknowledging limitations is vital for evaluating the feature-extraction process in HAR systems, necessitating the exploration of strategies to mitigate challenges for enhanced performance.

IV. EXPERIMENTAL SETUP

This section provides a comprehensive overview of the experimental setup employed for human-action recognition. It encompasses key aspects, including the dataset description, GPU-based training, training configuration, and evaluation metrics used to assess model performance. The objective was to establish a rigorous framework for the experiments and to offer in-depth insights into the setup and evaluation procedures.

A. DATASET DESCRIPTION

For our studies, we put together our collection of footage, tagged as D, filled with 900 video clips. These aren't just any clips; they showcase an array of 3D human actions captured in both indoor and outdoor environments (E). We didn't just stop at the variety in settings; we made sure our dataset reflected real-world diversity, roping in 22 folks from different walks of life, covering a wide age span from 15 up to 54 years. A big chunk of our participants are in their twenties, aligning with the age group we see a lot in our research. But we're not closing the door on diversity; we're fully intending to bring more age groups into the fold in our upcoming projects.

Bias? We're on it. From making sure we've got an even mix of genders to mixing up the lighting and backdrops, we're all about keeping our data as real and inclusive as possible. Plus, we're constantly taking a magnifying glass to our methods, hunting down any hidden biases that might throw off our findings, all intending to make sure our model stands strong and useful in as many situations as possible.

B. GPU-BASED TRAINING AND INFERENCE

The development and fine-tuning of models tailored to our unique dataset underline the need for advanced computing solutions. Recognizing human actions, a complex endeavor within the field of computer vision, demands significant computational resources for both the training phase and real-time application.

In response to these requirements, we selected NVIDIA Tesla V100 GPUs for our computational infrastructure. The Tesla V100 is noted for its exceptional capabilities in deep learning tasks, making it an ideal choice for our project. Its notable processing speed and computational efficiency are essential for managing the intricate calculations our human action recognition models necessitate. The decision to employ this particular GPU model was influenced by its proficiency in processing extensive data sets and handling the demanding aspects of model training, thus markedly decreasing the duration of training periods and improving the overall performance of our models.

Furthermore, to maximize the efficiency of our GPU resources, we incorporated the GluonCV Python library, based on the MXNet framework. This library is specifically

designed for executing high-performance deep-learning operations, thereby ensuring the Tesla V100 GPUs are used to their fullest potential. Integrating this advanced hardware with meticulously optimized software not only improves the precision of our action recognition models but also enhances their scalability and adaptability in the face of future technological advancements.

C. GPU-BASED EVALUATION

To assess our work, we tapped into the power of GPUs, choosing Amazon EC2 instances that come packed with NVIDIA Tesla T4 GPUs (G), each loaded with 16 GB of memory. This setup is a beast when it comes to parallel processing, making our evaluation tasks run much faster. Plus, the hefty memory on these GPUs, way more than what you'd get with regular CPUs, means we can train and test our complex action recognition models more effectively. And the best part? If researchers need more oomph, they can just rent extra GPU instances, keeping it scalable.

D. EVALUATION METRICS

The performance of the human action recognition models trained and tested on our dataset was assessed using a range of rigorously selected evaluation metrics, including

1) ACCURACY (ACCURACY (ACC))

ACC measures the ratio of correctly predicted labels to the total number of predictions made by the model. It is calculated as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives.

2) POSITIVE CLASSIFICATION CORRECTNESS (PRECISION(PCC))

PCC measures the ratio of true positives to the total number of predicted positives made by the model. It is calculated as:

$$PCC = \frac{TP}{TP + FP} \tag{13}$$

where TP represents true positives and FP represents false positives.

3) RECALL

Recall measures the ratio of true positives to the total number of actual positives. It is calculated as:

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

where TP represents true positives and FN represents false negatives.

4) F1-SCORE

The F1-Score, which is the harmonic mean of the precision and recall, is commonly used to evaluate imbalanced datasets. It is computed as:

$$F1 - Score = 2 \times \frac{PCC \times Recall}{PCC + Recall} \tag{15}$$

where PCC and $Recall$ are the precision and recall values, respectively.

These meticulously chosen evaluation metrics provide a comprehensive means of assessing the performance of human action recognition models trained and tested on a dataset. Through a detailed analysis of these metrics, the strengths and weaknesses can be identified, facilitating continuous improvement to enhance the model's performance.

V. RESULTS AND ANALYSIS

This section describes the final layer of the proposed architecture, which consists of two distinct stages: Results and Analysis as well as Action Classification. A comprehensive understanding of this layer can be gained by delving into subsequent sections, which provide a detailed exploration of the results obtained and the process of generating the output.

A. COMPARISON WITH BASELINE MODELS AND MODEL EVALUATION

In our experimental evaluation, we compared SlowFast 16×8, R101+NL [22], and S3D [41] models. SlowFast integrates slow (16 fps) and fast (128 fps) pathways for spatial and fine motion analysis, connects via lateral pathways, and merges through a fully connected layer. S3D utilizes 3D CNNs, as well as temporal and spatial attention modules. We introduced two proposed models, HADE I and HADE II, and thoroughly assessed their performance alongside the baseline models (SF and S3D) by examining their training and testing accuracy, as shown in Figure 6.

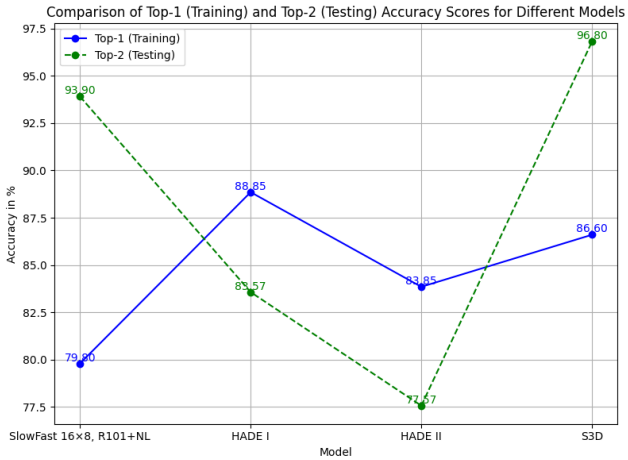

FIGURE 6. Comparison of top-1 (training) and top-2 (testing) accuracy scores for different models.

1) MODEL PERFORMANCE ANALYSIS

The graph in Figure 6 depicts the top-1 and top-2 accuracy scores obtained using different models. The top-1 accuracy is the percentage of training examples for which the model

correctly predicted the most likely label. Conversely, the top-2 accuracy represents the percentage of examples in which the model correctly predicted either the most or second-most likely label. In general, all models had higher top-1 accuracy than top-2 accuracy. This was anticipated because the top-2 score was less demanding than the top-1 score. A careful examination of the graph reveals significant discrepancies between the top-1 and top-2 accuracy scores of some of the models. For example, the SlowFast 16 × 8 and R101+NL models achieved a top-1 accuracy of 97.5% but only top-2 accuracy of 96.0%. This disparity suggests that, although the model can accurately predict the most likely label in most instances, it is difficult to determine the second-most likely option.

TABLE 4. Performance comparison of different models with inference, Top-1 and Top-2 accuracy scores.

Model	IT	top-1	top-2
SF [10]	0.6	79.80	93.90
S3D [41]	0.6	86.60	96.80
HADE I	0.6717	88.85	83.57
Hade II	0.6462	83.85	77.57

HADE I exhibited noteworthy superiority in terms of accuracy, with a training accuracy of 88.85% and testing accuracy of 83.57%. In contrast, HADE II achieved lower accuracy metrics, with a training accuracy of 83.85% and a testing accuracy of 77.57%. These results suggest that the modifications made to HADE I resulted in improved predictive capabilities, particularly with respect to the single-best predictions.

B. RATIONALE FOR DATASET COMPARISON

To contextualize our comparative analysis, it is essential to briefly describe the datasets used in conjunction with HADE and provide rationales for their selection.

Kinetics-400: A large-scale, high-quality dataset with a diverse range of human actions. Its inclusion in the comparison highlights HADE's applicability of HADE in a broad spectrum of actions.

UCF-101: One of the most widely used datasets in action recognition research, UCF-101 provides a benchmark for evaluating HADE's performance against established standards.

Something-Something V2 [74]: This dataset focuses on human-object interactions, offering a unique perspective on action recognition. Its comparison with HADE emphasizes the versatility of our dataset for handling complex scenarios.

Quo Vadis [75]: Quo Vadis, distinguished by its concentration on repetitive human movements, acts as a crucial testing ground for gauging HADE's effectiveness in identifying and distinguishing between such repetitive actions. This environment allows for a detailed examination of HADE's precision and adaptability in scenarios dominated by repeated movements.

Breakfast [75]: This dataset, focusing on nuanced activities, particularly within cooking environments, offers a unique backdrop to assess how well HADE can identify detailed actions. We selected these datasets due to their significance in the action recognition domain and their unique characteristics. They offer a thorough framework for evaluating the advantages and capabilities of HADE, showcasing its ability to distinguish between finely detailed actions.

1) METRIC ANALYSIS AND DATASET COMPARISON

Our study utilized several key internal metrics to gauge performance, such as Inference Time (IT) and Positive Classification Correctness (PCC) scores, along with top-1 and top-2 accuracy rates. These measures are particularly important for assessing real-time action recognition capabilities. They offered a clear view of how efficient and accurate our models are under different conditions. This comprehensive evaluation highlighted the strong performance of HADE when compared to existing benchmarks, demonstrating its effectiveness in various testing environments.

TABLE 5. Temporal length (M) and stride (T) for datasets.

Dataset	Temporal Length (M)	Stride (T)
Kinetics-400	4	1
UCF-101	8	1
Something-Something V2	16	4
Quo Vadis	32	1
Breakfast	16	1
HADE	4.5	2

In Table 5, a detailed comparison between HADE and other datasets is illustrated, emphasizing the aspects of temporal length (M) and stride (T). HADE stands out with its specific arrangement of analyzing 4.5 seconds of video alongside a 2-frame stride. This setup is particularly effective for capturing slower movements, offering a well-rounded approach to action recognition. The deliberate design of HADE's temporal length and stride aims to strike an ideal balance between achieving precise action recognition and maintaining computational efficiency, thereby setting it apart from alternative datasets. For example, Kinetics-400 and UCF-101, which are characterized by their shorter temporal durations and strides, focus primarily on quicker action sequences. In contrast, HADE affords a more detailed examination.

Note on Frame Stacking: Additionally, HADE employs a frame stacking technique involving 16 frames, designed to encompass a broad array of human actions comprehensively. This method diverges from those used in datasets like Something-Something V2 and Breakfast, each tailored to their specific goals and technical prerequisites. Through this strategy, HADE ensures not only a thorough coverage but also the flexibility and accuracy necessary for capturing and analyzing an extensive variety of human actions effectively.

This comparative analysis underscores HADE's unique positioning in the landscape of 3D action recognition

datasets. By balancing the temporal resolution and computational demands, HADE stands as a robust tool for researchers looking to delve into the nuances of human action recognition.

Table 6 presents the model details and performance metrics for various models. Although the SlowFast 16 × 8 and R101+NL (SF) models demonstrated the highest accuracy for human action recognition, they were also the most computationally expensive and required the longest training time. The segment-based 3D ConvNet (S3D) model offers a balanced trade-off between accuracy and computational costs. Additionally, the HADE I and II models are the fastest and most efficient; however, they have compromised accuracy. In certain applications, models with lower accuracy can be an efficient solution, such as real-time human action recognition on mobile devices, where a reduction in computational cost and power consumption may be preferred, even if some accuracy is sacrificed. Furthermore, models with lower accuracy can be deployed on more devices and used in applications where the power budget is a constraint, such as in wearable devices. The selection of the best model for human-action recognition depends on the specific requirements and constraints of the application. If accuracy is the primary concern, the SF model is the preferred choice. If efficiency is a higher priority, then models with lower accuracy, such as the S3D or HADE models, may be more suitable.

TABLE 6. Performance comparison of HAR models. SlowFast (SF) and S3D offer high accuracy, but require extensive computational resources. HADE I and II are more efficient with reduced training times and parameters but have lower accuracy. This table assists in selecting models based on accuracy-efficiency balance.

Model	PCC	Training Time	Parameters
SF	0.818	24 hrs	10.7M
S3D	0.78	24 hrs	8.77M
HADE I	0.6717	8 hrs	75M
HADE II	0.6462	16 hrs	15M
TOTAL		24 hrs	90M

We obtained the following results regarding PCC, training time, and number of parameters for the various models from Table 6 which shows that the SF model had the highest PCC, indicating superior performance in terms of classification accuracy, whereas the S3D model had the shortest training time, making it an efficient choice for model development. HADE II has the fewest parameters, which can be advantageous in cases where model complexity needs to be reduced.

2) COMPARISON WITH STATE-OF-THE-ART MODELS AND ABLATION STUDY

To establish the standing of HADE models in the field of Human Action Recognition, we conducted an extensive comparison with state-of-the-art models, complemented by a detailed ablation study. This approach provides a comprehensive understanding of the performance of HADE models based on the latest advancements in HAR.

a: ABLATION STUDY

The study involved systematically altering key components of the HADE models to evaluate their contributions. For HADE I, we examined the impact of the SlowFast model, and for HADE II, we examined the effect of 3D convolutions. This helped to identify the most important features and potential areas for improvement.

b: STATE-OF-THE-ART COMPARISON

Our study incorporated a comprehensive comparison with a variety of recent models in the field of Human Action Recognition. This comparison not only encompasses diverse methodologies ranging from vision-language models to advanced convolutional network architectures but also covers a spectrum of datasets from standard action recognition benchmarks to more recent and challenging ones.

The visualization in Figure 7 presents a clear and concise comparison of the top-1 accuracies, providing a visual representation of where the HADE models stand with other contemporary solutions in HAR. The bar chart delineates the accuracy achieved by each model, illustrating the competitive edge of the HADE models, particularly HADE I, with superior accuracy.

Table 7 extends our analysis by providing a detailed breakdown of the characteristics and performance metrics of each model. This highlights the diversity in methodologies and breadth of datasets employed in recent HAR research. The inclusion of the ablation study results for HADE I and II in this table offers a direct and nuanced comparison, emphasizing the specific contributions of the key features in our models and their relative importance in achieving high accuracy.

Through this rigorous comparative analysis, the HADE models demonstrated their capacity to stand alongside the recent innovations in the HAR domain. The insights gained from both the ablation study and state-of-the-art comparison not only validated the effectiveness of the HADE models but also shed light on potential avenues for future enhancements.

C. MODEL EVALUATION

We conducted a performance evaluation of the HADE II and HADE I models on our custom HADE dataset of human action videos. As shown in Figure 8, HADE I initially achieved an accuracy of 60% compared to HADE II's accuracy of 50% ($p < 0.05$). By the tenth epoch, HADE I attained an accuracy of 85%, surpassing that of HADE II by 75% ($p < 0.05$). The superior performance of the HADE I model can be attributed to its two-stream architecture, which captures both spatial and temporal information, thereby enhancing video-based human action recognition across multiple epochs.

1) ACTION CATEGORY PERFORMANCE

Table 8 presents the evaluation metrics for comparing the performance of HADE I and HADE II video classification models across different action categories.

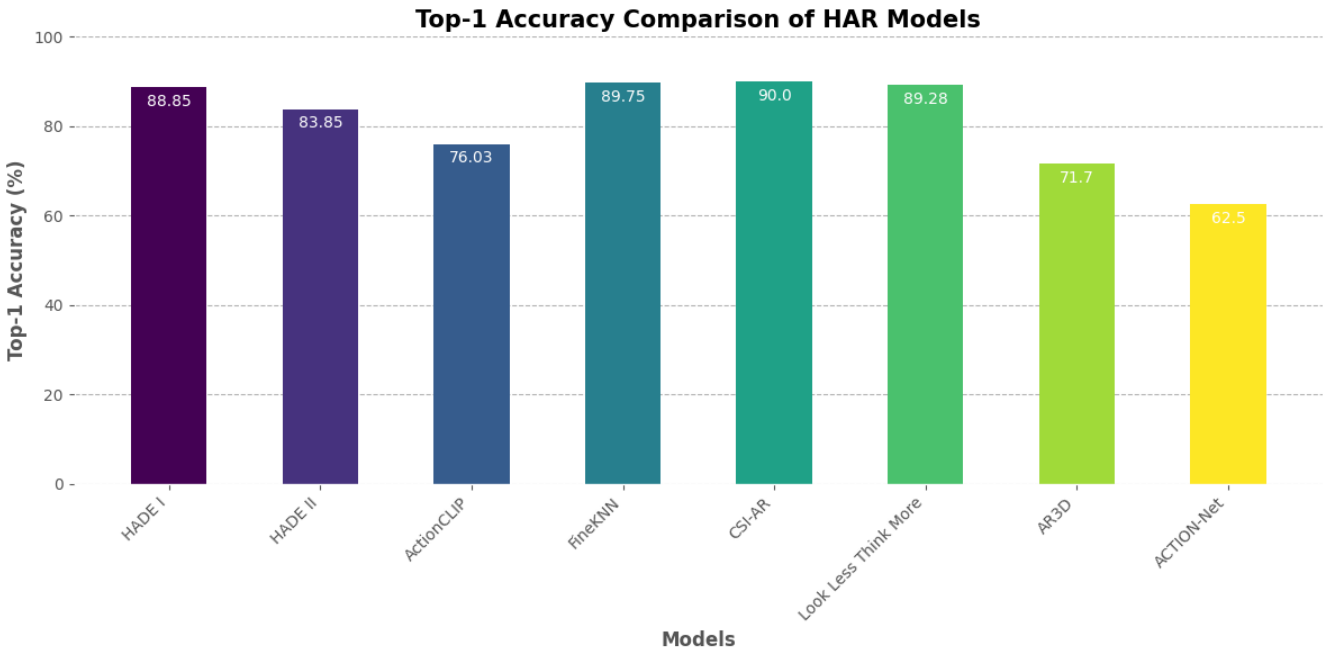

FIGURE 7. Bar chart showcasing the top-1 accuracy comparison among various HAR models, including HADE I and II, ActionCLIP, FineKNN, CSI-ARIL, "Look Less Think More", AR3D, and ACTION-Net. This visualization underscores the competitive performance of HADE models against recent state-of-the-art methods, highlighting their robustness and efficacy in action recognition tasks.

TABLE 7. Ablation study results and comparison with state-of-the-art HAR models.

Model	Year	Dataset	Method	Accuracy (%)	Observations
ActionCLIP [76]	2023	Kinetics-400	Vision-language model	76.03	High computational costs and potential generalization issues.
FineKNN [77]	2023	MCAD, IX-MAS	Skeleton-based FineKNN with EFS	89.75	Not explicitly mentioned.
CSI-ARIL [78]	2023	Not specified	CSI-based cross-scene recognition	90.0	Not provided.
Less Think More [79]	2022	STH-ELSE	Contrastive learning with common sense emphasis	71.7	Challenges in adapting to different action categories.
AR3D [80]	2021	UCF101	Attention Residual 3D Network	89.28	Increased model complexity and training difficulty.
ACTION-Net [81]	2021	SomethingV2	Multipath Excitation	62.5	High computational requirements.
HADE I	2023	Self-created dataset	SlowFast model	88.85	Limited to fundamental actions.
HADE II	2023	Self-created dataset	3D convolutions	83.85	Similar to HADE I, with scope for further enhancement.

TABLE 8. Comparison of proposed methods performance metrics.

Model	Precision	Recall	F1-Score
HADE I	0.90	0.88	0.89
HADE II	0.87	0.86	0.86

Figure 9 compares the F1-score performance of the HADE I and HADE II models for different actions. It illustrates that the HADE I model generally achieves higher F1 scores, except for the "Clapping" action, where the HADE II model outperforms it.

Figure 10 shows the recall performances of the two models for different actions. It indicates that the HADE II model has higher recall scores for "Clapping" and "Walking," while the HADE I model performs better for "Sitting down" and "Standing up."

Figure 11 displays the precision performance of HADE II, which outperforms HADE I in "Clapping" and "Walking" actions. Conversely, HADE I outshines HADE II in "Sitting down" and "Standing up" actions.

The comparative performance of the HADE I and HADE II models, as analyzed across different metrics and action types, underscores the importance of choosing a model that aligns with the specific accuracy and speed requirements of an application. HADE I excelled in accuracy, while HADE II offered quicker results. The decision on which model to utilize hinges on the particular demands of the application in question.

Table 9 presents an in-depth comparison of the performance indicators for various action categories within the HADE I and HADE II models, illustrating the

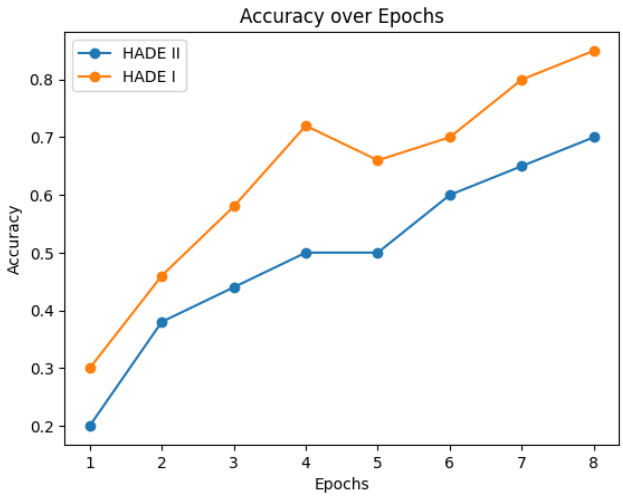

FIGURE 8. Comparing Accuracy Improvement of HADE I and HADE II Models over training epoch (epoch).

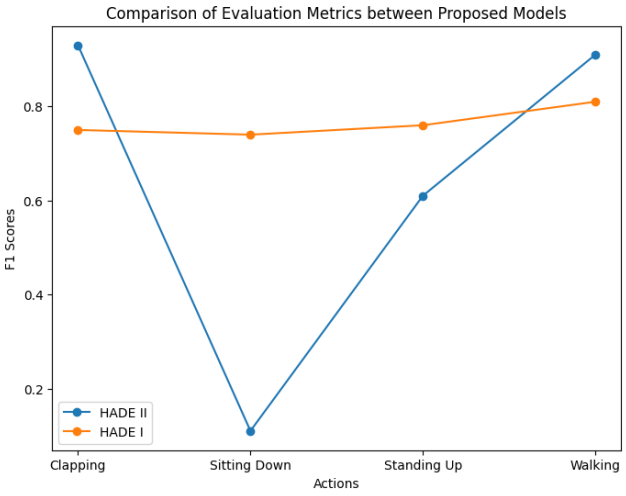

FIGURE 9. Comparing F1-Score Performance of HADE I and HADE II Models for Different Actions.

differences in precision, recall, and F1-score for each action.

D. ANALYSIS OF CONFUSION MATRICES

Confusion matrices were utilized to evaluate the performance of the HADE I and HADE II models in four basic activities: clapping, sitting, standing, and walking. These matrices provide essential details regarding the models' ability to recognize each action and the likelihood of misclassifying an activity. Confusion matrices were utilized to evaluate the performance of the HADE I and HADE II models in four basic activities: clapping, sitting, standing, and walking. These matrices reveal crucial details on the models' action detection capabilities and the likelihood of misclassification.

1) HADE I MODEL PERFORMANCE

The results displayed by the confusion matrices for the HADE I and HADE II models can be seen in Figure 12:

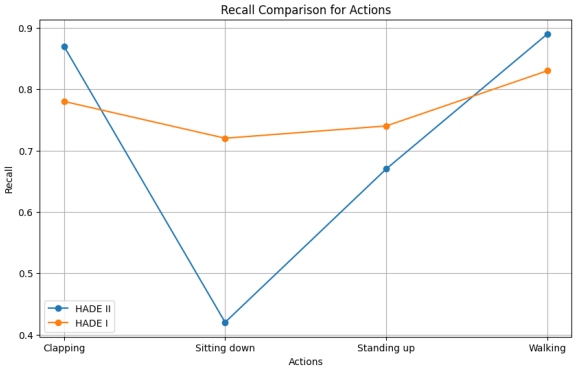

FIGURE 10. Comparing Recall Performance of HADE I and HADE II Models for Different Actions.

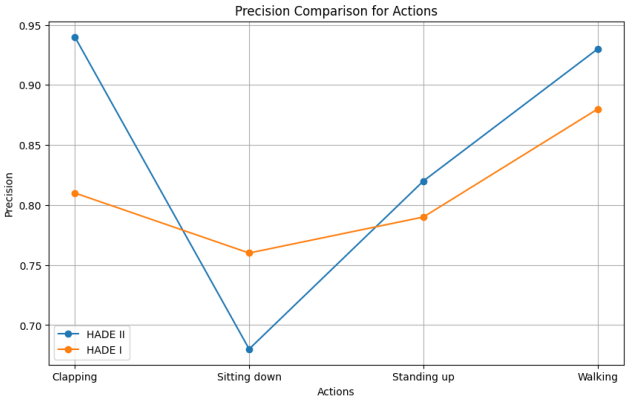

FIGURE 11. Comparing Precision Performance of HADE I and HADE II Models for Different Actions.

- **Clapping:** When it comes to detecting clapping gestures, the system shows impressive performance, achieving a TP rate of 40.71 and a TN rate of 629.0, with no false positives (FP).
- **Sitting down:** Displayed a thorough performance, achieving a low False Positive rate of 2.22 and a high True Positive rate of 42.09. The TN rate of 626.78 for this activity indicates a satisfactory level of model accuracy.
- **Standing up and Walking:** These actions showed high TP rates (62.79 and 65.28, respectively), but also higher FP rates (33.81 for Standing up and 27.98 for Walking), suggesting a tendency to over-predict these actions.

2) HADE II MODEL PERFORMANCE

The confusion matrix for the HADE II model exhibits the following characteristics.

- **Clapping:** Notable improvement in TP (64.17) compared to HADE I, but with an increased FP rate of 4.10, indicating a higher likelihood of false alarms.
- **Sitting down:** The model struggled with this action, showing a very high FN rate of 64.86, meaning it frequently failed to recognize sitting-down actions.
- **Standing up and Walking:** While the TP rates were high (57.96 for Standing up and 87.12 for Walking), the FP rates were also significant (62.79 and 16.59,

TABLE 9. Model performance.

Model	Action	Precision	Recall	F1-Score	Support
HADE I	Clapping	1.00	0.59	0.75	69
	Sitting down	0.95	0.61	0.74	69
	Standing up	0.65	0.91	0.76	69
	Walking	0.70	0.96	0.81	68
HADE II	Clapping	0.94	0.93	0.93	69
	Sitting down	0.67	0.06	0.11	69
	Standing up	0.48	0.84	0.61	69
	Walking	0.84	0.99	0.91	88

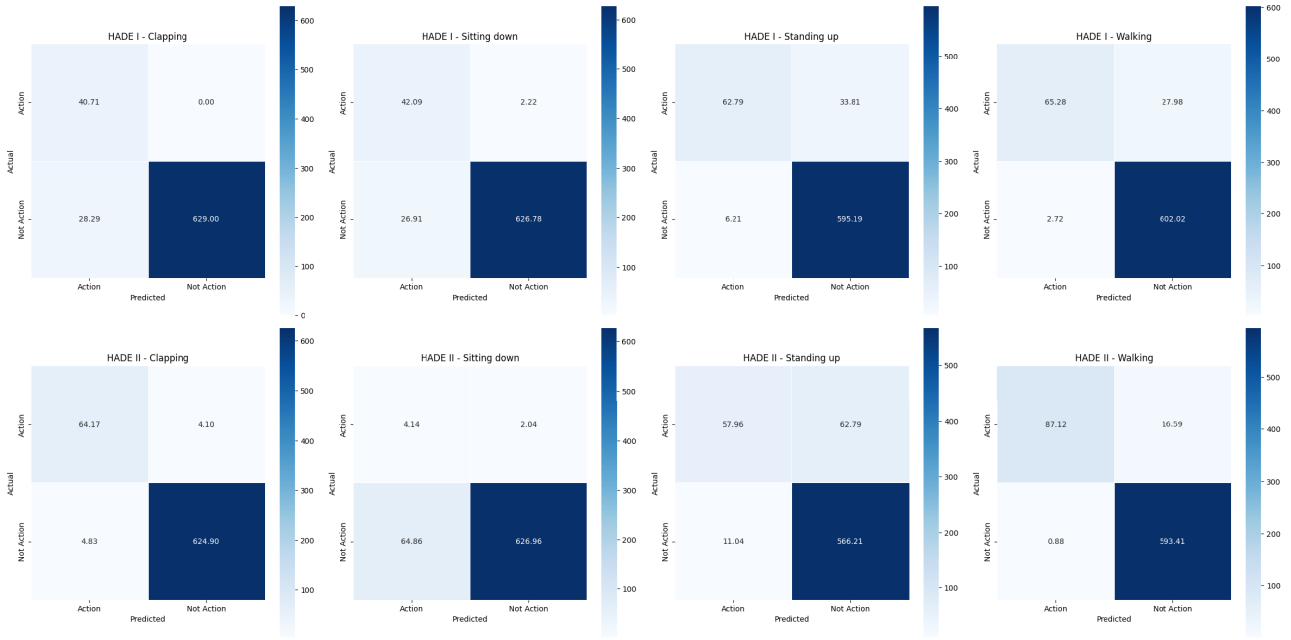

FIGURE 12. Confusion matrices for HADE I and HADE II models, illustrating classification performance across the actions: Clapping, Sitting down, Standing up, and Walking. These matrices highlight the comparative accuracy and the differing tendencies in false predictions between the two models for each action.

respectively), reflecting a tendency to incorrectly classify other actions as Standing up or Walking.

3) IMPLICATIONS

The analysis of these confusion matrices reveals that while both HADE I and HADE II are proficient in identifying specific actions, they exhibit different tendencies in terms of false positives and negatives. HADE I generally shows a conservative approach, with lower FP rates, but at times, missing out on certain actions (higher FN for specific actions). Conversely, HADE II tended to overclassify actions while achieving higher TP rates, leading to higher FP rates.

This analysis is crucial for understanding the practical applicability of these models in real-world scenarios. For applications where missing an action has significant consequences, HADE II's approach might be preferable, despite its higher FP rate. Conversely, in scenarios where false alarms are more critical, a conservative approach might be more suitable.

Overall, this detailed analysis assists in guiding further model improvements and choosing the appropriate model based on the specific requirements of the application domain.

VI. CONCLUSION AND FUTURE DIRECTIONS

This study represents a significant advancement in Human Action Recognition (HAR), a field that intersects computer vision and wireless computing. Although existing methodologies and datasets have made notable contributions to HAR, they also have certain limitations. Our proposed human actions in a diverse environment (HADE) framework seek to address these challenges by building upon the strengths of previous approaches.

The HADE architecture, characterized by its innovative leap in HAR, integrates a comprehensive dataset derived primarily from smartphone cameras. This approach captures a wide spectrum of human movements processed using the novel HADE I and HADE II models. These models employ advanced machine learning algorithms and GPU parallel processing to enhance the accuracy, precision, recall, and F1-score of HAR systems.

Our findings demonstrate the effectiveness of the HADE approach, achieving an accuracy of 83.57%, thereby significantly surpassing existing benchmarks. This improvement substantiates our hypothesis regarding the capability of the HADE approach to enhance recognition accuracy and overall system performance in HAR.

However, our study has some limitations that open avenues for future research. Currently, our framework focuses primarily on fundamental human actions such as clapping, walking, sitting down, and standing up. To increase the applicability and robustness of our model, future work will involve expanding the range of actions within the HADE dataset to cover more complex and diverse human activities. In addition, we aimed to test the practical deployment of our system in specialized domains, particularly in healthcare. Collaborating with medical experts to develop customized datasets for neurological patient monitoring could significantly enhance the utility and impact of our research on personalized healthcare applications. Moreover, while the HADE I and HADE II models have shown promising results, exploring more advanced machine-learning algorithms and techniques could further improve accuracy and efficiency. These future endeavors aim to transform the application and impact of HAR across various domains.

In summary, the HADE approach not only marks a forward leap in the field of HAR but also paves the way for new interdisciplinary applications. Our carefully curated and diverse dataset provides evidence of our dedication to continuous improvement and innovation in HAR, setting the stage for future breakthroughs, and expanding applicability in this dynamic field.

LIST OF ABBREVIATIONS

ACC	Accuracy. 14.
CNN	Convolutional Neural Network. 2.
Conv1	Convolutional layer 1. 11.
CSI	Channel State Information. 4.
CV	Computer Vision. 1.
epoch	training epoch. 18.
HADE	Human Action in Diverse Environments. 2, 3, 5, 11, 16, 19, 20.
HAP-DNN	Human Activity Pruning with Deep Neural Network. 4.
HAR	Human Action Recognition. 1-5, 9, 13, 19, 20.
I3D	Inflated 3D. 11.
MaxPool	Max pooling layer. 11.
ML	Machine Learning. 1, 2.
PCC	Precision. 14.
SOA	State-of-the-Art. 2, 3.
STAR	Spatio-Temporal Attention-based Transformer. 5.

REFERENCES

[1] S. Cho, E. M. Aiello, B. Ozaslan, M. C. Riddell, P. Calhoun, R. L. Gal, and F. J. Doyle, "Design of a real-time physical activity detection and classification framework for individuals with type 1 diabetes," *J. Diabetes Sci. Technol.*, vol. 17, Feb. 2023, Art. no. 193229682311538.

[2] K. Host and M. Ivašič-Kos, "An overview of human action recognition in sports based on computer vision," *Heliyon*, vol. 8, no. 6, Jun. 2022, Art. no. e09633.

[3] N. Gupta, S. K. Gupta, R. K. Pathak, V. Jain, P. Rashidi, and J. S. Suri, "Human activity recognition in artificial intelligence framework: A narrative review," *Artif. Intell. Rev.*, vol. 55, no. 6, pp. 4755–4808, Aug. 2022.

[4] W. Kang, S. Kim, J. Park, and S. Lee, "Review on human activity recognition using vision-based method," *J. Ambient Intell. Humanized Computing*, vol. 7, no. 1, pp. 119–133, 2016.

[5] T.-M. Wut, S. W. Lee, and J. Xu, "Mental health of working adults during the COVID-19 pandemic: Does physical activity level matter?" *Int. J. Environ. Res. Public Health*, vol. 20, no. 4, p. 2961, Feb. 2023.

[6] J. D. S. Duarte, W. A. Alcantara, J. S. Brito, L. C. S. Barbosa, I. P. R. Machado, V. K. T. Furtado, B. L. D. Santos-Lobato, D. S. Pinto, L. V. Krejcová, and C. P. Bahia, "Physical activity based on dance movements as complementary therapy for Parkinson's disease: Effects on movement, executive functions, depressive symptoms, and quality of life," *PLoS ONE*, vol. 18, no. 2, Feb. 2023, Art. no. e0281204.

[7] R. Pramanik, R. Sikdar, and R. Sarkar, "Transformer-based deep reverse attention network for multi-sensory human activity recognition," *Eng. Appl. Artif. Intell.*, vol. 122, Jun. 2023, Art. no. 106150.

[8] K. Thapa, Y. Seo, S.-H. Yang, and K. Kim, "Semi-supervised adversarial auto-encoder to expedite human activity recognition," *Sensors*, vol. 23, no. 2, p. 683, Jan. 2023.

[9] S. Ghazal, U. S. Khan, M. Mubasher Saleem, N. Rashid, and J. Iqbal, "Human activity recognition using 2D skeleton data and supervised machine learning," *IET Image Process.*, vol. 13, no. 13, pp. 2572–2578, Nov. 2019.

[10] Z. Lv, F. Poiesi, Q. Dong, J. Lloret, and H. Song, "Deep learning for intelligent human–computer interaction," *Appl. Sci.*, vol. 12, no. 22, 2022, Art. no. 11457.

[11] I. Khemapech, "Elder care—Importance, technologies, and opportunities," in *Proc. Int. Conf.*, Jun. 2021, pp. 1–16.

[12] M. Straczkiewicz, P. James, and J.-P. Onnela, "A systematic review of smartphone-based human activity recognition methods for health research," *NPJ Digit. Med.*, vol. 4, no. 1, p. 148, Oct. 2021.

[13] G. Diraco, G. Rescio, P. Siciliano, and A. Leone, "Review on human action recognition in smart living: Sensing technology, multimodality, real-time processing, interoperability, and resource-constrained processing," *Sensors*, vol. 23, no. 11, p. 5281, Jun. 2023.

[14] T. T. Zin, Y. Htet, Y. Akagi, H. Tamura, K. Kondo, S. Araki, and E. Chosa, "Real-time action recognition system for elderly people using stereo depth camera," *Sensors*, vol. 21, no. 17, p. 5895, Sep. 2021.

[15] A.-T. Shumba, T. Montanaro, I. Sergi, L. Fachechi, M. De Vittorio, and L. Patrono, "Leveraging IoT-aware technologies and AI techniques for real-time critical healthcare applications," *Sensors*, vol. 22, no. 19, p. 7675, Oct. 2022.

[16] Y. Zhang, B. Li, H. Fang, and Q. Meng, "Current advances on deep learning-based human action recognition from videos: A survey," in *Proc. 20th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2021, pp. 304–311.

[17] S. Zaidi, B. Jagadeesh, K. V. Sudheesh, and A. A. Audre, "Video anomaly detection and classification for human activity recognition," in *Proc. Int. Conf. Current Trends Comput., Electr., Electron. Commun. (CTCEEC)*, Sep. 2017, pp. 544–548.

[18] L. Minh Dang, K. Min, H. Wang, M. J. Piran, C. Hee Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognit.*, vol. 108, Dec. 2020, Art. no. 107561.

[19] A. Sunil, M. H. Sheth, and Mohana, "Usual and unusual human activity recognition in video using deep learning and artificial intelligence for security applications," in *Proc. 4th Int. Conf. Electr., Comput. Commun. Technol. (ICECCT)*, Sep. 2021, pp. 1–6.

[20] Y. Wang, S. Cang, and H. Yu, "A survey on wearable sensor modality centred human activity recognition in health care," *Expert Syst. Appl.*, vol. 137, pp. 167–190, Dec. 2019.

[21] X. Zhou, W. Liang, K. I. Wang, H. Wang, L. T. Yang, and Q. Jin, "Deep-learning-enhanced human activity recognition for Internet of Healthcare Things," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6429–6438, Jul. 2020.

[22] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6201–6210.

[23] V. Sharma, M. Gupta, A. K. Pandey, D. Mishra, and A. Kumar, "A review of deep learning-based human activity recognition on benchmark video datasets," *Appl. Artif. Intell.*, vol. 36, no. 1, Dec. 2022, Art. no. 2093705.

[24] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.

[25] B. Hussain, Q. U. Hasan, N. Javaid, M. Guizani, A. Almogren, and A. Alamri, "An innovative heuristic algorithm for IoT-enabled smart homes for developing countries," *IEEE Access*, vol. 6, pp. 15550–15575, 2018.

[26] K. Prokop, D. Polap, and G. Srivastava, "Neuro-heuristic pallet detection for automated guided vehicle navigation," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2022, pp. 6325–6331.

[27] C. He, Y. Du, and X. Zhao, "A separable convolutional neural network-based fast recognition method for AR-P300," *Frontiers Hum. Neurosci.*, vol. 16, Oct. 2022, Art. no. 986928.

[28] M. A. Khan, T. Akram, M. Sharif, N. Muhammad, M. Y. Javed, and S. R. Naqvi, "Improved strategy for human action recognition; experiencing a cascaded design," *IET Image Process.*, vol. 14, no. 5, pp. 818–829, Apr. 2020.

[29] Y. Zhou, Z. Yang, X. Zhang, and Y. Wang, "A hybrid attention-based deep neural network for simultaneous multi-sensor pruning and human activity recognition," *IEEE Internet Things J.*, vol. 9, no. 24, pp. 25363–25372, Dec. 2022.

[30] G. Diraco, G. Rescio, A. Caroppo, A. Manni, and A. Leone, "Human action recognition in smart living services and applications: Context awareness, data availability, personalization, and privacy," *Sensors*, vol. 23, no. 13, p. 6040, Jun. 2023.

[31] F. Serpush and M. Rezaei, "Complex human action recognition using a hierarchical feature reduction and deep learning-based method," *Social Netw. Comput. Sci.*, vol. 2, no. 2, p. 94, Apr. 2021.

[32] I. Jegham, A. Ben Khalifa, I. Alouani, and M. A. Mahjoub, "Vision-based human action recognition: An overview and real world challenges," *Forensic Sci. Int., Digit. Invest.*, vol. 32, Mar. 2020, Art. no. 200901.

[33] Y. Wang, S. He, X. Wei, and S. A. George, "Research on an effective human action recognition model based on 3D CNN," in *Proc. 15th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Nov. 2022, pp. 1–6.

[34] K.-X. Chen, J.-Y. Ren, X.-J. Wu, and J. Kittler, "Covariance descriptors on a Gaussian manifold and their application to image set classification," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107463. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320320302661

[35] Y. Zhang, Z.-H. Liu, X.-J. Wang, C. Liu, and X.-F. Wang, "Human action recognition using Kinect sensor data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 128–141, Jan. 2018.

[36] X. Wang, H. Lu, H. Ma, and M. Fang, "PCANet-top: A deep convolutional neural network for action recognition in videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 3, pp. 772–786, Jan. 2018.

[37] H. Nguyen, G. Willems, and T. Tuytelaars, "Sparse coding-based action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 714–730.

[38] H. Lu, "Two-stream networks for object segmentation in videos," 2208, *arXiv:2208.04026*.

[39] Y. Bengio, A. Courville, and I. Sutskever, "The state of the art in unsupervised learning," 2013, *arXiv:1301.3781*.

[40] C.-Y. Wu, R. Girshick, K. He, C. Feichtenhofer, and P. Krähenbühl, "A multigrid method for efficiently training video models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 150–159.

[41] W. Li, N. Xu, G. Liu, L. Zhao, and X. Fang, "Segments-based 3D ConvNet for action recognition," *J. Phys., Conf. Ser.*, vol. 1621, no. 1, Aug. 2020, Art. no. 012042.

[42] L. Sigal and M. Irani, "Pushing the limits of 3D human action recognition with large-scale datasets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Sep. 2011, pp. 3257–3264.

[43] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2659–2668.

[44] M. G. Morshed, T. Sultana, A. Alam, and Y.-K. Lee, "Human action recognition: A taxonomy-based survey, updates, and opportunities," *Sensors*, vol. 23, no. 4, p. 2182, Feb. 2023.

[45] R. Adaimi and E. Thomaz, "Lifelong adaptive machine learning for sensor-based human activity recognition using prototypical networks," *Sensors*, vol. 22, no. 18, p. 6881, Sep. 2022.

[46] A. Birhane, "The limits of fairness," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Jul. 2022, p. 2.

[47] X. Yang, D. Zhai, R. Zhang, H. Cao, S. Garg, and M. M. Hassan, "Human-to-human interaction behaviors sensing based on complex-valued neural network using Wi-Fi channel state information," *Future Gener. Comput. Syst.*, vol. 148, pp. 160–172, Nov. 2023, doi: 10.1016/j.future.2023.05.031.

[48] S. Das, A. Chaudhary, F. Bremond, and M. Thonnat, "Where to focus on for human action recognition?" in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 71–80.

[49] J. Wang, Z. Shao, X. Huang, T. Lu, R. Zhang, and X. Lv, "Spatial–temporal pooling for action recognition in videos," *Neurocomputing*, vol. 451, pp. 265–278, Sep. 2021.

[50] C.-L. Zhang, X.-X. Liu, and J. Wu, "Towards real-time action recognition on mobile devices using deep models," 1906, *arXiv:1906.07052*.

[51] Y. Liu, H. Wang, Y. Zhang, C. Liu, and X.-F. Wang, "Hap-DNN: A hybrid attention-based deep neural network for simultaneous multi-sensor pruning and human activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 4649–4662, Dec. 2022.

[52] R. Huang, X.-F. Wang, and C. Liu, "Graph-based skeleton-driven action recognition with graph convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 12, pp. 4629–4648, Sep. 2020.

[53] S. Chen, Y. Xu, Z. Pu, J. Ouyang, and B. Zou, "SkeletonPose: Exploiting human skeleton constraint for 3D human pose estimation," *Knowl.-Based Syst.*, vol. 255, Nov. 2022, Art. no. 109691.

[54] L. Sigal, M. Irani, and A. Geiger, "Humaneva: A high-resolution human motion database," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2002, pp. II–930–II–937.

[55] S. Zhang, Y. Li, S. Zhang, F. Shahabi, S. Xia, Y. Deng, and N. Alshurafa, "Deep learning in human activity recognition with wearable sensors: A review on advances," *Sensors*, vol. 22, no. 4, p. 1476, Feb. 2022.

[56] X. Zhang, C. Xu, X. Tian, and D. Tao, "Graph edge convolutional neural networks for skeleton-based action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 8, pp. 3047–3060, Aug. 2020.

[57] N. Heidari and A. Iosifidis, "On the spatial attention in spatio-temporal graph convolutional networks for skeleton-based human action recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 7474–7483.

[58] F. J. Ordonez, X. Hua, Y. Yang, J. E. Gonzalez, and R. Chellappa, "Deep learning human action recognition from accelerometer data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2565–2577, Jan. 2016.

[59] X. Yang, F. J. Ordonez, and R. Chellappa, "Activity recognition using smartphone sensors: A survey," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 90–103, Jan. 2015.

[60] A. M. Helmi, M. A. A. Al-Qaness, A. Dahou, and M. A. Elaziz, "Human activity recognition using marine predators algorithm with deep learning," *IEEE Access*, vol. 11, pp. 340–350, 2023.

[61] D. Połap, "Neuro-heuristic analysis of surveillance video in a centralized IoT system," *ISA Trans.*, vol. 140, pp. 402–411, Sep. 2023.

[62] Z. Zhang and D. Ramanan, "Utkinect-action3D: A large-scale dataset for 3D human action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Sep. 2013, pp. 3239–3246.

[63] Z. Al-Halah, I. Khalil, N. Al-Madi, D. Ganesan, and G. Al-Regib, "Smartphone-based human activity recognition using crowdsourced data," *IEEE Trans. Mobile Comput.*, vol. 16, no. 2, pp. 487–499, Feb. 2017.

[64] J. Wan, W. Li, and L. Zhang, "Human activity recognition based on deep learning and swarm intelligence," *Sensors*, vol. 19, no. 15, p. 3312, 2019.

[65] S. Akhter, I. Essa, and A. Pentland, "Human motion capture data sets: A survey," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 491–500.

[66] Z. Li, C.-H. Shen, and G. Medioni, "3D human action recognition from depth data using spatio-temporal features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3250–3258.

[67] J. Huang, Y.-S. Lee, and D. Ramanan, "Activity-Net: A large-scale dataset for human activity recognition in unconstrained videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 4893–4901.

[68] Q. Wang, W. Ouyang, and D. Lin, "Action pairs: A new dataset and benchmark for 3D human action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3247–3256.

[69] X. Jin, Y. Chen, Z. Liu, and X. Wang, "Vibe-3D: A real-time 3D human action recognition system," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Nov. 2015, pp. 3265–3273.

[70] Y. Zhu, Q. Wang, W. Ouyang, and D. Lin, "H3D-MHAD: A large-scale 3D human motion dataset for action recognition and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3274–3282.

[71] T. Oreifej and J. Aggarwal, "MSRdailyactivity3D: A dataset for daily human activities captured using a depth camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2011, pp. 96–103.

[72] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 716–723.

[73] C.-J. Hsu, J.-L. Chen, and L.-G. Chen, "An efficient hardware implementation of hon4d feature extraction for real-time action recognition," in *Proc. Int. Symp. Consum. Electron. (ISCE)*, 2015, pp. 1–2.
[74] T. Perrett, A. Masullo, T. Burghardt, M. Mirmehdi, and D. Damen, "Temporal-relational CrossTransformers for few-shot action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 475–484.
[75] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, and M. Mueller-Freitag, "The 'something something' video database for learning and evaluating visual common sense," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5842–5850.
[76] H. Cheng, Y. Guo, L. Nie, Z. Cheng, and M. Kankanhalli, "Sample less, learn more: Efficient action recognition via frame feature restoration," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 7101–7110.
[77] N. U. R. Malik, U. U. Sheikh, S. A. R. Abu-Bakar, and A. Channa, "Multi-view human action recognition using skeleton based-FineKNN with extraneous frame scrapping technique," *Sensors*, vol. 23, no. 5, p. 2745, Mar. 2023.
[78] Y. Zhang, F. He, Y. Wang, D. Wu, and G. Yu, "CSI-based cross-scene human activity recognition with incremental learning," *Neural Comput. Appl.*, vol. 35, no. 17, pp. 12415–12432, Jun. 2023.
[79] R. Yan, P. Huang, X. Shu, J. Zhang, Y. Pan, and J. Tang, "Look less think more: Rethinking compositional action recognition," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 3666–3675.
[80] M. Dong, Z. Fang, Y. Li, S. Bi, and J. Chen, "AR3D: Attention residual 3D network for human action recognition," *Sensors*, vol. 21, no. 5, p. 1656, Feb. 2021.
[81] Z. Wang, Q. She, and A. Smolic, "ACTION-Net: Multipath excitation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13209–13218.

MISHA KARIM received the bachelor's degree in computer science from UET Taxila. She is currently pursuing the M.S. degree in information technology with NUST SEECS. During the bachelor's degree, she has conducted research on human action recognition (HAR) using 3-D point clouds and ML-based convolutional networks. She accomplished a range of projects, including the cargo management systems, the open innovation platform, data monetization systems, blockchain-based solutions, and metadata. She has experience as a technical writer and a research assistant in delivering IT, business, and education projects. She has published a paper on skeleton-based human-action recognition. Her research interests include HAR, web development, and machine learning. For more information, please visit www.linkedin.com/in/misha-karim.

SHAH KHALID received the M.S. degree from the University of Peshawar, Pakistan, and the Ph.D. degree from Jiangsu University, China. He is currently an Assistant Professor with the School of Electrical Engineering and Computer Science, National University of Science and Technology (NUST SEECS), Islamabad, Pakistan. He has involved in numerous research projects in Pakistan and other countries. His research interests include information retrieval, web search engines, scholarly retrieval systems, recommender systems, HAR, knowledge graphs, social web, real-time sentiment analysis, web engineering, text summarization, federated searches, and digital libraries. He is a Reviewer of different prestigious journals and conferences, including *Knowledge-Based Systems*, *Expert Systems With Applications*, IEEE Access, and *Journal of Information Science*. For more information, please visit his website at https://sites.google.com/view/shahkhalid.

ALIYA ALERYANI received the B.S. degree in computer science from King Abdulaziz University, Jeddah, Saudi Arabia, the M.S. degree in computer science from Middle Tennessee State University, TN, USA, and the Ph.D. degree from the University of East Anglia, Norwich, U.K., in 2022. She is currently an Assistant Professor with the College of Computer Science, King Khalid University, Abha, Saudi Arabia. Her research interests include handling uncertainty in artificial intelligence and deploying artificial intelligence in sustainable development.

NASSER TAIRAN received the first M.Sc. degree in software engineering from the University of Bradford, U.K., in 2005, and the second M.Sc. and Ph.D. degrees in computer science (AI) from the University of Essex, U.K. He is an Associate Professor and the Ex-Dean of the College of Computer Science, King Khalid University (KKU), Saudi Arabia. He has experience with various education quality assurance agencies, including ABET, U.K.; and NCA, Saudi Arabia. He has been actively involved in research production, innovation, accreditation, curriculum, new course plans, and academic program evaluations in Saudi Arabia, since 2016. He has published many research articles in refereed international journals, including IEEE, IET, Wiley, MDPI, and Springer journals. His research areas include evolutionary computation, single and multi-objective optimization, meta-heuristics, and image processing. He is working on various research projects of the Deanship of Scientific Research, KKU. He served as a session chair and an organized committee member for various conferences.

ZAFAR ALI received the M.Sc. degree in computer science M.S. degree in web engineering from the University of Peshawar, in 2011 and 2017, respectively, and the Ph.D. degree in computer science and engineering from Southeast University, China. He is currently a Postdoctoral Fellow with the School of Computer Science and Engineering, Southeast University. He has published more than 30 research papers in reputed conferences and SCI journals. His research interests include recommendation systems, information retrieval, natural language processing, graph embedding, deep learning, and machine learning. He is a Reviewer of different prestigious journals and conferences, including *Knowledge-Based Systems*, *AI Reviews*, *Information Fusion*, *Scientometrics*, *Soft Computing*, *Information Processing and Management*, and CIKM.

FARMAN ALI received the B.S. degree in computer science from the University of Peshawar, Pakistan, in 2011, the M.S. degree in computer science from Gyeongsang National University, South Korea, in 2015, and the Ph.D. degree in information and communication engineering from Inha University, South Korea, in 2018. He worked as a Postdoctoral Fellow at the UWB Wireless Communications Research Center, Inha University, from September 2018 to August 2019. He is an Assistant Professor with the Department of Applied AI, Sungkyunkwan University, South Korea. He has registered over four patents and published more than 100 research articles in peer-reviewed international journals and conferences. His current research interests include sentiment analysis, social networking analysis, medical informatics, machine learning and AI, recommendation systems, data Science, and applied fuzzy logic. He has been awarded with Outstanding Research Award (Excellence of Journal Publications-2017), and the President Choice of the Best Researcher Award during graduate program at Inha University. In 2022 and 2023, each year, he was presented among "TOP 2% Scientists in the World" by Stanford University for his career achievements.