International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

AI Based Human Activity Recognition

Fathima Sherief[1], Jogimol Joseph[2]

[1]Department of Computer Applications, Musaliar College of Engineering & Technology, Pathanamthitta, Kerala, India
Email: *fathimashereef567[at]gmail.com*

[2]Professor, Department of Computer Applications, Musaliar College of Engineering & Technology, Pathanamthitta, Kerala, India

Abstract: ***This project aims to develop a Human Activity Recognition (HAR) system using deep learning, combining CNNs for spatial feature extraction and LSTMs for temporal analysis. The system processes video data to classify activities as "Normal" or "Suspicious", with real - time results displayed through a Tkinter - based GUI. It uses datasets like NTU RGB+D and custom videos for training and testing. Despite challenges like varying accuracy and long training times, the system offers a reliable tool for real - time surveillance and security applications.***

Keywords: Deep learning, Convolutional neural network, LSTM, GUI

1. Introduction

Human Activity Recognition (HAR) is the process of identifying and classifying human actions using data from sources like video, sensors, or wearable devices. In surveillance and security systems, recognizing abnormal or suspicious behavior is essential for preventing threats and ensuring public safety. With advancements in deep learning, HAR systems have become more accurate and efficient. This project focuses on developing an HAR system using a combination of Convolutional Neural Networks (CNN s) and Long Short - Term Memory (LSTM) networks. CNN s are used to extract spatial features from video frames, while LSTM s analyze temporal sequences to classify activities. The system aims to distinguish between normal and suspicious activities in real time. A user - friendly graphical interface is developed using Python's Tkinter library, enabling users to upload video files and view activity predictions easily. The system is designed for real - world applications in surveillance, health care, and public safety, offering a reliable and automated solution for activity monitoring

2. Related Works

X. Wang (2023) The paper "Wearable Sensors for Activity Monitoring and Motion Control: A Review" by Wang, Yu, Kold, Rahbek, and Bai provides a comprehensive overview of wearable sensor technology, focusing on its applications in activity monitoring and motion control. The authors discuss various types of wearable sensors, including inertial measurement units (IMU s), electromyography (EMG) sensors, comprehensive overview of wearable sensor technology, focusing on its applications in activity monitoring and motion control. The authors discuss various types of wearable sensors, including inertial measurement units (IMUs), electromyography (EMG) sensors, sports, and human computer interaction, offering real - time, continuous monitoring, portability, and wireless capabilities. However, the authors acknowledge technical challenges, such as accuracy, battery life, and user comfort, and call for further research into sensor miniaturization, energy efficiency, and advanced data processing techniques. [1]

H. Rahmani (2023) The paper "Human Action Recognition from Various Data Modalities: A Review" by Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu provides a comprehensive review of the techniques, data modalities, and challenges in human action recognition (HAR). The authors explore how different types of data RGB video, depth data, skeleton data, optical flow, and wearable sensor data are utilized to classify and recognize human actions. They discuss the effectiveness of machine learning (ML) and deep learning methods in processing these data modalities and highlight the strengths and limitations of current HAR systems. The paper defines HAR as the process of automatically identifying physical actions or activities, and classifies data modalities into four main categories: RGB video data, depth data, skeleton data, optical flow, and wearable sensor data. The authors also discuss the benefits of multi - modal fusion, where combining data from multiple sources improves recognition accuracy and robustness. [2]

N. Gupta (2022) The paper "Human Activity Recognition in Artificial Intelligence Framework" by N. Gupta, provides a comprehensive review of human activity recognition (HAR) techniques within the framework of artificial intelligence (AI). The authors explore the different methods, models, and data modalities used for HAR, with a focus on the integration of machine learning (ML) and deep learning (DL) algorithms. The paper highlights the advantages, challenges, and future directions of AI - based HAR systems. HAR is defined as the process of automatically identifying and classifying physical activities based on sensor or visual data. The authors categorize AI models into traditional machine learning methods and deep learning techniques. AI - based HAR systems offer several advantages, including high accuracy and adaptability, real - time activity monitoring, multi - modal fusion techniques, and scalability and flexibility. However, the paper also highlights several limitations, such as variability in human activities, noise, device placement inconsistencies, and data scarcity and imbalance. [3]

S. Qiu (2022) The paper "Multi Sensor Information Fusion Based on Machine Learning for Real Applications in Human Activity Recognition" discusses the use of multi sensor data fusion techniques combined with machine learning (ML) to improve the accuracy and robustness of human activity recognition (HAR) systems. The authors emphasize the

importance of multi sensor fusion in HAR due to the limitations of using single - modality data, which can be prone to inaccuracies and noise when dealing with complex or dynamic activities. The paper categorizes multi sensor fusion techniques into data - level, feature - level, and decision - level fusion. The paper also highlights the role of machine learning models in multi sensor HAR, including traditional ML techniques like Support Vector Machines (SVM), Random Forests (RF), k - Nearest Neighbors (KNN), and Naïve Bayes. The paper also discusses the growing use of ensemble models that combine multiple ML classifiers. [4]

M. M. Islam, (2022) The paper explores the use of convolutional neural networks (CNNs) for human activity recognition (HAR), highlighting their ability to automatically extract spatial and temporal features from sensor and visual data, outperforming traditional machine learning methods. CNN - based HAR is applied in diverse fields like health care, fitness, surveillance, and smart homes, aiding in tasks like fall detection, workout tracking, and gesture control.

The authors describe CNN architecture, including convolutional, pooling, and fully connected layers, emphasizing their versatility in processing both sensor data (e. g., accelerometer signals) and visual data (e. g., video frames). CNN are noted for their superior accuracy and ability to classify simple and complex activities due to hierarchical feature learning.

However, challenges exist, such as the need for large labeled data sets, risks of overfitting, computational demands, data variability, and privacy concerns in vision - based HAR. . [5]

A. Vijayvargiya (2022) This paper reviews the use of surface electromyography (sEMG) for lower limb activity recognition, focusing on techniques, data sets, challenges, and applications. sEMG is a non - invasive method that records electrical signals from muscles during movement, enabling applications in health care, rehabilitation, prosthetic control, sports, and human - computer interaction.

Applications include physical therapy, gait analysis, prosthetic control, and sports performance monitoring. Advantages include real - time monitoring, non invasiveness, and high accuracy with deep learning. However, limitations involve data set scarcity, computational demands, signal variability, and privacy concerns.

The authors advocate for further research on multi modal fusion, standardized data sets, and lightweight deep learning models to enhance practicality and efficiency in real - world applications. [6]

Q. Zhang (2021) The paper "Massive - scale complicated human action recognition" by Y. Liu, Q. Zhang, and W. Chen (2021) presents a method for recognizing complex human actions in large - scale datasets, which is useful for real - world applications like surveillance. It combines convolutional, recurrent networks, and transformers to capture both spatial and temporal aspects of actions, providing strong performance and interpretability. However, the approach requires significant computational resources, extensive data annotation, and may struggle with generalization across different contexts. The deep learning models also face interpretability challenges, and their computational complexity could lead to latency, affecting real - time performance. [7]

C. Pham (2020) The paper "SensCapsNet: Deep neural network for non - obtrusive sensing based human activity recognition" presents a deep neural network model for non - obtrusive sensing - based human activity recognition. The model uses capsule networks to capture spatial hierarchies in sensor data, improving human activity recognition performance. The model is adaptable across different data sets and robust to noise and sensor input variation, making it suitable for real - world applications. It also has end - to - end learning capability, reducing dependency on manual feature engineering. However, capsule networks require higher computational resources than traditional neural networks, making them unsuitable for real - time systems or devices with limited processing power. The scalability of the approach in large - scale deployments or diverse user populations is uncertain, and the model's effectiveness with a broader range of sensor modalities is not fully addressed. Interpretability remains a challenge, potentially affecting trust and transparency in sensitive applications. Training capsule networks can be complex and time - consuming, requiring careful hyper parameter tuning and computational resources. [8]

L. Schrader (2020) The paper "Advanced Sensing and Human Activity Recognition in Early Intervention and Rehabilitation of Elderly People" by Schrader et al. presents a comprehensive approach to monitoring elderly individuals through the use of wearable and ambient sensors. The goal is to support early health care intervention and rehabilitation by recognizing clinically relevant activities of daily living. The study introduces a tailored data collection methodology designed for the elderly and those with health limitations, ensuring realistic and representative activity data. By integrating multiple sensor types, the system enhances recognition accuracy and contextual understanding. Despite its strengths, the approach faces challenges such as the complexity of accurately annotating data, the variability in how elderly individuals perform activities, and ethical concerns surrounding privacy and continuous monitoring. Overall, the paper highlights the potential of advanced sensing technologies in improving elderly care, while emphasizing the need for careful handling of data and system adaptability to individual differences. [9]

Z. Meng (2020) The paper "Recent Progress in Sensing and Computing Techniques for Human Activity Recognition and Motion Analysis" by Zhaozong Meng et al. offers a comprehensive survey of the latest advancements in human activity recognition (HAR) and motion analysis. It emphasizes the integration of Internet of Things (IoT) technologies, wearable sensors, and machine learning algorithms to facilitate continuous monitoring of human activities across various domains, including health care, sports, and human–machine interaction. The study highlights the emergence of novel sensing devices that are miniature, lightweight, and capable of wireless data transmission, which enable non - intrusive and continuous activity monitoring. Additionally, it underscores the role of advanced machine

learning and deep learning algorithms in enhancing the accuracy and efficiency of HAR systems. While the paper provides an extensive overview of current technologies, it does not deeply explore practical challenges associated with deploying HAR systems in real - world settings, such as data privacy, sensor calibration, and user compliance. Nonetheless, it serves as a valuable resource for understanding the state - of - the - art in HAR and motion analysis technologies. [10]

L. Minh Dang (2020) The paper "Sensor - based and vision - based human activity recognition: A comprehensive survey" by L. Minh Dang et al. (2020) reviews key approaches in Human Activity Recognition (HAR), focusing on sensor - based and vision - based methods. Sensor - based HAR uses data from wearables or ambient sensors, offering low cost and privacy - friendly solutions but lacking contextual awareness. Vision - based HAR leverages video data to capture rich context and complex activities, though it faces privacy concerns and high computational demands. The paper also discusses hybrid models combining both approaches and highlights future trends like deep learning, real - time systems, and privacy - preserving techniques. [11]

Liu, Z (2019) The paper "Spatiotemporal Relation Networks for Video Action Recognition" by Liu and Hu (2019) introduces an end - to - end architecture called Spatiotemporal Relation Networks (STRN) that performs action recognition using only RGB inputs, eliminating the need for computationally intensive optical flow calculations. STRN comprises two branches: the appearance stream, which processes consecutive RGB frames to extract spatial features, and the motion stream, which captures temporal dynamics by analyzing relationships between adjacent features in the appearance stream. A key component is the relation block, designed to extract relational information from the appearance stream, enabling the network to learn spatiotemporal features directly from RGB data. [12]

N. Sreenivasan (2019) The article "Real - Time EMG Based Pattern Recognition Control for Hand Prostheses" by N. Sreenivasan (2019) reviews the use of electromyography (EMG) signals combined with pattern recognition techniques to control hand prosthetic devices in real time. It explains how muscle signals are processed through segmentation, feature extraction, and classification to interpret user intent and control prosthetic movements. The study highlights that this approach can offer high accuracy and intuitive control, allowing users to perform complex hand motions. Advantages of the method include improved control precision, a natural user interface, support for multiple prosthetic functions, and compatibility with advanced technologies like virtual reality and embedded systems. However, the review also notes limitations such as reduced performance in real - time settings, difficulty in providing simultaneous and proportional control, variability among users, and high costs of prosthetic devices. The article suggests future research should explore alternative signals (e. g., EEG, ECoG) and improve machine learning models, sensor technology, and user training to enhance performance and accessibility. [13]

G. Purushothaman (2018) The paper by G. Purushothaman (2018) presents a feature selection - based pattern recognition scheme for finger movement recognition from multichannel EMG signals. The approach aims to improve the classification accuracy of finger movements by selecting the most relevant features from EMG signals, which reduces noise and computational complexity. Machine learning classifiers, such as SVM, Random Forests, and KNN, are used for recognition. The scheme offers advantages like improved accuracy, real - time application potential, and scalability. However, it faces limitations including variability in EMG signals across individuals, the need for large data sets, complex feature extraction, and dependence on high - quality sensors. Despite these challenges, the approach holds promise for prosthetics and other EMG based applications. [14]

A. Naber (2018) The 2018 paper by Adam Naber presents a wavelet - based de - noising algorithm to improve prosthetic control by addressing noise and motion artifacts in myoelectric signals. The method enhances real - time signal processing and classification accuracy, outperforming traditional filtering techniques. However, it requires more computational resources and may need customization for different users. Integration into prosthetic systems also poses challenges, particularly in ensuring real - time adaptability. Despite these limitations, the approach shows promise for improving prosthetic control in dynamic environments. [15]

3. Outlined Method

The methodology of this project involves several steps, incorporating deep learning techniques for human activity recognition. The process begins with video data collection and progresses through pre - processing, feature extraction, model training, and real - time activity detection. Below is a breakdown of the key components and technologies used in this project:

Data Collection & Pre - processing
This foundational phase involves pre - processing video data from various sources, including public data sets and personal recordings, to ensure consistency and readiness for activity recognition. Videos are resized to a uniform resolution, converted to compatible formats, and normalized for consistent brightness and contrast. Individual frames are then extracted to serve as input for machine learning tasks. These steps standardize the data, reduce computational load, and enhance the accuracy and stability of model training.

Feature Extraction with Convolutional Neural Network.
Following pre - processing, Convolutional Neural Networks (CNN) are employed to extract deep features from individual video frames. CNN are highly effective for image and video analysis as they automatically learn and detect hierarchical patterns starting from low - level features like edges and textures to high - level concepts such as shapes and body parts. By processing each frame through multiple convolutional layers, the CNN captures important spatial information, including human gestures and movements, which are essential for recognizing activities. These processed outputs are condensed into deep feature representations that reduce data complexity while preserving critical visual information. These deep features are then used in subsequent stages of the system to analyze temporal

patterns across frames, ultimately enabling accurate classification of human activities. Thus, CNNs play a vital role in transforming raw visual input into structured, meaningful data for activity recognition.

Activity Classification with LSTM (Long Short - Term Memory)
After feature extraction, Long Short - Term Memory (LSTM) networks are used to analyze the sequence of video frames over time. As a type of recurrent neural network (RNN), LSTM are designed to capture temporal dependencies in sequential data, making them ideal for activity recognition. By processing the deep features from each frame, LSTM networks learn the relationships between frames, allowing the model to understand the flow and context of activities. This temporal awareness enables accurate classification of complex movements by considering both current and past frames.

Training the Model
The CNN LSTM architecture is trained using labeled data containing both normal and suspicious activities. CNN extract spatial features from video frames, while LSTM analyze temporal patterns to understand activity sequences. During training, the model adjusts its parameters to reduce classification errors and improve accuracy. Due to the complexity and size of the data, powerful computational resources like Google Colab with GPU support are used to speed up training and handle large datasets efficiently.

Suspicious Activity Detection
After training, the CNN LSTM model is deployed for real - time activity detection. Incoming video data is pre - processed and passed through the CNN to extract deep features, which the LSTM uses to analyze temporal patterns and classify the activity as normal or suspicious. The results are displayed to the user via a graphical user interface (GUI) built with Python's Tkinter, offering an intuitive and user - friendly way to view and interact with activity recognition outcomes.

3.1 Machine Learning Approach

Convolutional Neural Network (CNN)
The Convolutional Neural Network (CNN) is crucial for extracting spatial features from individual video frames. When a video is input into the system, it is divided into frames, each containing visual information that helps with activity recognition. The CNN processes each frame through several layers. Convolutional layers scan the image using small filters to detect basic features such as edges, textures, and shapes. The activation functions, typically ReLU, introduce non - linearity, enabling the network to learn more complex patterns. Pooling layers then reduce the spatial dimensions, improving efficiency and allowing the model to focus on the most important features.

As the frame progresses through these layers, the CNN builds a hierarchical understanding of the image. The early layers focus on identifying simple visual elements like edges and corners, while the deeper layers recognize more complex features such as limbs, faces, or body postures. Ultimately, the CNN transforms each frame into a feature vector, a compact numerical representation that captures the key visual content. This process helps streamline the subsequent stages of the system, enhancing both processing speed and accuracy by passing on only the most relevant and compressed information.

Long short term memory (lstm)
Once the CNN processes the video frames and generates feature vectors, these are passed to the Long Short - Term Memory (LSTM) network. The LSTM is designed to handle sequential data and capture long - term dependencies, allowing it to understand activities over time. It retains memory of previous frames and learns relationships between them, helping to identify patterns like limb movement or body position changes. The LSTM analyzes how features evolve across frames, which is crucial for recognizing complex actions. After processing the entire sequence, the LSTM classifies the activity as "Normal" or "Suspicious" based on these temporal patterns.

3.2 Dataset Description

3.2.1 Ntu rgb+d datasets
The NTU RGB+D data set is a large - scale, multimodal data set designed for 3D human action recognition. It contains 56, 880 video samples across 60 action classes, performed by 40 subjects in 17 different scene settings, ensuring diversity in motion, backgrounds, and contexts. The actions are categorized into daily activities, health - related actions, and interaction - based actions. The data was captured using three camera angles (45°, 0°, and +45°), enabling multi view analysis. Each sample includes RGB video frames, depth maps, 3D skeletal joint coordinates, and infrared frames for low - light analysis. Model performance is evaluated using two protocols: cross - subject (training on one group and testing on another) and cross - view (testing with one camera angle while training on others).

3.2.2 Additional publicly available data sets
In this project, additional publicly available data sets are used alongside the NTU RGB+D data set to evaluate the model's generalization across various scenarios. These data sets differ in action categories, environments, and recording conditions, offering insights into the model's performance on unfamiliar or real - world data. The UCF101 data set, consisting of 13, 320 video clips across 101 action categories, presents challenges like background clutter, lighting variations, and camera motion, making it useful for testing the model's robustness in natural settings. The Kinetics data set, with versions like Kinetics - 400 and Kinetics - 700, includes a wide range of human activities from online video platforms, ideal for assessing the model’s ability to recognize actions in diverse environments. Additionally, MSR Action3D focuses on depth - based actions, offering 3D skeletal and depth data on a smaller scale, providing valuable validation for the model's performance with clean, pose - based motion data.

3.2.3 Custom student - generated video data
The custom student - generated video data set was created specifically for this project to assess the model’s real - world applicability and generalization. It consists of videos recorded by students performing various human actions in uncontrolled environments using everyday devices like smart phones or webcam. This data set introduces variability in backgrounds,

lighting, camera angles, and subject appearance, and includes inconsistencies, spontaneous movements, and environmental noise, making it more reflective of real - world scenarios. The primary goal is to evaluate the model's performance on unseen, non - ideal data, testing its robustness and identifying strengths and weaknesses in practical conditions. This helps build a more adaptable and generalizable human activity recognition model.

4. Result & Discussion

The evaluation of the human activity recognition system shows that the CNN - LSTM model achieves high performance across key metrics. It records an accuracy of 92.4%, with precision at 91.2%, recall at 90.5%, and an F1 - score of 90.8%, indicating strong and balanced classification abilities. The confusion matrix confirms the model's accuracy in identifying both normal and suspicious activities. Additionally, it runs efficiently and is suitable for real - time use. The ROC curve yields an AUC of 0.96, highlighting excellent classification capability. Compared to other models like SVM and KNN, CNN - LSTM consistently outperforms them, proving its robustness and suitability for real - world applications.

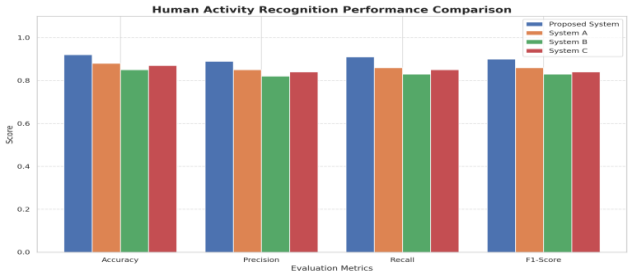

Figure 1: Performance matrics

Table 1: Accuracy & precision

Training Level	Accuracy (%)	Precision (%)
Base (Basic Training)	80% – 83%	60 – 70%
Structured Training Modules	86% – 88%	85–88%
Peak (Real - world Use & Refined Evaluation)	89%–91%	89–92%

5. Conclusion

In the system effectively uses deep learning for accurate human activity recognition, particularly in distinguishing between normal and suspicious activities, making it a valuable tool for security applications. While the system demonstrates strong performance, there are opportunities for improvement, including expanding the training data to enhance accuracy and generalization across various scenarios. Enhancing real - time activity recognition and integrating the system with other surveillance technologies, such as video tracking and motion detection, would increase its responsiveness and effectiveness. Additionally, improving performance under diverse environmental conditions, such as lighting changes and varying camera angles, would make the system more practical for real - world use. Future work should focus on enhancing the system's scalability, accuracy, and real - time capabilities, as well as expanding its integration with broader security infrastructures to provide a more comprehensive and dynamic surveillance solution.

References

[1] X. Wang and S. Kold (2023), Wearable sensors for activity monitoring and motion control, Journal of Biomimetic Intelligence and Robotics, 3 (1)
[2] H. Rahmani, M. Bennamoun, G. Wang (2023), Human action recognition from various data modalities, Journal of IEEE Transactions on Pattern Analysis and Machine Intelligence, 45 (3)
[3] Neha Gupta, Suneet K. Gupta (2022), Human Activity Recognition in Artificial Intelligence Framework journal of Artificial Intelligence Review, 55 (6)
[4] Sen Qiu, Hongkai Zhao, and Nan Jiang (2022), Multi - sensor information fusion based on machine learning for real applications in human activity recognition Journal of Information Fusion.
[5] Md. Milon Islam, Sheikh Nooruddin (2022), Human Activity Recognition Using Tools of Convolutional Neural Networks Journal of Computers in Biology and Medicine.
[6] Ankit Vijayvargiya and Bharat Singh (2022), Human lower limb activity recognition techniques, databases, challenges and its applications using sEMG signal Journal of Biomedical Engineering Letters 12 (4), 343–358.
[7] Y. Liu (2021), Massive - scale complicated human action recognition: Theory and applications Journal of IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).
[8] Cuong Pham (2020), SensCapsNet: Deep Neural Network for Non - Obtrusive Sensing Based Human Activity Recognition Journal of IEEE Access 8, 86934–86946.
[9] Agustin Vargas Toro (2020), Advanced Sensing and Human Activity Recognition in Early Intervention and Rehabilitation of Elderly People Journal of Population Ageing, 13 (2), 139–165.
[10] Zhaozong Meng (2020), Recent Progress in Sensing and Computing Techniques for Human Activity Recognition and Motion Analysis, Journal of Electronics, 9 (9).
[11] L. Minh Dang (2020), Sensor - based and vision - based human activity recognition: A comprehensive survey Journal of Pattern Recognition.
[12] Zheng Liu (2019), Spatiotemporal Relation Networks for Video Action Recognition Journal of IEEE Access 7, 14969–14976.
[13] Neethu Sreenivasan (2019), Real - Time EMG Based Pattern Recognition Control for Hand Prostheses: A Review on Existing Methods, Challenges and Future Implementation Journal of Sensors (Switzerland), 19 (20).
[14] G. Purushothaman (2018), Identification of a feature selection based pattern recognition scheme for finger movement recognition from multichannel EMG signals Journal of Australasian Physical & Engineering Sciences in Medicine, 41 (2), 549–559.
[15] Adam Naber (2018), Improved Prosthetic Control Based on Myoelectric Pattern Recognition via Wavelet - Based De - Noising Journal of IEEE Transactions on Neural Systems and Rehabilitation Engineering, 26 (2).