Human Activity Recognition with Hybrid Deep Learning: A Case Study Using CNN-LSTM and RFC-LSTM on the Human3.6M Dataset

Aun, Meng-Hing
[1] *Department of Computer Science and Information Engineering*
Chang Jung Christian University
Tainan, Taiwan
111b16885@mailst.cjcu.edu.tw

Yang, Pei-Ching*
[1]*Department of Computer Science and Information Engineering*
[2]*Artificial Intelligence Research Center*
Chang Jung Christian University
Tainan, Taiwan
0000-0002-2553-0695

***Abstract*—Human Activity Recognition (HAR) is essential for numerous real-world applications, including healthcare monitoring, intelligent surveillance, and innovative environments. This study compares two hybrid deep learning models—Convolutional Neural Network combined with Long Short-Term Memory (CNN-LSTM) and Random Forest Classifier combined with LSTM (RFC-LSTM)—using the Human3.6M dataset. The objective is to assess and compare their performance in terms of classification accuracy, robustness, and real-time monitoring capability. Experimental results reveal that CNN-LSTM outperforms RFC-LSTM in recognizing complex motion patterns, while RFC-LSTM demonstrates greater sensitivity to activity similarity, resulting in higher misclassification in overlapping actions. Both models, however, show limitations in consistent real-time performance due to data variability, overlapping behaviors, and architectural constraints. The study highlights the importance of incorporating attention mechanisms, multi-modal data fusion, and domain adaptation to enhance recognition stability. This work contributes to the advancement of HAR systems by analyzing the strengths and limitations of each hybrid model and offering insights for future improvements in spatial-temporal feature extraction and adaptive learning design.**

Keywords—Human Activity Recognition, Hybrid Deep Learning Models, Human3.6M Dataset

I. Introduction

Human activity recognition (HAR) has become a significant area of research in computer vision and machine learning due to its wide range of applications, including healthcare monitoring, surveillance systems, intelligent environments, and human-computer interaction. The ability to automatically detect and classify activities such as walking, sitting, smoking, greeting, and purchasing is crucial for developing intelligent systems capable of understanding and responding to human behavior in real time.

Traditional machine learning techniques have achieved moderate success in HAR tasks; however, the complexity and variability of human movements necessitate more advanced approaches. Hybrid deep learning models, which combine multiple learning paradigms, have emerged as a promising solution to enhance the accuracy and robustness of activity recognition systems. Among these, Random Forest Classifier combined with Long Short-Term Memory (RFC-LSTM) and Convolutional Neural Network combined with Long Short-Term Memory (CNN-LSTM) models have gained attention for capturing both spatial and temporal features of human activities.

This research aims to conduct a comparative analysis of RFC-LSTM and CNN-LSTM hybrid models in the context of human activity monitoring. By evaluating their performance on a diverse dataset of daily activities, this study seeks to determine each approach's effectiveness, strengths, and limitations. The findings of this research are intended to contribute to the development of more accurate and efficient HAR systems.

II. Literature Review

A. Potential Use Cases of Hybrid Models in HAR

- Healthcare: HAR systems using hybrid models can monitor patients with chronic conditions, such as Parkinson's disease, by recognizing specific movement patterns. For example, a CNN-BiLSTM model with undersampling achieved an accuracy of 98.5% on the MHEALTH dataset, making it suitable for medical emergency detection [4].
- Smart Homes: Hybrid models like ResBi-LSTM have been successfully applied in smart home environments to recognize activities such as eating, sitting, and walking, with an accuracy of 95.07% [5].
- Sports and Fitness: Wearable devices equipped with hybrid models can track and analyze athletic performance, providing insights into movement patterns and helping athletes optimize their training routines [3].
- Assistive Technology: HAR systems using hybrid models can assist individuals with disabilities by recognizing and interpreting their movements, enabling more independent living [1] [3].

B. RFC-LSTM and CNN-LSTM in HAR

The combination of CNNs and LSTMs is widely used in HAR. CNNs excel at extracting spatial features from sensor data, while LSTMs are adept at modeling temporal dependencies. For instance, a CNN-LSTM framework using surface EMG sensors achieved an accuracy of 87.18% in recognizing complex human activities [5]. While LRCN combines CNNs and LSTMs for sequential data processing. A ResBi-LSTM model achieved an accuracy of 95.07% in smart home environments, demonstrating its effectiveness in capturing long-term dependencies and reducing computational complexity [6]. The H36M dataset has been pivotal in evaluating these models, offering a comprehensive collection of 3D human poses across various activities such as walking, sitting, smoking, greeting, and purchasing [4].

979-8-3315-7517-5/25/$31.00 ©2025 IEEE

III. METHODOLOGY

This study adopts a systematic methodology to evaluate and compare the performance of CNN-LSTM and RFC-LSTM hybrid models in human activity recognition. The process starts with data collection and preprocessing, utilizing the Human3.6M (H36M) dataset [4]. The preprocessing stage involves cleaning, normalization, and preparing the 3D pose sequences to ensure they are suitable for model input.

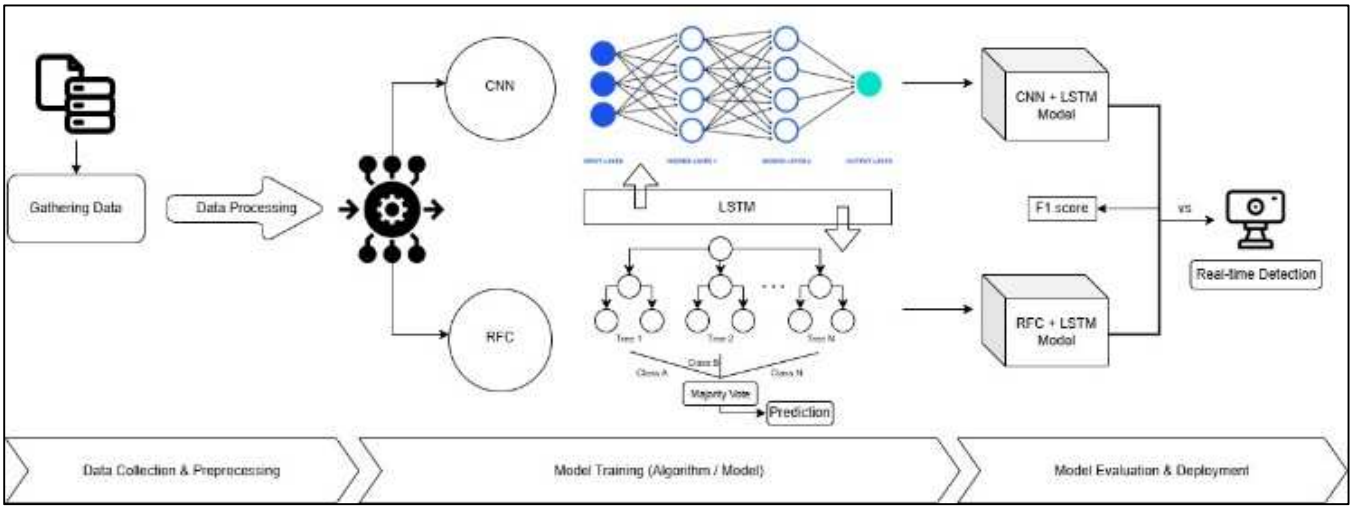

Fig. 1: Study Process: (a) Dataset Collection and Preprocessing; (b) Model Training; (c) Evaluation

A. *Dataset Collection and Preprocessing*

The Human3.6M dataset is selected due to its comprehensive coverage of daily human activities performed by 11 subjects in a controlled environment. This study uses data from six specific subjects (S1, S5, S6, S7, S8, S9 and S11) to ensure consistency and diversity in the training and evaluation process. The dataset comprises a broad range of activity categories (Table 1): Directions, Eating, Phoning, Posing, Sitting, Smoking, WalkDog, Walking, Discussion, Greeting, Photo, Purchases, SittingDown, Waiting, and WalkTogether. These categories provide a rich set of motion patterns that challenge the recognition capabilities of the hybrid models.

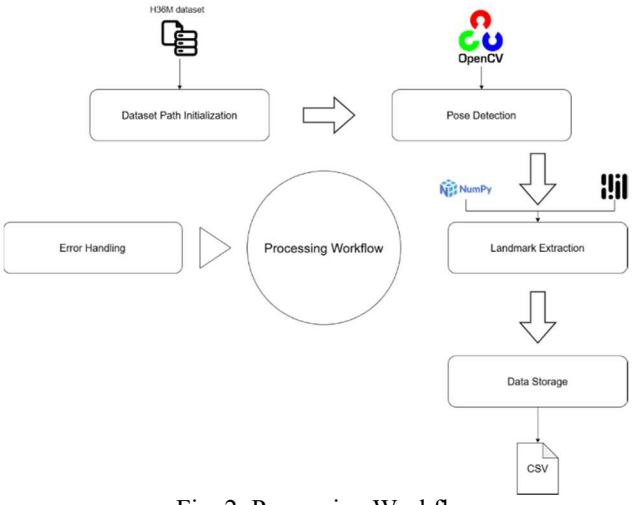

Fig. 2: Processing Workflow

The process begins with dataset path initialization, where the directories containing the raw data are identified and organized for sequential access. Pose detection was conducted using OpenCV integrated with the OpenPose library, which enables real-time detection of human skeletal keypoints. This facilitates consistent feature extraction across frames. This ensures the accurate localization of joints across all frames of an activity sequence.

TABLE I. ACTIVITIES AND STRUCTURED LABELS

Activity	Description
Directions	Actions involving giving or receiving directions.
Discussion	Poses and motions associated with conversational interactions.
Eating	Movements related to eating activities.
Greeting	Gestures like handshakes or waving.
Phoning	Actions depicting the use of a phone.
Photo	Poses related to taking or posing for photographs.
Sitting, SittingDown, Waiting	Variations of seated or waiting postures.
Purchases	Actions associated with shopping or transactions.
Walking, WalkDog, WalkTogether	Different walking-related scenarios.

It begins with dataset path initialization, organizing raw data directories for sequential access. Next, OpenCV performs pose detection, identifying key human joints in each frame using computer vision techniques. NumPy then extracts numerical joint coordinates, capturing spatial information and creating a high-dimensional feature vector. Normalization is applied to standardize joint coordinates, reducing scale and position variations.

To ensure data integrity, error handling addresses missing frames, inconsistent detections, and outliers. Issues are corrected via interpolation or flagged for exclusion.

Processed pose landmarks and activities are stored in CSV format, where each row represents a frame, and each column corresponds to a joint's X, Y, and Z coordinates. This

structured format enables efficient batch processing and seamless model training.

B. Model Training

The study aims to investigate and compare the performance of two hybrid deep learning models, CNN-LSTM and RFC-LSTM, in human activity recognition tasks. The Human3.6M dataset serves as the basis for the analysis, with the goal of evaluating these models in terms of classification accuracy, robustness, and real-time monitoring capabilities. In the CNN-LSTM model, CNN is responsible for extracting spatial features such as limb orientation and posture, while LSTM captures motion sequences over time. In the RFC-LSTM model, RFC pre-classifies static posture features before the LSTM analyzes their temporal dynamics.

- CNN-LSTM Model: This model integrates a Convolutional Neural Network (CNN) to extract spatial features from 3D pose sequences, and a Long Short-Term Memory (LSTM) network to capture temporal dependencies, thereby enhancing sequential pattern recognition.
- RFC-LSTM Model: In this model, a Random Forest Classifier (RFC) initially processes feature representations before feeding them into the LSTM network.
- Training Process: Both models undergo training using backpropagation with an adaptive optimizer to minimize loss functions. Hyperparameters, such as learning rate, batch size, and the number of LSTM units, are fine-tuned through cross-validation. Training continues until the models converge, ensuring optimal generalization for activity recognition tasks.

C. Evaluation

The Confusion Matrix and Accuracy will be used for model evaluation to assess the model's performance. The Confusion Matrix is a tool in supervised learning for analyzing classification models, offering a clear visualization of the model's predictive accuracy across different categories. It provides detailed insights into correct and incorrect classifications, helping to identify potential misclassifications and areas for improvement.

IV. Result

The confusion matrix showcases the results from the RFC-LSTM model. While this model still achieves high accuracy, it exhibits a broader spread of misclassifications compared to CNN-LSTM. For instance, activities like Directions (4978 correct classifications), Discussion (7859), and Eating (4844) show a reduction in classification performance compared to CNN-LSTM. The misclassification rates are higher for dynamic activities such as Walking, WalkDog, and Greeting, likely due to the RFC component's reduced ability to capture fine-grained spatial representations compared to convolutional layers.

A. CNN-LSTM Model Performance

The confusion matrix represents the results of the CNN-LSTM model. This model demonstrates high classification accuracy for most activity categories, particularly for frequently occurring activities such as Directions (5291 correct classifications), Discussion (8623), Eating (5935), and Phoning (6743). The strong diagonal presence in the heatmap confirms that the CNN-LSTM model effectively learns spatial and temporal dependencies of human activities. However, minor misclassifications occur in visually and kinematically similar activities such as Greeting and Posing, as well as Waiting and WalkTogether due to overlapping features in human motion. The overall classification accuracy achieved is 99.6 %, indicating a well-performing model.

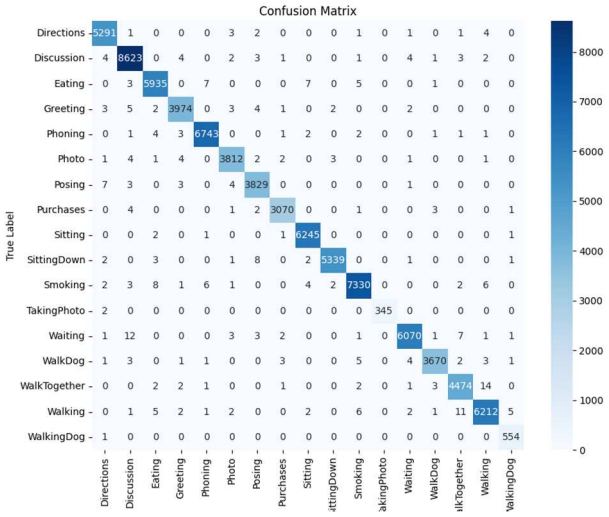

Fig. 3: Confusion Matrix – CNN-LSTM

B. RFC-LSTM Model Performance

The confusion matrix showcases the results from the RFC-LSTM model. While this model still achieves high accuracy, it exhibits a broader spread of misclassifications compared to CNN-LSTM. For instance, activities like Directions (4978 correct classifications), Discussion (7859), and Eating (4844) show a reduction in classification performance compared to CNN-LSTM. The misclassification rates are higher for dynamic activities such as Walking, WalkDog, and Greeting, likely due to the RFC component's reduced ability to capture fine-grained spatial representations compared to convolutional layers. The overall accuracy is relatively low, indicating potential issues with model performance, misclassifications, or dataset imbalance.

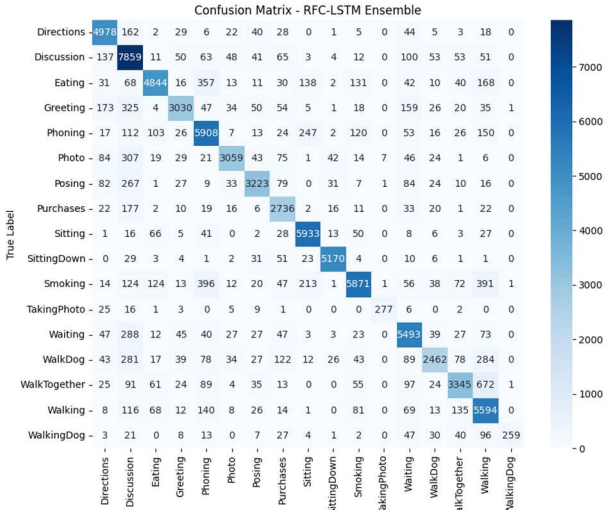

Fig. 4: Confusion Matrix – RFC-LSTM

The CNN-LSTM model exhibits higher classification accuracy across almost all activity categories. It particularly excels in distinguishing fine movements such as Phoning, Sitting, and Smoking, where precise spatial and temporal feature extraction is critical for accurate classification. Conversely, while the RFC-LSTM model remains effective, it demonstrates a higher misclassification rate. This suggests that the Random Forest classifier struggles to maintain the sequential integrity of human movements as effectively as CNN layers, leading to reduced classification precision in motion-dependent activities.

- Misclassification Trends: The RFC-LSTM model exhibits increased confusion between activities that share similar movement characteristics. For instance, Greeting and Discussion and Walking and WalkTogether show higher levels of misclassification. This pattern indicates that the RFC component has difficulty distinguishing between motions that involve subtle spatial variations. While the CNN-LSTM model performs better overall, it still struggles with minor misclassifications in activities involving similar body postures and hand movements, such as Smoking vs. Phoning. This suggests that further refinement—such as incorporating attention mechanisms—could improve differentiation between closely related activities even with convolutional feature extraction.
- Best Performing Model: The CNN-LSTM model emerges as the more accurate and reliable choice for human activity recognition. Its ability to extract spatial features through convolutional layers and model sequential dependencies with LSTMs ensures higher classification precision across diverse activity categories. In contrast, the RFC-LSTM model, while still valuable in specific contexts, struggles with motion continuity and spatial detail extraction. Its reliance on ensemble classification may limit its ability to discern fine-grained activity variations, making it less suitable for high-precision recognition tasks. However, it may still be advantageous when computational efficiency and model interpretability are critical.

V. CONCLUSION

This research compared CNN-LSTM and RFC-LSTM hybrid models for human activity recognition using the Human3.6M dataset. The study analyzed model performance through confusion matrices, classification accuracy, and real-world applicability, identifying key advantages and limitations of each approach.

The findings indicate that CNN-LSTM outperforms RFC-LSTM in classification accuracy, particularly for activities requiring detailed spatial-temporal feature extraction. Integrating convolutional layers enables CNN-LSTM to capture fine-grained spatial dependencies while the LSTM component effectively models sequential patterns. RFC-LSTM's relatively lower accuracy is attributed to the Random Forest component's inability to extract hierarchical spatial features, which are critical for distinguishing subtle differences in human posture across frames.

To enhance stability and accuracy, future research should explore attention mechanisms, multi-modal fusion, and domain adaptation techniques to capture contextual dependencies better and improve model robustness in dynamic settings. Additionally, real-time optimization strategies should be considered to facilitate practical deployment in applications such as healthcare monitoring, surveillance, and intelligent environments. While robustness was a stated objective, this study's evaluation is limited to the H36M dataset. Future work will involve testing on diverse datasets (e.g., NTU RGB+D, UTD-MHAD) to validate generalizability.

In conclusion, this study provides valuable insights into the trade-offs between deep learning and ensemble learning approaches for human activity recognition. While CNN-LSTM remains the more reliable model, advancements in hybrid architectures will be essential for achieving greater precision, adaptability, and real-world feasibility in human activity monitoring systems.

REFERENCES

[1] S. Abbaspour, A. Rezaei, H. Ghasemzadeh, and M. Akbari, "A comparative analysis of hybrid deep learning models for human activity recognition," Sensors, vol. 20, no. 19, p. 5707, 2020.

[2] N. A. Chandramouli, M. P. Khan, M. Sharma, and A. Joshi, "Enhanced human activity recognition in medical emergencies using a hybrid deep CNN and bi-directional LSTM model with wearable sensors," Sci. Rep., vol. 14, no. 1, p. 30979, 2024.

[3] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review," Comput. Vis. Image Underst., vol. 158, pp. 85–105, 2017.

[4] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 7, pp. 1325–1339, Jul. 2013.

[5] S. Mekruksavanich, T. Jitpattanakul, and W. Wuttisittikulkij, "Hybrid Attention with CNN-BiLSTM and CBAM for Efficient Wearable Activity Recognition," in Proc. 2024 Joint Int. Conf. Digit. Arts, Media Technol. & ECTI Northern Sect. Conf. Electr., Electron., Comput. Telecommun. Eng. (ECTI DAMT & NCON), 2024, pp. 572–576.

[6] S. Ostadabbas, A. Farzaneh, A. Banerjee, and T. Tasdizen, "Bayesian deep generative models for 3D human pose-based activity recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 6, pp. 1231–1244, Jun. 2021, doi: 10.1109/TPAMI.2019.2946361.

[7] S. V. Soman and J. R. A. Jeyaraj, "A hybrid approach to context-based human activity recognition in smart environments using ResBi-LSTM," Intell. Buildings Int., pp. 1–17, 2024.

[8] R. M. Vamsi, N. Adapa, D. Yelamanchili, N. A. Choudhury, and B. Soni, "An efficient and optimized CNN-LSTM framework for complex human activity recognition system using surface EMG physiological sensors and feature engineering," in Proc. Int. Conf. Smart Comput. Electron. Syst. (SCES), 2024, pp. 1