A Novel Approach for Human Activity Recognition Utilizing Modified Convolutional Neural Networks and Long Short-Term Memory Architectures

Hetal Z. Bhaidasna[1], Chirag Patel[2], Zubin C. Bhaidasna[3]

[1]Research Scholar,CE Dept, Parul University.
[2]HOD-CS Dept,Changa University.
[3]Asst. Prof., CSE Dept, GCET, CVM University, Vadodara, India.

Received: 10.04.2024 Revised : 12.05.2024 Accepted: 22.05.2024

ABSTRACT
The use of Human Activity Recognition in smart homes and health monitoring is gaining widespread acceptance as critical components of intelligent systems. State-of-the-art HAR approaches heavily rely on handcrafted features and less advanced learning techniques which might not capture human behavior's complex underlying patterns. This paper presents an approach for the HAR that integrates LSTMs and CNNs. In this model, LSTMs focus on disentangling timing, and ordering information within these features while CNNs learn to extract spatially distributed feature information in a fully automated manner. The recognition of different human activities is better with this fusion approach in terms of accuracy and reliability. Several experiments were conducted to verify the effectiveness of the proposed strategy using OWN and UCF-50 datasets.

Keywords: Machine Learning, Computer Vision, Deep Learning,CNN, RNN, LSTM

INTRODUCTION

Deep learning is a technique for realizing a subset of machine learning that imparts its efficiency to artificial neural networks having multi-layers for representing and extracting features from data. Deep Learning restrains neural networks towards multilayered structures (or deep networks) for automatic feature extraction and representation learning. These high-dimensional parameterized inter connectivities emulate the structure of the human neuron in the brain and are capable of directly capturing intricate patterns and representations from raw data. Feature engineering is not often required because deep learning algorithms learn the best way to represent input through intermediate representations learned as part of the network training process. It typically refers to the use of large amounts of labeled data to extract informative features at multiple scales with supervised classification, regression, unsupervised pertaining, or reinforcement learning objectives. The success of deep learning is ascribed to its scalability, versatility, and effective handling of high-dimensional data. Deep learning is a novel approach that uses neural networks to automatically recognize and comprehend patterns in human behavior, especially in the context of human activity detection. Through the analysis of raw data from several sources, including cameras, deep learning models are able to extract complex features and accurately classify varied behaviors.

Human Activity Recognition (HAR) is a branch of computer science and artificial intelligence that aims to spot actions or behaviors done by humans automatically. Human activity recognition tasks are of utmost importance with applications cutting across healthcare, surveillance, fitness tracking, smart homes, security, and robotics. The automatic identification and understanding of human activities from raw data streams has gained immense interest in the past few years. Even though traditional machine-learning approaches have recorded some success in this domain they often require handcrafted features not to mention their inability to adequately capture the complicated temporal dynamics which underlie human behavior. Spotting human activities (or HAR) has taken a "core-focused research area" tag borrowing from computer vision plus machine learning while steeling processing components.Its objective is to detect various movements and actions from video footage, encompassing a wide array of activities such as playing sports, studying, cooking, and more. It addresses the challenge of automatically identifying and interpreting human actions from raw data. HAR holds significant importance in various domains, particularly in healthcare where precise identification of activities like walking, sleeping, and exercising can aid in early health issue detection and improve patient care.

HAR is a fundamental type of research study for analyzing video data in surveillance applications by detecting unusual behavior for heightened security. Furthermore, HAR is a required type of research study for sports analytics to gain insights into performance, training systems, and injury prevention success rates. The accurate recognition of activities (e.g., running, jumping, throwing, etc.) will inform coaches and build a better training system. When seeing HAR, it commonly refers to machine learning-based techniques (e.g., supervised learning) for analyzing labeled datasets (e.g., sensor viewing experience) of data. Then, an algorithm detects and recognizes patterns and features of the data corresponding to the specific activity. After completing the supervised experience, the trained algorithm can classify unseen/new data for the appropriate activity. The end goal of HAR is to automate the ability to collect and analyze ongoing events of activity or behavior within video data and obtain insights into how humans may act or behave and interact with their environment. The challenges arise when you must recognize people and their spatial/temporally specific activities or course movements/interactions of themselves or with other objects/people. Typically, activities are organized by their functionality or complexity (1) single person player (e.g., distinct movements/actions) within the view area; (2) human-human/object interaction (e.g., punching, lifting); and (3) group activities (e.g., group dancing, capoeira, luxurious consumption). This organization provides the opportunity to study various human skills/behaviors and to develop machine learning algorithms that specialize and automatically recognize activities in video streams. HAR will expedite machines' ability to interpret and better understand human behavior to provide a new set of applications that we hope increase safety, convenience, and/or quality of life.

Background Study

Human activity recognition (HAR) has witnessed significant advancements driven by the integration of deep learning techniques, particularly through the application of transfer learning strategies. Sargano et al. [1] demonstrated the efficacy of transfer learning with deep representations in improving HAR accuracy by leveraging pre-trained convolutional neural networks (CNNs) on large-scale datasets. This approach allows models to transfer knowledge learned from one domain to another, thereby enhancing performance in new activity recognition tasks. Recent studies have further expanded the scope of HAR by integrating data from multimodal sensing devices, showcasing the benefits of combining diverse sensor modalities for improved recognition performance (Ihianle et al., 2020) [2].Architectural innovations such as LSTM-CNN hybrids have been pivotal in capturing both spatial and temporal dependencies from sensor data, facilitating more nuanced activity recognition (Xia et al., 2020) [3]. This approach combines the strengths of LSTM networks in modeling temporal sequences with CNNs' capability to extract spatial features, enabling robust performance across a variety of HAR tasks. Furthermore, advancements in sequential modeling using deep convolutional networks and extreme learning machines have contributed to the development of efficient HAR systems, particularly in wearable sensor applications (Sun et al., 2018) [4].

The application of recurrent neural networks (RNNs), including stacked gated recurrent units (GRUs), has proven effective in modeling complex human motion patterns (Wang et al., 2018) [5]. These networks excel in capturing long-term dependencies in sequential data, making them well-suited for HAR tasks where activities unfold over time. Context-aware recognition systems, such as hierarchical attention models, have addressed challenges in group activity recognition by incorporating contextual information and attention mechanisms (Kong et al., 2018) [6]. By focusing on relevant features and sequences, these models improve accuracy and robustness in identifying group dynamics and individual activities within them.Uddin et al. (2018) [7] explored the application of deep RNNs on translation and scale-invariant features for HAR, demonstrating their adaptability across different activity scales and environmental conditions. This approach enhances the generalizability of HAR models, enabling them to perform effectively in diverse real-world settings. Hassan et al. (2018) [8] extended HAR capabilities to smartphone-based systems, integrating deep learning models with smart sensors to achieve real-time activity detection and monitoring on portable devices.In sports and performance analysis, RNNs with LSTM architectures have been instrumental in real-time recognition and analysis of athletic movements (Wilton and Chen, 2018) [9]. These models support continuous learning and adaptation, crucial for capturing the nuances of dynamic sports activities. Structural recurrent neural networks (SRNNs) have provided insights into complex group dynamics, offering a structured approach to activity analysis and recognition (Biswas and Gall, 2018) [10]. By modeling interactions and dependencies between individuals in a group, SRNNs enhance the understanding of collective activities and behaviors.The integration of lightweight CNNs and efficient deep learning architectures has further optimized HAR systems for surveillance and real-world applications (Ullah et al., 2021; Jaouedi et al., 2020) [13], [14]. These models prioritize computational efficiency without compromising recognition accuracy, making them suitable for deployment in resource-constrained environments and real-time monitoring scenarios. Overall, these advancements underscore the transformative impact of deep learning methodologies on HAR, paving the way for more sophisticated and adaptive systems capable of understanding and responding to human actions across diverse environments.

Proposed Approach

In this study, we provide a unique hybrid model for human activity recognition (HAR) that makes use of both temporal and spatial information by combining Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs). Local dependencies and patterns are captured by the CNN component, which is made to efficiently extract spatial features from unprocessed sensor data. The LSTM component, which is skilled at understanding and modelling the long-term temporal dependencies present in sequential data, receives these features after that. A more reliable and precise recognition of intricate human behaviour is made possible by this combination. Benchmark UCF-50 & OWN datasets are used to train and assess the suggested model, which shows notable gains in accuracy over conventional techniques.

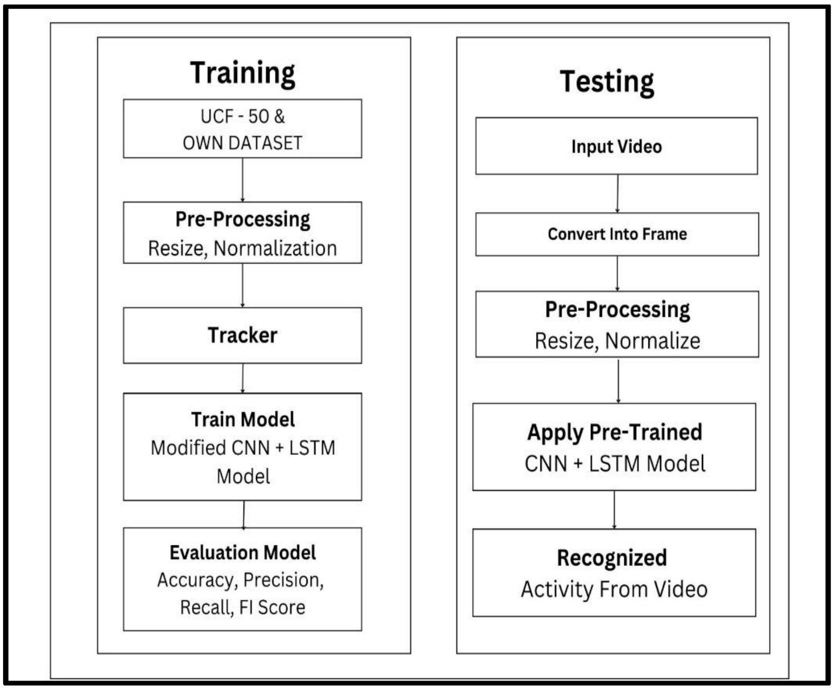


Fig 1. Proposed Model for HAR

The figure 1 shows the proposed diagram of human activity recognition. The research aims to improve the accuracy of human activity recognition using a hybrid model.In the first stage, the input to the model consists of sequences of sensor data or video frames that capture human activities. Each part of the system indicates a temporal segment of the activity recognition. A preprocessing step is first applied to the extracted frames. Preprocessing means reading each video frame one by one, transforming it into a common size, and then the video goes to tasks including further analysis or processing. Thus, the dimensions of the frames are kept uniform, which in turn allows effective processing of the video data in later phases of the workflow. A tracking algorithm can be used as a preprocessing step before sending the data through a Convolutional Neural Network (CNN) for feature extraction, allowing the input data to be converted into human activity recognition context.The frame is sent to the next model for feature extraction only if it contains an object detected by Region of Interest method. From a given image, locate all possible areas where objects are likely to be situated. The result should be several bounding boxes that may represent approximations of real object positions. These are often named region proposals or regions of interest. Tracking Algorithm in video sequences assigns unique identifiers to each person or object from one frame to another. By sustaining coherence and correspondence between frames about subjects, the algorithm makes sure their motion and interaction can be followed in time consistently.

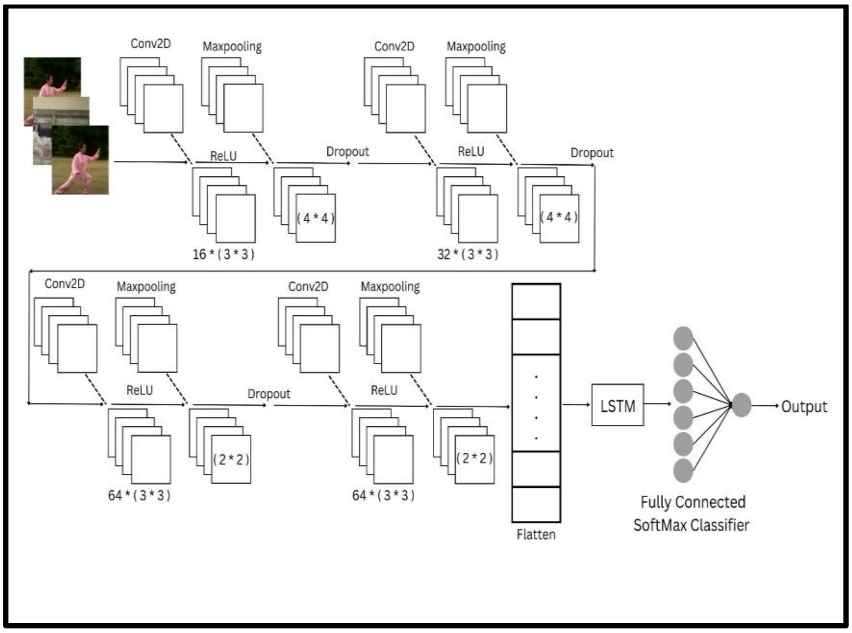


Fig 2. Modified CNN+LSTM model.

The figure 2 shows the modified CNN + LSTM model for human activity recognition, leveraging both spatial and temporal features. The model architecture begins with four convolutional and pooling layers designed to extract spatial features from input sequences, such as video frames or sensor data. These are an appropriate choice because the convolutional layers can effectively capture local patterns and dependencies for understanding the spatial aspects of human activities. In each convolution layer for regularization, we use the dropout technic. Following the convolutional layers, the extracted features are fed into a single LSTM layer that can learn well (specializes) from modeling temporal dependencies which enables the recognition of activities over time. This allows the model to handle complex and dynamic human activities since CNN is good at capturing local patterns (spatial information) while LSTM looks after temporal details. After the LSTM layer, a fully connected layer with a SoftMax activation function will be applied. This layer converts learned features into a probability distribution over possible activity classes— providing the final activity recognition output.The model is trained with the Adam optimizer. The Adam optimizer is known to be very efficient and effective when it comes to training deep learning models. The training parameters are a batch size of 20, a learning rate of 0.001, and 300 epochs. These parameters were selected to strike a balance between the time taken in model training and performance. For standardization of input dimensions, the input images are resized to 106 x 80 pixels. This step guarantees uniform sizes of input reaching the convolutional layers and thus more stable and efficient training in learning higher features. The presence of both convolutional layers for spatial feature extraction and LSTM for temporal feature learning makes the proposed model very robust and generalizable for human activity recognition tasks.

RESULT & DISCUSSION

This section describes the experimental results obtained on the UCF50 dataset and OWN dataset to show that the CNN + LSTM model is more effective for human activity recognition. The accuracy, precision, recall, and F1-scores of the model are calculated to prove its efficacy fully. The proposed CNN + LSTM model is applied to both the UCF50 and our custom OWN dataset to check its performance and generalization. A popular benchmark in human activity recognition, the UCF50 dataset has a large number of different activities that are very complex; thus, it is ideal for testing robustness. Our custom OWN dataset has activities like walking, reading, eating, and handshaking as well as upstairs and downstairs movements to further test applicability.

Realistic YouTube videos comprise the 50 action categories that make up the action recognition data set known as UCF50. The wide variations in camera motion, item appearance and posture, object scale, viewpoint, cluttered background, illumination conditions, etc. make this dataset extremely difficult to work with. The videos are divided into 25 groups, each with more than four action segments, for each of the 50 categories. Video clips belonging to the same group could have some things in common, such the same person, a similar background, similar opinions, and so on.

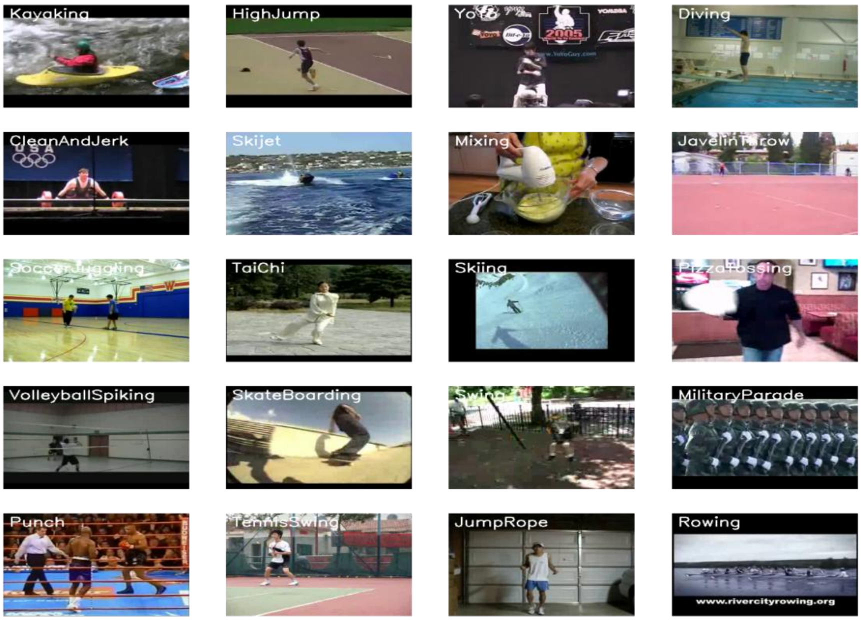


Fig 3. UCF-50 Dataset

We capture daily movements in great detail in our own dataset for human activity recognition. Every action is recorded in great detail, including eating, walking, reading, going upstairs or downstairs, and shaking hands. Every activity has an average 150 films, so we make sure that a variety of settings and situations are covered. Video records were used to catch every activity, guaranteeing a variety of human actions. Robust model training and evaluation are made possible by the carefully chosen dataset, which includes a balanced distribution of samples for each activity. Preprocessing techniques like segmentation and normalization were used to guarantee the quality and consistency of the data.

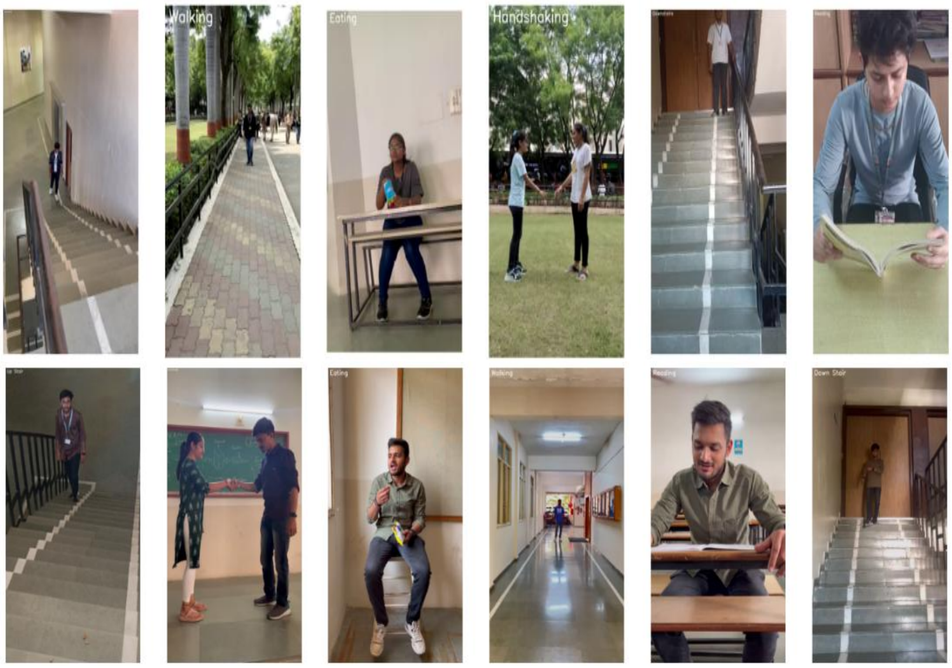


Fig 4. OWN Dataset

The figure shows the distribution of videos for each activity class in both the UCF-50 and custom datasets. Each bar indicates the number of videos per activity, clearly depicting the differences in class distribution across the two datasets.

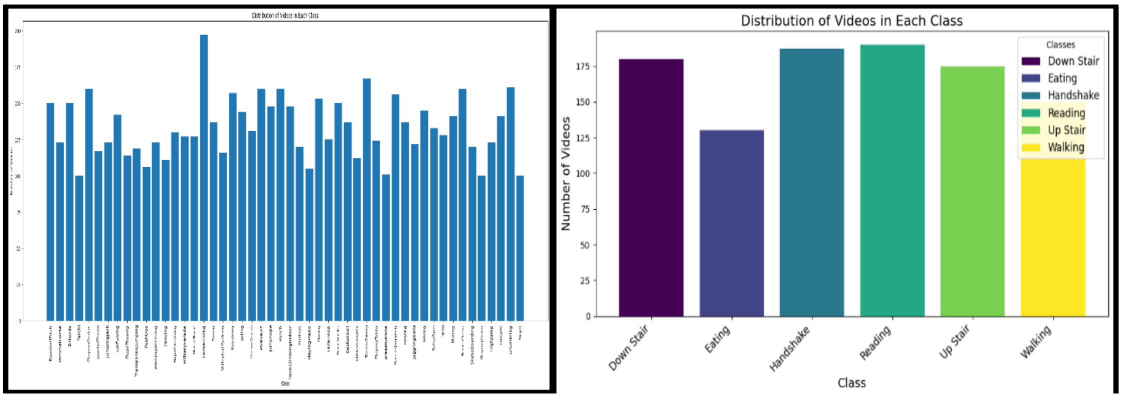


Fig 5. Distribution of Video per classes (Activity) in UCF-50 & OWN Dataset

Figure 6 shows the accuracy of the model over time for both the UCF-50 dataset and the custom dataset. On the y-axis, the accuracy is plotted, while the x-axis represents the number of epochs. The graph illustrates how the model's accuracy changes as training progresses through each epoch for both datasets.

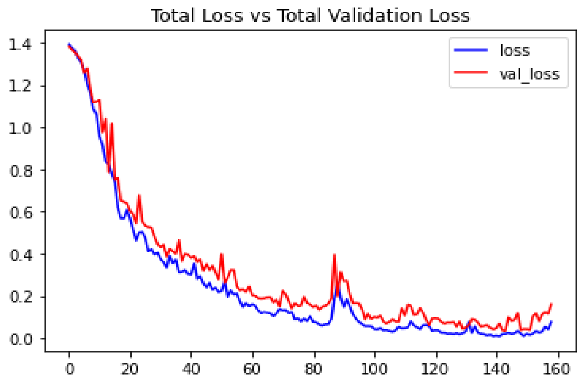


Fig 6. (a)Training vs. Validation Loss on UCF-50

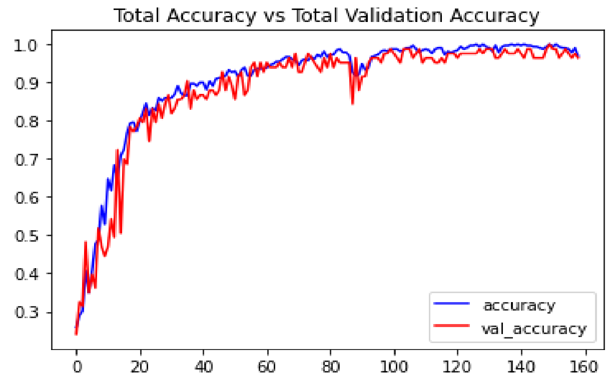


Fig 6.(b)Training vs. Validation accuracy on UCF-50

In our work, we divided this dataset into a training and testing set.The provided graphs illustrate the performance of a model trained on the UCF-50 dataset& OWN Dataset, with the first graph showing the training and validation loss and the second depicting accuracy. Initially, both loss curves decrease significantly, indicating effective learning, although the validation loss shows slight fluctuations later, possibly due to minor overfitting. The accuracy curves rise rapidly during the early epochs, eventually plateauing as the model stabilizes. The close alignment between training and validation accuracy suggests that the model generalizes well, a positive sign for its robustness on new data. The consistent performance, despite the complex nature of the UCF-50 dataset, underscores the model's capability in action recognition tasks.

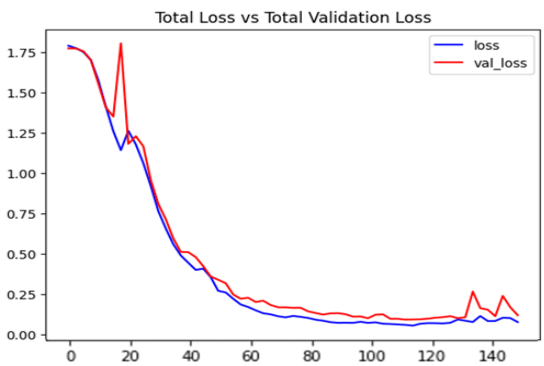


Fig 7.(a)Training vs. Validation Loss on OWN Dataset

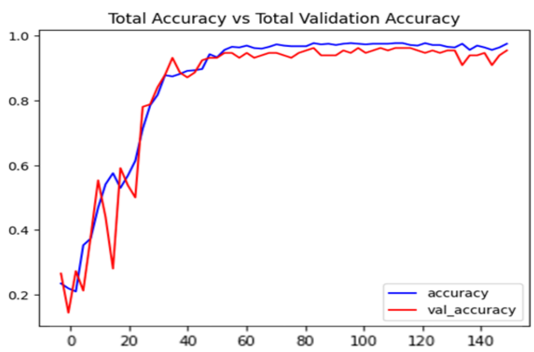


Fig 7.(b)Training vs. Validation accuracy on OWN Dataset

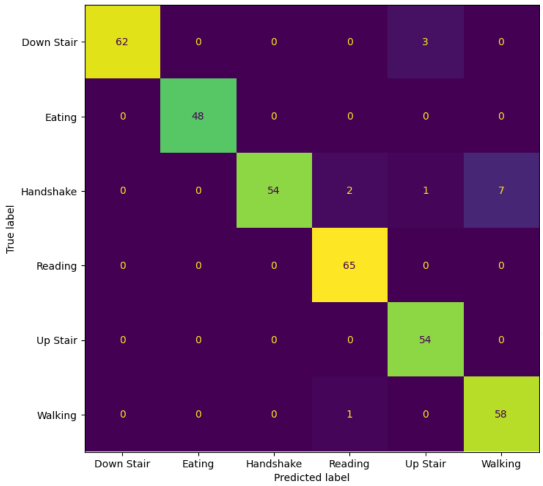
The study utilized a proprietary dataset comprising six distinct activities: Walking, Eating, Reading, Going Up Stairs, Going Down Stairs, and Handshake. The custom dataset, tailored to specific activities, provided additional insights into the model's adaptability. The model achieved an accuracy of 96.67% on OWN dataset. This performance underscores the effectiveness of the hybrid CNN + LSTM architecture in handling diverse and complex activities. The custom dataset results also revealed that the model could be further fine-tuned to improve recognition accuracy for less frequent or nuanced activities by incorporating more domain-specific training data.

Fig 8. Activity Recognition from Own Dataset

The UCF-50 dataset consists of videos depicting fifty different human actions or activities, encompassing a wide range of scenarios and movements. Each video clip in the dataset represents a specific activity, such as running, jumping, swimming, and dancing, among others. The dataset offers a diverse collection of activities captured from various perspectives and environments, including indoor and outdoor settings.On the UCF50 dataset, the proposed model achieved impressive results, with an average accuracy of 97.65%. The model demonstrated high precision and recall across most activity classes, indicating its capability to accurately recognize various human activities.

Fig 9. Video Recognition on UCF- 50

Below is the confusion matrix depicting the performance of the proposed model on our own dataset and the UCF-50 dataset. The rows represent the true labels, while the columns represent the predicted labels. Each cell in the matrix indicates the percentage of samples that were classified into the corresponding true label (row) and predicted label (column).



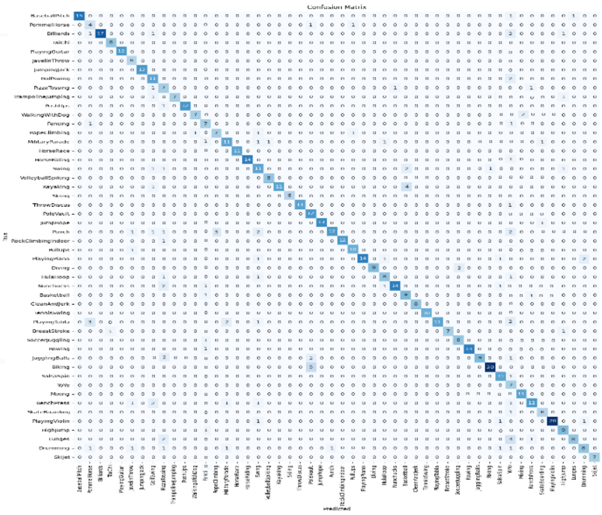


Fig 10. (a) Confusion Matrix of OWN Dataset **Fig 10.** (b) Confusion Matrix of UCF-50 Dataset

The proposed model's performance was evaluated on the UCF50 dataset and our custom dataset, with the results compared to existing models in the field. The table below summarizes the accuracy achieved by various models, including the proposed model.

Table 1. Comparison with state–of–the–art models with the proposed model

Sr. No	**Proposed Model**	**Dataset**	**Accuracy**
1	CNN+SVM+KNN[1],2017	UCF-50	91.47%
2	FlowNet2 CNN + LSTM [14], 2018	UCF-50	94.90%
3	CNN + LSTM [17] ,2021	UCF-50	95.20%
4	3D CNN + GRU [16] ,2023	UCF-50	87.78%
5	**Proposed Model** (Tracker + CNN + LSTM)	UCF-50	**97.65%**
6	**Proposed Model** (Tracker + CNN + LSTM)	OWN Dataset	**96.67%**

The proposed Tracker + CNN + LSTM model outperformed the existing models, achieving an accuracy of 97.65% on the UCF50 dataset. This significant improvement can be attributed to the effective combination of spatial feature extraction using CNNs, temporal sequence learning with LSTMs, and the inclusion of tracking mechanisms to better capture motion dynamics.On our custom dataset, which includes a diverse range of activities tailored to specific application scenarios, the model achieved an accuracy of 96.67%. This high accuracy indicates the model's robustness and ability to generalize well to new data. These results affirm the robustness and effectiveness of the model in accurately recognizing human activities across diverse datasets.the proposed Tracker + CNN + LSTM model demonstrates state-of-the-art performance in human activity recognition, with substantial improvements in accuracy on both the UCF50 and custom datasets.

CONCLUSION

In summary, our research effectively illustrated the use of a modified deep learning network for human activity recognition in video sequences. This network combines Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNNs) with the addition of a tracker. Utilizing this model on our own dataset and the UCF-50 dataset produced encouraging results, with 96.67% & 97.65% accuracy on our proprietary dataset, respectively. Our model is able to capture both spatial and temporal data from video sequences with robust activity recognition because to the fusion of CNNs and LSTMs. Furthermore, adding a tracker improves the model's capacity to track and identify people consistently across frames, which increases accuracy even further. We intend to expand our research in the future by investigating more datasets in order to confirm the generalizability and performance.

REFERENCES

[1] Sargano, Allah Bux, et al. "Human action recognition using transfer learning with deep representations." 2017 International joint conference on neural networks (IJCNN). IEEE, 2017.

[2] Ihianle IK, Nwajana AO, Ebenuwa SH, Otuka RI, Owa K, Orisatoki MO. A deep learning approach for human activities recognition from multimodal sensing devices. IEEE Access. 2020 Sep 30;8:179028-38.

[3] Xia K, Huang J, Wang H. LSTM-CNN architecture for human activity recognition. IEEE Access. 2020 Mar 20.

[4] Jian sun,Yongling FU, Shengguang Li, Jie He, Cheng Xu, Lin Tan , "Sequential Human Activity Recognition Based on Deep Convolutional Network and Extreme Learning Machine Using Wearable Sensors" ,Journal of Sensors,2018.

[5] Mingyang wang , Guolong cui, Xiaobo yang, Lingjiang kong , "Human body and limb motion recognition via stacked gated recurrent units network" IET Radar Sonar Navig, Vol. 12 Iss. 9, pp. 1046-1051,2018.

[6] Longteng kong, jie qin, di huang, yunhong wang and luc van gool , "Hierarchical attention and context modelling for group activity recognition", ICASSP-IEEE, 978-1-5386-4658-8/18 2018.

[7] Md. Zia Uddin, Weria Khaksar, Jim Torresen, "Activity Recognition Using Deep Recurrent Neural Network on Translation and Scale-Invariant Features", IEEE, 2018.

[8] Mohammed Mehedi Hassan, Md. Zia Uddin, Amr Mohamed, Ahmad Almogren , "A Robust Human Activity Recognition System using smart phone sensors and deep learning", Future Generation Computer Systems 81 (2018) 307–313-Elsevier,2018.

[9] Wilton, Loius CW, Carol Chen, "Artificial Intelligence for Sport Action and Performance Analysing using Recurrent Neural Network (RNN) with Long Short Term Memory(LSTM)", ACM,2018, ISBN 978-1-4503-6584-0/11.

[10] Sovan Biswas & Juergen Gall, "Structural Recurrent Neural Network (SRNN) for Group Activity Analysis", Winter Conference on Applications of Computer Vision-IEEE, 978-1-5386-4886-5,2018.

[11] Masaya Inoue, Sozo Inoue,Takeshi Nishida , "Deep Recurrent Neural Network for Mobile Human Activity Recognition with high Throughput", Artificial Life and Robotics-Springer,2017.

[12] Chandrashekar M Patil, Jagadeesh B, Meghana M N , "An Approach of Understanding Human Activity Recognition and Detection for Video Surveillance using HOG Descriptor and SVM Classifier" , ICCTCEEC-IEEE;978-1-5386-3243-7,2017.

[13] Li Wei & Shishir K. Shah, " Human Activity Recognition using Deep Neural Network with Contextual Information",In Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications-IEEE, pages 34-43,ISBN: 978-989-758-226-4,2017.

[14] Ullah, Amin, et al. "Activity recognition using temporal optical flow convolutional features and multilayer LSTM." IEEE Transactions on Industrial Electronics 66.12 (2018): 9692-9702.

[15] Haanvid Lee, Member, Minju Jung, and Jun Tani, "Recognition of visually perceived compositional human actions by multiple spatio-temporal scales recurrent neural network", IEEE Transactions On Cognitive And Developmental Systems, February 2017

[16] Vrskova, R., Kamencay, P., Hudec, R., & Sykora, P. "A new deep-learning method for human activity recognition. Sensors" Sensors 23.5 (2023): 2816

[17] Ullah, Amin, et al. "Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications." Applied Soft Computing 103 (2021): 107102.

[18] Jaouedi, N., Boujnah, N., & Bouhlel, M. S. ,"A new hybrid deep learning model for human action recognition. "Journal of King Saud University-Computer and Information Sciences, 2020

[19] Hui, Tak-Wai, Xiaoou Tang, and Chen Change Loy. "Liteflownet: A lightweight convolutional neural network for optical flow estimation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[20] Bolme, David S., et al. "Visual object tracking using adaptive correlation filters." 2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, 2010.

[21] Le, Manh-Cuong, and My-Ha Le. "Human detection and tracking for autonomous human-following quadcopter." 2019 International Conference on System Science and Engineering (ICSSE). IEEE, 2019.

[22] Oluwalade, Bolu, et al. "Human Activity Recognition using Deep Learning Models on Smartphones and Smartwatches Sensor Data." arXiv preprint arXiv:2103.03836 (2021).

[23] Zhuang, Zhendong, and Yang Xue. "Sport-related human activity detection and recognition using a smartwatch." Sensors 19.22 (2019): 5001.

[24] Bhaidasna, Mr Zubin C., and Priya R. Swaminarayan. "A SURVEY ON CONVOLUTION NEURAL NETWORK FOR FACE RECOGNITION." Journal of Data Acquisition and Processing 38.2 (2023): 4364.

[25] Joshi, Arpit, et al. "Human Activity Recognition Using Fusion of Features from Accelerometer and Gyroscope Data." Proceedings of the 2020 IEEE International Conference on Industrial Technology (ICIT). IEEE, 2020.

[26] Modi, Jhinal, Hetal Bhaidasna, and Jubin Bhaidasna. "Human Activity Recognition Using Deep Learning Methods: A Survey." (2020).
[27] Han, Junseok, and Minwoo Lee. "A Hybrid Deep Learning Framework for Human Activity Recognition with Local Features." Sensors 20.3 (2020): 850.
[28] Hetal Bhaidasna, et al. "A Survey on Different Deep Learning Model for Human Activity Recognition Based on Application". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 11, no. 10, Nov. 2023, pp. 1496-08, doi:10.17762/ijritcc.v11i10.8700.
[29] Lee, Kyuyeon, and Jun Mo Kim. "A Performance Comparison of 3D-CNN and LSTM for Human Activity Recognition." Proceedings of the 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019.
[30] Bhaidasna, Z. C. ., P. R. . Swaminarayan, and H. Z. . Bhaidasna. "Enhancing Face Recognition With Deep Learning Architectures: A Comprehensive Review". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 11, no. 9, Oct. 2023, pp. 164-80, doi:10.17762/ijritcc.v11i9.8331.
[31] Kautz, Tim, et al. "Activity Recognition in Working Environments with Wearable Sensors Using Deep Learning." Proceedings of the 2017 ACM International Symposium on Wearable Computers. ACM, 2017.