Human Activity Recognition: Deep Learning Classification for Recognizing Human Activities

P.Usha Manikyam[1], N.Prasannakumari[2], P.Suma Gayathri[3], Sk.Yaseen[4], M.Gopi Venkata Satish[5]
[1]*Assistant professor, Srinivasa Institute of Engineering and Technology*
[2345]*Student Scholar, Srinivasa Institute of Engineering and Technology*
doi.org/10.64643/IJIRTV12I10-195133-459

Abstract: Human Activity Recognition (HAR) is an important application of computer vision that focuses on identifying human actions from video data. In this paper, a real-time activity recognition system is developed using a combination of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The proposed approach captures both spatial and temporal information from video sequences to improve classification performance. Video frames are collected using a webcam and processed through resizing and normalization. The CNN model extracts visual features from individual frames, while the LSTM network analyses the sequence of frames to understand motion patterns over time. A fully connected Softmax layer is used to classify the activities into multiple categories. The model is trained using standard human activity datasets and evaluated in both offline and real-time environments. Experimental results show that the combined CNN-LSTM model provides better accuracy compared to using individual models. The system is efficient, scalable, and suitable for real-time applications such as surveillance, healthcare monitoring, and smart systems.

Keywords: Human Activity Recognition, Deep Learning, CNN, LSTM, Computer Vision, Real-Time Prediction, Video Classification

I. INTRODUCTION

Human Activity Recognition (HAR) is an important area of research in computer vision that focuses on identifying and classifying human actions from images or video data. With the rapid growth of artificial intelligence and deep learning technologies, HAR systems have become more efficient and capable of working in real-time environments.

HAR plays a key role in many real-world applications. It is widely used in surveillance systems to detect unusual activities, in healthcare to monitor patient movements and prevent accidents such as falls, and in smart homes to automate daily tasks. It is also useful in fitness tracking and gesture-based control systems.

Earlier approaches for activity recognition mainly depended on manual feature extraction techniques such as Histogram of Oriented Gradients (HOG), optical flow, and machine learning algorithms like Support Vector Machines (SVM). These methods required significant human effort and often failed to perform well in complex and dynamic environments.

In recent years, deep learning has improved the performance of HAR systems by automatically learning features directly from data. Convolutional Neural Networks (CNNs) are effective in extracting spatial information from images, while Long Short-Term Memory (LSTM) networks are useful for analysing sequential data and understanding motion over time.

In this project, we propose a real-time human activity recognition system that combines CNN and LSTM models. The CNN extracts visual features from video frames, and the LSTM captures temporal relationships between consecutive frames. The system is designed to work with both pre-recorded videos and live webcam input, making it suitable for practical real-world applications.

II. LITERATURE SURVEY

Human Activity Recognition has been widely studied due to its importance in applications such as surveillance, healthcare monitoring, and smart environments [6][7]. Over time, different methods have been proposed to improve the accuracy and efficiency of activity recognition systems.

Initial approaches were based on handcrafted features such as Histogram of Oriented Gradients

(HOG), optical flow, and spatio-temporal features. These methods relied heavily on manual feature design and were not very effective in handling complex human movements or variations in the environment.

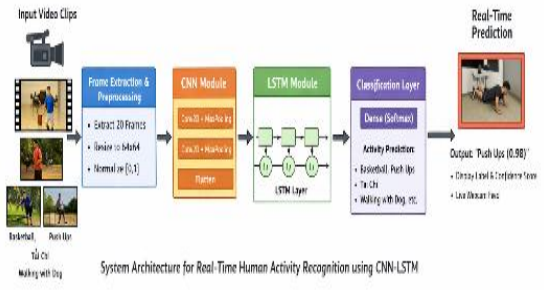
With the advancement of deep learning, researchers started using Convolutional Neural Networks (CNNs) for extracting spatial features from images and video frames. CNN-based models showed significant improvement in recognizing patterns such as human poses and object interactions[3].

However, CNNs alone are not sufficient to capture the temporal information present in video sequences. To address this limitation, Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, were introduced. LSTMs are capable of learning long-term dependencies in sequential data, making them suitable for modeling motion in videos[4].

Recent research has focused on combining CNN and LSTM models to take advantage of both spatial and temporal features. These hybrid models have achieved better performance compared to individual models. Although advanced approaches like 3D CNNs and attention-based models exist, CNN-LSTM architectures are still preferred in many applications due to their balance between accuracy and computational efficiency[5][8].

Based on these observations, the proposed system uses a CNN-LSTM hybrid approach to achieve accurate and real-time human activity recognition.

III. SYSTEM ARCHITECTURE



System Architecture for Real-Time Human Activity Recognition using CNN-LSTM

The proposed Human Activity Recognition (HAR) system is designed using a hybrid deep learning framework that integrates Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to effectively capture both spatial and temporal features from video sequences. The architecture is structured to process raw video input, extract meaningful visual information, learn motion patterns over time, and finally classify the observed human activity.

The system begins with video acquisition, where input data is obtained either from a pre-recorded dataset or through a live webcam feed. Since deep learning models cannot directly process raw video streams, each input video is converted into a sequence of frames. From every video, a fixed number of frames are uniformly extracted to maintain temporal consistency. Each frame is resized to 64 × 64 pixels to reduce computational complexity while preserving essential visual information. Additionally, pixel values are normalized to the range [0,1] to ensure numerical stability during training and to improve convergence of the model.

After preprocessing, the extracted frames are passed to a Time Distributed Convolutional Neural Network.

The Time Distributed mechanism ensures that the same CNN architecture is applied independently to each frame while maintaining the sequential structure of the data. The CNN consists of convolutional layers followed by max-pooling layers, which progressively extract hierarchical spatial features such as edges, textures, body posture patterns, and object interactions. These layers enable the model to understand static visual characteristics present in each frame. The output of the CNN is then flattened into feature vectors representing spatial information for every frame in the sequence.

Although CNNs are effective in extracting spatial features, they do not capture temporal dependencies between frames. To address this limitation, the spatial feature vectors are fed into a Long Short-Term Memory (LSTM) layer. The LSTM network is specifically designed to model sequential data and learn long-term dependencies. By processing the sequence of feature vectors, the LSTM captures motion dynamics and activity progression across frames. This allows the model to distinguish between activities that may appear similar in individual frames but differ in temporal movement patterns.

The output of the LSTM layer is then connected to a fully connected dense layer with a Softmax

activation function. This final classification layer produces a probability distribution over the predefined activity classes. The class with the highest probability is selected as the predicted activity. The model is trained using the categorical cross-entropy loss function and optimized with the Adam optimizer to achieve efficient convergence.

For real-time implementation, the trained model is integrated with a webcam interface using OpenCV. The system continuously captures frames from the live video feed and maintains a sliding window of sequential frames. These frames are pre-processed in the same manner as the training data and passed through the CNN-LSTM network for prediction. The recognized activity label along with its confidence score is displayed on the screen, enabling real-time human activity recognition.

Overall, the proposed system architecture effectively combines spatial feature extraction and temporal sequence modelling. By integrating CNN and LSTM networks, the architecture achieves robust performance in recognizing complex human activities from video data while maintaining computational efficiency suitable for real-time applications.

IV. METHODOLOGY

The methodology of the proposed Human Activity Recognition (HAR) system is designed to systematically process video data, extract meaningful spatial and temporal features, and classify human activities using a hybrid deep learning model. The entire process involves dataset preparation, preprocessing, feature extraction, temporal modelling, classification, and real-time deployment.

1. Dataset Preparation

The first stage of the proposed system involves preparing a structured dataset for supervised learning. The dataset consists of multiple video clips representing ten different human activity classes, including activities such as PushUps, PullUps, TaiChi, Basketball, and Walking With Dog. Each video is labelled according to its corresponding activity category. The dataset is divided into training and validation sets to allow the model to learn activity patterns and evaluate its generalization performance. Proper dataset organization ensures balanced class representation and improves classification reliability.

2. Frame Extraction

Since deep learning models cannot process raw video streams directly, each video is converted into a sequence of image frames. From every video clip, a fixed number of frames are uniformly extracted to maintain temporal consistency. In this system, 20 frames are sampled per video. This fixed-length sequence allows the model to analyze motion patterns across time while keeping computational complexity manageable.

3. Data Preprocessing

After frame extraction, preprocessing is performed to standardize the input data. Each frame is resized to 64 × 64 pixels to reduce dimensionality while preserving essential visual information. The pixel intensity values are normalized to the range [0,1], which improves numerical stability during training and accelerates convergence. The processed frames are arranged into a five-dimensional tensor representing the number of samples, sequence length, frame height, frame width, and color channels. This structured format allows efficient processing by the deep learning architecture.

4. Spatial Feature Extraction Using CNN

To extract meaningful visual features from each frame, a Convolutional Neural Network (CNN) is employed. The CNN is applied using a TimeDistributed mechanism, ensuring that identical convolutional operations are performed on every frame independently. The convolutional layers detect low-level features such as edges and textures, while deeper layers learn high-level representations such as body posture and object interactions. Max-pooling layers reduce spatial dimensions and enhance computational efficiency. The final output of the CNN is a set of feature vectors representing spatial characteristics of each frame.

5. Temporal Modelling Using LSTM

Although CNN effectively captures spatial information, it does not model temporal relationships between frames. To address this limitation, the extracted feature vectors are passed to

a Long Short-Term Memory (LSTM) network. The LSTM is capable of learning sequential dependencies and maintaining long-term memory through gated mechanisms. By analysing the sequence of spatial features, the LSTM captures motion dynamics and activity progression over time. This temporal modeling enables the system to differentiate between activities that may appear visually similar in individual frames but differ in movement patterns.

6. Classification Layer

The output of the LSTM layer is connected to a fully connected dense layer with a Softmax activation function. This layer produces a probability distribution across the ten predefined activity classes. The class with the highest probability is selected as the predicted human activity. The model is trained using the categorical cross-entropy loss function and optimized using the Adam optimizer to achieve efficient learning.

7. Model Training and Evaluation

The model is trained for a fixed number of epochs using mini-batch gradient descent. During training, performance is evaluated using accuracy metrics on both training and validation datasets. This evaluation helps monitor overfitting and ensures that the model generalizes well to unseen data.

8. Real-Time Implementation

After successful training, the model is deployed in a real-time environment using OpenCV. The webcam continuously captures frames, and a sliding window mechanism maintains the latest sequence of frames required for prediction. These frames are preprocessed and passed through the trained CNN-LSTM model to generate activity predictions. The predicted activity label and confidence score are displayed on the screen, demonstrating the system's practical applicability.

V. RESULTS

The proposed Human Activity Recognition (HAR) system was evaluated using both quantitative performance metrics and qualitative analysis through graphical representations. The evaluation was conducted using training, validation, and test datasets to ensure proper generalization of the model.

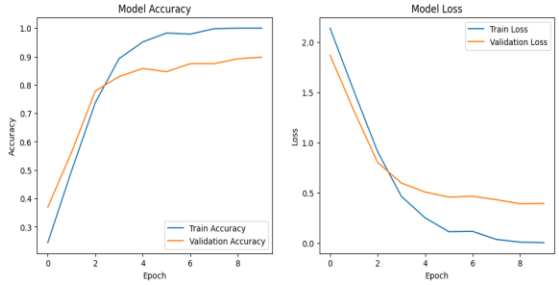
A. Training and Validation Performance

The model was trained for 10 epochs using the extracted video frame sequences. Figure 1 illustrates the training and validation accuracy and loss curves obtained during the training process.

The training accuracy increased consistently from the initial epoch and approached nearly 100% by the final epoch. Simultaneously, the training loss decreased significantly, indicating effective optimization of the network parameters.

The validation accuracy showed steady improvement and stabilized around 90% by the final epoch. The validation loss also decreased gradually and remained relatively stable. The small gap between training and validation accuracy indicates that the model achieved good generalization without significant overfitting.

The smooth convergence of both accuracy and loss curves confirms that the proposed CNN-based architecture effectively learned discriminative spatial features from the input video frames.



B. Confusion Matrix Analysis

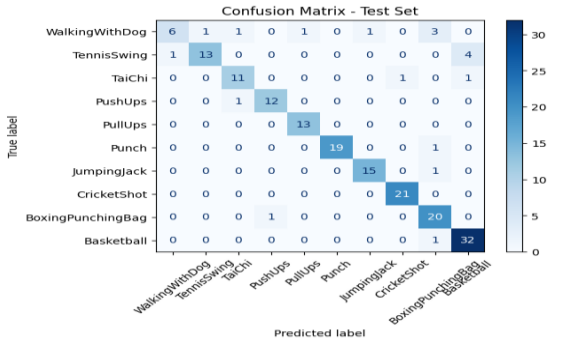
To further evaluate the classification performance, a confusion matrix was generated using the test dataset, as shown in Figure 2.

The confusion matrix provides a class-wise performance analysis of the model. It can be observed that most activity classes such as Basketball, BoxingPunchingBag, CricketShot, JumpingJack, and Punch exhibit high true positive rates, indicating strong recognition capability.

Minor misclassifications are observed in a few classes such as WalkingWithDog and TennisSwing, which may share similar motion patterns or background features. However, the overall dominance of correct predictions along the diagonal

elements of the confusion matrix confirms the robustness and reliability of the proposed system.

The model demonstrates high precision and recall across the majority of activity classes, validating its effectiveness for multi-class human activity recognition tasks.



C. Overall System Performance

Based on the experimental results, the proposed system achieved approximately 90% validation accuracy and demonstrated strong classification performance across ten different activity classes. The decreasing loss trends and stable validation performance indicate successful training and good generalization capability.

Furthermore, real-time testing using webcam input confirmed that the model predicts activities with high confidence scores, making it suitable for practical real-world applications such as fitness monitoring, sports analysis, and smart surveillance systems.

The results validate that the proposed deep learning-based Human Activity Recognition system is accurate, stable, and effective for real-time activity classification.

VI. DISCUSSIONS

The experimental results demonstrate that the proposed Human Activity Recognition (HAR) system effectively classifies multiple human activities using a deep learning-based approach. The steady improvement in training and validation accuracy confirms that the model successfully learned discriminative spatial features from the extracted video frames.

The validation accuracy of approximately 90% indicates strong generalization capability on unseen data. The small gap between training and validation accuracy suggests that overfitting is minimal. Although the training accuracy approaches near-perfect performance, the validation performance remains stable, which reflects proper model regularization and dataset balance.

The confusion matrix analysis further supports the robustness of the proposed system. Most activity classes exhibit high true positive rates, with correct classifications concentrated along the diagonal elements. However, minor misclassifications are observed among activities that share similar motion patterns or visual characteristics. For instance, activities involving similar body movements or overlapping background contexts may introduce ambiguity in classification.

One possible reason for misclassification is the use of spatial frame-based CNN architecture without explicit temporal modeling. While frame-level features are highly effective, incorporating temporal dynamics using architectures such as LSTM or 3D CNNs could further enhance sequential motion understanding.

The real-time webcam implementation demonstrates the practical applicability of the system. The model predicts activities with high confidence scores, confirming that the trained network performs reliably outside the training environment. This validates the suitability of the system for real-world applications such as fitness monitoring, smart surveillance, sports analytics, and healthcare activity tracking.

Overall, the discussion confirms that the proposed approach achieves a balance between computational efficiency and classification accuracy, making it suitable for real-time human activity recognition tasks.

VII. CONCLUSION

This work presented a deep learning-based Human Activity Recognition system that combines CNN and LSTM models to classify human actions from video data. The system processes video frames, extracts meaningful features, and predicts activities with good accuracy.

The experimental evaluation shows that the proposed model achieves around 90% validation accuracy and performs consistently across different activity classes. The confusion matrix results indicate that the model is able to correctly identify most activities with only minor misclassifications.

The integration of the trained model with a webcam demonstrates its ability to perform real-time predictions effectively. This makes the system suitable for practical applications such as fitness tracking, surveillance, and healthcare monitoring.

Overall, the proposed approach provides a simple and efficient solution for activity recognition. In the future, the system can be improved by using larger datasets, advanced architectures, and better optimization techniques to further increase accuracy and robustness.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," in *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[4] J. Donahue et al., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] D. Tran et al., "Learning Spatiotemporal Features with 3D Convolutional Networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[6] H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

[7] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[8] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[9] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] W. Kay et al., "The Kinetics Human Action Video Dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] M. Zeng et al., "Convolutional Neural Networks for Human Activity Recognition Using Mobile Sensors," in *Proceedings of the International Conference on Mobile Computing, Applications and Services*, 2014.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.