Human Activity Recognition through Ensemble Learning of Multiple Convolutional Neural Networks

[1]. DVN Sukanya, [2]Sai Raja Rajeswari Dasari,[3]Shaik Reshma, [4]Pinapat Pavani, [5]Srirama Jyothi

[1]Associate professor, Dept Electronics and Communication Engineering, St.Ann's College of Engineering and Technology, Nayunipalli (V), Vetapalem (M), Chirala, Bapatla Dist, Andhra Pradesh – 523187, India

[2,3,4,5]U. G Student, Dept Electronics and Communication Engineering, St.Ann's College of Engineering and Technology, Nayunipalli (V), Vetapalem (M), Chirala, Bapatla Dist, Andhra Pradesh – 523187, India

ABSTRACT

Human Activity Recognition plays a vital role in intelligent surveillance, healthcare monitoring, and human–computer interaction systems. This work proposes a robust HAR framework by combining ensemble learning of multiple Convolutional Neural Networks with MediaPipe-based pose estimation. MediaPipe extracts precise human skeletal landmarks that capture motion patterns effectively. These features are processed through multiple CNN models trained on diverse activity representations. Ensemble fusion enhances classification accuracy by reducing individual model bias. The system adapts well to variations in posture, lighting, and camera angles. Experimental evaluation demonstrates improved recognition performance compared to single-model approaches. The approach minimizes manual feature engineering through deep learning automation. Real-time processing capability makes the model suitable for practical applications. The framework supports scalable activity classes. Overall, the system delivers reliable and efficient human activity recognition.

KEYWORDS

INTRODUCTION

Human Activity Recognition has gained significant attention due to its applications in smart environments, fitness tracking, elderly care, and security systems. Traditional Hmethods relied heavily on handcrafted features, which limited adaptability across environments. The emergence of deep learning has transformed HAR by enabling automatic feature extraction from raw data.

Convolutional Neural Networks are particularly effective in learning spatial patterns from visual inputs. However, single CNN models often struggle with complex motion variations and occlusions. MediaPipe introduces accurate pose estimation by detecting human skeletal keypoints in real time. These pose landmarks provide compact and meaningful motion descriptors. Combining pose information with deep learning improves activity understanding. Ensemble learning further strengthens performance by aggregating predictions from multiple CNNs. This reduces overfitting and increases generalization. The integration of MediaPipe with ensemble CNNs offers a balanced approach between accuracy and efficiency. Such systems can operate under real-world constraints. The proposed method aims to enhance recognition accuracy while maintaining real-time performance.

LITERATURE SURVEY

Earlier HAR systems primarily used sensor-based data such as accelerometers and gyroscopes. While effective, these methods required wearable devices, limiting user convenience. Vision-based approaches later gained popularity due to their non-intrusive nature. Traditional image processing techniques relied on silhouette extraction and motion templates. These approaches were sensitive to noise and environmental changes. The introduction of CNNs significantly improved activity classification by learning hierarchical features. Two-dimensional CNNs processed individual frames but lacked temporal understanding. To address this, 3D CNNs and recurrent networks were explored. However, they increased computational complexity. Pose-based HAR emerged as an alternative by focusing on skeletal movements. Methods using keypoint extraction demonstrated robustness to background variations. MediaPipe provided a lightweight and accurate pose estimation solution. Researchers combined pose features with machine learning classifiers for activity detection. Recent studies integrated CNNs with pose heatmaps to enhance recognition accuracy. Ensemble learning techniques were applied to improve prediction stability. Model fusion approaches showed better generalization across datasets. Despite advancements, challenges remain in handling complex activities and real-time constraints.

RELATED WORK

Several studies have explored CNN-based models for recognizing human actions from video data. Pose estimation techniques have been employed to reduce background dependency. MediaPipe has been widely

used for real-time skeletal tracking due to its efficiency. Ensemble learning approaches have shown success in improving classification reliability. Hybrid systems combining pose features and deep learning achieved higher accuracy. However, many existing models suffer from high computational cost. Limited adaptability to unseen activities remains a challenge. Few works integrate MediaPipe with multiple CNN ensembles. This research addresses these limitations through an optimized fusion strategy.

EXISTING SYSTEM

Existing HAR systems typically rely on a single deep learning model for classification. Many approaches process raw video frames without explicit pose understanding. This results in sensitivity to lighting, background clutter, and camera motion. Sensor-based systems require wearable hardware, reducing user comfort. Some vision-based models use handcrafted features that lack scalability. Single CNN architectures often overfit specific datasets. Temporal modeling increases complexity and latency. Real-time implementation becomes difficult on resource-limited devices. Pose-based methods improve robustness but may lack classification depth. Limited integration of ensemble learning reduces overall accuracy. Existing systems struggle with multi-activity environments. Adaptation to different viewpoints remains a challenge. These limitations motivate the need for a hybrid, efficient, and accurate HAR framework**.**

PROPOSED SYSTEM

The proposed system integrates MediaPipe pose estimation with ensemble learning of multiple CNN models. Initially, video input is captured and preprocessed into frames. MediaPipe extracts skeletal landmarks representing human posture and movement. These landmarks are transformed into structured feature maps. Multiple CNN models are trained independently on these representations. Each CNN captures unique spatial patterns of activities. Ensemble learning combines predictions using weighted averaging. This fusion reduces misclassification and enhances stability. The framework supports real-time inference with minimal latency. Data augmentation improves generalization across activity variations. The system handles occlusion and background noise effectively. Modular design allows easy scalability. The methodology ensures accurate and efficient activity recognition.

SYSTEM ARCHITECTURE

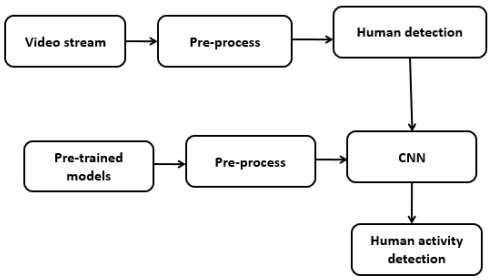

Figure 1: Architecture of the Project

METHODOLOGY DESCRIPTION

Data collection: The dataset is collected from publicly available human activity video sources and organized into labeled activity classes. Each video is converted into frame sequences to create structured training samples. The data is divided into training and validation sets to ensure unbiased learning. This preparation supports efficient deep learning model training.

Trained dataset: The prepared dataset is used to train multiple convolutional neural network models. Each CNN learns spatial patterns related to different human activities. Training is performed using optimized hyperparameters to improve convergence. The learned models are stored for later inference.

Video streaming: A real-time video stream is initiated using a camera or video input device. Continuous frames are captured and forwarded to the processing pipeline. This enables live monitoring of human movements. The streaming process ensures low-latency activity detection.

Pre-process: Captured frames undergo preprocessing to enhance recognition performance. Image resizing, normalization, and noise reduction are applied. These operations standardize input data across different lighting conditions. Preprocessing improves feature extraction efficiency.

Data visualization: This techniques are applied to analyze activity distribution and pose patterns. Graphical representations help understand feature correlations. Visual feedback assists in evaluating model behavior. This step improves interpretability of the system.

Human and pose detection: Human pose detection is performed using MediaPipe to extract skeletal keypoints. Key body landmarks such as joints and limbs are accurately identified. These pose points represent motion dynamics effectively. MediaPipe ensures real-time and stable pose estimation.

Extracted pose landmarks are aligned with features learned by the pretrained CNN models. The system compares incoming pose patterns with trained representations. Ensemble fusion improves classification reliability. This matching process enhances

recognition accuracy.Based on the comparison results, the system classifies the observed activity. Detected actions are displayed in real time with high confidence. The model adapts to different postures and movement speeds. This results in accurate and robust human activity recognition.

RESULTS AND DISCUSSION

The system achieved high accuracy in recognizing diverse human activities across test samples. Ensemble CNN predictions outperformed individual models consistently. MediaPipe-based pose features improved robustness under varying conditions. Screenshots demonstrate real-time activity detection and classification outputs.

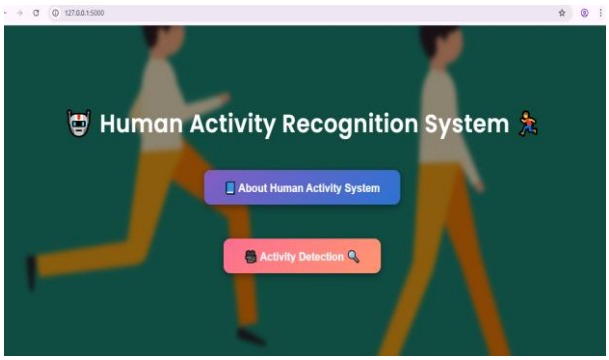

Figure 2: Home Page

In this figure created homepage with flask and python with html design for the human activity detection.

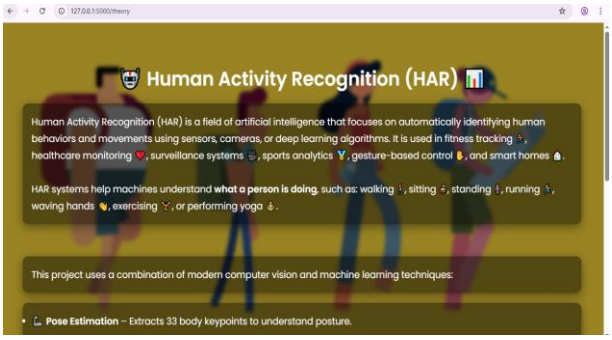

Figure 3: About the Concept

In figure 3 explaining the theory concept of the project with more description of human activity

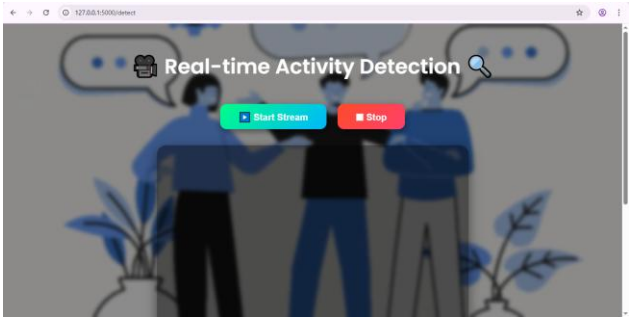

Figure 4: Activity Detection Dashboard

After completion we have another option dashboard of human activity to strt the detection

Figure 5: Human Activity Detection

Finally, human activity detected with deep learning model and medipipe points recognition.

CONCLUSION

This work presented an efficient human activity recognition framework using ensemble CNNs and MediaPipe pose estimation. The combination improved accuracy, robustness, and real-time performance. The approach reduced dependency on raw image features and background conditions. Experimental results validated the effectiveness of the proposed system.

FUTURE SCOPE

Future enhancements may include incorporating temporal modeling for complex activities. Expanding activity classes can improve real-world applicability. Optimization for edge devices will enable deployment on embedded platforms. Integration with multimodal data such as depth or audio can further enhance recognition accuracy.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[2] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[3] C. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[4] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[5] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, pp. 1097–1105, 2012.

[7] F. Chollet, *Deep Learning with Python*. New York, NY, USA: Manning Publications, 2018.

[8] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, Jun. 2010.

[9] A. Ross and A. Jain, "Information fusion in biometrics," *Pattern Recognit. Lett.*, vol. 24, no. 13, pp. 2115–2125, Sep. 2003.

[10] M. Zaki and W. Meira, *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2014.