HumanActivityRecognition UsingDeep Learning

JohnWesley[1*],AishwaryaDashrath[2],AishwaryaSomanath[3]and,AkshataYenkapalli[4]
DepartmentofComputerScienceandEngineering,
GuruNanakDevEngineeringCollegeBidar,Karnataka585403,India
johnwesley@gndecb.ac.in

Abstract

Human Activity Recognition (HAR) has become a vital area of research in computer vision and artificial intelligence. The primary goal is to enable machines to interpret and categorize human actions from video data. This paper presents a robust HAR system utilizing the R(2+1)D Convolutional Neural Network architecture, which factorizes 3D convolutions into separate spatial and temporal components for improved learning efficiency. The model was trained and validated on the UCF-101 and HMDB-51 datasets, achieving a validation accuracy of 87.34%. To demonstrate practical utility, we developed a full-stack web applicationusingReact.jsforthefrontendandFastAPIforthebackend,allowinguserstoperformreal-timevideoclassification.

Keywords—HumanActivityRecognition(HAR),DeepLearning,R(2+1)DCNN,ComputerVision,SpatiotemporalFeatures, PyTorch, FastAPI.

I. Introduction

HumanActivity Recognition (HAR) is an important area in artificial intelligence and computer vision that focuses on enabling machines to automatically identify human actions from video or sensor data. It has wide applications in healthcare monitoring, security surveillance, human-computer interaction, and smart environments. By understanding human behavior, HAR systems can improve safety,automation,anduserexperience,andareincreasingly being integrated into real-time applications [5][8].

TraditionalHARmethodsreliedonhandcraftedfeaturessuch asSIFTandHOG,whichoftenfailedtoperformwellinreal-world conditions due to variations in lighting, background, and motion. These approaches required significant manual effort and domain expertise, and lacked scalability when appliedtolargeanddiversedatasets[2].Consequently,their performance was limited in complex and dynamic environments.

With the advancement of deep learning, models can now automatically learn features directly from raw video data, significantly improving accuracy and robustness. Deep learning approaches such as CNNs and LSTMs have been widelyusedforHARtasks,astheyeffectivelycapturespatial andtemporaldependenciesinsequentialdata[1][4].Among these approaches, 3D CNNs are particularly effective in modeling spatiotemporal information; however, they are computationally expensive and require high processing power [18].

To overcome this limitation, this research focuses on the R(2+1)D architecture, which separates spatial and temporal feature extraction into 2D and 1D operations. This decomposition reduces computational complexity while preserving the model's ability to learn motion dynamics effectively. As a result, the proposed approach achieves improvedefficiencyandaccuracy,makingitsuitableforreal-world human activity recognition applications.

II. PROBLEMSTATEMENT

Recognizinghumanactivitiesfromvideosischallengingdueto variations in environmental conditions such as lighting, background, camera angle, and speed of movement, all of which can significantly affect model accuracy. Traditional machine learningapproachesoftenstruggletohandlesuchcomplexityand variability in real-world scenarios [2][5].

This project aims to develop a deep learning-based system capableofaccuratelyandefficientlydetectingandclassifying human activities from short video clips through a web-based interface. Deep learning models have demonstrated strong performanceinhandlingcomplexvisualpatternsandtemporal dependencies, making them suitable for real-time HAR applications [1][4].

III. LITERATUREREVIEW

Human Activity Recognition (HAR) has gained significant attention in recent years due to its wide range of applications in healthcare, surveillance, and smart environments. Early approaches to HAR primarily relied on handcrafted features combined with traditional machine learning algorithms such as Support Vector Machines (SVM) and Decision Trees. However, these methods were limited in their ability to capture complex spatial and temporal patterns in data [2][5].

Withtheadvancementofdeeplearning,ConvolutionalNeural Networks(CNNs)havebeenwidelyadoptedforextractingspatial features from images and video frames. CNN-based models have shown improved performance by automatically learning hierarchicalfeaturerepresentations.However,CNNsalonearenot sufficient for modeling temporal dependencies in sequential data [1][4].

To addressthislimitation,RecurrentNeuralNetworks(RNNs), particularly Long Short-Term Memory (LSTM) networks, were introduced to capture temporal dynamics in human activities. Hybrid architectures combining CNN and LSTM have further enhancedperformancebyintegratingspatialandtemporalfeature learning. These models have demonstrated significant improvementsinrecognizing complexhumanactivities[9][10].

In addition, advanced frameworks such as DeepSense have been proposed for processing time-series sensor data, providing a unified approach for activity recognition [3]. More recent studies have explored Graph Convolutional Networks(GCNs)andattention-basedmechanismstomodel relationships between different body joints and improve recognition accuracy [12].

Despitetheseadvancements,challengessuchasreal-time processing, occlusion handling, and variability in environmental conditions still remain. The proposed work addresses these challenges by utilizing the R(2+1)D architecture,whichefficientlyseparatesspatialandtemporal feature extraction, leading to improved performance and computational efficiency [18].

IV. METHODOLOGY

A.SystemArchitecture

TheproposedHumanActivityRecognition(HAR)systemis designed to process video input and accurately classify human activities using a deep learning-based approach.The overall system architecture consists of multiple stages, includingdataacquisition,preprocessing,featureextraction, model inference, and output generation, similar to modern deep learning-based HAR frameworks [18].

A. InputModule

The system begins with the input of video data, which can be uploaded by the user through a web-based interface.The inputvideocontainssequencesofhumanactivitiesthatneed to be analyzed and classified.

B. PreprocessingModule

The input video is converted into a sequence of frames.A fixednumberof**16consecutiveframes**areextractedfrom each video clip. Each frame isresized to **112 × 112 pixels** and normalized to ensure consistency and improve model performance, as commonly followed in video-based HAR systems [18].

C. FeatureExtractionModule

The preprocessed frames are passed to the R(2+1)D ConvolutionalNeuralNetwork.Thismodelseparates:

- **Spatialfeatures**using2Dconvolution(capturing shapes, textures, and objects)
- **Temporalfeatures**using1Dconvolution (capturing motion across frames)

Thisseparationenhancesthemodel'sabilitytolearn meaningfulrepresentationsefficientlyandimproves performancecomparedtotraditional3DCNNapproaches [18].

D. ClassificationModule

The extracted features are fed into a fully connected layer followedbya**Softmaxfunction**,whichconvertstheoutput intoprobabilityscoresforeachactivityclass.Theclasswiththe highest probability is selected as the predicted activity.

E. OutputModule

The final output is displayed to the user through the interface, showingthepredictedhumanactivity.Thesystemsupportsreal-timepredictions,makingitsuitableforpracticalapplicationssuch as smart monitoring systems [7].

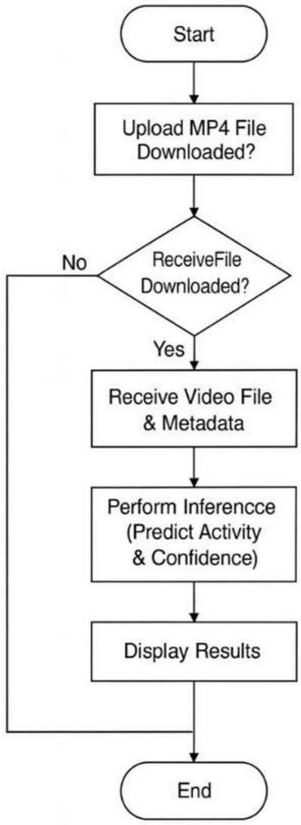

Figure1:SystemFlowDiagram

A. SystemImplementation

TheproposedHumanActivityRecognition(HAR)systemis implemented using Python and the PyTorch deep learning framework. The R(2+1)D model is trained on benchmark datasets such as UCF-101 and HMDB-51, where videos are converted into sequences of 16 frames, resized, andnormalizedbeforebeingfedintothemodel[16][17].

The backend is developed using FastAPI, which handles video upload, preprocessing, and model inference. The frontendisbuiltusingReact(Vite),providingauser-friendly interface for uploading videos and displaying predictedactivities.

Theoverallworkflowinvolvesvideoinput,frameextraction, preprocessing,modelprediction,and resultdisplay,enabling efficientandreal-timeactivityrecognition,similartomodern HAR systems [7][15].

B. WorkflowOverview

TheoverallworkflowoftheproposedHumanActivityRecognition(HAR)systembeginswiththeuseruploadinga video through the web interface. The input video is then processed by extracting a fixed sequence of frames, which are resized and normalized during preprocessing.

The processed frames are fed into the R(2+1)D deeplearning model,whichextractsspatialand temporalfeatures toclassifytheactivity.Finally,thepredictedactivitylabelis returned and displayed to the user. This workflow ensures efficientandaccuratereal-timeactivityrecognition,aligning with recent advancements in HAR systems [7].

V. USERINTERFACE

The proposed system provides a simple and user-friendly interface developed using React (Vite). Users can easily upload video files through the interface and initiate the activityrecognitionprocess,supportinginteractiveandreal-time applications [15].

Once the video is processed, the system displays the predicted human activity as the output. The interface also provides basic feedback such as upload status and processingprogress,ensuringasmoothuserexperience.The results are generated quickly, enabling near real-time interaction with the system, similar to modern real-time HAR frameworks [7].

VI. CONCLUSION

“HumanActivityRecognitionusingDeepLearning” demonstrates the effectiveness of artificial intelligence andcomputervisiontechniquesinaccuratelyclassifying human activities from video data. Using the R(2+1)D CNN model, the system successfully captures spatial andtemporalfeatures,achievinganaccuracyof**87.34%** on benchmark datasets .

The system provides a user-friendly web-based interface and supports real-time predictions, making it suitable for practical applications such as healthcare, surveillance, and smart systems . Future improvementscanfocusonreal-timeprocessing,multi-person detection, and advanced models like Vision Transformers to further enhance performance.

VII. REFERENCE

[1] F. J. Ordóñez and D. Roggen, “Deep convolutional and LSTM recurrent neural networks for multimodal wearableactivityrecognition,”*Sensors*,vol.16,no.1,p. 115, 2016.

[2] C. A. Ronao and S. B. Cho, “Human activity recognitionusingsmartphonesensorswithdeeplearning models,” *Expert Systems with Applications*, vol. 59, pp. 235–244, 2016.

[3] S. Yao *et al.*, “DeepSense: A unified deep learning frameworkfortime-seriesmobilesensingdataprocessing,”in *Proc. ACM MobiCom*, 2017.

[4] Y.GuanandT.Plötz,“EnsemblesofdeepLSTMlearners for activity recognition using wearables,” *Proc. ACM IMWUT*, vol. 1, no. 2, pp. 1–28, 2017.

[5] H. F. Nweke *et al.*, “Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: A review,” *IEEE Access*, vol. 6, pp. 5356–5389, 2018.

[6] Z. Xiao *et al.*, “A deep learning method for complex human activity recognition using virtual wearable sensors,” *IEEE Sensors Journal*, 2020.

[7] Y. Cheng *et al.*, “Real-time human activity recognition using conditionally parameterized convolutional neural networks on mobile devices,” *IEEE Transactions on Mobile Computing*, 2020.

[8] J. Wang and Y. Liu, “Wearable sensor-based human activity recognition using hybrid deep learning techniques,” *Information Sciences*, vol. 561, pp. 137–154, 2020.

[9] S. W. Pienaar and R. Malekian, “Human activity recognition using LSTM-RNN deep neural networks,” in *Proc. IEEE Int. Conf. Industrial Informatics*, 2019.

[10] R. Zhao, D. Wang, and R. Yan, “Deep residual bidirectional LSTM for human activity recognition using wearable sensors,” *Applied Sciences*, vol. 9, no. 5, p. 984, 2019.

[11] P. Wang *et al.*, “DeepHAR: Human activity recognition fromvideosusingdepthinformation,” *IEEETransactionson Circuits and Systems for Video Technology*, 2018.

[12] S. Yan *et al.*, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *Proc. AAAI Conf. Artificial Intelligence*, 2018.

[13] S. Ha *et al.*, “Multi-modal sensor-based deep learning framework for human activity recognition,” *IEEE Sensors*, 2020.

[14] N. Y. Hammerla*et al.*, “Self-supervised learning for human activity recognition,” *Journal of Machine Learning Research*, 2021.

[15] C. Xu *et al.*, “EfficientHAR: Lightweight CNN for real-time human activity recognition on edge devices,” *IEEE Internet of Things Journal*, 2021.

[16] K. Soomro, A. R. Zamir, and M. Shah, “UCF101: A datasetof101humanactionclassesfromvideosinthewild,” 2012.

[17] H. Kuehne *et al.*, “HMDB:Alarge video database for human motion recognition,” in *Proc. IEEE ICCV*, 2011.

[18] D. Tran *et al.*, “A closer look at spatiotemporal convolutions for action recognition,” in *Proc. IEEE CVPR*, 2018.