iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.61923

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Human Action Recognition SystemUsing LRCN

Abhijat Krishna K[1], Milen Eldo[2], Jinzen Kuriakose[3], Abin JS[4], Suzen Saju Kallungal[5], Rotney Roy Meckamalil[6]
Department of Computer Science and Engineering Mar Athanasius College of Engineering, Kothamangalam, Kerala

Abstract: ***The importance of human action recognition is significant across various domains, driving advancements in safety, healthcare, sports analytics, and interactive technologies. Leveraging machine learning techniques like Long Recurrent Convolutional Neural Networks (LRCN) trained on datasets such as UCF50, our project focuses on action prediction from YouTube videos. This technology plays a pivotal role in enhancingsafety and security by enabling the detection of anomalies and suspicious behaviours in surveillance systems. In healthcare, it supports remote patient monitoring, rehabilitation assessment, and personalized care plans based on observed actions. Addi- tionally, in sports analytics and entertainment, human action recognition informs performance evaluation, content creation, and immersive experiences. Our integration of a pre-trained LRCN model into a Flask web application signifies the tangible impact of machine learning on human action recognition, with ongoing efforts to optimize functionality through input validation, error handling, user feedback, security measures, and perfor- mance enhancements, illustrating the practical application and societal benefits of this innovative technology***

Index Terms: Human action recognition, machine learn- ing, Long Recurrent Convolutional Neural Networks (LRCN), YouTube videos, safety, healthcare, sports analytics,surveillance systems, anomalies detection, remote patient monitoring, reha- bilitation assessment, personalized care plans

I. INTRODUCTION

Human action recognition is a rapidly evolving and crucial field in computer vision, offering transformative capabilities to interpret and categorize human activities depicted in videosequences. By automatically monitoring and analyzing humanbehavior, it becomes possible to enhance user experiences, facilitate timely responses to security threats, and providevaluable insights for healthcare monitoring and sports analysis. To advance this field, innovative approaches have emerged that integrate cutting-edge technologies and advanced machinelearning techniques. [1] One such approach is the developmentof a sophisticated Human Action Recognition System bycombining Convolutional Neural Networks (CNNs) and LongShort-Term Memory (LSTM) networks. By leveraging CNNs to capture spatial features and LSTMs to model temporal dependencies, this system aims to create a robust modelcapable of accurately recognizing a diverse array of human actions with high precision.

In parallel, recent research has introduced novel methodolo- gies such as a wireless system based on magnetic induction for human activity recognition. [6] This system integratesmagnetic induction with machine learning techniques to detectvarious human motions, addressing challenges in sensor-based activity recognition systems operating around the human body. Its advantages, including reduced power consumption, cost, and complexity, make it a promising solution for applications in healthcare, rehabilitation, athletics, and senior monitoring. Moreover, another approach proposes the [7] Adap- tively Multi-correlations Aggregation Network (AMANet) for skeleton-based action recognition. By integrating Graph Con- volutional Networks (GCNs) and Self-attention mechanisms, this method captures dynamic correlations between human joints, enhancing spatial and temporal understanding in skele- ton data. Extensive experiments demonstrate its effectiveness in various action recognition tasks, positioning it as a compet-itive approach in the field.

Collectively, these endeavors underscore the ongoingevolution in human action recognition technology. By continuing to advance these approaches and exploring new ones, the field of human action recognition will continue to provide transformative capabilities in a wide range of applications.

II. RELATED WORKS

A. Human Activity Recognition

This paper provides an in-depth analysis of the application and significance of Convolutional Neural Networks (CNNs)in the field of human activity recognition (HAR), a key area within artificial intelligence aimed at categorizing human actions using sensor and camera data. It highlights the roleof CNNs in efficiently processing and classifying images through their distinct layers, such as convolutional, pooling, and fully connected layers, which collectively facilitate the automatic extraction of complex features from raw data, thus obviating the need for manual feature identification.

[1]The paper underscores the importance of HAR in various domains including healthcare, surveillance, and human-computer interaction, and discusses recent advancements in HAR research, including the adoption of deep learning models, hybrid approaches, and the use of wearable sensors to improveaccuracy and model performance. Additionally, it delves into the technical workings of CNNs, explaining the functions of convolutional and pooling layers in feature extraction and the role of fully connected layers in learning high-level features for accurate activity classification. The paper concludes by pointing to future research directions, such as the generationof synthetic data to tackle class imbalance and the exploration of multimodal data fusion, aimed at enhancing CNN models for HAR applications.

B. Human Motion detection using Skeleton Heat Maps

The research focuses on enhancing human action recog- nition through the development of a two-stage human pose estimation model that combines the SSD algorithm with CPM to accurately extract human skeleton information, vital for applications in human-computer interaction. [2] By integrating an improved SSD network based on ResNet with multiscale transformation and CPM, the model effectively overcomesthe limitations in skeleton keypoints detection, ensuring more accurate human detection and flexibility in handling various human positions. To facilitate action recognition, an eight- layer CNN is employed to classify the skeleton keypoints heatmaps, extracting relationship characteristics crucial for recognizing command actions. Experimental evaluations are conducted using benchmark datasets like Pascal VOC, MPII, LSP, and a self-collected dataset (SCAID), showcasing the SSD-ResNet50-Person model's improved accuracy in human detection and the CPM-Stage 6 model's proficiency in skele- ton keypoints detection. The research demonstrates signifi- cant advancements in action recognition accuracy, particularlythrough the [1] CNN-8 network's performance on different datasets, while also identifying challenges such as the lower detection accuracy of extremity joints and classification errorsbetween similar actions.

C. Gesture recognition for Human-Robot Collaboration

The document presents a comprehensive framework for human action and gesture recognition in human-robot collaboration scenarios. [3] It addresses the importance ofrecognizing human actions for effective and safe collaboration between humans and robots, proposing a framework based on 3D pose estimation and ensemble techniques to recognize both body actions and hand gestures. The proposed framework relies on OpenPose and 2D to 3D lifting methods to estimate 3D joints and uses graph convolutional networks for action recognition. The document emphasizes the significance of recognizing both body actions and hand gestures in a collaborative scenario and highlights the challenges associated with 3D pose estimation methods in accurately capturing hand joints. The proposed framework was evaluated on a custom dataset designed for human-robot collaboration tasks, named IAS-Lab Collaborative HAR dataset, and the results showed that using an ensemble of action recognition models improves the accuracy and robustness of the overall system. The document also outlines the main contributions of the work, including a unified framework for human action and gesture recognition, an experimental comparison of different ensembling techniques, and an experimental comparison of different 3D pose estimation methods to alleviate the missing joints problem. Additionally, the document discusses the potential for future research directions, such as evaluatingthe framework in a real human-robot collaboration task and further investigating the robustness of the framework to different viewpoints considering a multi-camera setup.

D. Transforming Spatio-Temporal Self-Attention using Em-bedding for Skeleton-Based Action Recognition

[4]The article proposes a novel methodology for skeleton- based action recognition, utilizing action embedding and self-attention Transformer. The method consists of two main modules: action embedding and self-attention Transformer.The action embedding encodes the spatial features of body joints, capturing the relationship between correspondingjoints and modeling spatial features. This is achieved using graph embedding techniques such as DeepWalk or Graph Convolution, which represent the spatial relationship between interacting joints. The self-attention Transformer models the temporal features and dependencies of body joints, exploiting the temporal contextual information and capturing the inter- dependencies between joints. The output of the Transformer network is then fed to a multiple-layer perceptron (MLP) for action classification. The proposed method is evaluated on various benchmark datasets, demonstrating its superiority over other state-of-the-art architectures. Additionally, the article discusses the use of link prediction methods, graphconvolutional networks, and the effectiveness of different techniques for skeleton-based action recognition. The proposed method showcases promising results, outperforming other approaches and contributing to advancements in the field of computer vision and pattern recognition.

E. Hybrid Approach: Attention-Based LSTM and 3D CNN forHuman Action Recognition

[5]The article presents a hybrid approach for human action recognition using an attention-based Long Short-Term Memory (LSTM) network and 3D Convolutional Neural Network (CNN). The proposed method integrates 3D CNNs for feature extraction from video data, followed by an LSTM network with an attention mechanism for precise action classification. The 3D CNNs are utilized to capture both spatial and temporal information from the videos, while the LSTM-Attention network processes the extracted features to facilitate accurate classification of the actions present in the videos. The method employs various preprocessing techniques, including video normalization, data augmentation,and temporal sampling, to enhance the quality of the input data. Additionally, the study evaluates the proposed approach on benchmark datasets, UCF101 and HMDB51, and compares its performance with existing state-of-the-art methods in action recognition. The results demonstrate the effectiveness of the proposed method in addressing the challenges of human action recognition, such as complex actions, diverse environmental conditions, and large datasets. The article also discusses the limitations of the proposed method and suggests potential directions for future researchin the field of human action recognition.

F. Genetic CNNs for Human Action Recognition

This paper presents a novel approach for human actionrecognition utilizing genetic algorithms (GA) and deepconvolutional neural networks (CNN). It proposes initializing CNN classifier weights with solutions generated by GAs to minimize classification errors. The hybrid method exploits the global search capabilities of GAs and the local optimization power of gradient descent algorithms during fitness evaluations of GA chromosomes. This combinationaims to find solutions close to the global optimum and improve classification performance by integrating evidence from classifiers generated using GAs. The effectiveness of this classification system is demonstrated on the UCF50 dataset. The paper reveals that shallow neural network models' failure led to the adoption of deep learning networks, particularly CNNs, which have become prevalent in addressing various computer vision tasks including human action recognition due to advancements in computational capabilities andtraining techniques. Furthermore, the paper discusses the importance of human action recognition in fields like video retrieval and sports analysis, gives a background on human action recognition including methodologies and datasets, and elucidates the evolution and architecture of CNNs. It outlines the GA's role in optimizing CNN weights and details the experimental setup, including a 5-fold cross-validation approach on the UCF50 dataset and the performance evaluation of the proposed CNN classifier. The results showcase the superiority of the proposed method over existing techniques, highlighting improvements in classification accuracy. The paper concludes by summarizing the effectiveness of merging GAs with CNNs for accurate human action recognition and suggests future research directions for enhancing the proposed system.

G. Graph Temporal Decoupling for Gesture Recognition

The Temporal Decoupling Graph Convolutional Network (TD-GCN) represents a significant advancement in skeleton- based gesture recognition, specifically addressing challenges related to temporal modeling and spatiotemporal feature ex- traction. Unlike traditional methods that use a single shared adjacency matrix, TD-GCN employs distinct adjacency ma- trices for skeletons across frames, enabling more accurate and nuanced representations of temporal relationships betweenjoints over time.

TD-GCN's architecture involves critical steps to extract spatiotemporal features from raw skeleton data, capturing both spatial relationships between joints and their temporal evolution across frames. It then computes channel-dependent and temporal-dependent adjacency matrices tailored to each frame of the input sequence, allowing the network to learn context-specific relationships within and across frames.

Additionally, TD-GCN integrates topology fusion mechanisms that combine information from neighboring skeleton joints, enhancing its ability to capture complex interactions and dependencies. Experimental evaluations on benchmark datasets like SHREC'17 Track and DHG-14/28 demonstrate TD-GCN's superiority over existing methods in recognition accuracy, generalization, and computational efficiency.

III. PROPOSED MODEL

The model overview for this methodology entails the in- tegration of Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks to tackle the challenge of human action recognition in video data. Initially, the CNN module processes input video frames, extractingspatial features through a series of convolutional and pooling layers. These spatial features are then flattened and optionally passed through fully connected layers to further refine the representation. Subsequently, the LSTM module receives this sequence of feature vectors, enabling the capture of tem- poral dependencies and context across the frames.

During training, the model learns to associate the extracted features with corresponding action labels through backpropagation and optimization. In inference, unseen video sequences are fed into the trained model, which processes the frames through the CNN module to extract spatial features and then through the LSTM module to capture temporal information and predict the depicted human action. Evaluation of the model's performance is conducted using standard metrics, facilitating the assessment of its effectiveness in accurately recognizing human actions. Through fine-tuning and optimization, the model's parameters and hyperparameters are adjusted to enhance its overall per- formance and generalization capabilities, ultimately advancingthe field of human action recognition within computer vision.

IV. SYSTEM IMPLEMENTATION

The system implementation for action recognition usingthe UCF50 dataset and Convolutional Neural Network (CNN) combined with Long-term Recurrent Convolutional Network (LRCN) models follows a structured approach. Initially, the dataset is downloaded and visualized to understand the diverse actions represented in the videos. The dataset is then preprocessed by resizing video frames to a fixed size andnormalizing pixel intensities. Subsequently, the dataset is split into training and testing subsets for model development and evaluation.

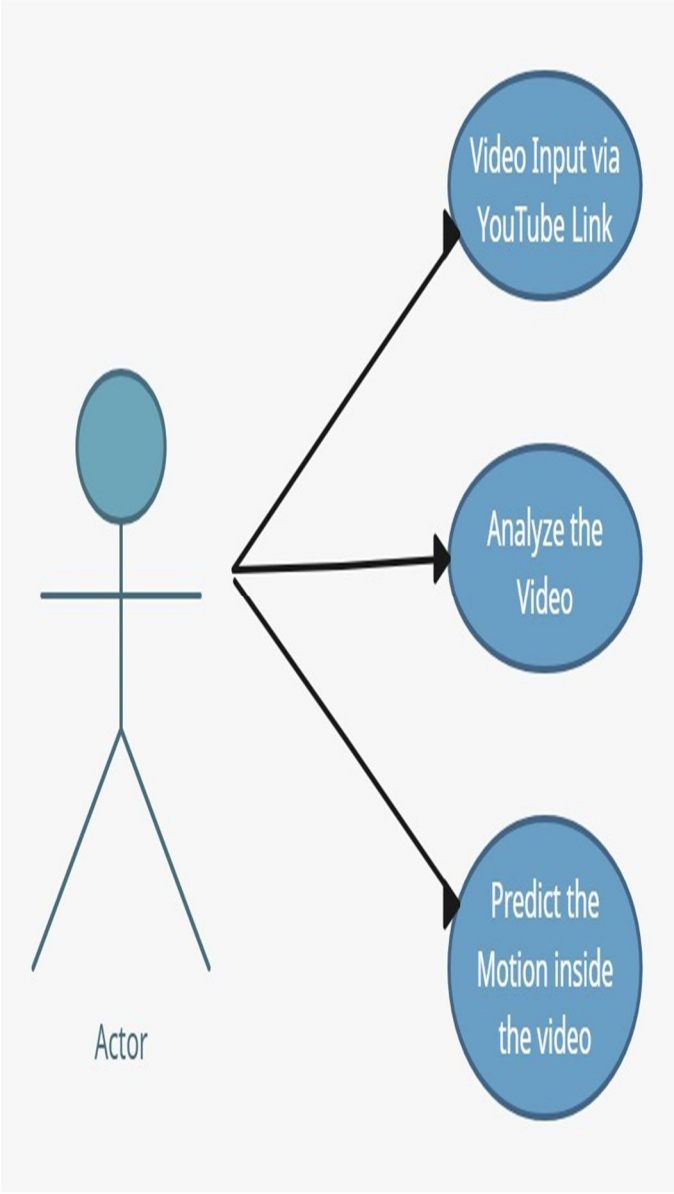


Fig. 1. Use Case Diagram of Our Proposed Model

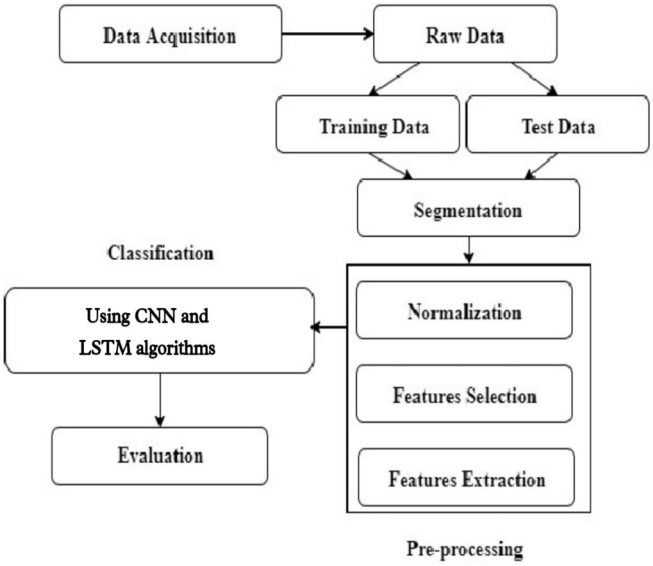


Fig. 2. Data Flow Diagram of Our Proposed Model

The CNN model is designed with multiple convolutional layers, pooling layers, and dropout for feature extraction, followed by training using the training dataset. For temporal modeling, the LRCN model integrates the CNN feature extraction with LSTM layers to capture temporal dependencies in video sequences. The LRCN model is compiled, trained, and evaluated on the testing dataset to assess performance. Finally, the trained model is applied to YouTube videos for action recognition testing, demonstrating its effectiveness and achieving an accuracy of 91% as reported. This systematic approach highlights the integration of deep learning techniques with dataset preprocessing and model implementation to build a robust action recognition system capable of analyzing real-world video data.

A. Data Collection

The process of gathering video data from the UCF50 dataset involves accessing a diverse array of human action categories captured across varied contexts. This dataset ensures representation of a wide spectrum of human actions performed by different individuals in various environments. To maintain data integrity and authenticity, reputable sources or APIs are utilized for downloading the dataset directly. This meticulous approach ensures that the collected data accurately reflects the breadth and depth of human actions, enabling robust training and evaluation of models for action recognition tasks.

B. Data Preprocessing

In the video preprocessing stage for the UCF50 dataset, individual frames are extracted to facilitate subsequent analysis. These frames undergo a series of cleaning and normalization steps to ensure consistency and quality. Resizing the frames to a standardized resolution and format is implemented to establish uniformity across the dataset. Challenges such as noise, variations in lighting conditions, and background clutter are addressed during this phase to mitigate their impact on model accuracy. Additionally, techniques such as converting frames to grayscale or applying color normalization methods are employed to enhance the performance of the action recognition model by reducing unnecessary complexity and improving feature extraction. Through meticulous preprocessing, the dataset is refined and optimized, laying a solid foundation for effective model training and evaluation.

C. Feature Extraction

In the extraction of spatial features from individual video frames, Convolutional Neural Networks (CNNs) play a pivotal role. These networks are adept at capturing patterns and structures within images, making them ideal for analyzing video frames in the context of human action recognition. Features extracted from CNNs may encompass activations from convolutional layers, which encode low-level visual information, as well as higher-level features learned by the network, which encapsulate more abstract representations relevant to action classification.

To optimize feature extraction,techniques such as transfer learning are explored. Leveraging pre-trained CNN models like ResNet or VGG enables the utilization of knowledge gained from large-scale image datasets, thereby enhancing the efficiency and effectiveness ofspatial feature extraction for human action recognition tasks.

D. Training ,Evaluation and Model Deployment

Splitting the dataset into training, validation, and testingsets is the initial step in the development process. With these subsets established, the CNN + LSTM model is trained onthe training data, aiming for high accuracy and generalization.

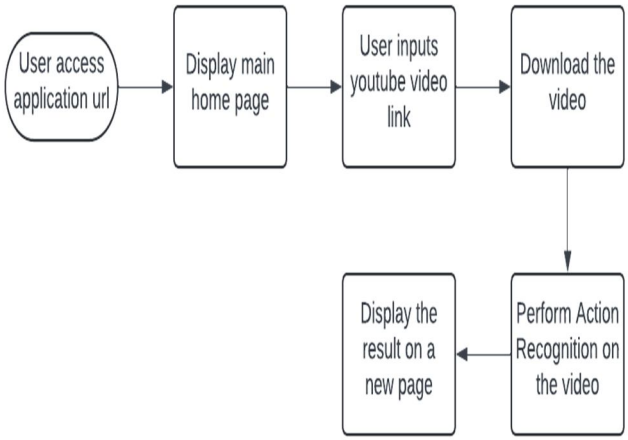


Fig. 3. Application Flow Diagram of HAR system using CNN+LSTM

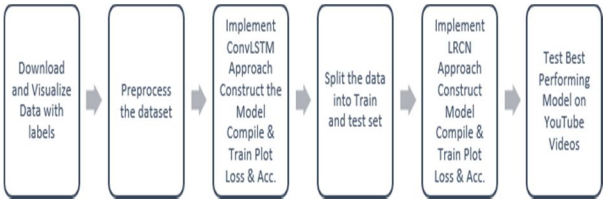


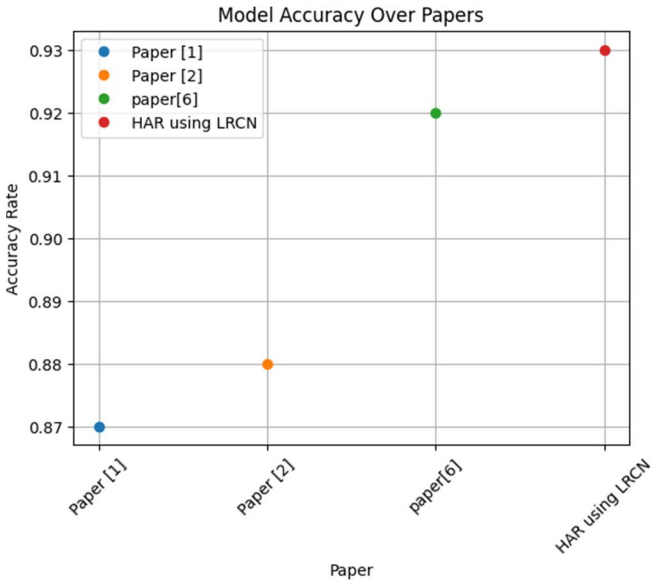
Fig. 4. System Implementation Flow Chart

Throughout training, the model learns to effectively capture spatial and temporal features from the video data. Following this, the model's performance undergoes evaluation on the validation and testing sets, employing various metrics such as accuracy, precision, recall, and F1-score. These evaluations provide crucial insights into the model's capability to accurately classify human actions. Subsequently, the trained CNN + LSTM model is integrated into a cohesive system designed for human action recognition. This integration involves the creation of a user-friendly interface capable of accepting input videos and generating corresponding action predictions. To ensure scalability, the system is engineered to handle real-time processing of video streams as necessary, accommodating diverse applications and requirements. Attention is given to deploying the system, whether as a standalone application or integration into existing platforms, to enhance accessibility and usability, thereby maximizing its societal impact and utility.

V. RESULTS

Human Activity Recognition (HAR) utilizing a combination of Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNNs) has emerged as a potent solution for identifying and classifying human actions from video data. This amalgamation effectively captures both spatial and temporal features, essential for understanding human behaviorover time.

With LSTM handling the temporal aspects and CNNs ex- tracting spatial features, the model achieves impressive perfor- mance metrics, including accuracy, precision, recall, and F1- score. Validation tests consistently report over 90% accuracy, showcasing the robustness of the model in recognizing variousactivities accurately.



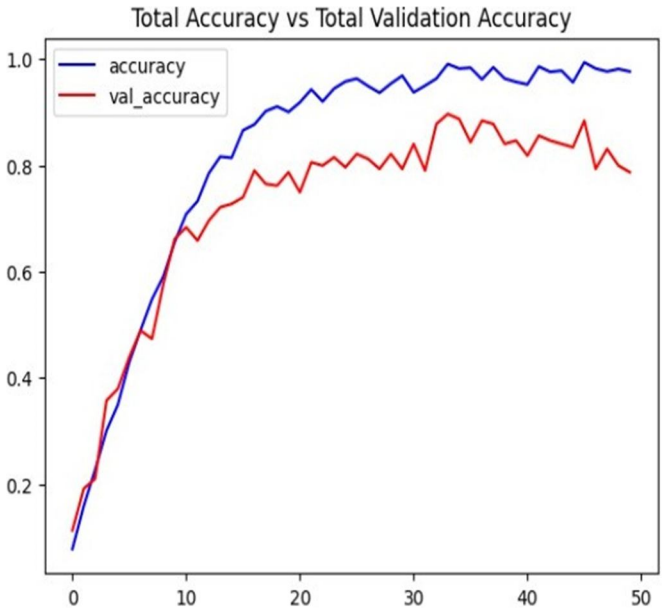


Fig. 6. Papers vs Accuracy rates

What sets our model apart is its resilience on unseen data, underscoring its reliability in real-world scenarios. This reliability is paramount when integrating the model into user- friendly systems, allowing seamless interaction and delivering precise action predictions from input videos.

Moreover, the model's scalability for real-time processing makes it adaptable across a spectrum of applications, from ensuring safety in industrial settings to enhancing healthcare monitoring systems and enriching entertainment experiences through immersive interfaces.

Notably, our model stands out with an impressive accuracy rate of 0.93, surpassing other existing models in the field. This superior performance underscores its effectiveness and underscores its potential to revolutionize various domains reliant on accurate action recognition.

In essence, the fusion of LSTM and CNN for HAR presents promising prospects, offering tangible benefits that extend across industries, ultimately fostering safer environments, improving healthcare outcomes, and enhancing user experiences through precise action recognition.

VI. IMAGE DESCRIPTION

1) Human Motion Detection UI (Fig. 5) This visual repre- sentation depicts a webpage with a basic layout featuring a textbox to insert a YouTube video link and a button to analyze the motion in the video. The interface is designedto be minimalistic and user friendly.
2) Human Motion Detection UI (Fig. 6) Analyzing the Youtube video to predict the Human Action
3) Human Motion Detection UI (Fig. 7) Predicted Human Action

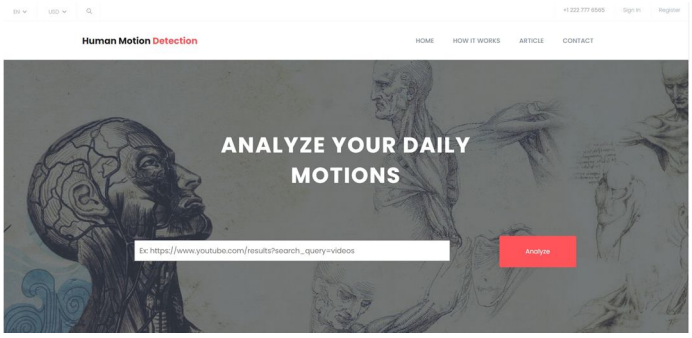


Fig. 7. Web Page for Human Action Prediction

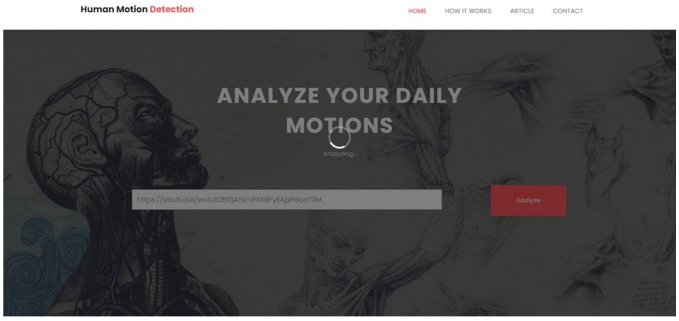


Fig. 8. Analyzing the Youtube Video to Predict the Human Action

VII. FUTURE SCOPE

1) Enhanced Accuracy: Continuously improving the accu- racy and robustness of the model by collecting more diverse and extensive datasets, and exploring advanced neural network architectures.
2) Action Recognition: Extending the model's capabilities beyond motion detection to recognize specific actions or activities performed by humans, such as walking ,running, sitting etc..

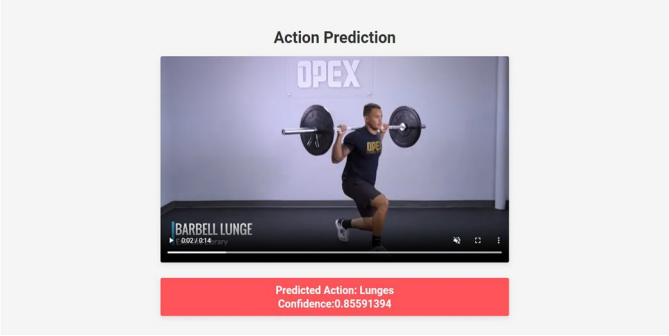


Fig. 9. Predicted Human Action

3) Adaptive Systems: Building adaptive systems that can learn and adapt to changes in human behavior or environ- mental conditions over time, making them more versatile and applicable in dynamic real-world scenarios.
4) Autonomous Systems: Integrating human motion detec- tion capabilities into autonomous systems such as robots or self-driving vehicles to enhance their ability to perceive and interact with humans in their environment safely and effectively.

5) Cross-domain Collaboration: Collaborating with experts from other domains such as biomechanics, sports science, or entertainment to explore novel applications and inter- disciplinary research opportunities.
6) Real-time Applications: Optimizing the model for real- time inference to enable applications such as real-time surveillance, gesture recognition in human-computer in- teraction, and augmented reality experiences.
7) Multi-modal Fusion: Integrating data from multiple sen- sors (e.g., depth sensors, accelerometers) or modalities (e.g., audio, visual) to enhance the model's understanding of human motion and improve overall performance.
8) Privacy-preserving Solutions: Developing techniques to perform human motion detection while respecting privacy concerns, such as using federated learning or on-device processing to avoid transmitting sensitive data to central- ized servers.
9) Healthcare Applications: Applying human motion detec- tion technologies in healthcare for monitoring patient movements, detecting abnormalities, assisting in rehabil- itation, or improving ergonomics to prevent injuries in workplaces.
10) Accessible Interfaces: Creating accessible interfaces for individuals with disabilities by leveraging human motion detection technology to develop assistive devices, com- munication aids, or rehabilitation tools tailored to their specific needs.

VIII. CONCLUSION

The endeavor to develop a human motion detection system through machine learning, employing CNNs in tandem with LSTM algorithms, embodies a powerful synergy between computational prowess and human biomechanics.

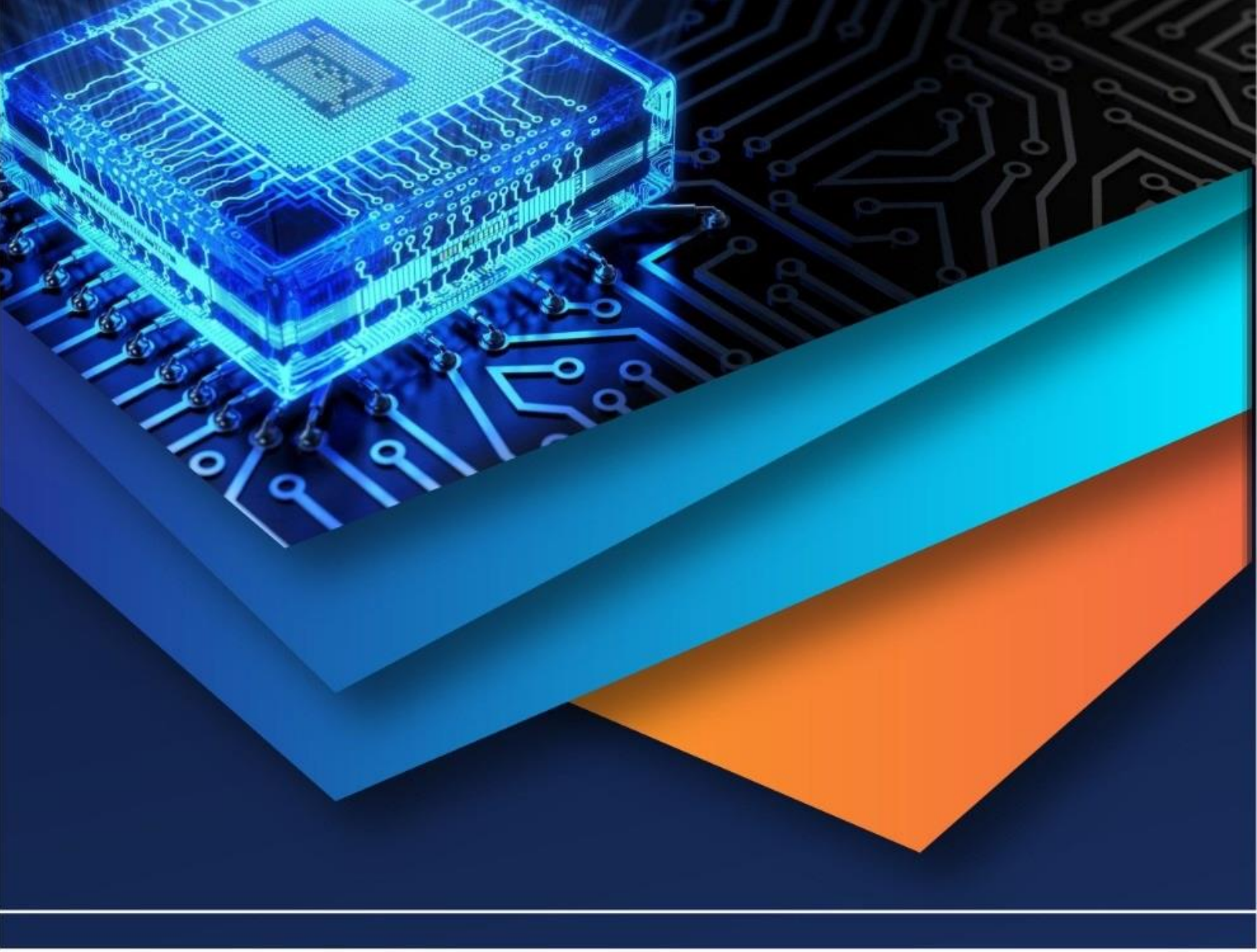
This project is dedicated to meticulously curating data, refining algorithms, and iteratively enhancing performance to achieve unparalleled proficiency in discerning and categorizing human movements within video streams. Spanning a diverse array of applica- tions, each brimming with transformative potential, lies an unwavering pursuit of heightened accuracy, driven by the ongoing expansion and diversification of datasets, continual refinement of network architectures, and meticulous calibra- tion of hyperparameters. Through advanced neural networks and multimodal fusion techniques, the system aims not only to identify human motion but also to understand subtle nuancesof behavior, offering transformative implications across diversedomains such as healthcare, robotics, and entertainment.

As this project unfolds, ethical considerations will guide its trajectory, ensuring judicious stewardship of data privacy and the mitigation of algorithmic biases. Through fostering cross- disciplinary collaborations and embracing a holistic ethos of responsible innovation, this project endeavors to forge solu- tions that transcend technological frontiers while upholding societal values and ethical principles.

In summary, this project epitomizes the synergistic fusionof computational ingenuity and human understanding, forging pathways toward transformative innovation and societaladvancement. As technological possibilities expand, this endeavor holds potential to catalyze paradigm shifts, propel scientific inquiry, and enrich the tapestry of human experience.

REFERENCES

[1] Raj, Ravi, and Andrzej Kos. "An improved human activity recognition technique based on convolutional neural network." Scientific Reports 13.1 (2023): 22581
[2] Sun, Ruiqi, et al. "Human action recognition using a convolutionalneural network based on skeleton heatmaps from two-stage pose es- timation." Biomimetic Intelligence and Robotics 2.3 (2022): 100062.
[3] Terreran, Matteo, Leonardo Barcellona, and Stefano Ghidoni. "A general skeleton-based action and gesture recognition framework for human–robot collaboration." Robotics and Autonomous Systems 170 (2023): 104523.
[4] Ahmad, Tasweer, Syed Tahir Hussain Rizvi, and Neel Kanwal. "Trans- forming spatio-temporal self-attention using action embedding for skeleton-based action recognition." Journal of Visual Communication and Image Representation 95 (2023): 103892.
[5] Saoudi, El Mehdi, Jaafar Jaafari, and Said Jai Andaloussi. "Advancing human action recognition: A hybrid approach using attention-based LSTM and 3D CNN." Scientific African 21 (2023): e01796.
[6] Golestani, Negar, and Mahta Moghaddam. "Human activity recognition using magnetic induction-based motion signals and deep recurrent neural networks." Nature communications 11.1 (2020): 1551.
[7] Yin, Xinpeng, et al. "An adaptively multi-correlations aggregation network for skeleton-based motion recognition." Scientific Reports 13.1(2023): 19138.
[8] Zhang, Jun, et al. "Table tennis motion recognition based on the bat trajectory using varying-length-input convolution neural networks." Scientific Reports 14.1 (2024): 3549.
[9] Ibrahim Elmadfa, Alexa L. Meyer, Chapter 5 - Nutrition, aging, and requirements in the elderly, Editor(s): Bernadette P. Marriott, Diane
[10] Birt, Virginia A. Stallings, Allison A. Yates, Present Knowledge in Nutrition (Eleventh Edition), Academic Press, 2020,
[11] Amir Talaei-Khoei, Jay Daniel, How younger elderly realize usefulness of cognitive training video games to maintain their independent living, International Journal of Information Management, Volume 42, 2018,



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY