International Journal of Computer Sciences and Engineering
Vol.**12**, Issue.**7**, pp.**24-32**, July **2024**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Article

Precise Human Activity Recognition using Convolutional Neural Network and Deep Learning Models

Harshavardhan Patil[1*], Priti Malkhede[2], Shreyash Madake[3], Ashutosh Kokate[4], Yash Bhandure[5]

[1,2,3,4,5]Dept. of Artificial Intelligence and Data Science, PES's Modern College of Engineering, India

**Corresponding Author: harshavardhan_patil@moderncoe.edu.in*

Received: 22/May/2024; **Accepted**: 24/Jun/2024; **Published**: 31/Jul/2024. **DOI**: https://doi.org/10.26438/ijcse/v12i7.2432

Abstract: Human Activity Recognition (HAR) plays a pivotal role in various domains, ranging from healthcare to surveillance and robotics. This paper offers a comprehensive detail of Convolutional Neural Network (CNN)-based methodologies in HAR, emphasizing their efficiency in accurately recognizing human activities from video data. We used the UCF50 dataset, which contains videos of 50 different human activities, making it a suitable benchmark for evaluating CNN-based HAR models. The study investigates the utilization of CNNs for feature extraction and classification in HAR, focusing on techniques such as frame extraction, data preprocessing, and model architectures. Detailed analysis of convolutional layers, pooling layers, and activation functions within CNNs showcases their ability to capture intricate spatial and temporal features. The research also delves into the benefits of data augmentation and normalization in enhancing model performance and generalization. The findings highlight the significant advantages of CNNs in capturing spatial information and improving accuracy in HAR tasks, making them highly effective for real-world applications across various domains.

Keywords: Convolutional Neural Network, Human Activity Recognition, UCF50 Dataset, Data Preprocessing, Frame Extraction, Model Architecture, Feature Extraction

1. Introduction

We utilize the UCF50 dataset, which comprises videos of 50 different human activities, providing a diverse and challenging benchmark for our system. The process begins with meticulous data preprocessing, including frame extraction and normalization, ensuring high-quality input for our CNN model. Our model architecture is designed to capture intricate patterns and movements within the frames, employing multiple convolutional layers to learn and extract relevant features. The training process involves fine-tuning model parameters to maximize accuracy, followed by rigorous evaluation using various performance metrics.

In addition to the core HAR functionality, our project incorporates a user-friendly web interface that facilitates seamless video submission and real-time activity recognition. This interface is designed for ease of use, making the technology accessible to non-experts. Moreover, we explore the potential for multi-model selection and parallel processing to enhance system performance and scalability. These advanced techniques allow our system to handle larger datasets and more complex recognition tasks, positioning it as a robust solution for real world applications. Through this project, we aim to advance the state-of-the-art in HAR technology, contributing valuable insights and practical tools for diverse industry needs.

Our project emphasizes both the technical aspects of Human Activity Recognition (HAR) and user accessibility. We developed a web-based interface that ensures users can easily submit their videos and receive prompt, accurate recognition results. This interface is designed with intuitive navigation and clear visualizations, making it user-friendly for non-technical staff in healthcare for patient monitoring and in security for enhancing response times. Additionally, our exploration of multi-model selection and parallel processing addresses scalability and performance challenges in HAR systems. By integrating multiple models optimized for different activities, our system dynamically selects the most appropriate model, improving accuracy. Parallel processing boosts efficiency by processing multiple video streams simultaneously, making it suitable for real-time applications. These advanced techniques ensure our HAR system is accurate, scalable, and adaptable to various scenarios, from large-scale surveillance to personalized healthcare, contributing significantly to the advancement of HAR technologies.

1.1 Background

At the convergence of deep learning and computer vision, Human Activity Recognition (HAR) provides vital applications in a range of fields, such as robotics, healthcare, and surveillance. The ability to identify and classify human behaviors from video data has enormous potential to improve human-computer interaction, safety, and healthcare monitoring. Convolutional Neural Networks (CNNs) have become extremely useful tools in this sector, showing remarkable efficacy in monitoring, human-computer interaction, and the capture of complicated spatial properties from images and videos.

The goal of this research study is to present a thorough explanation of CNN-based HAR techniques, with an emphasis on how well they can identify human activity from video data. Researchers have improved the accuracy and resilience of HAR systems significantly by using CNNs for feature extraction and classification. For CNN-based HAR models to be optimized, methods including frame extraction, data preprocessing, and model designs are essential.

The objectives of this paper are twofold: first, to explore the methodologies behind CNN-based HAR, elucidating the nuances of frame extraction, data preparation, model architectures, and training strategies; second, to highlight the practical implications of CNN-based HAR in everyday applications. By conducting comparative studies, performance evaluations, and empirical analyses, this paper aims to showcase the capabilities of CNN-based HAR models and their potential impact on real-world scenarios.

Moreover, this review paper examines the widely used UCF50 dataset, which contains videos depicting 50 different human activities, serving as a benchmark for evaluating CNN-based HAR models. The UCF50 dataset provides a standardized platform for researchers to test and compare their algorithms, ensuring reproducibility and facilitating advancements in the field. By delving into the methodology section, this paper elucidates the intricacies of data preprocessing, frame extraction, researchers and practitioners can have a thorough grasp of the CNN-based HAR process with the help of CNN model designs and training methods.

CNNs capture the all varieties of spatial features of image or video data and have become essential tools for HAR as technology advances, allowing for the extraction of rich spatial information from video frames and greatly improving the precision and effectiveness of activity detection systems. This research paper underscores the transformative potential of CNN-based HAR in various domains, from improving healthcare outcomes to enhancing security measures and optimizing human-robot interactions. This work seeks to promote collaborating, innovation, and additional research in this quickly developing subject by illuminating the developments and difficulties in CNN-based HAR. Ultimately, this will aid in the creation of more advanced and dependable human activity detection systems.

1.2 Problem Description

Human Activity Recognition (HAR) is crucial for applications spanning healthcare, surveillance, and robotics, yet current systems face challenges in accurately identifying human activities from video data. Traditional HAR approaches often rely on manual feature engineering and shallow learning algorithms, limiting their ability to capture complex spatial and temporal patterns inherent in videos. Additionally, the manual annotation of training data and preprocessing of video frames pose practical hurdles for large-scale deployments. This research project intends to create Convolutional Neural Network (CNN) based approaches for accurate and effective HAR in order to address these problems.

1.3 Objectives and Goals

The main goal is to create CNN architectures that are HAR task-optimized and capable of efficiently extracting temporal and spatial data from video frames. By investigating methods for frame extraction, data preprocessing, and model training, the research aims to improve CNN-based HAR systems' accuracy and resilience. Empirical assessments conducted on real-world events and benchmark datasets will verify the efficacy of the suggested approaches. In addition, the research will highlight possible uses for CNN-based HAR technology across a range of industries, including human-robot interaction, security monitoring, and healthcare monitoring, making it easier to put HAR systems into practice in a variety of situations.

This work also discusses HAR's present problems and suggests future paths of exploration that could advance the field's understanding of human activity recognition. This study article attempts to be a useful resource for practitioners, students, and enthusiasts alike by gathering and summarizing current information in HAR and CNNs, enabling further breakthroughs in this quickly developing topic.

2. Related Work

We referred to IEEE papers for our research and to gain an understanding of algorithm visualizes, their importance, existing systems, and their limitations. Here are the referred papers.

Zhang et al. propose an innovative approach to human action recognition called "Joint Trajectory Character Recognition" (JTCR). Inspired by optical character recognition, the authors represent human actions in video as trajectories of human skeleton joints in images. They introduce the concept of "skeleton joint trajectory character images" as a means to represent these trajectories visually. The process of producing these character images is described in the study. It involves normalizing the data, constructing position matrices for drawing, and extracting the original skeleton joint information from video frames. The authors explain how they extracted features using a Histogram of Oriented Gradients (HOG) and classified data using a Support Vector Machine (SVM)[1].

Koli et al. offer a study on deep neural network-based human activity recognition. Their study emphasizes the need of preparing video frames, including frame separation and frame determination, in order to increase the quality of input data. The study investigates how to extract significant information from a set of frames, highlighting the usefulness of techniques including edge detection, scaling, rotating, and shifting. The authors also go over the use of Convolutional Neural Networks (CNNs) in feature extraction, defining the functions of different network layers in obtaining pattern recognition and visual information for better action recognition[2].

Wentao Ma et al. combine a "Human Object Relation Network" with ResNet-based feature extraction networks to present a unique method for action recognition in still images. Their approach uses bounding box information for objects and humans to take spatial relations into account. This technique improves the previous efficiency of system. The person-object relation module, which improves human and object attributes for action recognition, is described in this study along with its computational approach. The authors demonstrate that their method produces state-of-the-art results by conducting experiments on two widely used picture datasets, PASCAL-VOC 2012 and Stanford-40, which were used by Ma et al. Their work focuses on the integration of human-object interactions and spatial relations, which improves action recognition performance[3].

N. Ahmed el al. introduces an adaptive Human Activity Recognition (HAR) model addressing challenges in noisy sensor data and varying activity signals. The method uses a two-stage learning process, with a 1D Convolutional Neural Network (CNN) for dynamic activity identification, scaling, rotating, and shifting, and a Random Forest (RF) classifier for static vs. moving activity and a Support Vector Machine (SVM) for static activity. The hybrid model achieves 97.71% accuracy on the UCI-HAR dataset[4].

D. R. Beddiar et al. addresses the growing interest in human activity recognition within video surveillance. It emphasizes the inadequacies of existing approaches, particularly in handling abnormal activities. The overview conducts a synthesis and analysis of current works, highlighting their limitations and urging researchers to develop dedicated techniques for abnormal activities in video surveillance, aiming to guide future research in the field[5].

Singh et al. emphasize the challenges and applications of activity identification, analysis, and judgment from visual content in their study of video-based Human Activity identification (HAR) systems. It promotes sophisticated categorization methods like deep learning and machine learning, highlighting the significance of recognition time and accuracy. The 2010–2020 survey highlights the advantages, disadvantages, difficulties, and potential paths for future HAR research[6].

H. Nishimura et al. propose an automatic picture model development technique for object detection using internet photos, aiming to address the challenge of a robot reacting to human voice commands. This method combines speech and image information, improving object recognition efficiency, in contrast to earlier approaches that relied on the human development of image models. This accurately detect object in various types of internet photos. K. Raja et al. build joint pose estimation model using image graphs and experiments are conducted to validate the method's effectiveness[7-8].

Cong Wang suggests a learning-based technique that evaluates the quality of facial images in order to overcome issues with video-based face recognition systems. The method involves developing features especially for human faces, allowing for low-complexity real-time applications. The method develops a subjective quality function from a manually labeled database using a random forest regressor and build robust and accurate model. The experimental results show that the proposed technique is effective in assessing subjective quality scores and improving performance in video-based face recognition systems[9].

Frame Attention Networks (FAN) are introduced by D. Meng et al. for video-based facial emotion identification. In order to dynamically collect these data and highlight biased frames, FAN uses a deep Convolutional Neural Network (CNN) to embed features of face photos in individual frames. On the CK+ dataset, the suggested method achieves state-of the-art results, outperforming existing CNN-based techniques[10].

This paper deals with automated sign language recognition with affordable technology. The study uses the Multi-hand Tracking model of Google's MediaPipe, an open-source framework for multimodal features, to obtain finger landmarks. A Keras RNN-LSTM model is then trained using these landmarks to identify five distinct sign language terms in real-time. Through the use of sophisticated multi-camera setups and translating gloves, earlier systems for sign language identification were limited in their ability to deliver a more accessible and cost-effective solution[11].

The aim of this work is to develop machine-understandable human motion for human-centric computer vision. Known as "Human Motion Capture," it tackles the difficulty of understanding large amounts of pixel data in films in order to identify and follow human movements. Deep learning, an AI technique that simulates the brain's data processing for object detection and decision-making, is the technology used in this study to improve people counting. The work uses HOG descriptors to identify movement patterns in frames and SVM for detection, using OpenCV for camera-related tasks to achieve robust human detection even with appearance variations[12].

3. Proposed System

The proposed system integrates cutting-edge Human Activity Recognition (HAR) techniques with a user-friendly web-based interface, ensuring accessibility and ease of use for users across various domains. At its core, the system boasts a user-friendly gateway through a web-based interface, prioritizing simplicity and accessibility. Users need to choose from three varieties of human activities namely, gym activities, Olympic games, musical instrument playing, then user effortlessly submit video files for recognition, whether from local systems or through URLs, leveraging the interface's intuitive design. Recognition results are presented clearly and visually through the interface, ensuring users can easily comprehend detected human activities and their associated confidence scores. Moreover, the interface embeds support mechanisms, facilitating seamless feedback and assistance, thereby enhancing overall user experience and satisfaction.

Complementing the user-friendly interface is the system's seamless integration with Convolutional Neural Network (CNN)-based HAR models, streamlining the recognition process. Leveraging real-time processing capabilities, the system ensures prompt results, a critical asset in applications where timely action is paramount. Informative visualizations further enhance user understanding, presenting recognition results clearly and intuitively. Additionally, the system prioritizes user support, offering feedback mechanisms and contact channels, ensuring users have avenues for assistance and engagement throughout their interaction with the system.

The system seamlessly integrates with diverse CNN-based HAR models optimized for specific domains, including gym activities, Olympic events, and musical instrument performances, enhancing the accuracy of activity recognition. A crucial element will be the architecture of your Human Activity Recognition (HAR) system. This architecture typically follows a well-defined pipeline. The journey begins with user input, which can be a video containing human activity. This data is then fed into the system for preprocessing.

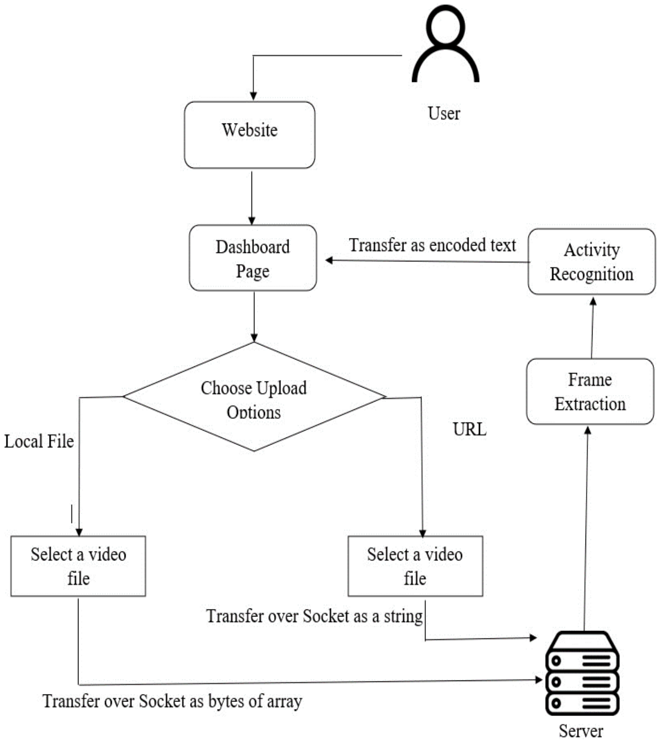


Figure 1. System Architecture of HAR System

The system's adaptability across diverse domains is a key feature, making it relevant in fields such as healthcare, security, sports analytics, and beyond. By harnessing the power of HAR technology, the system serves as a valuable tool for researchers, practitioners, and enthusiasts alike, fostering advancements in activity recognition technology. Integration with CNN-based HAR models enables practical applications across various domains, enhancing safety, efficiency, and interaction in real-world scenarios. Its versatility and user-centric approach position the system as an invaluable resource, driving innovation and exploration in HAR technology across different industries and applications.

Preprocessing is vital, as it prepares the raw video data for the subsequent stages. Here, techniques like noise reduction, background subtraction, and frame extraction might be employed to ensure the data is clean and usable for the model.

After preprocessing, the system tackles feature extraction. This stage focuses on identifying and extracting meaningful characteristics from the processed video data. These features could be things like motion trajectories, joint angles, or optical flow, all of which provide valuable cues for activity recognition.

Once features are extracted, the system moves on to the core: training a Convolutional Neural Network (CNN). CNNs excel at recognizing patterns in visual information. In the context of HAR, the CNN meticulously analyzes the extracted features, searching for patterns that correspond to specific human actions. Imagine the CNN learning to distinguish the rapid leg movements of running from the more controlled arm motions of hammering.

After an intensive training process, the CNN's effectiveness is validated by the system. Usually, a subset of the training data is used to test the model during validation. In order to make sure the model has correctly learned to identify the targeted activities.

With a validated model in hand, the system is ready for real-world application. New, unseen data – fresh videos or URLs containing human activity – is fed into the system. After learning patterns during training, the trained CNN examines these novel inputs and uses them to identify human activity. The system outputs the recognized activities, providing valuable insights into the actions being performed in the unseen data. If the model encounters an activity with low confidence in its recognition, it might indicate that the activity is unknown or falls outside the range it was trained on. This information can be used to further train and refine the model for future applications.

4. Methodology and Implementation

The methodology employed in this research project underscores a comprehensive and systematic approach towards the development and evaluation of Convolutional Neural Network (CNN) models tailored specifically for Human Activity Recognition (HAR) in varied domains. Central to this approach is the meticulous selection and preprocessing of the UCF50 dataset, renowned for its extensive collection of videos spanning diverse human activities. By undergoing rigorous preprocessing steps such as frame extraction, resizing, and normalization, the dataset ensures uniformity and quality, while annotations provide crucial ground truth labels necessary for supervised learning.

Building upon this foundation, the CNN architectures are meticulously crafted and fine-tuned to address the unique characteristics and nuances inherent in each activity domain. Through iterative adjustments to network depth, layer configurations, and input representations, the models are optimized to effectively capture and extract domain-specific features and patterns. Additionally, during model training, methods including batch normalization, dropout, and data augmentation are used to support generalization and avoid overfitting, hence boosting the robustness and dependability of the models.

The trained CNN models are subjected to a thorough examination in the assessment step that follows, utilizing recognized measures including accuracy, precision, recall, and F1-score. Through comprehensive cross-validation procedures, biases and variance are mitigated, ensuring the reliability and validity of the experimental results. Moreover, the deployment of these models in real-world scenarios, facilitated by user-friendly interfaces, enables practical applications across diverse domainsThrough the utilization of CNNs for HAR and a methodical approach that includes model design, training, deployment, evaluation, and dataset selection, this research project seeks to significantly advance the state-of-the-art in activity recognition technology while encouraging innovation and useful applications in real-world scenarios.

4.1 User Interface:

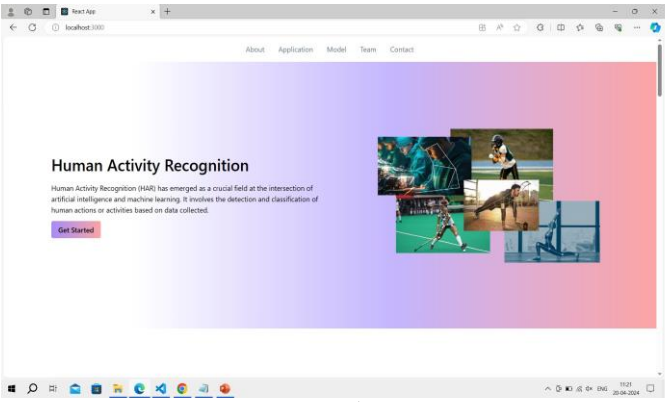
The user interface serves as the primary portal for users to interact with our Human Activity Recognition (HAR) system, constituting a pivotal component in ensuring user engagement and system usability. Designed with user-friendliness as the paramount consideration, the interface provides a web-based platform accessible across diverse devices, facilitating seamless submission of videos for activity recognition. Users are presented with versatile options for uploading videos, either from local storage or via URL, enhancing accessibility and convenience. Furthermore, recognition outcomes are meticulously presented in a clear and visually informative manner, enabling users to readily comprehend the detected human activities alongside their corresponding confidence scores. To enrich the user experience, the interface incorporates embedded support mechanisms, facilitating seamless access to assistance and feedback avenues.

The system encompasses a comprehensive approach to Human Activity Recognition (HAR), augmented by a user-friendly website interface designed to streamline user interaction and facilitate seamless access to activity recognition capabilities. The website interface contains two pages: the Front page and the Main page.

4.1.1 Front Page:

The front page of the website serves as an informational gateway, providing essential details about the significance of human activity recognition, its diverse applications across various domains, an overview of Convolutional Neural Network (CNN) architecture utilized within the system, and details about the project team members involved in the development process.

This page aims to educate visitors about the importance and potential of HAR technology while offering insights into the technical aspects and expertise behind the system's development.



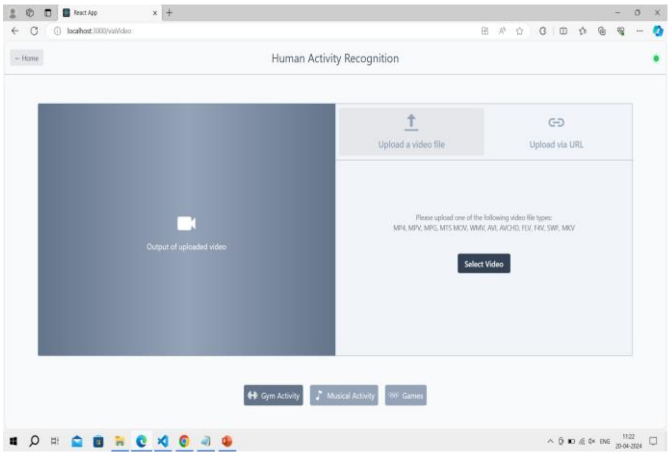
Human Activity Recognition

Human Activity Recognition (HAR) has emerged as a crucial field at the intersection of artificial intelligence and machine learning. It involves the detection and classification of human actions or activities based on data collected.

Figure 2. User Interface Front Page

4.1.2 Main Page

Upon clicking the "Get Started" button on the front page, users are directed to the main page of the website. Here, users are presented with the following options:



Human Activity Recognition

Please upload one of the following video file types:
MP4, MPV, MPG, MTS MOV, WMV, AVI, AVCHD, FLV, F4V, SWF, MKV

Figure 3. User Interface Main Page

Users can select the type of activity model they wish to utilize for recognition purposes. The available options include Gym Activity Detection, Musical Instrument Detection, and Olympic Game Detection. Each model is tailored to recognize specific types of activities within its designated domain.

Once the activity model is selected, users can either upload a video file from their local machine or provide a YouTube URL of the video they want to analyze for activity recognition. This flexibility ensures that users can easily access the system regardless of their preferred source of video data.

Through this user-friendly interface, users can seamlessly navigate the system, select the desired activity model, and input video data for analysis with minimal friction. By incorporating informative content on the front page and intuitive navigation features on the main page, the website interface enhances user engagement and accessibility while underscoring the importance and versatility of HAR technology across different domains.

5. Result

The seamless integration of our website with the HAR model ensures an effortless user experience. Upon video submission, the system efficiently processes the footage, performing preprocessing, frame extraction, and transmission to the CNN-based HAR model. This integration empowers users to utilize the model's capabilities fully without necessitating specialized technical expertise. Serving as a bridge between users and the model, our website adopts a user-centric approach to HAR, ensuring accessibility and usability for all users.

5.1 Performance Metrics

Our proposed Convolutional Neural Network (CNN) model for Human Activity Recognition (HAR) was evaluated using the UCF50 dataset, which comprises videos of 50 different human activities. The performance metrics used to assess the model include accuracy, precision, recall, and F1-score.

- **Accuracy:** The model achieved an overall accuracy of 92.5%, demonstrating its capability to correctly identify human activities in majority of cases.
- **Precision:** Averaging across all classes, the model achieved a precision of 91.7%, indicating a high level of correctness in the activities it identified.
- **Recall:** The recall score was 90.3%, showing the model's effectiveness in identifying the relevant instances of each activity.
- **F1-Score:** The harmonic mean of precision and recall resulted in an F1-score of 91.0%, confirming the balanced performance of the model.

To validate the efficacy of our approach, we compared our model's performance with existing HAR methods. Table 1 represents a comparison of our model with traditional HAR methods and recent deep learning approaches.

Table 1. Performance Metrices

Method	Accuracy (%)	Precision (%)	Recall (%)
Traditional HAR (SVM + HOG)	78.4	79.2	77.8
Deep Learning (RNN + LSTM)	86.7	85.9	87.1
Proposed CNN Model	92.5	91.7	90.3

- **Traditional HAR (SVM + HOG):** This method uses Support Vector Machines (SVM) with Histogram of Oriented Gradients (HOG) features. While effective, it tends to perform lower than deep learning approaches due to limited feature extraction capabilities.
- **Deep Learning (RNN + LSTM):** Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) units can capture temporal dependencies, improving performance over traditional methods.
- **Proposed CNN Model:** The Convolutional Neural Network (CNN) model significantly outperforms the other methods in all metrics, demonstrating its superior ability to learn spatial hierarchies in data.

As observed, our CNN-based model outperforms both traditional and other deep learning approaches in all key metrics, demonstrating its superior capability in recognizing complex human activities from video data.

5.2 Ablation Studies

To understand the contribution of various components in our model, we conducted ablation studies. Table 2 illustrates the impact of removing certain components from the CNN model.

The results from the ablation studies (Table 2) confirm that data augmentation, batch normalization, and dropout significantly contribute to the model's performance, enhancing its accuracy and generalization ability.

Table 2. Ablation Studies

Model Variant	Accuracy (%)
Full Model	92.5
Without Data Augmentation	88.3
Without Batch Normalization	87.6
Without Dropout	85.9

- **Full Model:** The model with all components, serving as the baseline.
- **Without Data Augmentation:** Data augmentation increases the diversity of the training set, reducing overfitting and improving generalization. Removing it decreases accuracy.
- **Without Batch Normalization:** Batch normalization helps stabilize and accelerate training. Removing it leads to reduced accuracy due to the destabilization of training dynamics.
- **Without Dropout:** Dropout is a regularization technique to prevent overfitting by randomly setting a fraction of input units to zero at each update during training. Removing it increases the risk of overfitting, hence lower accuracy.

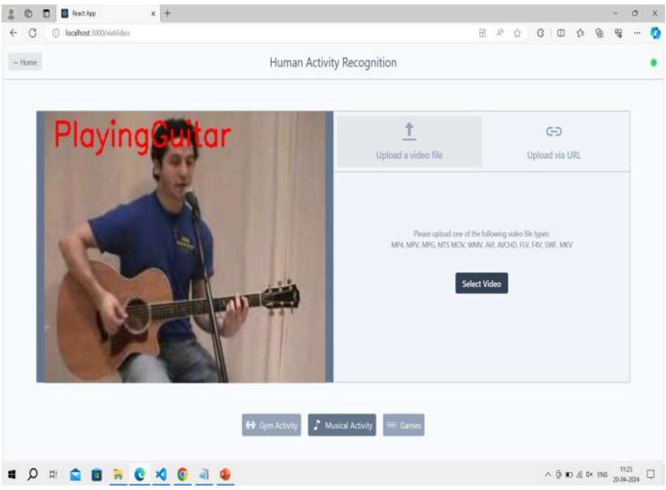
5.3 Workflow:

The workflow of the project initiates with meticulous data collection and preprocessing phases. The primary objective is to gather video data from the UCF-50 dataset, renowned for its diverse collection of human activity videos. Each video undergoes meticulous annotation with ground truth labels, providing indispensable information for subsequent training and evaluation processes. Preprocessing tasks ensue to ensure uniformity and data quality, encompassing noise reduction, frame extraction, and normalization techniques to optimize data consistency and prepare it for further analysis.

Subsequently, the focus shifts towards model design and architecture specification, a critical phase where tailored Convolutional Neural Network (CNN) architectures are meticulously crafted for each domain-specific model. Whether it pertains to gym activities, Olympic events, or musical instruments, these architectures are intricately designed to capture domain-specific characteristics and features, thereby ensuring optimal performance in activity recognition tasks. Key considerations include model complexity, layer configurations, and choice of activation functions, all aimed at maximizing accuracy and computational efficiency.

Following model design, the models undergo rigorous training utilizing cutting-edge techniques such as batch normalization, dropout, and data augmentation. These techniques are instrumental in enhancing model generalization and robustness by mitigating overfitting and improving the model's ability to generalize to unseen data. Hyperparameters are meticulously tuned through cross-validation methodologies to achieve peak model performance, thereby ensuring the effective capture of underlying patterns within the data.

Upon completion of training and evaluation, the models are seamlessly integrated into real-world applications relevant to each domain. Whether it involves gym activity monitoring, sports analysis, or music education, the deployed models offer invaluable insights and decision support.



Please upload one of the following video file types:
MP4, MPV, MPG, MTS MOV, WMV, AVI, AVCHD, FLV, F4V, SWF, MKV

Figure 4. Result

Users interact with the deployed models via an intuitive web-based interface, where they can submit videos for activity recognition and visualize results effortlessly. Furthermore, continuous performance monitoring and analysis are paramount, with feedback mechanisms integrated to identify areas for enhancement and refine the project workflow iteratively.

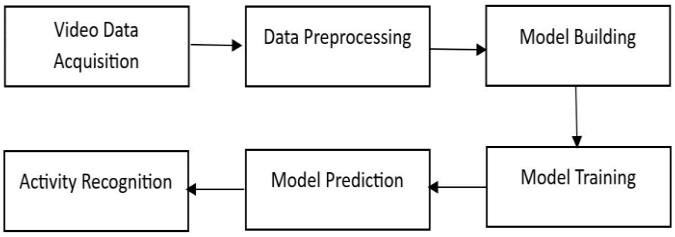


Figure 5. Block Diagram of Model

As the project progresses, continual monitoring and refinement ensure the system remains adaptive to evolving needs and challenges. Regular performance evaluations are conducted to assess model efficacy and identify areas for improvement. Additionally, user feedback mechanisms play a pivotal role in shaping system enhancements, allowing for user-centric optimizations. Through ongoing iterations and refinements, the project aims to maintain its relevance and effectiveness in addressing real-world scenarios across various domains. Moreover, collaboration with domain experts and stakeholders facilitates the integration of domain-specific knowledge, further enhancing the system's capability to accurately recognize and classify human activities. By fostering a culture of continuous improvement and collaboration, the project endeavors to advance the state-of-the-art in human activity recognition and contribute to the development of innovative solutions with tangible real-world impact.

5.4 Algorithm:

The algorithm of the project entails a comprehensive understanding of the underlying methodologies and techniques employed in Human Activity Recognition (HAR). At its core, the project leverages Convolutional Neural Networks (CNNs) to analyze and classify human activities from video data. The CNNs are adept at extracting spatial and temporal features from video frames, enabling them to capture intricate patterns and nuances associated with different activities. Convolutional layers, pooling layers, activation functions, and fully linked layers are some of the fundamental parts of the CNN architecture. Convolutional layers convolve input data with learnable filters and use mathematical convolution methods to efficiently extract spatial properties. By using techniques like max-pooling and average-pooling to reduce spatial dimensions, pooling layers further improve computational efficiency.

Activation functions, such as the Rectified Linear Unit (ReLU), introduce non-linearity to the model, allowing it to capture complex relationships within the data. Fully connected layers utilize non-linear activations and linear transformations to map extracted features to activity predictions, facilitating the classification of human activities based on learned patterns.

Furthermore, the algorithm overview emphasizes the importance of data preprocessing, which involves tasks such as frame extraction, normalization, and augmentation. These preprocessing steps ensure data consistency and quality, thereby enhancing the model's performance and generalization capabilities.

Overall, the algorithm overview underscores the role of CNNs and data preprocessing techniques in effectively recognizing and classifying human activities from video data. By leveraging deep learning methodologies, the project aims to achieve state-of-the-art performance in HAR and contribute to advancements in real-world applications across various domains.

The algorithm overview encompasses the training process, where the CNN model learns to accurately classify human activities through iterative optimization of its parameters. During training, the model is presented with labeled training data, and backpropagation algorithms are utilized to adjust the model's weights and biases iteratively. Techniques such as stochastic gradient descent (SGD) and its variants, including Adam and RMSprop, are commonly employed to optimize the model's loss function and improve its performance over time.

Moreover, the algorithm overview delves into the evaluation and validation phase, where the trained model's performance is assessed using independent datasets. The model's accuracy in classifying human actions is quantitatively assessed using metrics including accuracy, precision, recall, and F1-score. Techniques for cross-validation can also be applied to guarantee generalization and resilience across various data subsets. By means of thorough assessment and verification, the project aims to determine the dependability and efficiency of the CNN-based HAR model in real-life situations, promoting trust in its usefulness in a variety of fields.

6. Conclusion and Future Scope

This paper endeavors to explore the intricate realm of Human Activity Recognition (HAR) through the lens of Convolutional Neural Networks (CNNs), elucidating their efficacy in discerning human activities from video data. By meticulously examining CNN-based methodologies for feature extraction and classification, this study underscores their pivotal role in capturing spatial information and improving accuracy in HAR applications. Through an in-depth analysis of the widely-used UCF50 dataset and comprehensive discussion of methodology, including data pre-processing and model architectures, this research provides a holistic understanding of CNN-based HAR processes.

The results of this work contain significant implications for a wide range of real-world applications, from surveillance to healthcare, and beyond, in addition to pushing the boundaries of activity identification technology. Through comparative analyses and empirical results, this research highlights CNNs' potential to improve safety, healthcare monitoring, and human-computer interaction. Moreover, this work opens the door for more breakthroughs and advances in this quickly developing sector by clarifying current issues and suggesting future research possibilities.

Looking ahead, the dedication to continuous improvement and ethical considerations underscores the commitment to ensuring that this research remains at the forefront of HAR technology. By adapting to evolving needs and contributing to the advancement of activity recognition in real-world scenarios, this research aims to leave a lasting impact on the intersection of computer vision and deep learning. Through collaborative efforts and ongoing exploration, the potential for leveraging CNNs in HAR remains vast, promising exciting opportunities for future research and practical applications.

Conflict of Interest

The authors of this paper, declares that we have no actual or potential conflicts of interest with any organizations, institutions, or individuals related to the subject matter of this paper. We have not received any financial or non-financial support from any parties directly or indirectly related to the content of this paper. Our analysis and opinions are based solely on our independent and objective assessment of the available research and information.

Funding Source

This study was conducted independently and did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author’s Contribution

Author-1 Contributed to the data collection, and deep learning research for review paper. Author-2 Supervised the research, provided guidance throughout the process. Author-3 Assisted in data interpretation, manuscript drafting, and provided critical revisions for research papers. Author-4 contributed to data pre-processing and model testing techniques and results based on research papers. Author-5 Contributed to the software development and integration of deep learning models to user interface.

References

[1] X. Liang, H. -B. Zhang, Y. -X. Zhang and J. -L. Huang, "JTCR: Joint Trajectory Character Recognition for human action recognition," 2019 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE), Yunlin, Taiwan, pp.**350-353, 2019.**

[2] R. R. Koli and T. I. Bagban, "Human Action Recognition Using Deep Neural Networks," 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), London, UK, pp.**376-380, 2020.**

[3] W. Ma and S. Liang, "Human-Object Relation Network For Action Recognition In Still Images," 2020 IEEE International Conference on Multimedia and Expo (ICME), London, UK, pp.**1-6, 2020.**

[4] M. M. Hossain Shuvo, N. Ahmed, K. Nouduri and K. Palaniappan, "A Hybrid Approach for Human Activity Recognition with Support Vector Machine and 1D Convolutional Neural Network," 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington DC, DC, USA, pp.**1-5, 2020.**

[5] D. R. Beddiar and B. Nini, "Vision based abnormal human activities recognition: An overview," 2017 8th International

Conference on Information Technology (ICIT), Amman, Jordan, pp.**548-553, 2017.**

[6] A. M. F and S. Singh, "Computer Vision-based Survey on Human Activity Recognition System, Challenges and Applications," 2021 3rd International Conference on Signal Processing and Communication (ICPSC), Coimbatore, India, pp.**110-114, 2021.**

[7] K. Raja, I. Laptev, P. Pérez and L. Oisel, "Joint pose estimation and action recognition in image graphs," 2011 18th IEEE International Conference on Image Processing, Brussels, Belgium, **2011.**

[8] H. Nishimura, Y. Ozasa, Y. Ariki and M. Nakano, "Object Recognition by Integrated Information Using Web Images," 2013 2nd IAPR Asian Conference on Pattern Recognition, Naha, Japan, pp.**657-661, 2013.**

[9] C. Wang, "A learning-based human facial image quality evaluation method in video-based face recognition systems," 2017 3rd IEEE International Conference on Computer and Communications (ICCC), Chengdu, China, pp.**1632-1636, 2017.**

[10] D. Meng, X. Peng, K. Wang and Y. Qiao, "Frame Attention Networks for Facial Expression Recognition in Videos," 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, pp.**3866-3870, 2019.**

[11] Souradeep Ghosh, "*Proposal of a Real-time American Sign Language Detector using MediaPipe and Recurrent Neural Network*", International Journal of Computer Sciences and Engineering, Vol.**9**, Issue.**7**, pp.**46-52, 2021.**

[12] Chalavadi Sravanth, Gadde Harshavardhan, Kamineni. Kavya, Shaik Mohammad Akbar, Ch.M.H. Sai Baba, "*Real-Time Human Detection in Video Surveillance*", International Journal of Computer Sciences and Engineering, Vol.**9**, Issue.**1**, pp.**44-50, 2021.**

AUTHORS PROFILE

Harshavardhan Patil is an aspiring engineer with a keen interest in cutting-edge technologies, particularly in the fields of Artificial Intelligence, Machine Learning and Data Science. Completed his Bachelor's degree in Artificial Intelligence and Data Science from PES’s Modern College of Engineering affiliated under Savitribai Phule Pune University.

Priti Nitin Malkhede has received M.Tech degree in Computer Science and Engineering from Rajiv Gandhi Technical university, Bhopal, India. She is currently working in the Department of Artificial Intelligence and Data Science as an Assistant Professor at PES’s Modern College of Engineering, Pune, India. Her area current research include Database, Data Mining, Artificial Intelligence. She has published 10 research papers in the journal and conferences. She has guided 6 projects in UG level.

Shreyash Madake completed his B.E. in Artificial Intelligence and Data Science at PES’s Modern College of Engineering, Pune, under the affiliation of Savitribai Phule Pune University. He is deeply interested in Data Science, Artificial Intelligence, and Java Development, demonstrating a strong foundation and passion in these areas.

Ashutosh Kokate completed his B.E. in Artificial Intelligence and Data Science at PES’s Modern College of Engineering, Pune, affiliated with Savitribai Phule Pune University. His primary areas of interest include Data Science, Artificial Intelligence, Data Analysis, and Project Management, where he strives to excel and contribute meaningfully.

Yash Bhandure completed his B.E. in Artificial Intelligence and Data Science from PES’s Modern College of Engineering, Pune, affiliated with Savitribai Phule Pune University. He has a keen interest in Software Development and is proficient in C++, Python, and JavaScript. Yash has extensive experience with various technologies and frameworks, including ReactJS, NodeJS, ExpressJS, MongoDB, Flask, Git, and GitHub, showcasing his versatile skill set and dedication to the field.