AUTOMATIC ACTION RECOGNIZATION USING DEEP LEARNING

Mr. Deepanshu Nirvan Student, Department of Computer Science and Technology (AI & ML) ABES Engineering College, Ghaziabad Deepanshu.20b1531052@abes.ac.in
Ms. Dolly Bindal Student, Department of Computer Science and Technology (AI & ML) ABES Engineering College, Ghaziabad Dolly.20b1531058@abes.ac.in
Mr. Vikas Choudhary Assistant Professor Computer Science & Engineering-AIML ABES Engineering College, Ghaziabad Vikas.choudhary@abes.ac.in

ABSTRACT

Action recognition is an important task in area of computer vision and has wide applications in areas such as surveillance, human, computer interaction, and video analysis. Deep learning has revolutionized the field by providing state-of-the-art solutions for analysing and understanding motion in video. This article provides a review and analysis of deep learning machine learning techniques. It highlights the important role of deep learning in solving the challenge of complex recognition operations on large amounts of video data. This review explores in detail the development of deep learning models for action recognition and their differences, including models such as convolutional neural networks (CNN), recurrent neural networks (RNN). Additionally, the summary introduces the basic techniques and strategies used in deep learning-based recognition, including feature extraction and tracking techniques. It also highlights the important role of large data sets in training efficient cognitive models. It concludes with a critical analysis of current issues and future research directions in this field. It highlights the need to address issues related to model interpretation, generalization across different contexts, and integration of multimodal data to improve understanding of action. In summary, this article provides an overview of the progress, challenges, and future prospects of models for deep learning-based systems, providing a valuable resource for researchers, practitioners, and enthusiasts working in computer vision and artificial intelligence**.**

Keywords:
Action recognition, Deep learning, Convolutional, Neural Networks (CNNs), Recurrent Neural Networks (RNNs),Feature extraction, Attention mechanisms, Video analysis Computer vision.

I. Introduction

Human action recognition referred as HAR is a difficult but crucialtask in the field computer vision, with applications ranging from video surveillance and human and computer interaction to healthcare and sports analytics. The ability to automatically recognize and interpret human actions from videos holds immense potential for various domains. Traditionally, HAR has relied on handcrafted features and machine learning algorithms, which have shown limited success in capturing the complex temporal and spatial dynamics of human actions.In recent years, deep learning has emerged as a powerful tool for HAR, revolutionizing the field with its ability to learn complex representations from raw data. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have become the dominant architectures for HAR, demonstrating remarkable performance in various datasets and scenarios. CNNs excel at extracting spatial features fromimages or frames of videos, while RNNs effectively capture temporal dependencies in action sequences.

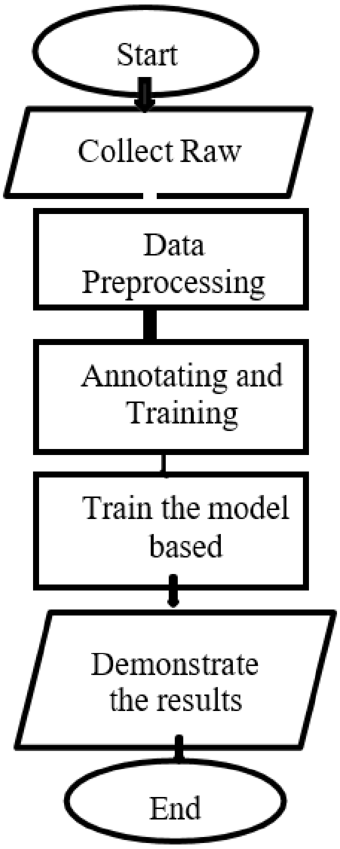


Figure 1: Basic Steps in System

The combination of CNNs and RNNs has led to the development of sophisticated HAR architectures, such as two-stream networks and convolutional long short-term memory (ConvLSTM) networks. These architectures leverage the strengths of both CNNs and RNNs, achieving state-of-the-art performance in HAR.

Despite significant advances, HAR remains a challengingtask due to variations in appearance, lighting, and background conditions. Additionally, the complexity of human actions and the need for real-time processing posefurther challenges. Ongoing research is focused on addressing these challenges and developing more robust,efficient, and real-time HAR systems.

This paper delves into the application of deep learning for HAR, exploring the underlying principles, architectures, andrecent advancements. We present a comprehensive overviewof deep learning-based HAR methods, highlighting their strengths, limitations, and applications. We also discuss the challenges and future directions of HAR research.

II. Literature

Deep learning, a technology enabling machines to learn patterns from data, has greatly influenced action recognition in videos. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) stand as prominent architectures in this domain. CNNs excel in image understanding, while RNNs specialize in sequence analysis.

Methods in Action Recognition:

2.1 CNN-based Approaches

CNNs, adapted for video analysis, demonstrate proficiency in extracting spatial and temporal features from frames, enabling detailed understanding of motion patterns.

Benefits: Efficient at capturing fine-grained details and motion characteristics.

Challenges: High computational demands, limitations in processing prolonged video sequences.

2.2 RNN-based Approaches:

RNNs, particularly Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRUs), excel in modeling temporal dependencies, beneficial for recognizing actions evolving over time.

Benefits: Proficient in understanding temporal sequences and long-range dependencies.

Challenges: Complexity in training, difficulty in capturing extensive temporal contexts.

2.3 Hybrid Architectures

Both spatial and temporal information, achieving comprehensive action understanding.

Benefits: Combined strengths in capturing static and dynamic features.

Challenges: Increased complexity, potential informationredundancy, and intricate model optimization.

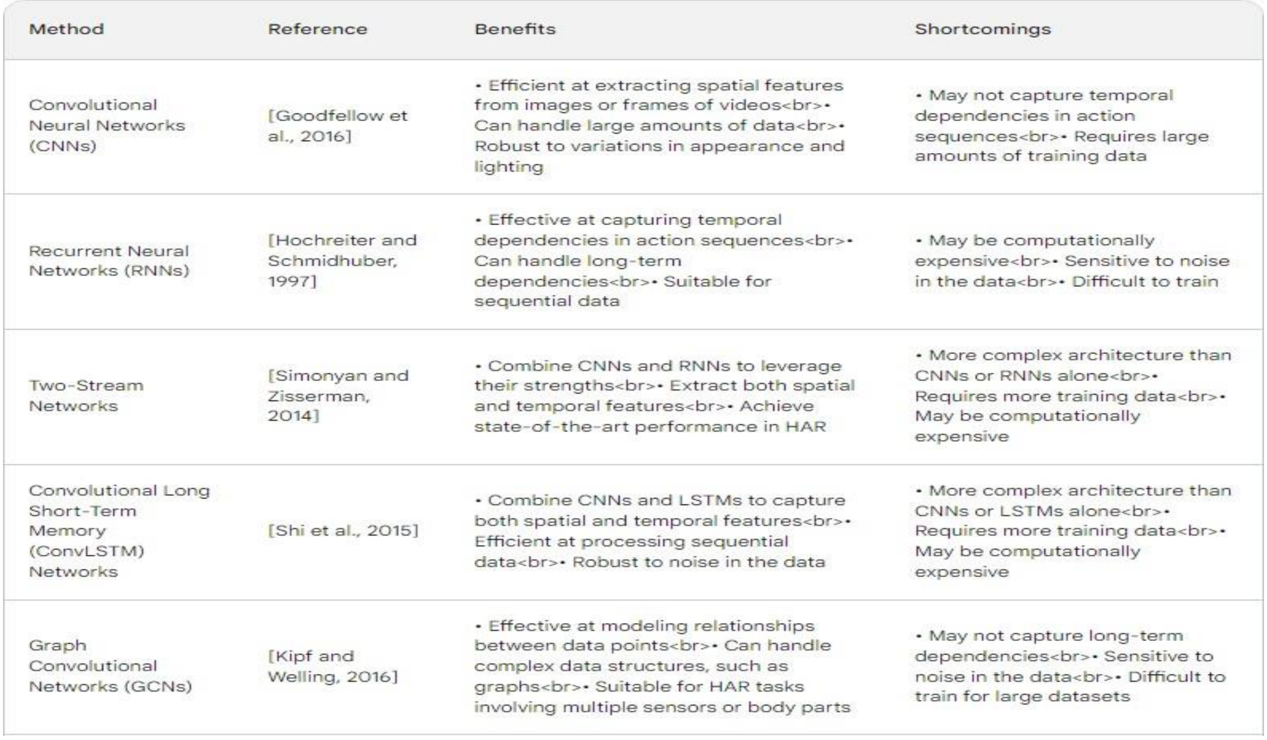
Challenges and Limitations:

Varied Perspectives: Models encounter difficulty when actions exhibit diverse viewpoints or get obscured, impactingrecognition accuracy.

Environmental Adaptation: Some models struggle with generalizing across different settings or recognizing novel actions due to limited training data.

Computational Overhead: Deep learning models demand substantial computational resources, hindering real-timedeployment in resource-constrained scenarios

Table 1 Comparison of different methods



Method	Reference	Benefits	Shortcomings
Convolutional Neural Networks (CNNs)	[Goodfellow et al., 2016]	• Efficient at extracting spatial features from images or frames of videos • Can handle large amounts of data • Robust to variations in appearance and lighting	• May not capture temporal dependencies in action sequences • Requires large amounts of training data
Recurrent Neural Networks (RNNs)	[Hochreiter and Schmidhuber, 1997]	• Effective at capturing temporal dependencies in action sequences • Can handle long-term dependencies • Suitable for sequential data	• May be computationally expensive • Sensitive to noise in the data • Difficult to train
Two-Stream Networks	[Simonyan and Zisserman, 2014]	• Combine CNNs and RNNs to leverage their strengths • Extract both spatial and temporal features • Achieve state-of-the-art performance in HAR	• More complex architecture than CNNs or RNNs alone • Requires more training data • May be computationally expensive
Convolutional Long Short-Term Memory (ConvLSTM) Networks	[Shi et al., 2015]	• Combine CNNs and LSTMs to capture both spatial and temporal features • Efficient at processing sequential data • Robust to noise in the data	• More complex architecture than CNNs or LSTMs alone • Requires more training data • May be computationally expensive
Graph Convolutional Networks (GCNs)	[Kipf and Welling, 2016]	• Effective at modeling relationships between data points • Can handle complex data structures, such as graphs • Suitable for HAR tasks involving multiple sensors or body parts	• May not capture long-term dependencies • Sensitive to noise in the data • Difficult to train for large datasets

III. Proposed System Model

The proposed system for action recognition utilizes a combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to effectively capture both spatial and temporal features from camera data, enabling accurate recognition of actions.

CNN for Capturing Visual Details:

We'll use CNNs to examine each frame of the video. CNNs are great at capturing visual details like shapes, patterns, and textures. They'll help our system understand what's happening in each video frame, identifying essential visual that define different actions.

LSTM for Understanding Sequences:

LSTMs will work alongside CNNs to understand the sequence of actions. LSTMs are good at remembering information over time, making them perfect for recognizing how actions unfold in a sequence of video frames. They'll help the system make sense of the order in which actions occur, improving recognition accuracy.

Integration of CNN and LSTM:

By combining CNNs and LSTMs, our system will benefit from the strengths of both. CNNs will capture visual details in each frame, while LSTMs will analyze these frames in a sequence, allowing the system to recognize actions based on both visual information and their order of occurrence.

Classification Layer:

The fused features are then passed through a classification layer, which utilizes a trained algorithm to classify the actionbased on the learned patterns. The classification layer assignsa probability to each possible action, indicating the likelihood that the observed action matches the corresponding label.

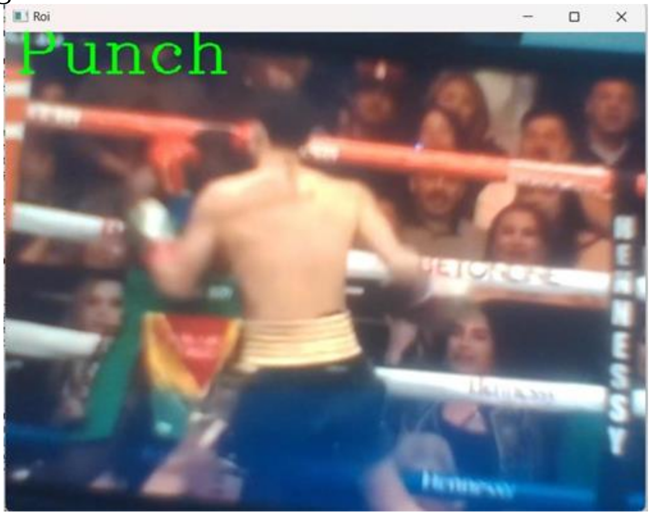


Figure 2 Output 1

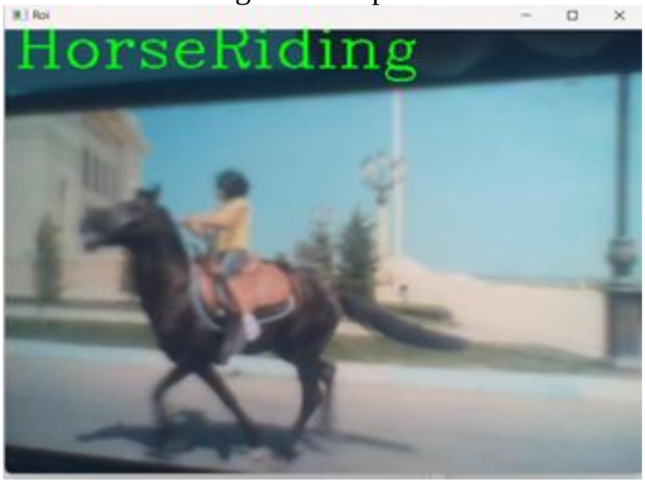


Figure 2 Output 1

IV. Discussion and Future Scope

Healthcare: In the healthcare sector, action recognition systems using CNN+LSTM can play a pivotal role in patient monitoring, rehabilitation assessment, and personalized treatment plans. By analyzing patient movements and activities, these systems can provide valuable insights into their physical condition, progress, and potential risks. This information can be used to tailor rehabilitation exercises, monitor recovery from injuries or surgeries, and identifyearly signs of neurological disorders.

Sports Analytics: For athletes and sports enthusiasts, action recognition systems powered by CNN+LSTM can revolutionize sports analytics and performance enhancement. By analyzing the movements of athletes during training and competitions, these systems can provide real-time feedback, identify areas for improvement, and optimize training strategies. This data-driven approach can lead to enhanced performance, reduced injury risks, and improved overall athletic outcomes.

Human-Computer Interaction: Action recognition systems using CNN+LSTM can transform human-computer interaction by enabling natural and intuitive interactions between humans and machines. By understanding human gestures, postures, and movements, these systems can control devices, navigate interfaces, and respond to user commands without the need for traditional input methods. This can revolutionize the way we interact with computers, making them more user-friendly and accessible.

Surveillance and Security: In surveillance and security applications, action recognition systems using CNN+LSTM can play a crucial role in monitoring environments, detecting suspicious behavior, and preventing unauthorized access. By analyzing the movements of individuals in real-time, these systems can identify potential threats, track individuals of interest, and alert security personnel. This can enhance security measures in public spaces, airports, and other critical areas.

Robotics and Automation: Action recognition systemsusing CNN+LSTM can empower robots and autonomoussystems to interact with the world in a more human-like manner. By understanding human actions and intentions, robots can assist humans in various tasks, collaborate in complex environments, and provide personalized services. This can lead to advancements in robotics, automation, and human-robot collaboration.

Virtual Reality and Augmented Reality: Action recognition systems using CNN+LSTM can enhance virtual reality (VR) and augmented reality (AR) experiences by enabling natural and intuitive interactions within virtual environments. By tracking user movements and gestures, these systems can allow users to manipulate virtual objects, navigate virtual spaces, and interact with virtual avatars. This can create more immersive and engaging VR/AR experiences.

Accessibility and Assistive Technologies: Action recognition systems using CNN+LSTM can improve accessibility for individuals with disabilities by enabling alternative input methods for controlling devices and interacting with technology. By recognizing gestures, facial expressions, and body movements, these systems can providepersonalized solutions for communication, navigation, and control. This can enhance the quality of life for individuals with physical or cognitive limitations.

V. Conclusion

People are really interested in recognizing actions, like gestures or movements, because it's useful for lots of things, such as security cameras, how we interact with computers, or even in healthcare. But there are still many problems that haven't been fixed yet.

In our study, we looked at different ways people are trying to solve these problems. One big issue is that actions can look different from different angles or when things get in the way. Some methods work okay with this, but none are perfect. Also, when the camera moves or when there are lots of things happening in the background, it's hard for computers to understand actions. Some methods try to fix this, but they still have limits.

There's hope for new ways to fix these problems, like making better systems to recognize actions or creating new sets of data to test these systems. But right now, there's no one solution that solves all these issues. We need to explore new ideas and areas to make a system that can handle all these problems. Our study shows the problems that still need fixing. This can guide researchers to focus on these problems and hopefully create a system that's really good at recognizing actions no matter what challenges come up.

References

[1] M. Ryoo and J. Aggarwal, “Human activity analysis: A Review,” ACM Computing Surveys, Article 16, vol. 43, pp. 16:1 – 16:43, April 2011.

[2] D. Siewiorek, A. Smailagic, and A. Dey, “Architecture and applications of virtual coaches,” Proceedings of the IEEE, Invited Paper, vol. 100, pp. 2472–2488, August 2012.

[3] T. Kanade and M. Hebert, “First-person vision,” Proc. of the IEEE, Invited Paper, vol. 100, pp. 2442–2453, August 2012.
[4] T. Shibata, “Therapeutic Seal robot biofeedback medical device: qualitative and quantitative evaluations of robot therapy in dementia care,” Proceedings of the IEEE, Invited Paper, vol. 100, pp. 2527–2538, August 2012.
[5] K. Yamazaki, R. Ueda, S. Nozawa, M. Kojima, K. Okada,
K. Matsumoto, M. Ishikawa, I. Shimoyama, and M. Inaba, “Home-assistant robot for an aging society,” Proceedings of the IEEE, Invited Paper, vol. 100, pp. 2429–2441, August 2012.
[6] P. Kelly, A. Healy, K. Moran, and N. E. O’Connor, “A virtual coaching environment for improving golf swing technique,” in ACM Multimedia Workshop on Surreal Media and Virtual Cloning, pp. 51 – 56, October 2010.
[7] L. Palafox and H.Hashimoto, “Human action recognition using wavelet signal analysis as an input in 4W1H,” in IEEE Intl. Conf. on Industrial Informatics, pp. 679 – 684, July 2010.
[8] R. Poppe, “A survey on vision-based human action recognition,” Image and Vision Computing, vol. 28, pp. 976 990, June 2010.
[9] Y. Li and Y. Kuai, “Action recognition based on spatio- temporal interest points,” in Intl. Conf. on BioMedical Engineering and Informatics, pp. 181 – 185, October 2012.
[10] X. Ji and H. Liu, “Advances in view-invariant human motion analysis - A Review,” IEEE Trans. on Systems, Man, and Cybernetic Part C: Applications and Reviews, vol. 40, pp. 13 – 24, January 2012.
[11] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, “The i3DPost multi-view and 3D human action/interaction database,” in Conf. for Visual Media Production, pp. 159 – 168, November 2009.
[12] A. F. Bobick and J. W. Davis, “The recognition of human movement using temporal templates,” IEEE Trans. onPattern Analysis and Machine Intelligence, vol. 23, pp. 257 –267, March 2001.
[13] T. Darell and A. Pentland, “Space-time gestures,” in IEEE Conf. on Computer Vision and Pattern Recognition, pp.335 – 340, 1993.
[14] G. Rogez, J. Guerrero, and C. Orrite, “View-invariant human feature extraction for video-surveillance applications,”in IEEE Conf. on Advanced Video and signal based Surveillance, pp. 324 – 329, 2007.
[15] H. Ragheb, S. Velastin, P. Remagnino, and T. Ellis, “Human action recognition using robust power spectrum features,” in IEEE Conf. of Image Processing, pp. 753–756, October 2008.
[16] Y. Lu, Y. Li, Y. Chen, F. Ding, X. Wang, J. Hu, and S. Ding, “A Human action recognition method based on Tchebichef moment invariants and temporal templates,” in Intl. Conf. on Intelligent Human-Machine Systems and Cybernetics, pp. 76–79, August 2012.
[17] A. Iosifidis, A. Tefas, and I. Pitas, “Neural representation and learning for multi-view human action recognition,” in IEEE World Congress on Computational Intelligence, pp. 1–6, June 2012. [18] K. Huang, D. Tao, Y. Yuan, X. Li, and T. Tan, “View-independent behavior analysis,” IEEE Trans. on Systems, Man and Cybernetics - Part B: Cybernetics, vol. 39, pp. 1028 – 1035, August 2009.