Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 1, No.11: 2024
ISSN : **1906-9685**

DEEP LEARNING-BASED RECOGNITION OF HUMAN ACTIONS FROM DEPTH MAPS AND POSTURES

[1]Dr S Dhanalaxmi, [2]Dr K Niranjan Reddy, [3]Dr K Pradeep Reddy, [4]RadhaKrishna Karne
[1,2,3,4]CMR Institute of Technology, Hyderabad.

Abstract — Deep learning-based human action recognition using depth maps and postures is a difficult but exciting area of study. While posture and depth maps can yield useful insights into human movement, developing a system that can recognize between various motions based only on this data remains challenging. Since deep learning can extract intricate patterns from massive data sets, it is a potential method for recognizing human actions. The study looks on a deep learning approach to human activity recognition. It examines data from two sources: posture data and depth maps. While posture data indicates the relative locations of the body, depth maps record the distance between the camera and objects in the image. Compared to current depth-based techniques for human action recognition, this method has advantages. The study shows how well CNN works to extract characteristics from posture data and depth maps, improving recognition accuracy.

Keywords— Convolution Neural Network (CNN), Deep Learning, Accuracy, Human Action Recognition (HAR),

I. INTRODUCTION

A vital area of computer vision, human action recognition (HAR) finds applications in robotics, human-computer interaction, and surveillance. It has proven difficult to reliably recognize complicated human activities because of things like occlusions, clothing, and viewpoint alterations. The goal of this project is to overcome this difficulty by introducing a novel HAR technique that makes use of the advantages of two different data sources: postures and depth maps. Although posture data records the arrangement of bodily joints, depth maps offer detailed information about the three-dimensional structure of a scene. The research intends to achieve significant increases in human action recognition accuracy by merging multiple sources and leveraging the capability of deep learning, namely Convolution Neural Networks (CNNs). The introduction probably goes into more detail about the drawbacks of conventional techniques and encourages the usage of has become an effective instrument for deciphering intricate patterns in data, such as postures and depth maps. Researchers have made impressive progress in a variety of computer vision applications, such as object recognition, scene understanding, and human activity detection, by utilizing deep learning approaches. Given the variety of human appearance, mobility, and surrounding conditions, human action recognition (HAR) is an extremely difficult problem. Postures and depth maps offer important insights into human movement. Since deep learning can extract intricate patterns from massive data sets, it is a promising method for handling HAR. Data collection, model design, model training, and model deployment are challenges in deep learning-based HAR from depth maps and postures. Video surveillance, human-computer interaction, healthcare, gaming and entertainment, and robots are among the industries that could benefit from the use of HAR derived from depth maps and postures through deep learning. The project's goal is to automate the process of recognizing and categorizing human actions by combining two variable data sources: depth maps and body postures. Deep learning will be used to detect and classify human actions. The goal of this study is to integrate these

two different forms of information to create a more robust and improved understanding of human actions.

II, RELATED WORK AND ITS ALGORITHMS

The Action-Fusion system is currently in use and was created by scientists from the Universities of Leeds and Manchester. With one stream for posture data and another for depth maps, it employs a two-stream CNN architecture. The final prediction is then generated by combining the outputs from the two streams. Microsoft Research researchers created the Deep Kinect technology. It classifies activities by using a deep learning algorithm to extract characteristics from posture data and depth maps.

Dense Pose: Google AI researchers developed this method. It employs depth maps and a deep learning model to determine the human body's 3D posture. Actions are then categorized using the estimated pose.

A general overview of how such a system might operate includes data splitting, preprocessing, data collection, training, testing, preprocessing, deployment, etc. The proposed system will probably use deep learning techniques to recognize and understand human actions based on depth maps and postures. The solutions that have been suggested offer the following benefits: enhanced privacy; real-time processing potential; and effective gesture recognition. as well as less sensitivity to changes in color. The Figure.1. shows that System Architecture of Video-Based Human Action Recognition. For classification or regression tasks, fully connected layers and output layers are frequently included as well. Because 3D CNNs have more parameters than 2D CNNs and their input data is more dimensional, training them might need a lot of computing power. But because to developments in optimization methods and hardware acceleration (GPUs), it is now possible to train and use 3D CNNs efficiently. 3D CNNs are used in many different fields, including as medical image analysis for illness diagnosis, 3D object recognition, activity recognition in films, and spatiotemporal anomaly detection. In tasks requiring comprehension of both spatial and temporal contexts, 3D CNNs have proven to perform better by utilizing the temporal information stored in the input data. The Figure.2. Shows that the Flow Diagram of Human Action Recognition

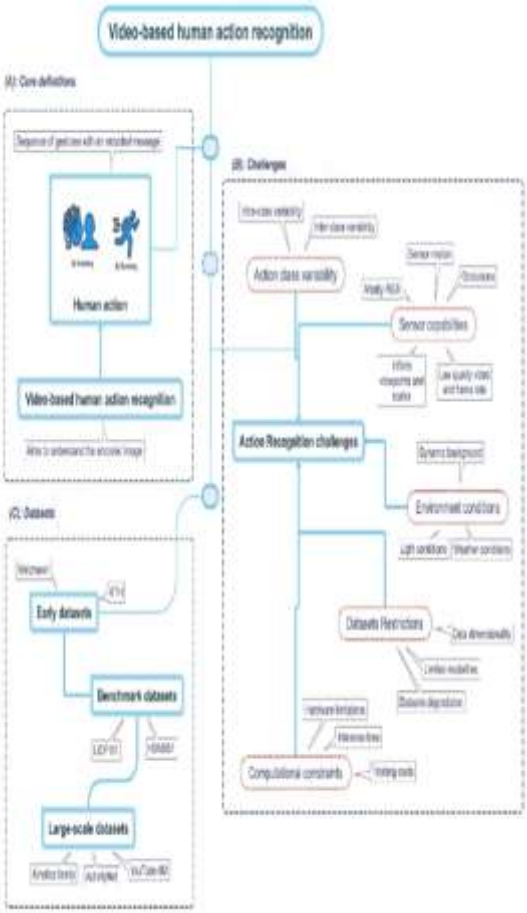

Fig.1. System Architecture of Video-Based Human Action Recognition

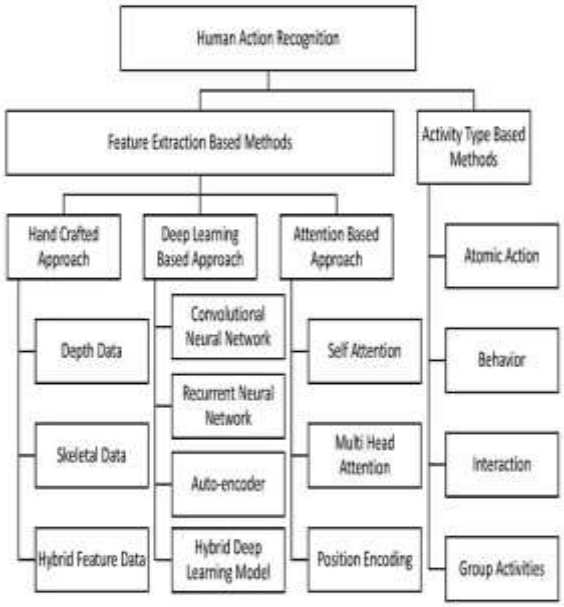

Fig.2. Human Action Recognition Flow Diagram

A few popular deep learning architectures and methods used for this task; 1) Two stream networks, 2) temporal convolution networks (TCNs), and 3) three-dimensional CNNs

1. Convolution Neural Networks in 3D

Convolution neural networks, or 3D CNNs, are an extension of 2D CNNs that were created with the express purpose of handling spatiotemporal data, including volumes of medical imaging, movies, and 3D sensor data. In order to account for the temporal element of the input, 3D CNNs add a third

dimension—depth or time—to the two dimensions that 2D CNNs use to process images. Convolution filters are applied in three dimensions (height, breadth, and depth) in three-dimensional neural networks (3D CNNs), which enables the network to concurrently collect spatial and temporal data. As a result, the model can gradually acquire intricate correlations and patterns across several frames or data slices.

Multiple convolution layers are usually the first layer in a 3D CNN architecture, followed by pooling layers for feature extraction and spatial down sampling. For classification or regression tasks, fully connected layers and output layers are frequently included as well. Because 3D CNNs have more parameters than 2D CNNs and their input data is more dimensional, training them might need a lot of computing power. But because to developments in optimization methods and hardware acceleration (GPUs), it is now possible to train and use 3D CNNs efficiently. 3D CNNs are used in many different fields, including as medical image analysis for illness diagnosis, 3D object recognition, activity recognition in films, and spatiotemporal anomaly detection. In tasks requiring comprehension of both spatial and temporal contexts, 3D CNNs have proven to perform better by utilizing the temporal information stored in the input data.

2. Two-Stream Network - A well-known method in computer vision and deep learning, the Two-Stream Network algorithm was created especially for action recognition tasks in videos. It is made up of two distinct but complimentary streams: temporal and spatial. Like classic image classification problems, the spatial stream relies on extracting static appearance information from individual video frames. Convolution Neural Networks (CNNs) are usually used to extract spatial data separately from each frame. The temporal stream, on the other hand, highlights motion information by taking the frame sequence across time into account. It frequently makes use of methods like optical flow or 3D Convolution Neural Networks (3D CNNs) to record the video's temporal dynamics and motion patterns. These two streams function independently, handling temporal and geographic data simultaneously. To arrive at final predictions regarding the action class, the temporal and spatial information that were collected from both streams are fused, or integrated at a higher level. Numerous techniques, such as concatenation, element-wise addition, or multi-layer fusion networks, can be used to accomplish this fusion phase.

The Two-Stream Network technique greatly enhances action recognition performance by taking use of the complementary nature of spatial and temporal information. It can efficiently capture complicated spatiotemporal patterns present in movies by taking into account both appearance and motion signals at the same time, producing more accurate action identification results. This method has been extensively used and has produced state-of-the-art results in a number of action recognition benchmarks and applications, such as gesture recognition, video surveillance, and human activity recognition.

3. Network with Temporal Convolution - A deep learning architecture called the Temporal Convolution Network (TCN) method was created especially for modeling sequential data, including time series or sequential text data. TCNs are able to efficiently capture temporal patterns and long-range relationships because they use one-dimensional convolution layers that function throughout the temporal dimension of the input sequence. TCNs have a number of benefits over recurrent neural networks (RNNs), such as parallelization, stability, and ease of training. RNNs have the drawback of vanishing or bursting gradients and are computationally expensive to train.

The convolution layers in a TCN are usually followed by residual connections or dropout for regularization and better gradient flow, and optionally by non-linear activation functions like ReLU (Rectified Linear Unit). Using a series of convolution layers stacked at progressively larger receptive field sizes, TCNs are able to record intricate and hierarchical temporal patterns at various time scales. Furthermore, TCNs frequently use dilated convolutions, in which the convolution filters include gaps between their elements, to expand the receptive field without adding more parameters, allowing them to effectively capture long-range relationships. When it comes to a variety of sequential data modeling tasks, such as language modeling, speech recognition, music production, and time series forecasting, TCNs have proven their superior performance.

Their capacity to preserve temporal relationships throughout lengthy durations while When it comes to a variety of sequential data modeling tasks, such as language modeling, speech recognition, music production, and time series forecasting, TCNs have proven their superior performance. They are an effective tool for evaluating and modeling sequential data in a variety of fields because of their capacity to capture temporal dependencies over lengthy sequences while retaining computing efficiency. Additionally, TCNs improve interpretability since the convolution filters can shed light on significant temporal patterns that the model has learnt, which can help with comprehending and interpreting the underlying dynamics of the data.

IV. RESULTS

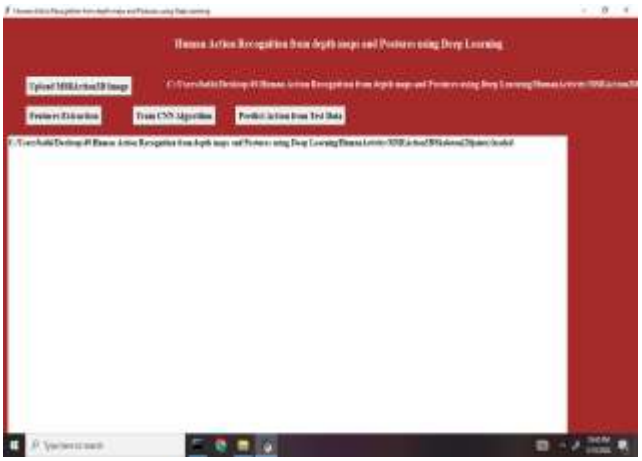

Fig.3. MSRAction3D Image Ddataset – Feature Extraction

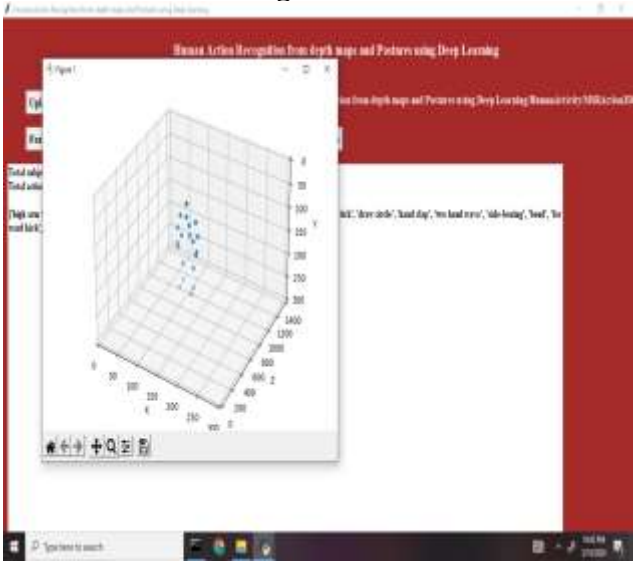

Fig.4. Dataset with Train CNN

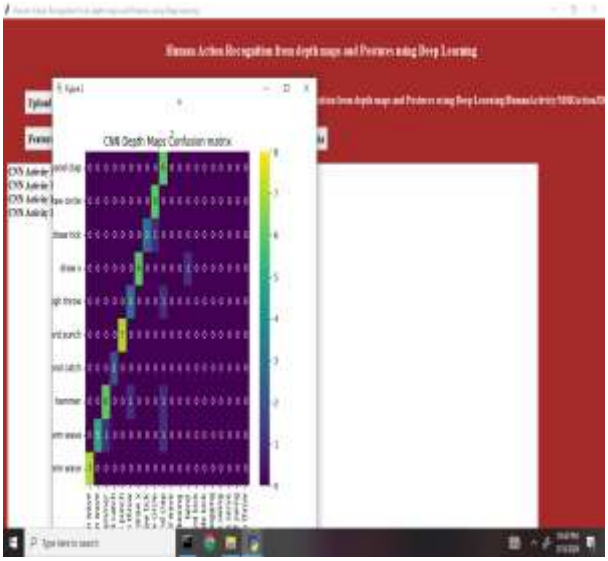

Fig.5. Action Recognition in Accuracy with Confusion Matrix

Fig.6. Skeleton Values with Activity Recognized

V. CONCLUSION

Depth maps and postures were used in the experiment to investigate the possibilities of deep learning for human activity detection. The results indicate that there is potential with this strategy. Achieving good accuracy in identifying different actions, the deep learning model combined posture data—which records important joint positions in the body—with depth data, which yields rich three-dimensional information. This clears the path for the continued creation of reliable and effective action recognition technologies. It's always possible to get better, though. The deep learning architecture may be improved in the future, along with the inclusion of new data sources like bone tracking and performance testing of the model on even more complicated datasets. All things considered, this effort opens up new avenues for breakthroughs in computer vision applications such as robots, human-computer interaction, and surveillance by showcasing the efficacy of deep learning for human action recognition using depth maps and postures. All things considered, the deep learning approach to human action identification marks a significant leap in computer vision and its uses for deciphering and interpreting human activities from postures and depth maps. The project is anticipated to yield insights and approaches that will propel additional innovation and growth in the field of deep learning, ultimately leading to improved human-computer interaction systems and a deeper knowledge of human behavior.

VI REFERENCES

[1] Schulz, Hannes; Behnke, Sven (1 November 2012). "Deep Learning". KI - Künstliche Intelligenz. 26 (4): 357–363. doi:10.1007/s13218-012-0198-z. ISSN 1610-1987. S2CID 220523562.

[2] ^ LeCun, Yann; Bengio, Yoshua; Hinton, Geoffrey (2015). "Deep Learning". Nature. 521 (7553): 436–444. Bibcode:2015Natur.521..436L. doi:10.1038/nature14539. PMID 26017442. S2CID 3074096.

[3] Ciresan, D.; Meier, U.; Schmidhuber, J. (2012). "Multi-column deep neural networks for image classification". 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3642–3649. arXiv:1202.2745. doi:10.1109/cvpr.2012.6248110. ISBN 978-1-4673-1228-8. S2CID 2161592.

[4] Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey (2012). "ImageNet Classification with Deep Convolutional Neural Networks" (PDF). NIPS 2012: Neural Information Processing Systems, Lake Tahoe, Nevada. Archived (PDF) from the original on 2017-01-10. Retrieved 2017-05-24.

[5] "Google's AlphaGo AI wins three-match series against the world's best Go player". TechCrunch. 25 May 2017. Archived from the original on 17 June 2018. Retrieved 17 June 2018.

[6] Tanzeem Choudhury, Gaetano Borriello, et al. The Mobile Sensing Platform: An Embedded System for Activity Recognition. Appears in the IEEE Pervasive Magazine – Special Issue on Activity-Based Computing, April 2008.

[7] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, Michael Littman. Activity Recognition from Accelerometer Data. Proceedings of the Seventeenth Conference on Innovative Applications of Artificial Intelligence (IAAI/AAAI 2005).

[8] Senthilkumar D,"Blockchain and its integration as a disruptive technology",2019, AI and Big Data's Potential for Disruptive Innovation, pp. 261-290.

[9] Dhanalakshmi S., Charles Babu G, "An examination of big data and blockchain technology", 2019, International Journal of Innovative Technology and Exploring Engineering, volume 8, Issue 11, pp. 3118-3122

[10] Yogitha Lakshmi K., Dhanalakshmi S., Obula Reddy B.G, "An overview of data management in cloud computing", 2019, International Journal of Recent Technology and Engineering, Volume 7, Issue 5C, pp. 61-64.

[11] Dhanalakshmi S., Obula Reddy B.G., Yogitha Lakshmi K, "Building a blockchain approach with hyperledger transaction flow and distributed consensus algorithms",2018, International Journal of Innovative Technology and Exploring Engineering, Volume 8, Issue 2S, pp. 423-426.

[12] Dhanalakshmi, S., & Ravichandran, D. T. (2012). A new method for image segmentation. International Journal of Advanced Research in Computer Science and Software Engineering, 2(9), 293-299.

[13] Senthilkumar, D. (2020). Cross-industry use of Blockchain technology and opportunities for the future: Blockchain technology and aritificial intelligence. In *Cross-industry use of blockchain technology and opportunities for the future* (pp. 64-79). IGI Global.

[14] Senthilkumar, D. (2021). Data confidentiality, integrity, and authentication. In *Research Anthology on Blockchain Technology in Business, Healthcare, Education, and Government* (pp. 459-487). IGI Global.

[15] Vijayasekar, D., Dhivya, S., Dhanalakshmi, S., & Karthik, D. S. (2015). Survey on Detection of Glaucoma in Fundus Image by Segmentation and Classification. Int. J. Eng. Res, 4(09), 529-532.

[16] Dhanalakshmi, S., & Ravichandran, T. (2012). A modified approach for image segmentation in information bottleneck method. Inter-national Journal of Advanced Research in Computer Engineering & Technology, 1(7), 59-63.

[17] Dhanalakshmi, S., Reddy, B. O., & Lakshmi, K. Y. (2018). Building a blockchain approach with hyperledger transaction flow and distributed consensus algorithms. Int. J. Innov. Technol. Explor. Eng, 8(2S), 423-426.

[18] Dhanalakshmi, S., & Kishore, T. P. D. K. (2017). Content delivery networks—A survey. Int. J. Adv. Res. Comput. Sci. Softw. Eng, 7(7), 228-230.

[19] Senthilkumar, Dhanalakshmi. "Blockchain and Its Integration as a Disruptive Technology." AI and Big Data's Potential for Disruptive Innovation. IGI Global, 2020. 261-290.

[20] Lavanya, S., Kumar, B. N., Obuliraj, R., & Dhanalakshmi, S. (2014). Gradient watershed transform based automated cell segmentation for THG microscopy medical images to detect skin cancer. The International Journal of Science and Technoledge, 2(3), 98.

[21] Reddy, Kumbala Pradeep, Gullipalli Apparao Naidu, and Bulusu Vishnu Vardhan. "View-Invariant Feature Representation for Action Recognition under Multiple Views." International Journal of Intelligent Engineering & Systems 12.6 (2019).

[22] Pradeep Reddy, K., T. Raghunadha Reddy, G. Apparao Naidu, and B. Vishnu Vardhan. "Term weight measures influence in information retrieval." Int J Eng Technol 7, no. 2 (2018): 832-836.

[23] Ramesh, A., Reddy, K. P., Sreenivas, M., & Upendar, P. (2022, April). Feature Selection Technique-Based Approach for Suggestion Mining. In Evolution in Computational Intelligence: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021) (pp. 541-549). Singapore: Springer Nature Singapore.

[24] Kodati, Sarangam, Kumbala Pradeep Reddy, Sreenivas Mekala, PL Srinivasa Murthy, and P. Chandra Sekhar Reddy. "Detection of Fake Profiles on Twitter Using Hybrid SVM Algorithm." In E3S Web of Conferences, vol. 309, p. 01046. EDP Sciences, 2021.

[25] Reddy, K. Niranjan, and P. V. Y. Jayasree. "Low power process, voltage, and temperature (PVT) variations aware improved tunnel FET on 6T SRAM cells." Sustainable Computing: Informatics and Systems 21 (2019): 143-153

[26] Reddy, Kallem Niranjan, and Pappu Venkata Yasoda Jayasree. "Low Power Strain and Dimension Aware SRAM Cell Design Using a New Tunnel FET and Domino Independent Logic." International Journal of Intelligent Engineering & Systems 11, no. 4 (2018).

[27] *Yang, Yang; Leung, Howard; Shum, Hubert P. H.; Li, Jiao; Zeng, Lanling; Aslam, Nauman; Pan, Zhigeng (2018). "CCESK: A Chinese Character Educational System Based on Kinect". IEEE Transactions on Learning Technologies. 11 (3): 342–347. doi:10.1109/TLT.2017.2723888. S2CID 52899136.*

[28] *Ho, Edmond S. L.; Chan, Jacky C. P.; Chan, Donald C. K.; Shum, Hubert P. H.; Cheung, Yiu-ming; Yuen, P. C. (2016). "Improving Posture Classification Accuracy for Depth Sensor-based Human Activity Monitoring in Smart Environments". Computer Vision and Image Understanding. 148: 97–110. doi:*10.1016/j.cviu.2015.12.011. *S2CID 207060860.*

[29] W. Chi, J. Wang, and M. Q.-H. Meng, "A gait recognition method for human following in service robots," IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017.

[30] J. Yu and J. Sun, "Multiactivity 3-d human pose tracking in incorporated motion model with transition bridges," IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017.

[31] G. Liang, X. Lan, J. Wang, J. Wang, and N. Zheng, "A limb-based graphical model for human pose estimation," IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017.

[32] D. Kim, W.-h. Yun, H.-S. Yoon, and H.-S. Jaehong, "Action recogntion with depth maps using hog descriptors of multi-view motion," in The Eighth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies. UBICOMM, 2014, pp. 2308–4278.