International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 08 | Aug - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Pose Based Human Action Recognition System

Vinay Patel G L [1] Ruchitha P B [2]

[1] *Assistant Professor, Department of MCA, BIET, Davanagere*

[2] *Student, 4th Semester MCA, Department of MCA, BIET, Davanagere*

ABSTRACT

In the rapidly evolving field of computer vision, Recognition of Human Action (HAR) aims to recognize and categorize human behaviors from visual data, such as pictures or videos. Despite significant advancements in video-based HAR systems due to the availability of temporal information, action detection of static images remains a difficult issue due to the lack of movement tips. The process is made significantly more difficult by real-world elements including attire, body type, lighting, and disparities in viewpoints. Notwithstanding these problems, the image-based HAR has recently drawn interest due to its wide range of applications in fields including simulation, behavior analysis, and surveillance. With dependable data collection and a wealth of benchmarks for comparison, smartphones and wearable sensors have also shown themselves to be effective instruments for activity recognition. However, many systems require users to carry or use equipment, which can be restrictive and inconvenient. However, a less intrusive option that combines environmental data is offered by vision-based Har Systems with camera hardware. Without the user's direct involvement, these systems rely primarily on posture estimation to identify practical features of the human body, which serve as a solid foundation for action recognition. The method leverages the Open Library, which employs coevolutionary neural networks (CNNS), to identify human attitude by detecting 18 key locations in the body of input photos. The crucial observed activity is then determined by adding points to a supervised machine learning model.

Keyword: Vision-Based Recognition, Image Processing, User Interface, Action Labeling.

I. INTRODUCTION

Human Action Recognition (HAR) has emerged as a crucial area of research in computer vision due to its wide range of real-world applications, including surveillance systems, simulation, behavior analysis, healthcare monitoring, and human-computer interaction. The primary objective of a HAR system is to identify and label human actions from visual input such as images or video sequences. While action recognition from videos has been extensively explored and continues to evolve with deep learning techniques, action recognition from still images remains a relatively underdeveloped area. The main challenge of image-based action recognition lies in the absence of temporal information, such as motion and sequence, which are readily available in videos. This makes it difficult to distinguish between similar poses representing different actions.

Image-based HAR must overcome several complex challenges. Factors such as body shape, clothing, lighting conditions, occlusions, and varying camera angles significantly impact the accuracy of pose

estimation and subsequent action classification. In addition, recognizing human poses when the subject is not directly facing the camera introduces further complexity. Due to these constraints, techniques that work well for video-based HAR systems are often ineffective in image-based scenarios, thereby demanding innovative approaches.

An alternative to vision-based systems has been the use of wearable sensors and mobile devices for activity recognition. These devices, which include accelerometers, gyroscopes, and GPS modules, provide accurate data for tracking physical activities like walking, running, or cycling. Numerous benchmarks and datasets have been established to standardize and evaluate the performance of these sensor-based HAR methods. However, the major drawback of this approach is the requirement for users to carry or wear the sensors, which may be uncomfortable or impractical in real-world scenarios. To overcome these limitations, vision-based HAR systems offer a more natural and unobtrusive solution. They rely on cameras installed in the environment to monitor and analyze human actions without requiring users to wear any equipment. One of the most promising vision-based approaches involves extracting pose information using keypoint detection methods. Pose estimation provides valuable spatial data about human joint positions, which can be effectively used to understand actions.

In this paper, we propose a human action recognition system that utilizes pose-based features extracted from still images. We employ the OpenPose library to extract 18 key body joint locations from a 2D image, which are then processed using a Convolutional Neural Network (CNN) architecture. The extracted pose data is fed into a supervised machine learning model to classify different human actions accurately. Our system includes two primary modules: Admin and User. The admin can manage the application, including user information and FAQs, while Users can register, log in, access the camera for real-time action prediction, and view support materials.

II. RELATED WORK

This paper is to study and evaluate different techniques used for recognizing human activities from still images and videos. It focuses on helping computer systems automatically identify what action a person is doing, which is useful in many areas like healthcare, surveillance, and human-computer interaction. To achieve this, the paper reviews and categorizes various existing methods. It explains how features are taken from images using three types of representations: global features, local features, and body modeling. It also divides activity recognition approaches into template-based, generative, and discriminative methods. Special focus is given to pose-based recognition, where both traditional and deep learning methods are discussed. Lastly, the paper highlights the common challenges in this field, such as occlusion and background clutter, and suggests ways to improve future research[1].

This paper aims to give an overview of different approaches used in Human Action Recognition (HAR), with a special focus on pose-based attention using video data. The main goal is to study how various algorithms process videos to detect and understand human actions by analyzing body poses and attention focus. To support this, the paper lists popular HAR datasets that offer a wide range of video samples for testing and evaluation. It reviews both

local and global feature extraction techniques and explores some of the most widely used deep learning methods, including Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and Generative Adversarial Networks (GAN). All the discussed methods are aimed at improving the recognition of human posture and attention within recorded videos[2].

This paper is to develop computational methods for detecting human activities from still images, which is a challenging area in computer vision, especially when compared to action detection in videos that have time-based features. The focus is on fine-grained activity detection where actions are recognized from a single image, particularly those involving objects. To achieve this, the study breaks down complex human activities into smaller parts based on their meaning (semantics) and examines the role of each part in recognizing the overall action. The approach involves tasks such as object detection and pose estimation to understand what a person is doing in a single frame. A custom dataset with specific activity classes is created using images from various sources, ensuring the system can recognize actions related to object interaction effectively[3].

This paper is to provide a comprehensive review of recent developments in human action recognition and posture prediction, which are important areas in computer vision. These technologies help machines understand and respond to human actions, making them essential for intelligent systems like human-computer interaction and self-driving vehicles. The paper first explains the background and recent progress in the field, especially the impact of deep learning methods. It then introduces widely used datasets, common feature extraction techniques, and advanced algorithms used for recognizing actions and predicting postures. Finally, the paper discusses current challenges and highlights key research areas, showing how these technologies play a major role in improving interactive systems, with self-driving vehicles as a practical example[4].

This paper aims to develop a multitask framework that can handle both human action recognition from videos and 2D/3D human pose estimation from still images using a single architecture. Unlike most studies that treat these tasks separately, this approach combines them efficiently within one model. The proposed method uses an end-to-end training process, which improves accuracy compared to training the tasks separately. It can also be trained on different types of data at the same time without difficulty. The effectiveness of this model is proven through experiments on four well-known datasets—MPII, Human3.6M, Penn Action, and NTU—showing that it performs strongly in both action recognition and pose estimation[5].

This paper aims to review recent progress in human pose estimation and how it is used for action recognition, with a focus on 2D skeleton-based approaches using regular RGB images. Since human actions are often linked to body movements, understanding body poses is important for recognizing actions accurately. The paper surveys both bottom-up and top-down deep learning models for pose estimation and explains how these models contribute to recognizing human actions. Unlike other surveys that focus on 3D skeleton data from depth sensors like Kinect, this work concentrates on pose estimation from standard color images. It also

compares the performance of recent action recognition methods based on 2D pose data and highlights the need for further improvements in this area[6].

This paper review recent advancements in human pose recognition, which involves estimating human body poses from single images or video frames. Human pose recognition is important for many applications such as activity recognition, human tracking, gaming, animation, medical gait analysis, and human-computer interaction. Despite its usefulness, it remains a challenging task due to issues like poor lighting, body occlusion, outdoor settings, and variations in clothing. The study focuses on analyzing recent methods that use computer vision for feature extraction and machine learning for classifying poses. It also highlights current research gaps and suggests possible directions for future improvements in the field[7].

This paper aims to recognize human actions using pose primitives, making it suitable for both videos and still images. In the learning phase, it extracts pose and activity features from video data. During actual use, the system can process either video sequences or single images. The method improves traditional Histogram of Oriented Gradient (HOG) descriptors to better handle complex body movements and messy backgrounds. Human actions are identified by comparing histograms of pose primitives, and for video sequences, local motion patterns are captured using n-gram models. Unlike many other approaches, this method does not depend on background subtraction or motion detection, making it more flexible and effective for recognizing actions even in static images[8].

This paper is to develop a multi-task framework that can jointly perform human pose estimation (in 2D or 3D) and action recognition using a single architecture. Since both tasks rely on analyzing human body movements, the proposed model efficiently combines them to improve performance. It processes monocular color images for pose estimation and video clips for action recognition within the same system. The method is designed to run at high speed—over 100 frames per second—and supports end-to-end training by separating key prediction parts, which improves accuracy. It also allows training on different types of data at the same time. The model achieves strong results on four benchmark datasets—MPII, Human3.6M, Penn Action, and NTU RGB+D—proving its effectiveness, and both the code and pre-trained models are made publicly available[9].

This paper aims to develop a privacy-friendly and robust method for human pose estimation (HPE) using mm-wave radar technology, especially for applications like public surveillance and gaming. Traditional vision-based HPE systems may struggle in poor lighting or bad weather and also raise privacy concerns. To address these issues, the proposed method uses mm-wave radar combined with the inverse synthetic aperture radar (ISAR) algorithm to produce high-resolution radar images of moving individuals. These binarized ISAR images are then used as input to a pose estimation model, which is trained using labels generated by a vision-based model like AlphaPose. The system is effective at estimating human poses from distances of 4–12 meters and works not only with advanced MIMO radar systems but also with low-cost SISO radars,

making it a practical and affordable solution for behavior analysis in real-world environments[10].

III. METHDODLOGY

The proposed Human Action Recognition (HAR) system fig1 shows that the a structured multi-phase methodology that combines advanced pose estimation techniques with machine learning algorithms to classify human actions from still images. The methodology is designed to ensure accuracy, efficiency, and real-time performance while maintaining user-friendliness through a well-integrated interface.

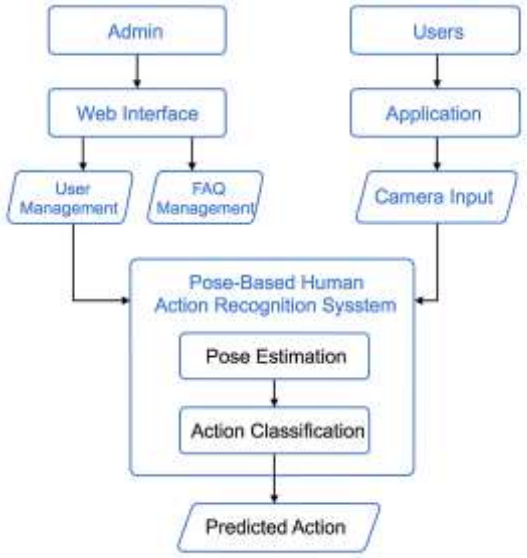

Fig 3. Architecture Diagram

The process can be broken down into the following key phases:

3.1. Pose Estimation using OpenPose: The initial phase of the system involves extracting human pose information from still images. This is achieved using the OpenPose library, which employs Convolutional Neural Networks (CNNs) to detect 18 key body joints on a 2D plane. These joints include essential human articulation points such as the neck, shoulders, elbows, wrists, hips, knees, and ankles. By accurately localizing these keypoints, the system creates a skeletal structure representing the subject's pose, which serves as the foundation for subsequent analysis.

3.2. Feature Extraction and Pre-processing: Once the pose keypoints are extracted, the x and y coordinates of each body joint are compiled to form a structured dataset for classification. These features are then passed through several pre-processing steps, such as:

- Normalization of coordinates to standardize input scale.
- Noise reduction to eliminate outlier points or errors in keypoint detection.
- Data augmentation techniques to increase the diversity of pose samples and improve model generalization.

This pre-processing ensures the dataset is clean, stable, and consistent, thereby improving the performance of the classification algorithm.

3.3. Action Classification using Machine Learning: The normalized pose data is used as input to a supervised machine learning model that classifies the human action. Algorithms such as Support Vector Machine (SVM) and Random Forest are used for training the model on labeled pose data. These algorithms are selected for their ability to handle high-dimensional data and nonlinear decision boundaries, making them suitable for recognizing complex pose-based action patterns. The model is trained on a diverse dataset of human poses representing various actions such as walking, sitting, jumping, waving, etc.

3.4. System Interface and User Interaction: A user-friendly web interface is designed to enable seamless interaction with the system. Key functionalities include:

- User registration and login using secure credentials.
- Image capture or upload functionality through the user's device camera.
- Action prediction triggered upon image submission.

In parallel, an Admin module provides access to system management features, allowing administrators to oversee registered users and manage frequently asked questions (FAQs) to assist users.

3.5. Real-Time Prediction and Output Display: After the image is captured or uploaded, the system performs real-time action recognition by passing the image through the Open Pose-based pose extraction and trained classification model. The predicted action label is then displayed on the interface, providing immediate feedback to the user in an intuitive manner.

3.6. Evaluation and Testing: To ensure robustness and reliability, the system is rigorously tested using a diverse set of human action images captured under different lighting conditions, body orientations, clothing styles, and camera angles. Evaluation metrics such as accuracy, precision, recall, and F1-score are used to assess the system's performance and generalization capability across various real-world scenarios.

IV. TECHNOLOGIES USED

4.1. Python

Python is a high-level, general-purpose programming language known for its clean syntax and ease of readability. It is widely used across various domains such as web development, automation, data science, artificial intelligence, and software development. One of Python's greatest strengths is its extensive library support, which allows developers to work faster and more efficiently. The language is platform-independent and has a large and active community, making it easier for developers to find support, resources, and solutions to problems. Python's simple structure and dynamic typing make it an ideal choice for both beginners and professionals working on complex projects.

Python libraries are OpenCV, TensorFlow and NumPy.

4.2. OpenCV

OpenCV is an open-source library designed for computer vision and image processing tasks. It provides a wide range of functionalities such as object detection, facial recognition, edge detection, motion tracking, and real-time video analysis. Written primarily in C++ with bindings for Python, OpenCV is optimized for performance and supports various platforms like Windows, Linux, macOS, and mobile systems. It is frequently used in robotics, security systems, augmented reality, and automated inspection. OpenCV can process images and videos quickly and works well in combination with other libraries like NumPy and TensorFlow to build complex computer vision applications.

4.3. TensorFlow

TensorFlow is an open-source framework developed by Google for machine learning and deep learning applications. It enables developers to build and train neural networks for tasks such as image recognition, natural language processing, and time series forecasting. TensorFlow offers flexibility with both low-level control for research and experimentation, and high-level APIs such as Keras for quick and easy model development. It supports CPU and GPU acceleration, allowing efficient computation on large datasets. TensorFlow is widely adopted in industries and academia due to its scalability, robustness, and tools for deployment in production environments, including mobile and cloud platforms.

4.4. NumPy (Numerical Python)

NumPy is a core library in Python for numerical computing, particularly useful for working with large arrays and matrices. It provides a powerful N-dimensional array object along with a collection of mathematical functions to perform operations such as linear algebra, statistics, and Fourier transforms. NumPy is highly optimized for performance and is considered the foundation for many scientific computing libraries in Python, including pandas, SciPy, and scikit-learn. It is often used in data preprocessing tasks and integrates seamlessly with OpenCV and TensorFlow, making it essential for image processing, machine learning, and data science projects.

V. RESULT AND DISCUSSION

The displayed output demonstrates a real-time human pose estimation system in action. The system is actively detecting the subject's body posture, highlighting key body joints and forming a skeletal structure with green lines and red points. The interface provides options for camera control, pose estimation, multi-person tracking, and behavioral recognition. The current frame rate (FPS) is visible, indicating the processing speed. The subject stands on a staircase, and the system accurately overlays pose estimation markers, confirming its functionality in a real-world environment.

VI. CONCLUSION

In conclusion, the proposed Human Action Recognition system presents an efficient and user-friendly approach to identifying human activities from still images by integrating pose estimation and supervised machine learning techniques. By leveraging the OpenPose library. The use of pre-processing and normalization enhances the reliability of the input data, while machine learning algorithms such as Support Vector Machine (SVM) or Random Forest ensure robust performance in recognizing various human actions. Unlike traditional wearable sensor-based systems, this vision-based solution does not require any physical equipment to be worn by the

user, making it more comfortable and practical for real-world applications. The web-based interface facilitates seamless interaction, allowing users to register, log in, capture images, and receive real-time action predictions. Through extensive testing under diverse conditions, the system has demonstrated strong accuracy and adaptability, making it suitable for applications in surveillance, behavioral analysis, and smart environments. Overall, this paper contributes to the advancement of image-based action recognition and provides a foundation for further enhancements in human behavior understanding through computer vision.

VII. REFERENCES

[1]. S. N. Boualia and N. Essoukri Ben Amara, "Pose-based Human Activity Recognition: a review," *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, Tangier, Morocco, 2019, pp. 1468-1475, doi: 10.1109/IWCMC.2019.8766694.

[2]. D. Nikolova, I. Vladimirov and Z. Terneva, "Human Action Recognition for Pose-based Attention: Methods on the Framework of Image Processing and Deep Learning," *2021 56th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)*, Sozopol, Bulgaria, 2021, pp. 23-26, doi: 10.1109/ICEST52640.2021.9483503.

[3]. B. Snehitha, R. S. Sreeya and V. M. Manikandan, "Human Activity Detection from Still Images using Deep Learning Techniques," *2021 International Conference on Control, Automation, Power and Signal Processing (CAPS)*, Jabalpur, India, 2021, pp. 1-5, doi: 10.1109/CAPS52117.2021.9730709.

[4]. Ma, N., Wu, Z., Cheung, Y., Guo, Y., Gao, Y., Li, J., & Jiang, B. (2022). A survey of human action recognition and posture prediction. Tsinghua Science & Technology, 27(6), 973–1001. https://doi.org/10.26599/tst.2021.9010068.

[5]. D. C. Luvizon, D. Picard and H. Tabia, "2D/3D Pose Estimation and Action Recognition Using Multitask Deep Learning," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 5137-5146, doi: 10.1109/CVPR.2018.00539.

[6]. Song, L., Yu, G., Yuan, J., & Liu, Z. (2021). Human pose estimation and its application to action recognition: A survey. Journal of Visual Communication and Image Representation, 76, 103055. https://doi.org/10.1016/j.jvcir.2021.103055

[7]. F. Sajjad, A. F. Ahmed and M. A. Ahmed, "A Study on the Learning Based Human Pose Recognition," *2017 9th IEEE-GCC Conference and Exhibition (GCCCE)*, Manama, Bahrain, 2017, pp. 1-8, doi: 10.1109/IEEEGCC.2017.8448200.

[8]. Thurau, C., & Hlavac, V. (2008). Pose primitive based human action recognition in videos or still images. 2009 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2008.4587721

[9]. D. C. Luvizon, D. Picard and H. Tabia, "Multi-Task Deep Learning for Real-Time 3D Human Pose Estimation and Action Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 8, pp. 2752-2764, 1 Aug. 2021, doi: 10.1109/TPAMI.2020.2976014.

[10]. S. H. Javadi, A. Bourdoux, N. Deligiannis and H. Sahli, "Human Pose Estimation Based on ISAR and Deep Learning," in IEEE Sensors Journal, vol. 24, no. 17, pp. 28324-28337, 1 Sept.1, 2024, doi: 10.1109/JSEN.2024.3426030.