


# Gait-Driven Pose Tracking and Movement Captioning Using OpenCV and MediaPipe Machine Learning Framework <sup>†</sup>

Malathi Janapati \* , Leela Priya Allamsetty, Tarun Teja Potluri and Kavya Vijay Mogili

Department of Artificial Intelligence and Data Science, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada 520007, Andhra Pradesh, India; 218w1a5429@vrsec.ac.in (L.P.A.); 218w1a5447@vrsec.ac.in (T.T.P.); 218w1a5433@vrsec.ac.in (K.V.M.)

\* Correspondence: malathi.j@vrsiddhartha.ac.in

<sup>†</sup> Presented at the 11th International Electronic Conference on Sensors and Applications (ECSA-11), 26–28 November 2024; Available Online: <https://sciforum.net/event/ecsa-11>

**Abstract:** Pose tracking and captioning are extensively employed for motion capturing and activity description in daylight vision scenarios. Activity detection through camera systems presents a complex challenge, necessitating the refinement of numerous algorithms to ensure accurate functionality. Even though there are notable characteristics, IP cameras lack integrated models for effective human activity detection. With this motivation, this paper presents a gait-driven OpenCV and MediaPipe machine-learning framework for human pose and movement captioning. This is implemented by incorporating the Generative 3D Human Shape (GHUM 3D) model which can classify human bones while Python can classify the human movements as either usual or unusual. This model is fed into a website equipped with camera input, activity detection, and gait posture analysis for pose tracking and movement captioning. The proposed approach comprises four modules, two for pose tracking and the remaining two for generating natural language descriptions of movements. The implementation is carried out on two publicly available datasets, CASIA-A and CASIA-B. The proposed methodology emphasizes the diagnostic ability of video analysis by dividing video data available in the datasets into 15-frame segments for detailed examination, where each segment represents a time frame with detailed scrutiny of human movement. Features such as spatial-temporal descriptors, motion characteristics, or key point coordinates are derived from each frame to detect key pose landmarks, focusing on the left shoulder, elbow, and wrist. By calculating the angle between these landmarks, the proposed method classifies the activities as “Walking” (angle between  $-45$  and  $45$  degrees), “Clapping” (angles below  $-120$  or above  $120$  degrees), and “Running” (angles below  $-150$  or above  $150$  degrees). Angles outside these ranges are categorized as “Abnormal”, indicating abnormal activities. The experimental results show that the proposed method is robust for individual activity recognition.

**Keywords:** activity recognition; gait analysis; human movement; machine learning; movement captioning; pose tracking



**Citation:** Janapati, M.; Allamsetty, L.P.; Potluri, T.T.; Mogili, K.V. Gait-Driven Pose Tracking and Movement Captioning Using OpenCV and MediaPipe Machine Learning Framework. *Eng. Proc.* **2024**, *6*, 0. <https://doi.org/>

Academic Editor:

Published: 26 November 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Human movement is more than just a series of physical actions, it also reflects emotional and psychological dimensions. The way people walk, move their hands, and engage their bodies can convey a great deal about their emotions, intentions, and motivations. In recent years, researchers have increasingly recognized the subtle ways in which our movements serve as a form of non-verbal emotional expression. Xu et al. have shown how advanced video analysis systems can help identify emotions based on how we walk [1]. Karg and colleagues further highlighted that specific gait patterns can be used to detect emotional states [2]. Bhattacharya and his team added to this by applying hierarchical attention methods to better understand the connection between walking styles and emotional expressions [3]. The growing interest in how movement reflects emotion has led to

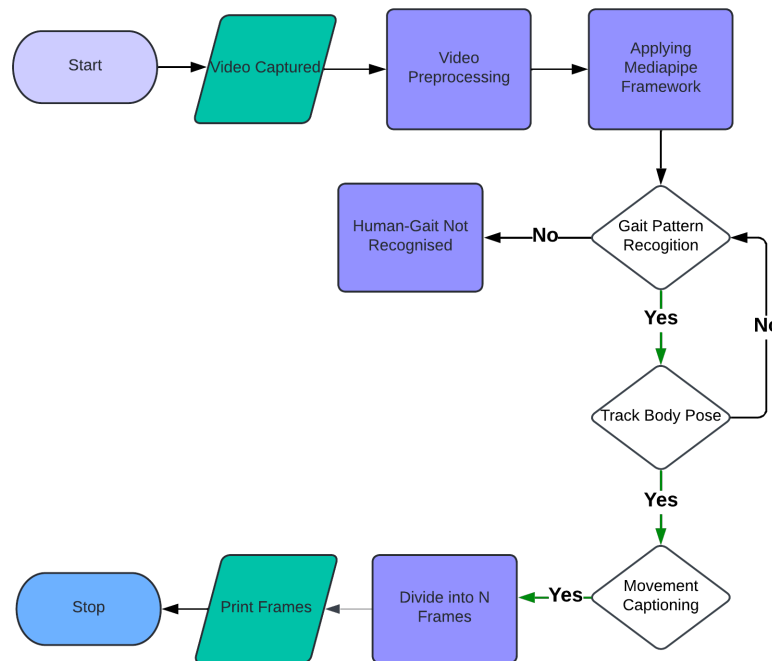
the development of sophisticated real-time video analysis systems that can interpret and categorize human movements [1,4]. This proposed work builds upon these advancements by combining state-of-the-art artificial intelligence with Python programming. Using the GHUM 3D model for detailed human body reconstruction, our system accurately classifies the skeletal structure and identifies whether movements are typical or atypical [4]. This innovative approach allows us to achieve new levels of accuracy in detecting and understanding human movement. This system offers in-depth insights into posture and motion, with potential applications in areas such as healthcare, sports, surveillance, and behavioral analysis. By merging computer vision, machine learning, and the study of human motion this work contributes to a rapidly evolving field with far-reaching implications [5,6]. Further, this research pushes the limits of how we interpret movement, offering exciting opportunities for both scientific exploration and practical use.

## 2. Literature Survey

Fan et al. introduced CM-YOLOv8, a lightweight version of the YOLO object detection model, optimized for use in coal mine fully mechanized mining face monitoring [7]. Singh et al. provided a comprehensive survey on vision-based gait recognition, detailing its methods, challenges, and future directions for human identification systems [8]. Zhang, Vogler, and Metaxas discussed early advancements in human gait recognition, emphasizing the use of visual data for identifying individuals [9]. Bashir, Xiang, and Gong explored gait recognition techniques that work without the active cooperation of subjects, useful in surveillance applications [10]. Liu and Sarkar proposed an improved approach to gait recognition by normalizing gait dynamics, enhancing recognition accuracy across different walking conditions [11]. Bobick and Johnson presented a gait recognition method that uses static, activity-specific parameters to identify individuals based on their walking patterns [12]. Shakhnarovich, Lee, and Darrell proposed an integrated face and gait recognition system that uses multiple views for enhanced biometric identification [13]. Xu et al. introduced a matrix representation method for human gait recognition, improving recognition performance by capturing spatial and temporal features [14]. Gafurov provided a survey on biometric gait recognition, discussing various approaches, security issues, and associated challenges in the field [15]. Gonçalves dos Santos et al. reviewed deep learning-based approaches to gait recognition, offering insights into current advancements and future trends in this emerging area of biometric research [16]. Dudekula et al. presented a system for physiotherapy assistance that utilizes human pose estimation on Raspberry Pi, offering an affordable and effective solution for patient movement analysis and rehabilitation monitoring [17]. Mangone et al. reviewed advancements in gait analysis technology, emphasizing its rehabilitation benefits and exploring its potential applications in forensic science for identifying movement abnormalities and personal identification [18]. Xu et al. proposed a novel method for human gait pattern recognition, offering valuable insights for its application in both sports performance analysis and clinical gait assessments [19]. Fang et al. presented a multi-module sensing and bi-directional human-machine interface (HMI) that integrates interaction, recognition, and feedback mechanisms for enhancing the capabilities of intelligent robots [20].

## 3. Methodology

The general philosophy of the gait-driven pose tracking and movement captioning system is to develop this method in several important stages so that the final system will not only bring the expected effectiveness but also be simple in its operation. A detailed description of the proposed methodology is provided in the next subsections namely the involvement of GHUM 3D for skeleton recognition, Python for movement categorization, HAR for activity recognition, the chosen video analysis techniques, and finally the utilization of Streamlit for UI development. Including a frame division methodology is explained in Figure 1.



**Figure 1.** Workflow of the proposed methodology.

### 3.1. GHUM 3D for Skeleton Recognition

GHUM 3D is a combination of technology and software components that allow us to precisely measure and identify the bones of a skeleton in gait-driven pose tracking. The process used for deploying the GHUM 3D system is to get a look at the bones. This proposed methodology is implemented using the GHUM 3D trained model that is shown in Figure 1. GHUM 3D, a new medium-level software incorporating a network of specialists employing advanced vision techniques, is intended to be used for tracking people from video data and coming up with exact estimations of posture and skeleton patterns. Using advanced computer vision algorithms and deep learning, it determines the correct human skeletons in real-time reliably. Support for GHUM 3D does not mean you install and set it up beforehand but rather facilitate the process. GHUM 3D is an input video data-based method, which suppresses human versions whose skeletons are supposed to be recognized by it. Various sources of video data can be used, such as surveillance camera footage, recorded videos, or live streams. Before starting the preprocessing of input video data, it is aimed to improve the image quality and the dataset usability. This includes tasks like video stabilization, noise reduction, and frame alignment. It applies deep learning models, trained on large datasets collected by video, to identify the human skeleton in video information. These models perform feature extraction and body part identification tasks with the help of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to obtain high accuracy in finding joints and limbs. Following the recognition of human skeletons, GHUM 3D finds each element of the human body in space in each frame of the video. It starts with locating key landmarks and uses a skeleton model to calculate its moving pose. The final product will be the skeletons of humans rebuilt from scratch and the posed estimate of the subjects in the video images. Such results can be shown in real mode or saved for subsequent examination and further processing. The GHUM 3D integrates into the workflow as a component of Python programming for joint project development, bringing its functionality inside the Python environment for automatic movement classification and analysis to achieve intended data collection and streamlined processing. Beyond 3D skeleton recognition, the conformity and reliability of the calculated skeletons need to be considered, including comparing discovered skeletons to ground truth data or manual annotations to establish system accuracy. Optimization and polishing of

GHUM 3D will improve performance and reduce energy costs, involving methods for adjusting model parameters, fine-tuning algorithms, and incorporating feedback from real-life experiences.

### 3.2. Python for Categorizing Movements

In this work, it is explored how Python can categorize human movements into common and unusual categories. This section focuses on Python’s role in this process and other tools involved. The initial approach involves preprocessing the data to prepare it for analysis. Data may come from recorded videos or live streams. The motion-capture data is then segmented into individual moves or specific scenes and processed. Python retrieves essential features from motion data, focusing on identifying significant criteria for building classifiers that differentiate normal and abnormal patterns. Features include type of movement, speed, acceleration, joint space, or spatial-temporal descriptors. Python libraries used for machine learning are ideal for building categorization models. Algorithms like decision trees, SVM, random forests, or deep neural networks are used to train models for movement classification. The dataset is labeled with ground truth data indicating normal or abnormal movements. Python scripts extract features and preprocess data for training machine learning models. The models learn to recognize movement patterns and relationships to maximize classification accuracy by minimizing errors.

### 3.3. Human Activity Recognition

Human Activity Recognition (HAR) aims to recognize specific activities in video captures. HAR involves machine learning and pattern recognition algorithms to classify and identify human activities from a camera. Activities range from simple actions like walking and running to more complex actions like gait or item manipulation. It starts with collecting video data showing the activities to be identified, sourced from surveillance cameras, or video recordings. Data are annotated with labels representing activity classes and timestamps. Feature extraction provides features that represent individual activities, including spatial-temporal descriptions, motion patterns, or frequency-domain representations. Python scripts are used to extract features from tagged video data for training. Various machine learning algorithms, including decision trees, SVM, k-nearest neighbors (KNN), and deep learning algorithms like CNNs and RNNs, are available in Python. GHUM 3D model is trained on a labeled dataset to improve accuracy, followed by validation to check performance and generalization ability. Once trained and validated, HAR is integrated into the Python-based system for real-time identification of human activities, with Python scripts overseeing communication between the HAR model and system components. It automatically reviews processed video data, detects activities, and provides feedback or initiates actions. The HAR system’s performance is measured using accuracy, precision, recall, and F1-score, with Python scripts highlighting areas for improvement. Human Activity Recognition using Gait is observed in Figure 2.

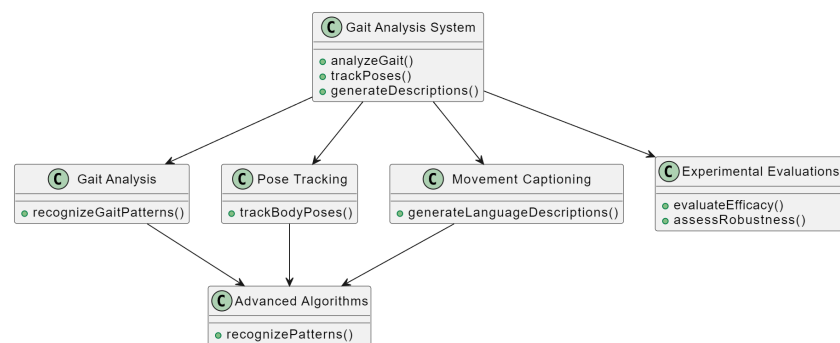


Figure 2. Human activity recognition using Gait.

### 3.4. Video Analysis by Dividing Video Data into 15 Frames

This phase emphasizes the diagnostic ability of video analysis by dividing the data collected from real-time camera-captured videos into 15 frames for detailed examination. Each segment represents a time frame with detailed scrutiny of human movement. Features such as spatial-temporal descriptors, motion characteristics, or keypoint coordinates are derived from each frame. Python scripts analyze each segment separately using computer vision and machine learning algorithms to modify patterns, find anomalies, or classify actions. Integration of frame analysis with HAR and pose tracking is a system component, with Python scripts aligning these parts for comprehensive human movement understanding. Real-time analysis allows automated documentation, with Python scripts providing high operational efficiency even with large video data. Frame data are processed into visual plots and reports for interpretation, with iterative improvements based on feedback.

### 3.5. Streamlit for UI Development

Streamlit is used to develop the UI part of the system, providing a user-friendly interface for interaction and visualization. The initial setup involves installing required dependencies, including Python, Streamlit, and package managers like pip or conda. Streamlit offers an intuitive interface for incorporating UI elements such as buttons, sliders, charts, and tables. It integrates backend components like data processing modules, machine learning models, and visualization libraries. Python scripts create communication systems between the front end and backend, enabling real-time updates and interactions. Streamlit's interaction widgets allow users to manipulate parameters, explore data, and visualize results. Python scripts define widget behavior and functionality, creating interactive plots and charts. Streamlit renders immediate visualizations following modifications or user interactions, with Python scripts managing event triggers and data updates. The deployment of the UI involves hosting Streamlit web applications on platforms like Heroku or AWS, with Python scripts ensuring stability and performance. User feedback is continuously incorporated into the UI design, making it accessible and usable for all users, including those with disabilities. The scalable and extendable UI framework allows for future improvements and functionality enhancements.

### 3.6. Integration of Gait Analysis, Pose Tracking, and Movement Captioning

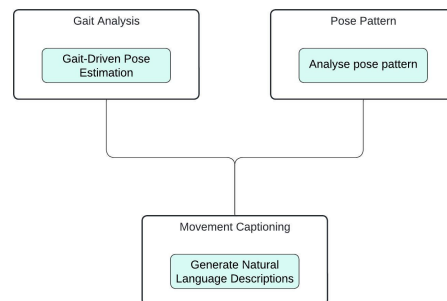
The work provides a holistic approach to gait analysis, body pose tracking, and motion interpretation, integrating these elements into a concise framework. GHUM 3D detects and segments body joints during gait analysis using real-time pose-tracking algorithms. Movement recognition systems communicate with movement classification systems, where Python scripts turn abnormal actions into normal ones and draw useful conclusions. Data integration and synchronization create a complete image of gait and pose data, fitting into machine learning algorithms for activity recognition and anomaly detection. Real-time online data streams provide immediate input and visualizations of human movement. The user-friendly interface, created with Streamlit, allows data entry and meaningful interpretation, with continuous enhancement based on user feedback to ensure the system remains vigorous, prompt, and dynamic.

### 3.7. Feature Extraction

Feature extraction focuses on abstracting meaningful characteristics of human gait movements and gestures along with pose data are shown in Figure 3. This section clarifies the procedure involved in extracting valuable features. Feature extraction involves using Python scripts to process and derive significant characteristics from video data, such as joint angles, movement speeds, accelerations, and motion trajectories. These features are extracted from GHUM 3D data and segmented video frames, focusing on variables that can discriminate between different movements and postures. The features are used to develop classification models and assess the model's performance for effective movement recognition. Data preprocessing, including normalization and scaling, ensures the extracted



features are in an appropriate format for model input. The performance of the feature extraction process is continuously monitored, and improvements are made to enhance the feature set and overall system efficiency. Iterative validation ensures that the extracted features support accurate recognition and classification.



**Figure 3.** Process of feature extraction.

#### 4. Results

The proposed work utilizes OpenCV and MediaPipe machine learning frameworks for real-time activity recognition through human pose estimation. It processes video frames to detect key pose landmarks, focusing on the left shoulder, elbow, and wrist. By calculating the angle between these landmarks, the system classifies activities as “Crawling” for angles below  $-120$  or above  $120$  degrees which is observed in Figure 4. The activity “Walking” when the angle is between  $-45$  and  $45$  degrees is observed in Figure 5. Any angles falling outside these specified ranges are categorized as “Abnormal”, indicating that the detected movement does not match predefined activity patterns. The system displays pose landmarks with connected skeletal structures and the identified activity on each frame, demonstrating its effectiveness for applications in sports analytics, physical therapy, and interactive systems. The real-time activity analyzed on the system is shown in Figure 6.

The results provide a detailed analysis of various movement patterns, showcasing the differences between walking, running, and clapping, with an emphasis on activity recognition through varying angles and utilization. Every activity parameter is differentiated and calculated using gait angles, landmarks, and utils. Activity recognition is enhanced by analyzing these patterns from different angles, which helps in accurately distinguishing between the movements based on their biomechanical properties and dynamic characteristics. The clapping graphs display frequent, short bursts of motion, with notable peaks corresponding to each clap, illustrating the high-frequency, repetitive nature of this action. Clapping activity analysis is shown in Figure 7. Running, on the other hand, shows more pronounced variations in speed and shorter, faster strides, emphasizing increased intensity and energy expenditure compared to walking, with variations depending on the angle of observation. Running analysis is shown in Figure 8. Analysis of video divided into frames and the correctness of frames and crawling is shown in Figure 9. For walking, the images reveal a consistent and rhythmic pattern with a moderate stride length and steady pace, observed from a standard angle. The walking activity analysis is shown in Figure 10.



**Figure 4.** Crawling recognised.



Figure 5. Walking recognised.

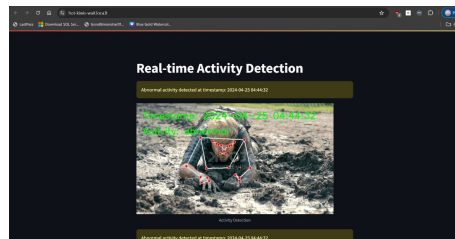


Figure 6. Real-time activity analysis.

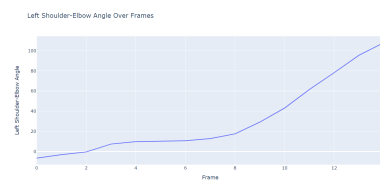


Figure 7. Graph of clapping.

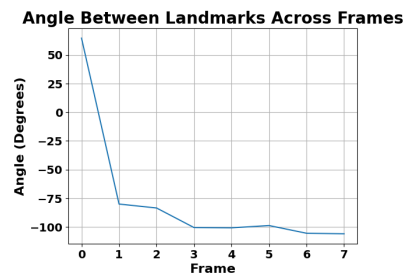


Figure 8. Graph of running.

Figure 9. Graph of crawling.

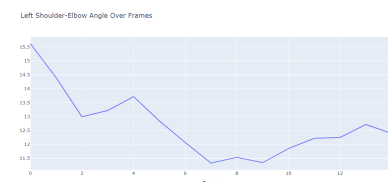


Figure 10. Graph of walking.

## 5. Conclusions and Future Work

The development of the gait-driven pose tracking and movement captioning system has successfully integrated technologies like GHUM 3D, Python-based machine learning, and HAR algorithms for accurate human movement analysis and achieved the purpose. The system allows users to upload videos, view results, and interpret movement captions, offering valuable insights for sectors such as healthcare, surveillance, and entertainment.

Future enhancements include incorporating emotion detection through gait analysis, refining machine learning models for better accuracy, and improving the user interface with features like real-time feedback and personalized recommendations.

#### *Ethical Considerations*

The study adhered to established ethical principles, ensuring that participants provided informed consent and that their privacy and confidentiality were maintained throughout the research. All procedures were conducted in compliance with relevant legal and ethical standards.

**Author Contributions:** Conceptualization, L.P.A. and T.T.P.; formal analysis, M.J.; investigation, T.T.P. and K.V.M.; software, L.P.A. and T.T.P.; supervision and Validation, M.J.; writing—original draft, L.P.A. and K.V.M.; writing—review and editing, M.J. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:**

**Informed Consent Statement:**

**Data Availability Statement:**

**Conflicts of Interest:** The authors declare no conflict of interest.

#### References

- Xu, S.; Fang, J.; Hu, X.; Ngai, E.; Wang, W.; Guo, Y.; Leung, V.C.M. Emotion recognition from gait analyses: Current research and future directions. *arXiv* **2020**, arXiv:2003.11461.
- Karg, M.; Kühnlenz, K.; Buss, M. Recognition of affect based on gait patterns. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2010**, *40*, 1050–1061.
- Bhattacharya, U.; Roncal, C.; Mittal, T.; Chandra, R.; Kapsaskis, K.; Gray, K.; Bera, A.; Manocha, D. Take an emotion walk: Perceiving emotions from gaits using hierarchical attention pooling and affective mapping. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2020; pp. 145–163.
- Sheng, W.; Li, X. Multi-task learning for gait-based identity recognition and emotion recognition using attention enhanced temporal graph convolutional network. *Pattern Recognit.* **2021**, *114*, 107868.
- Roether, C.L.; Omlor, L.; Christensen, A.; Giese, M.A. Critical features for the perception of emotion from gait. *J. Vis.* **2009**, *9*, 15.
- Cui, L.; Li, S.; Zhu, T. Emotion detection from natural walking. In *Proceedings of the Human Centered Computing: Second International Conference, HCC 2016, Colombo, Sri Lanka, 7–9 January 2016; Revised Selected Papers 2*; Springer International Publishing: Cham, Switzerland, 2016; p. 2333.
- Fan, Y.; Mao, S.; Li, M.; Wu, Z.; Kang, J. CM-YOLOv8: Lightweight YOLO for Coal Mine Fully Mechanized Mining Face. *Sensors* **2024**, *24*, 1866.
- Singh, J.P.; Jain, S.; Arora, S.; Singh, U.P. Vision-based gait recognition: A survey. *IEEE Access* **2018**, *6*, 70497–70527.
- Zhang, R.; Vogler, C.; Metaxas, D. Human gait recognition. In *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop*, Washington, DC, USA, 27 June–2 July 2004; IEEE: Piscataway, NJ, USA, 2004; p. 18.
- Bashir, K.; Xiang, T.; Gong, S. Gait recognition without subject cooperation. *Pattern Recognit. Lett.* **2010**, *31*, 2052–2060.
- Liu, Z.; Sarkar, S. Improved gait recognition by gait dynamics normalization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 863–876.
- Bobick, A.F.; Johnson, A.Y. Gait recognition using static, activity-specific parameters. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. CVPR 2001, Kauai, HI, USA, 8–14 December 2001; IEEE: Piscataway, NJ, USA, 2001; Volume 1, pp. I-I.
- Shakhnarovich, G.; Lee, L.; Darrell, T. Integrated face and gait recognition from multiple views. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. CVPR 2001, Kauai, HI, USA, 8–14 December 2001; IEEE: Piscataway, NJ, USA, 2001; Volume 1, pp. I-I.
- Xu, D.; Yan, S.; Tao, D.; Zhang, L.; Li, X.; Zhang, H.J. Human gait recognition with matrix representation. *IEEE Trans. Circuits Syst. Video Technol.* **2006**, *16*, 896–903.
- Gafurov, D. A survey of biometric gait recognition: Approaches, security, and challenges. In *Proceedings of the Annual Norwegian Computer Science Conference*, Norway, 2007; pp. 19–21.
- dos Santos, C.F.G.; de Souza Oliveira, D.; Passos, L.A.; Pires, R.G.; Santos, D.F.S.; Valem, L.P.; Moreira, T.P.; Santana, M.C.S.; Roder, M.; Papa, J.P. Gait recognition based on deep learning: A survey. *ACM Comput. Surv. (CSUR)* **2022**, *55*, 1–34.



17. Dudekula, K.V.; Chalapathi, M.M.V.; Kumar, Y.V.P.; Prakash, K.P.; Reddy, C.P.; Gangishetty, D.; Solanki, M.; Singhu, R. Physiotherapy assistance for patients using human pose estimation with Raspberry Pi. *ASEAN J. Sci. Technol. Rep.* **2024**, *27*, e251096. <https://doi.org/10.55164/ajstr.v27i4.251096>
18. Mangone, M.; Marinelli, E.; Santilli, G.; Finanore, N.; Agostini, F.; Santilli, V.; Bernetti, A.; Paoloni, M.; Zaami, S. Gait analysis advancements: Rehabilitation value and new perspectives from forensic application. *Eur. Rev. Med. Pharmacol. Sci.* **2023**, *27*, 3–12.
19. Xu, D.; Zhou, H.; Quan, W.; Jiang, X.; Liang, M.; Li, S.; Ugbolue, U.C.; Baker, J.S.; Gusztav, F.; Ma, X.; et al. A new method proposed for realizing human gait pattern recognition: Inspirations for the application of sports and clinical gait analysis. *Gait Posture* **2024**, *107*, 293–305.
20. Fang, P.; Zhu, M.; Zeng, Z.; Lu, W.; Wang, F.; Zhang, L.; Chen, T.; Sun, Li. A Multi-Module Sensing and Bi-Directional HMI Integrating Interaction, Recognition, and Feedback for Intelligent Robots. *Adv. Funct. Mater.* **2024**, *34*, 2310254.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.