Proceedings of the Second International Conference on Applied Artificial Intelligence and Computing (ICAAIC 2023)
IEEE Xplore Part Number: CFP23BC3-ART; ISBN: 978-1-6654-5630-2

Computer Vision & Deep Learning based Realtime and Pre-Recorded Human Pose Estimation

Milind Shah *[1], Kinjal Gandhi [2], Bhagyesha M Pandhi [3], Priyanka Padhiyar [4], Sheshang Degadwala [5]

[1 2] *Assistant Professor, Department of Computer Science & Engineering, Krishna School of Emerging Technology & Applied Research, Drs. Kiran & Pallavi Patel Global University (KPGU),*
Vadodara, Gujarat, India
[3] *Assistant Professor, Department of Computer Science & Engineering, Parul Institute of Engineering & Technology (PIET), Parul University (PU)*
Vadodara, Gujarat, India
[4] *Assistant Professor, Department of Computer Science & Engineering, Depstar Charusat University,*
Anand, Gujarat, India
[5] *Associate Professor, Head of Computer Engineering Department, Sigma Institute of Engineering,*
Vadodara, Gujarat, India

Email: *[1] milindshahcomputer@gmail.com , [2] kinjal445@gmail.com , [3] bhagyeshapandhi@gmail.com ,
[4] priyankapadhiyar31@gmail.com , [5] sheshang13@gmail.com

2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC) | 978-1-6654-5630-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAAIC56838.2023.10141279

***Abstract* – This research study utilizes computer vision to estimate multi-person human pose from real-time and pre-recorded video. Computer vision examines human posture detection from RGB images. The proposed method works well for gesture control, gaming, human tracking, action detection, and action tracking. Tracking semantic important points is pose estimation. Two-dimensional human pose estimation predicts the spatial placement of key human body points from images and videos. Several anatomical areas use hand-crafted feature extraction methods to estimate two-dimensional human position. Visual input data and human body component locations are used to estimate human pose. OpenCV and Mediapipe detected 33 posture landmarks in our research. Estimating human body state requires modeling. Model-based methods are used to describe and infer human posture in 2D or 3D. This research uses the BlazePose GHUM 3D Pose Landmark Model for 2D human pose estimation. Output poses include x, y, z, and visibility. Image width and height are X and Y landmark coordinates. landmark depth Z. Image shows landmark. Image or video resolution, the number of persons in the scene, and anomalies might impact visibility point accuracy, which was 0.969%.**

Keywords – Computer Vision, Deep Learning, OpenCV, Human Pose Estimation, RGB Camera, Single person and Multi person pose estimation.

I. INTRODUCTION

Human posture estimation determines the body's position in an image or video. Locating skeletal joints and integrating them to generate a stick figure representing a person's position is conventional. It gives critical points for each part (arm, head, body, etc.) to determine a person's position. The relationship between these points is known as a pair. The relationship between points must be significant, which means that not all points can be paired. HPE's initial goal was to create bone-like images of the human body and then process them for specific applications. Human posture estimation approaches are mainly in the field of computer vision and are used to understand the geometry and motion data of the human body, which can be quite complex. The general flow of human pose estimation system starts with initial data acquisition and loading for system operation. When dealing with motion detection, we need to analyze a series of images instead of images. We need to know how the main points change during the movement.

After the image is uploaded, the HPE system will find and track key points needed for analysis. In short, different software modules are responsible for tracking 2D points, creating body images, and converting them into 3D space. In general, when we talk about creating a body position estimation model, we consider applying two different modules for 2D and 3D planes.

Assessing human posture during exercise is an example in the field of fitness. Some models can also identify key points in the face and track head positions, which can be used for entertainment applications such as Snapchat masks [1].

This paper is broken down into six sections. The first section is the Introduction, the second portion is Why Computer Vision is Important for Human Pose Estimation, the third portion is Related Work, the fourth section is Methodology, the fifth section is Comparative Analysis, the sixth section is Results, Analysis, and Discussion, and the final seventh section is the Conclusion.

II. IMPORTANCE OF COMPUER VISION IN HUMAN POSE ESTIMATION

High-performance real-time pose detection and tracking is driving some of the biggest trends in computer vision. For

example, real-time tracking of human posture allows computers to understand human behavior in a more accurate and natural way. It also allows computers to better understand and predict pedestrian behavior and enables more natural driving.

One of the key challenges in human pose estimation is to accurately detect and track the position of hidden or semi-hidden parts of the body. Computer vision techniques such as image processing, machine learning, and deep learning can be used to analyze and interpret visual data to accurately locate and track body parts even in challenging situations. In general, computer vision is an essential tool for human pose estimation. It enables machines to automatically analyze and interpret visual data to understand and track human gestures in images and videos. This is important for a wide range of applications and helps to improve the accuracy and reliability of human gesture estimation systems.

III. RELATED WORK

In [1] Yongtao Zhang et al, this research aimed to provide a solution for 3D human location based on a multi-step regression deep network and a 2D to 3D point mapping algorithm. Initially, we take RGB photos as input and use a heat map and multistep regression to continually improve the locations of human touch points. The 2D joint points are then inserted into the mapping mesh for calculation, and the 3D human joint point coordinates are acquired in order to conduct the 3D human posture estimation job. 40.7 MPJPE method for the M dataset Adam3.6. The analysis of our method's benefits via the use of a database demonstrates its evident advantages. For the joint position in two dimensions, we suggest using a multistage regression network and providing a heat map. Heatmaps are used to communicate position information of frequent points between two locations, so that the network pays greater attention to information around common points and effectively extracts characteristics. Our method outperforms the Martinez method on the Human3.6 M dataset and decreases the MPJPE by 17%. Using just 2D composite data, the 2D to 3D mapping panel may produce 3D composite coordinates. This approach is simpler than the algorithm that employs several kinds of videos and photos, and testing indicate that the algorithm is more effective. Since the method consists of two components, we plan to enhance the performance of both components in the future. For two-dimensional joint point placement challenges, a joint heatmap is preferable to a human anatomy heatmap (joint line) for point placement. In the meanwhile, there exist 2D and 3D mapping algorithms that address the issue of human self-occlusion. Optimized to address this issue. This is due to the fact that the method suggested in this research is primarily concerned with performance.

In [2] Hui Tang et al, this research aims to develop a novel technique for human 3D posture estimation, color extraction, and depth imaging using RGBD cameras. Using a convolutional neural network to predict the 2D human position and 3D joint point information, extract the joint point coordinates from the color picture and transfer the output to the appropriate inner image. Faster-RCNN and Resnet50 residual structure increase hourglass network gathered as human target extract for 2D pose estimation. The mapping procedure determines images color and depth calibration values using SURF-based sparse feature point matching. Three important developments. Utilize an RGBD camera to estimate the position of a 2D human in a color image, then use the depth image to fill in depth. Faster-human RCNN's detector utilizes Resnet50 to extract human characteristics and increase target detection. Sparse feature point matching can color match the depth image and appropriately match map color points to depth image points. These two approaches accurately estimate 3D human position, according to research.

In [3] Marko Linna et al, using convolutional neural networks, we propose a technique for real-time estimate of distinct human activities from video. Our approach is designed for situations where fine precision is required, and backdrop and position changes are minimized. It permits the employment of an accurate, quick, and versatile network design. The problem is divided into two phases: (1) preliminary preparation and (2) fine-tuning. Prepare the network in preparation using input data from a public database, then change the train using operational data collected from Kinect. Our technique is distinct from current practices because it incorporates a human detector automated registration system, a state estimator, and cleaning-specific training materials. Our approach is much quicker than the majority of methods. Our approach is the most sophisticated alternative to Kinect and may be utilized for high-level activities, including gesture control, gaming, personality tracking, motion detection, and behavior recognition. Our technique obtained 96.8% accuracy (PCK@0.2).

In [4] Miniar Ben Gamra et al, the objective of this research is to analyze advanced approaches that have been offered to address the issue of 2D and 3D location estimation. Two primary pipelines are developed based on the number of persons in the image: single-person and multiple-person techniques. According to the suggested architecture, each of these categories is split into two groups. In addition, we briefly outline the present data set and the metrics used to assess the method's success. Finally, we analyze the benefits and drawbacks of the preceding technique. The goal of this article is to examine the state-of-the-art methodologies and research offered to tackle 2D and 3D state estimation issues. Two primary pipelines are developed based on the number of persons in the image: single-person and multiple-person techniques. According to the proposed architecture, each of these categories is split into two groups. In addition, we briefly describe the present data set and the metrics used to analyze the method's success.

In [5] Ali Rohan et al, this research utilised CNNs to classify normal and abnormal gaits from extracted skeletal pictures. CPU and GPU simultaneous processing reduces real-time CNN computation time. One frame and batch processing takes 47 milliseconds. This method yields 20 FPS. CNN on ResNet50 was used to compare network computation time and FPS. ResNet50 takes 12 MS longer per frame than the suggested CNN at 15 fps. The proposed method classifies gait disorders by body posture evaluation. The proposed approach addresses earlier analysis method issues by offering an appropriate CNN data gathering and training strategy. Experimental results provide ways to overcome motion analysis system constraints. The proposed

method classified normal and pathological gaits with 97.3% accuracy, proving its applicability. By adding classifier training details, this system can distinguish various human behaviors.

In [6] Yalin Cheng et al, the purpose of this research is RGB image-based modular interactive framework. This framework addresses the disadvantages of human-robot interaction (HRI) frameworks based on the human body, which are dependent on human cameras and unsuitable for long distances. An optical camera instead of a depth camera improves HRI frame adaption at different distances. Tests indicate an interactive framework. Indoor cameras work well at different distances, lighting, costumes, customers, and settings and are safer than frame-based. Our experimental human posture estimation algorithm can only detect one individual. Therefore, if there are several people in the image captured during the interaction, it is more important that the combined information (close to the center or the difference in the background color) is captured for the interaction. Then, after modeling static class interaction, the interaction mechanism here only has one effect of the same type, but with a different movement size. Therefore, the next task is to consider the full interaction with the presence of many people and ask the robot to perform appropriate response actions in different ranges of motion to empower the robot with stronger intelligence.

In [7] Chen Wang et al, this research aims to present a learning-aware learning (CAL) approach that explains the two primary drawbacks of current offset learning methods: asynchronous training and testing, distributed heatmaps, and offset learning. In particular, CAL picks the heat map based on the ground truth and the most dependable estimations and determines the statistical significance of the model developed using small batch learning. Extensive testing on the COCO indicator demonstrates that our technique for low-resolution human posture estimation greatly beats state-of-the-art methods. This research's primary contribution is a thorough analysis of current approaches for low-resolution issues, which demonstrates that offset learning is a successful technique. By including heavy masks into the thermal and offset training processes, CAL addresses the two most significant drawbacks of earlier offset-based techniques. This methodology improves the consistency of learning and assessment, bringing these two learning objectives closer together. In addition, a two-stage training technique was created to enhance the performance of this method. Extensive experiments on generic COCO indicators demonstrate that our CAL technique for low-resolution location estimation beats current state-of-the-art methods. Finally, intriguing research was done to examine the efficacy of our method's various components. Future research will use other approaches, such as lightweight meshes, sparse solutions, mesh truncation, and quantization, to expand our method and deploy it in various computing environments, such as CAL. Our objective is to decrease the computational complexity of neural architecture research and approaches.

In [8] Jinbao Wang et al, the objective of this research is to present a comprehensive analysis of available deep learning algorithms for 3D pose estimation, to highlight the strengths and shortcomings of various methods, and to get a greater knowledge of the subject. In addition, we present a complete overview of the most widely utilized large databases. Our research highlights the current status of 3D human posture estimate research and gives insights that will aid in the construction of future models and algorithms. We can also notice that this issue is increasingly prevalent in the computer vision field and that it has performed well on the Human3.6M, Human Eva, and MPI-INF-3DHP databases. However, it remains challenging to generalize to real-world settings. Taking into account the multiplayer mode, the one-step technique is undeveloped, indicating that human 3D posture estimate is impractical in real-world circumstances. Recent emphasis has been placed on gaining a complete comprehension of settings and postures. In addition, deep learning is quite successful at resolving this issue, so we may anticipate several improvements in the future year, particularly if new deep learning methods are used in this sector. In addition, we feel that future research paths on 3D human pose estimation reliability, safety, and federated learning are very promising.

In [9] Diogo C Luvizon et al, the purpose of this research is to propose a multi-objective framework for joint estimation of human motion classification from 2D or 3D monocular color images and video sequences. The proposed method combines the processing of images and video clips in a single pipeline, exploiting the sharing of advanced parameters between the two tasks. This allows you to train your model using different data sets at the same time. In addition, the isolation of the critical forecast section provides important insights for further training of the proposed multivariate model. This improves both jobs. Our target task technique works on MPII, Human3.6M, Penn Action, and NTU RGB+D datasets. The proposed CNN architecture and pose regression algorithms allow multi-dimensional posture and motion modification and re-injection for tight control. Manually annotating "wild" photos and 2D postures with accurate 3D data is possible with this strategy. 3D posture estimation improves significantly. The method processes frames and video clips concurrently.

In [10] Inho Chang et al, the aim of this research is to propose autonomous 3D pose estimation without 3D interpretation. Instead, we use various images and camera parameters to train the network to learn 3D human poses based on geometric sequences. Testing verifies the method. This research provides self-learning image-based 3D human posture estimation. We use a loss function to reduce the depth ambiguity and dependence on 3D data, and experimentally prove a convincing improvement.

In [11] Manuel Palermo et al, this research's objective is to analyze a real-time body evaluation framework for a rehabilitation smart walking device using two RGB+D camera channels with non-overlapping views. A two-step neurological procedure is used to determine a person's focal point. The fundamental contribution of this research is the introduction of a unique patient monitoring and human control technique in the context of intelligent walkers. It can build a complete, compact body representation using authentic, low-cost sensors and serve as a common base for sub-index extraction and human-robot interaction applications. To evaluate the effectiveness of rehabilitation technology in real-world situations, it is necessary to collect further data on disabled users irrespective optimistic results.

In [12] Weijian Chen et al, it uses the MobileNet framework and TensorFlow Lite launched by Google to

properly optimize complex network models and deploy them on Android smartphones. It avoids the latency problems caused by cloud computing solutions and can quickly achieve people's condition assessment using only local computing. In this paper, we propose an Android-based multiplier gesture estimation system and prove that it performs well in complex environments. The network structure is optimized based on this guarantee. It allows the Android system to quickly and accurately estimate the human skeleton.

In [13] Van-Thanh Hoang et al, the objective of this research is to improve human location using the RCNN Mask. MobileNetV3 and deep-decomposable convolution are recommended to improve model size, FLOPs, and search time. This model performs well at 25 FPS. The proposed technique utilizes the Coco database as its foundation. In the future, we must improve the model's efficiency. For speed, this model will be re-optimized.

In [14] Qiuhui Chen et al, new Status of Human Dataset (SHPD). SHPD serves two functions, unlike human mode database keypoint-based fine-grained segmentation: range of human things, the main goal of real-world open detection programs; SHPD collects images from live security cameras that show individuals in various outdoor spaces. Human control worldwide. It has relevant rich features. Global pose estimation utilizing a few SHPD-based deep learning networks indicates that identification accuracy may be improved. Four popular deep learning networks showed that most models without objective design do not fulfil actual control application accuracy criteria. Hence, subject position accuracy can be improved.

In [15] Tewodros Legesse Munea et al, the aim of this research is to fill the gap in knowledge and inform research on 2D human posture assessment. After a brief introduction, we classify them as single or multi-person assessments depending on the number of people following. In the next step, we will step-by-step discuss the approach used in human pose estimation and list some applications and shortcomings encountered in pose estimation. This article serves as a foundation for new introductions and guides researchers to find new models by finding gaps in existing research architectures and procedures.

V. METHODOLOGY

In this research, we have implemented human pose estimation using Mediapipe and OpenCV and then analyzed human poses by extracting a pre-recorded video or doing a real-time recording and, at last, identifying and tracking all of the elements and actions of those poses.

A) OpenCV Library

OpenCV (Open Source Computer Vision Library) is a well-known open-source computer vision library that comprises various image and video processing features and capabilities. OpenCV's potential to assist with human posture estimation is one of its most valuable features. In this research, we utilized a Skeleton-based strategy, which estimates the human position using a specified skeleton model. OpenCV includes a human skeleton model implementation that may be utilized for motion estimation.

OpenCV offers a comprehensive collection of tools and methods for human pose estimation. Its algorithms and methodologies may be used to reliably identify and estimate the location and configuration of a person's biological components, making it beneficial for a variety of computer vision applications.

B) Mediapipe Library

Google's Mediapipe is a robust open-source toolkit that provides machine learning solutions to a wide range of tasks, including human pose estimation. It provides a collection of pre-built models, APIs, and tools to assist developers and researchers in creating real-time computer vision applications.

Based on a deep neural network model, Mediapipe's human pose estimation solution can recognize 33 2D landmarks of a human body from an input image or video. These points of reference include the shoulders, elbows, wrists, hips, knees, and ankles.

OpenCV and Mediapipe can offer a more comprehensive solution for human pose estimation when used together. OpenCV may be used, for instance, to preprocess the incoming images or videos in order to recognize the human outline, or to modify the output of Mediapipe's landmark identification model. Mediapipe may be utilized to do accurate landmark identification, which can then be integrated with OpenCV's algorithms to reliably estimate a person's pose.

To use this model with OpenCV you can follow these steps:

1. Install mediapipe and OpenCV on the machine.
2. Load the Pose Landmark Model using the Mediapipe Library into a Python script.
3. OpenCV is used to read images and videos.
4. Process every video frame using the Pose Landmark Model.
5. Display the output keypoints using OpenCV drawing functions on the input picture.

VI. COMPARATIVE ANALYSIS

Literature	Algorithm Used	Accuracy / Visibility	Dataset	Findings	Limitation & Future Scope
Research on 3D Human Pose Estimation using RGBD Camera [2]	Faster-RCNN	0.266	MPII and Coco2007	Several research efforts have centered on retrieving 3D human body information from color	-

				images, which is inaccurate, slow, and imprecise.	
Realtime Human Pose Estimation from Video with Convolutional Neural Network [3]	CNN	96.8%	MPII, Fashion Pose, Leeds Sports Pose	We consider the full system, including human detector, posture estimator, and an automated mechanism to capture application-specific training information for finetuning, unlike other state-of-the-art methods.	Further work might improve accuracy by considering numerous factors. Use the present model as a coarse estimator and another network for pose estimation.
Human Pose Estimation based Real time gait analysis using CNN [5]	CNN	97.3%	ImageNet	Recent developments in Artificial Intelligence (AI), particularly deep learning, could now create a framework where a deep learning tool like Convolutional Neural Network can translate the data captured by embedded specialized devices. Previously, a researcher, usually a medical expert, had to translate the data (CNN).	By adding additional data to the classifier's training, the system may identify gait.
Multitask deep learning for Realtime 3D Huma pose estimation and Action recognition [9]	CNN architecture	49.5% & 48.6% for single and multi-task, 98.2% & 98.7% for single and multi-clip, 89.9%	MPII, Human 3.6M, Penn Action, NTU RGB+D	Provide essential information for end-to-end training the proposed multi-task model by differentiating important prediction components,	Ultimately, a single training method may cut our multi-task model at several levels for pose and action predictions, making it very scalable.

				which consistently increases accuracy on both tasks.	

VII. RESULTS, ANALYSIS & DISCUSSION

Human pose estimation is an active and emerging field of research in the fields of deep learning and computer vision. The reason why so many machine learning, deep learning, and computer vision enthusiasts are attracted to human pose estimations is because of its wide variety of applications and usefulness. And it's usually implemented by locating key points on an object. Based on these key points, different movements and situations can be compared and insights can be drawn. In this research, we will be using the Pose Landmark Model (BlazePose GHUM 3D) for realtime & pre-recorded human pose estimation, and it will be used to recognize human gestures and extract key points. This model in Mediapipe Pose predicts the location of 33 pose landmarks.

A media pipeline library may implement this model efficiently. Media Pipeline is a cross-platform, open-source multimodal machine learning system. It can apply complex models including human face recognition, multi-hand tracking, hair segmentation, object detection, and tracking. MediaPipe Pose can estimate a two-class, full-scale segmentation mask (person or background).

Google's BlazePose (full body) pose recognition model calculates the skeleton's 33 main points (x, y, and z). BlazePose has a detector and estimator machine learning models. The detector removes the human image from the input image, and the estimator recovers key points from the 256x256 human image.

The main components of BlazePose GHUM are a Convolutional Neural Network (CNN) - based model and tracking and analysis algorithms. The model is trained with millions of frames of human motion data to accurately predict and track human motion. Tracking and analysis algorithms use this data to create motion visualization and motion analysis in various ways. It includes calculating the speed, direction, and path of movement and tracking individual body parts. It can also detect subtle differences in movement, which helps analyze more complex activities such as dance and martial arts.

The system was tested on an Intel Core i5 6th gen, 8GB DDR3 Ram, 1TB HDD, AMD Radeon Graph CS 4GB Graphic Card, Windows 10 Operating System etc.

Figure 1 and 2 shows human pose estimation process for pre-recorded video.

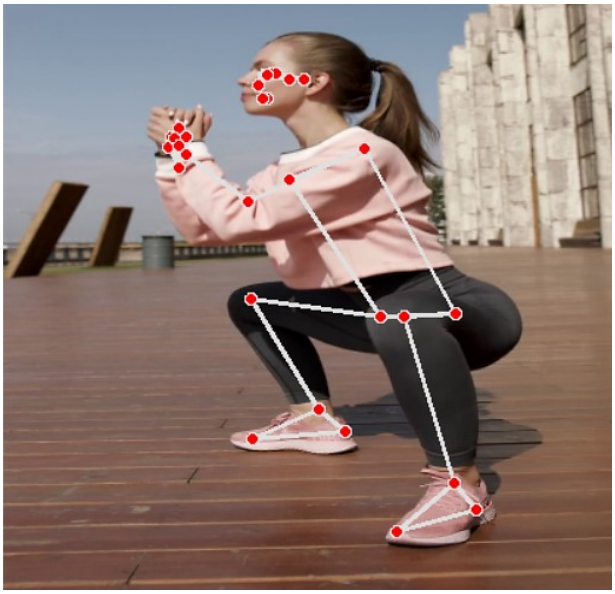

Fig 1. Human Pose Estimation for Pre-Recorded Video

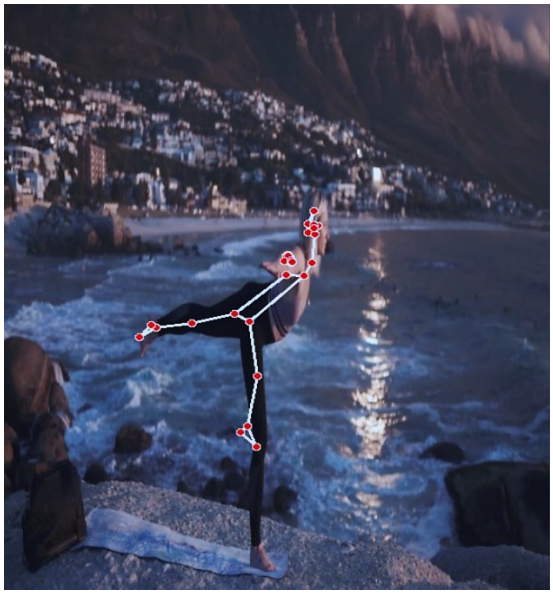

Fig 2. Human Pose Estimation for Pre-Recorded Video

In [1] propose a new method for 3D human posture estimation. Solve the 3D human assembly location problem in two steps. First find the 2D common point coordinates, then map the 2D common point coordinates to the 3D common point coordinates. A 2D to 3D mapping system can generate 3D composite coordinates using only 2D composite data. This method is simpler than the algorithm that uses several types of movies and images, and experimental tests show that the algorithm does better.

In [2] by combining color and depth images, it proposes a new way to assess a person's 3D state. There are three major advancements. (1) Utilize an RGBD camera to estimate the 2D location of a person in a color image, and use a depth image to fill in the missing depth information. (2) Improving human accuracy by utilizing Fast-RCNN as a human detector and Resnet50 to extract human characteristics. Feature detection (3) matches picture color and depth utilizing sparse feature point matching to precisely map

color spots to their associated depth image. The experimental findings demonstrate that our two techniques can correctly estimate the 3D human position.

In [3] method is considered for video input. The approximate steps of a video frame in the test are 1. Person detection, 2. Person-centered image cropping, and 3. Sending the person image to a pose estimation network. We use another object detector to identify the individual bounding box from the input frame. Pose estimation is done individually for each person. Location of body key points in mesh output as a result of pose estimation .

CONCLUSION

This research study discussed about human pose estimation and the importance of computer vision in human pose estimation, and implemented a single & multiperson human pose estimation using MediaPipe and the OpenCV library. We got the accuracy in terms of visibility points at **0.969%**, and it can be affected by factors such as image or video resolution, the number of people in the scene, and the presence of occlusions. This task is important in computer vision and has many applications in virtual reality, robotics, and human-computer interaction. Despite these challenges, significant progress has been made in recent years with the development of powerful deep learning models that can accurately estimate human posture. Single and multi-person pose estimation is an important problem in computer vision and has many applications, including virtual reality, robotics, and human-computer interaction.

Although human position estimation with OpenCV and Mediapipe is an interesting research area, it also has its limitations. The accuracy of the model depends on the quality of the input video and the specific pose being analyzed. Additionally, models may struggle to identify poses in difficult lighting conditions or when the subject is wearing clothing that hides important areas.

Finally, this research study concludes that, human pose estimation using OpenCV and Mediapipe is a valuable tool for analyzing and understanding human behavior. However, it is important to consider their limitations and use them in conjunction with other methods and techniques for a more complete analysis of human behavior.

REFERENCES

[1] Y. Zhang, S. Li, and P. Long, "3D human pose estimation in motion based on multi-stage regression," Displays, vol. 69, no. August, 2021, doi: 10.1016/j.displa.2021.102067.

[2] H. Tang, Q. Wang, and H. Chen, "Research on 3D human pose estimation using RGBD camera," *ICEIEC 2019 - Proc. 2019 IEEE 9th Int. Conf. Electron. Inf. Emerg. Commun.*, pp. 538–541, 2019, doi: 10.1109/ICEIEC.2019.8784591.

[3] M. Linna, J. Kannala, and E. Rahtu, "Real-time human pose estimation with convolutional neural networks," VISIGRAPP 2018 - Proc. 13th Int. Jt. Conf. Comput. Vision, Imaging Comput. Graph. Theory Appl., vol. 5, pp. 335–342, 2018, doi: 10.5220/0006624403350342.

[4] M. Ben Gamra and M. A. Akhloufi, "A review of deep learning techniques for 2D and 3D human pose estimation," *Image Vis. Comput.*, vol. 114, 2021, doi: 10.1016/j.imavis.2021.104282.

[5] A. Rohan, M. Rabah, T. Hosny, and S. H. Kim, "Human pose estimation-based real-time gait analysis using convolutional neural network," *IEEE Access*, vol. 8, pp. 191542–191550, 2020, doi: 10.1109/ACCESS.2020.3030086.

[6] Y. Cheng, P. Yi, R. Liu, J. Dong, D. Zhou and Q. Zhang, "Human-robot Interaction Method Combining Human Pose Estimation and Motion Intention Recognition," 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Dalian, China, 2021, pp. 958-963, doi: 10.1109/CSCWD49262.2021.9437772.

[7] C. Wang, F. Zhang, X. Zhu, and S. S. Ge, "Low-resolution human pose estimation," *Pattern Recognit.*, vol. 126, p. 108579, 2022, doi: 10.1016/j.patcog.2022.108579.

[8] J. Wang *et al.*, "Deep 3D human pose estimation: A review," *Comput. Vis. Image Underst.*, vol. 210, no. August 2020, 2021, doi: 10.1016/j.cviu.2021.103225.

[9] D. C. Luvizon, D. Picard, and H. Tabia, "Multi-Task Deep Learning for Real-Time 3D Human Pose Estimation and Action Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2752–2764, 2021, doi: 10.1109/TPAMI.2020.2976014.

[10] I. Chang, M. Park, J. Kim, and J. H. Yoon, "Multi- View 3D Human Pose Estimation with Self-Supervise d L e a rn in g," pp. 255–257, 2021.

[11] M. Palermo, S. Moccia, L. Migliorelli, E. Frontoni, and C. P. Santos, "Real-time human pose estimation on a smart walker using convolutional neural networks," Expert Syst. Appl., vol. 184, no. February, 2021, doi: 10.1016/j.eswa.2021.115498.

[12] W. Chen, S. Wang, and J. Wang, "Realtime Multi-Person Pose Estimation Based on Android System," *Proc. - 2020 Int. Conf. Intell. Comput. Autom. Syst. ICICAS 2020*, pp. 286–289, 2020, doi: 10.1109/ICICAS51530.2020.00065.

[13] V. T. Hoang, V. D. Hoang, and K. H. Jo, "Realtime Multi-Person Pose Estimation with RCNN and Depthwise Separable Convolution," *Proc. - 2020 RIVF Int. Conf. Comput. Commun. Technol. RIVF 2020*, 2020, doi: 10.1109/RIVF48685.2020.9140731.

[14] Q. Chen, C. Zhang, W. Liu, D. Wang, and S. Jiao, "SHPD : Surveillance Human Pose Dataset And Performance Evaluation For Coarse-Grained Pose Estimation School of Electronic Information and Electrical Engineering , Shanghai Key Lab of Digital Media Processing and Transmission , Shanghai 200240 , China Corr," pp. 4088–4092, 2018.

[15] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang, and C. Yang, "The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation," *IEEE Access*, vol. 8, pp. 133330–133348, 2020, doi: 10.1109/ACCESS.2020.3010248.

[16] D. Mehta *et al.*, "XNect: Real-time Multi-Person 3D Motion Capture with a Single RGB Camera," *ACM Trans. Graph.*, vol. 39, no. 4, 2020, doi: 10.1145/3386569.3392410.