ROOT BASED 3D RECONSTRUCTION OF MULTIPLE HUMAN POSE

Ayushya Rao[1] and Sumer Raravikar[2]

[1]Makers Lab, Tech Mahindra
[2]Rajiv Gandhi Infotech Park, Hijewadi, Pune, Maharashtra, India

ABSTRACT

This paper explores the current landscape of 3D pose estimation methods, pivotal in virtual reality, computer-aided design, and motion capture. Focusing on transforming estimated 3D poses for virtual environments, the emphasis lies in converting pose coordinates to align with virtual avatars. A novel pipeline is proposed, converting 2D pose images into 3D humanoids in the virtual realm. Evaluation metrics include accuracy, speed, and scalability, comparing techniques to state-of-the-art methods. The paper aims to summarize findings, showcasing the potential of proposed techniques to advance 3D pose estimation in virtual environments. It serves as a valuable resource for researchers, developers, and practitioners in computer vision, AI, and virtual reality by providing a comprehensive review and experimental evaluation of 3D pose estimation and representation techniques.

KEYWORDS

Pose Tracking, Instance Segmentation, Bio vision format, Extraction.

1. INTRODUCTION

The convergence of 3D Motion Capture (Mocap) technology with immersive 3D environments marks a transformative phase in human-computer interaction. This project is dedicated to exploring the dynamic realm of 3D Multi Person Pose Estimation and the integration of Mocap data within virtual worlds. Our goal is to unravel the possibilities and challenges inherent in this groundbreaking fusion, shedding light on its intricacies.

As we delve into the project, a particular focus lies on innovative human-computer interaction. We aim to investigate how the amalgamation of 3D Multi Person Pose Estimation and Mocap data goes beyond traditional interaction paradigms, opening new avenues for user engagement and responsiveness.

Furthermore, our project introduces a cost-effective alternative to traditional motion capture expenditures. Instead of relying on expensive Motion Capture Cameras, our approach leverages readily available RGB cameras, which can be as low as $50 to $100, minimizing costs while maintaining robust performance. The need for specialized Motion Capture Suits or Markers is circumvented through advanced computer vision algorithms that extract pose information directly from the RGB camera feed, eliminating the need for additional hardware.

In place of high-cost Optic Systems, our project employs streamlined software solutions, reducing the financial burden without compromising functionality. The computational demands are met by optimizing existing computer systems, avoiding the need for high-performance workstations that

David C. Wyld et al. (Eds): SEC, NLPML, CYBI, SIGL, CRIS, AIAPP, CoSIT- 2024
pp. 49-56, 2024. CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.140705

can range from $2,000 to $5,000 or more. Additionally, we explore innovative methods to utilize existing studio and location resources, ensuring a budget-friendly approach.

Comparatively, our project requires only a standard PC with a decent processor (typically $500 to $1,500) and a standard RGB webcam (ranging from $50 to $100), making it significantly more accessible from a financial standpoint. This streamlined approach not only provides a more cost-effective entry point for researchers and practitioners but also underscores our commitment to democratizing the integration of motion capture technology in virtual environments. As we navigate these cost-conscious strategies, our research continues to contribute to the evolution of human-computer interaction, ushering in new possibilities without the prohibitive financial constraints associated with traditional motion capture methodologies.

2. STUDIES AND FINDINGS

In our exploration of 3D multi person pose estimation and its integration into a 3D environment, we draw upon various innovative approaches and techniques in the realm of computer vision and human pose estimation. Here, we offer a summary of pertinent research initiatives. that have contributed to the development of methodologies in this domain.

2.1. Traditional 2D Human Pose Detection

2.1.1. Top-Down Approach

Techniques such as Histograms of Oriented Gradients (HOG) and Deformable Parts Models (DPM) were employed for initial human presence detection.

2.1.2. Bottom-UpApproach

Pioneering part-based models like pictorial structures and graphical models focused on Identifying distinct body parts and assembling them into complete poses. [3]

2.2. Evolutionto3DHumanPoseEstimation

Early techniques involved mapping 2D key points to 3D space using methods like stereo cameras and multi-view geometry to obtain more detailed information about human poses.

2.3. Inverse Kinematics

Inverse kinematics became pivotal for 3D pose estimation, with methods focusing on computing joint angles based on the desired positions of end-effectors. This approach found applications in robotics and motion capture.

2.4. Combined Approaches

With the advent of deep learning, a fusion of top-down and bottom-up approaches emerged:

2D Pose Estimation with CNNs: Convolutional Neural Networks (CNNs) excelled in 2D key point estimation. Real-time Multi-Person System: OpenPose implemented a bottom-up approach for detecting body parts and associating them with individuals.

2.4.1. 3D Pose Estimation

SMPL (SMPLify): The Skinned Multi-Person Linear model combined 2D joint detections with a parameterized 3D human shape and pose model.

2.5. Techniques for Pose Modification

Techniques emerged for modifying and recreating poses:

2.5.1. BVH (Biovision Hierarchy)

Originally developed for animation, BVH provided a file format for storing skeletal motion data and was used for pose modification and recreation, especially in computer graphics and virtual reality.

2.5.2. PoseGANs

Generative Adversarial Networks were employed to modify and generate realistic human poses.

2.6. Recent Advances

2.6.1. Graph Neural Networks(GNNs)

GNNs were applied to model spatial dependencies between body parts in a structured manner, significantly improving accuracy in pose estimation.
Self-Supervised Learning:

Recent research has explored self-supervised learning methods, where models learn from unlabeled data, reducing the reliance on large annotated datasets.

In our pursuit of enhancing 3D multiperson pose estimation within a 3D environment, we draw inspiration from a range of innovative methodologies in computer vision and human pose estimation. Leveraging insights from traditional 2D human pose detection, we consider both top-down approaches, employing techniques like Histograms of Oriented Gradients and Deformable Parts Models, and bottomup approaches, such as pictorial structures and graphical models. Evolution into 3D human pose estimation involves mapping 2D key points to 3D space using stereo cameras and multi-view geometry. The incorporation of inverse kinematics becomes crucial, allowing computation of joint angles based on desired end-effector positions. Combined approaches, particularly integrating deep learning with topdown and bottom-up strategies, have shown promise, exemplified by OpenPose's real-time multi-person system. Techniques like SMPLify, BVH, and PoseGANs contribute to pose modification and recreation, while recent advances in Graph Neural Networks and self-supervised learning offer structured spatial modeling and reduced reliance on annotated datasets, respectively. By synthesizing these diverse approaches, we aim to enrich our 3D multiperson pose estimation project with comprehensive and effective techniques.

3. METHODOLOGY

Our research employed a comprehensive methodology that encompassed several key steps and utilized various cutting-edge tools and techniques. This section outlines the methodology we followed:

3.1. Instance Segmentation with YOLOv7

We initiated our research by utilizing YOLOv7, a state-of-the-art real-time object detection system renowned for its exceptional speed and accuracy. YOLOv7 excels in detecting objects at frame rates ranging from 5 FPS to 160 FPS. Trained anew using the MS COCO dataset. YOLOv7 offers superior performance without relying on pre-trained weights. We used YOLOv7 to perform instance segmentation and extract individual instances from the input video frames.

3.2. Clustering with K-Means Algorithm

Following the instance segmentation, we encountered a need to organize the extracted frames into individual clusters based on similarity. To accomplish this, we applied the K-means clustering algorithm. This algorithm enabled us to group frames with similar characteristics, facilitating the subsequent analysis and processing of individual instances. [12]

3.3. UtilizingVGG16 Convolutional NeuralNetwork

For the clustering process, we leveraged the VGG16 Convolutional Neural Network (CNN). VGG-16 is a deep neural network architecture with 16 layers that excels in feature extraction from images. We employed VGG16 to compute image similarities, a critical step in the K-means clustering process. This allowed us to group frames effectively based on their visual content. [11]

3.4. Video Compilation with FFmpeg

With frames divided into distinct clusters, we employed FFmpeg, a powerful multimedia framework, to compile individual video sequences from these cluster folders. FFmpeg's capabilities in handling video manipulation, encoding, and decoding ensured the creation of separate videos for each cluster of frames.

3.5. Generating BVH Files using MocapNet

To obtain accurate 3D human body pose estimations, we utilized MocapNet, a neural network ensemble designed for this specific task. MocapNet derives 3D Bio Vision Hierarchy (BVH) skeletons from estimated 2D human body joint projections, making it directly compatible with various 3D graphics engines. The BVH format, originally developed by Biovision for motion capture data, provides the necessary skeletal hierarchy information for animation. MocapNet's ability to estimate 3D poses from RGB images allowed us to generate BVH files for each of the individual videos. [1]

3.6. Integration into Blender

The BVH files obtained from MocapNet served as a crucial bridge between our 3D pose estimations and their representation in a virtual 3D environment. These BVH files were directly imported into Blender and were transformed into realistic humanoid skeletons, complete with animations. Additionally, we fine tuned the visual realism of these representations by applying textures, shaders, and lighting effects.

Through adherence to this thorough methodology, we successfully seamlessly transition from 3D pose estimation to the recreation of these poses in a virtual 3D environment. Each step in the process, from instance segmentation with YOLOv7 to the final integration into Blender, played a role in the achievement of our research outcomes 3D multi person pose estimation and its visualization within a dynamic 3D space. [8]

length in 10 pt. Times New Roman italics. The text must be fully justified, with a 12 pt. paragraph spacing following the last line.

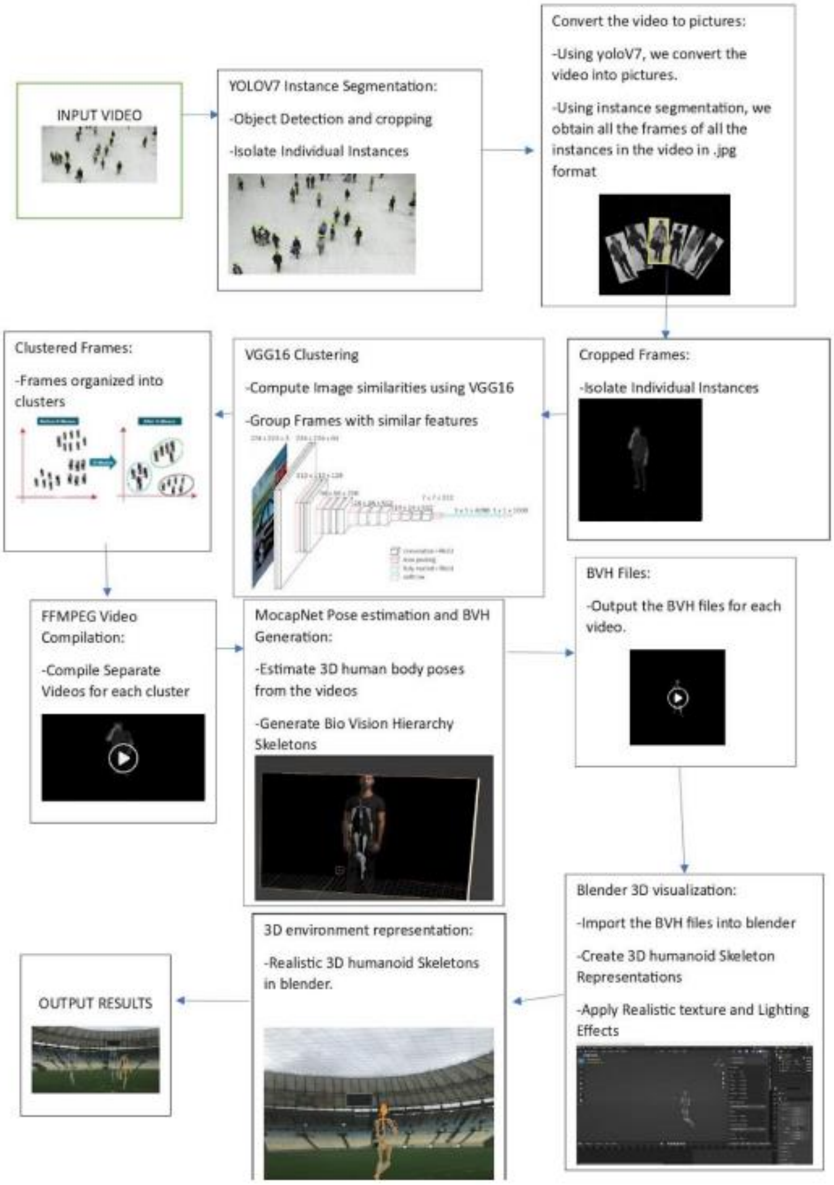

INPUT VIDEO
YOLOV7 Instance Segmentation:
-Object Detection and cropping
-Isolate Individual Instances
Convert the video to pictures:
-Using yoloV7, we convert the video into pictures.
-Using instance segmentation, we obtain all the frames of all the instances in the video in .jpg format
Cropped Frames:
-Isolate Individual Instances
VGG16 Clustering
-Compute Image similarities using VGG16
-Group Frames with similar features
Clustered Frames:
-Frames organized into clusters
FFMPEG Video Compilation:
-Compile Separate Videos for each cluster
MocapNet Pose estimation and BVH Generation:
-Estimate 3D human body poses from the videos
-Generate Bio Vision Hierarchy Skeletons
BVH Files:
-Output the BVH files for each video.
Blender 3D visualization:
-Import the BVH files into blender
-Create 3D humanoid Skeleton Representations
-Apply Realistic texture and Lighting Effects
3D environment representation:
-Realistic 3D humanoid Skeletons in blender.
OUTPUT RESULTS

Fig. 1: Block Diagram of the pipeline of the mode

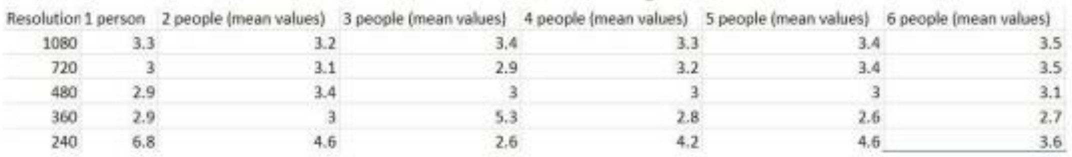
4. RESULTS

The outcomes presented in this context are derived from an exploration influenced by the research detailed in "MocapNET: Ensemble of SNN Encoders for 3D Human Pose Estimation in RGB Images." Our quantitative experiments, conducted on the H36M dataset, underscore the commendable accuracy of MocapNET. Notably, the method, trained exclusively on the CMU dataset, exhibits robustness, particularly evident in its ability to navigate diverse scenarios. This adaptability is evaluated through mean per joint position error (MPJPE) after Procrustes alignment.

In terms of evaluation protocols, our alignment with industry standards, specifically BP1 and BP2, consistently reveals accuracy metrics across various actions and viewing angles. While grounded in the H36M dataset, these results seamlessly bridge to the unique demands of our application in the mocap technology domain.

A distinctive characteristic of MocapNET lies in its focus on rotation regression, setting it apart from traditional 3D position regression. In comparison to state-of-the-art techniques, MocapNET emerges as a robust contender. The observed limitations, intricately tied to its specialized design and intended application, add nuance to its performance evaluation.

Table 1. MPJPE Evaluation Of output of the model



Resolution	1 person	2 people (mean values)	3 people (mean values)	4 people (mean values)	5 people (mean values)	6 people (mean values)
1080	3.3	3.2	3.4	3.3	3.4	3.5
720	3	3.1	2.9	3.2	3.4	3.5
480	2.9	3.4	3	3	3	3.1
360	2.9	3	5.3	2.8	2.6	2.7
240	6.8	4.6	2.6	4.2	4.6	3.6

Throughout our model evaluation encompassing one to six individuals, we observed a commendable level of robustness in pose detection. Despite a slight variation in MPJPE readings as the number of individuals increased, the overall consistency remained noteworthy. Notably, at lower resolutions such as 240P and 360P, we observed a marginally increased variation, attributed to the inherent distortion in video quality associated with reduced resolutions. However, even in these challenging scenarios, our model consistently delivered MPJPE readings within the confined range of 2.6 to 6.8 for each resolution, signifying its reliability across diverse settings. The observed resilience and accuracy, especially considering lower resolutions, highlight the model's adaptability and efficacy in handling scenarios involving multiple individuals, reinforcing its applicability to real-world applications demanding precise and consistent human pose detection across varying complexities.

5. CONCLUSIONS

We proposed a model which is able to perform 3-Dimensional positional estimation on multiple people and represent these 3D poses in a virtual environment. In this pipeline we decided to go with a different approach since there were no other models which created multiple human representations in a virtual environment

To attain this objective, we incorporated an instance segmentation layer into existing single-person pose estimation algorithms. This enhancement enables our model to execute 3D pose estimation for multiple individuals.

Additionally, while most of the current systems limit themselves to static poses, our model is able to incorporate dynamic poses with movement around the root base.

We are using BioVision Hierarchy(bvh) files for representation in 3D environments. Our model is the first model which directly obtains the 3D poses as an output from the video input.

While the pipeline was mainly designed for sports scene analysis, we also foresee diverse potential uses in other fields such as Crime scene analysis, Animation, game scene reconstruction, Crowd control, etc.

We can also look into certain advancements in our methodology such as implementing an algorithm to estimate the exact positions of the objects relative to other objects on the flat plane of the ground in order to generate an even more accurate representation of the captured motion.

Therefore, we can conclude by saying that Sports & Recreation can generate poses in a virtual environment. It is a versatile platform capable of performing 3D reconstruction that is root based and for multiple people. This technology has diverse applications within the realms of motion capture and presents opportunities for further refinement and optimization. We hope that this work might spark a future interest in the field of motion capture without the use of any additional equipment.

ACKNOWLEDGEMENTS

We would like to express my sincere gratitude to the Tech Mahindra Makers Lab for their invaluable support and resources throughout the course of this research. The collaborative environment and access to cutting-edge facilities provided by the organization have greatly contributed to the success of this project. I appreciate the commitment to innovation and the conducive atmosphere that the Tech Mahindra Makers Lab has fostered, enabling us to achieve meaningful results in our research endeavours.

REFERENCE

[1] Qammaz, Ammar, and Antonis A. Argyros. "MocapNET: Ensemble of SNN Encoders for 3D Human Pose Estimation in RGB Images." BMVC. 2019.

[2] Sárándi, István, et al. "Metrabs: metric-scale truncation-robust heat maps for absolute 3d human pose estimation." IEEE Transactions on Biometrics, Behavior, and Identity Science 3.1 (2020): 16-30.

[3] Dabral, Rishabh, et al. "Multi-person 3d human pose estimation from monocular images." 2019 international conference on 3D vision (3DV). IEEE, 2019.

[4] Yanjun Ma,Dianhai Yu,Tian Wu,Haifeng Wang. PaddlePaddle: An Open-Source Deep Learning Platform from Industrial Practice[J]. Frontiers of Data and Computing, 2019, 1(1): 105-115.

[5] Sengupta, Arindam, et al. "mm-Pose: Real-time human skeletal posture estimation using mmWave radars and CNNs." IEEE Sensors Journal 20.17 (2020): 10032-10044.

[6] Nakano, Nobuyasu, et al. "Evaluation of 3D markerless motion capture accuracy using OpenPose with multiple video cameras." Frontiers in sports and active living 2 (2020): 50.

[7] Kendall, Alex, Matthew Grimes, and Roberto Cipolla. "Posenet: A convolutional network for realtime 6-dof camera relocalization." Proceedings of the IEEE international conference on computer vision. 2015.

[8] Li, Jia, Wen Su, and Zengfu Wang. "Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 07. 2020.

[9] Cheng, Bowen, et al. "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2020.

[10] Scataglini, Sofia, et al. "Moving statistical body shape models using blender." Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume V: Human Simulation and Virtual Environments, Work With Computing Systems (WWCS), Process Control 20. Springer International Publishing, 2019.

[11] Qassim, Hussam, Abhishek Verma, and David Feinzimer. "Compressed residual-VGG16 CNN model for big data places image recognition." 2018 IEEE 8th annual computing and communication workshop and conference (CCWC). IEEE, 2018.

[12] Sinaga, Kristina P., and Miin-Shen Yang. "Unsupervised K-means clustering algorithm." IEEE access 8 (2020): 80716-80727.

AUTHORS

Ayushya Rao, Software developer with half a decade worth of experience in extended reality development and vision AI.

Sumer Raravikar, Final-year Electronics and Communication Engineering student specializing in Data Science. Passionate about machine learning and data science, dedicated to leveraging technology for innovative solutions.