3D Human Motion Prediction Based Deep Learning

Nivedita Wani[1] and Shailesh Bendale[2]
Student, Department of Computer Engineering[1]
Professor and Head, Department of Computer Engineering[2]
NBN Sinhgad School of Engineering, Pune Maharashtra, India

Abstract: *A growing number of established methods utilising deep neural networks to handle the problem of human motion prediction have been developed as a result of deep learning's success in a wide range of computer vision and computer graphics jobs. Modern motion prediction techniques concentrate on resolving numerous problems to forecast accurately and regular human movement across time. We give a thorough analysis of deep learning-based human motion prediction techniques in this paper. The goal of this work and the human motion prediction challenge are first defined. On the basis of our suggested classification, we then present pertinent background information and a thorough list of motion prediction techniques. Next, we give a thorough analysis of the traits frequently employed in the literature and describe the evaluation procedures. We concludedby providing a quantitative comparison.*

Keywords: Action Prediction, Path Integral, Future Motion, Deep Learning

I. INTRODUCTION

Future human motion prediction by machines is advantageous since it opens up a wide range of potential applications. A perfect human motion prediction model can be used to further enhance the animation of human characters in virtual environments and video games. More crucially, knowing how people behave when carrying out specific behaviours is necessary for many computer vision tasks in order to create a prediction model that is equivalent to human behaviour. Applications for this information include sports, surveillance, robots, autonomous vehicles, and user-friendly interfaces. The area of autonomous driving is one example of a field where it might be applied. Although this sector has bright future prospects, there are risks because conducting car tests has been linked to multiple deadly incidents.

If cars in urban areas could forecast the movement of people on the sidewalk, accidents like these would have been prevented. An accident with the autonomous vehicle may occur if a pedestrian decided to cross the street abruptly without scanning for oncoming traffic. If the autonomous vehicle had a competent human motion prediction system, the accident might have been prevented since the vehicle would have slowed down or attempted to avoid hitting the pedestrian.

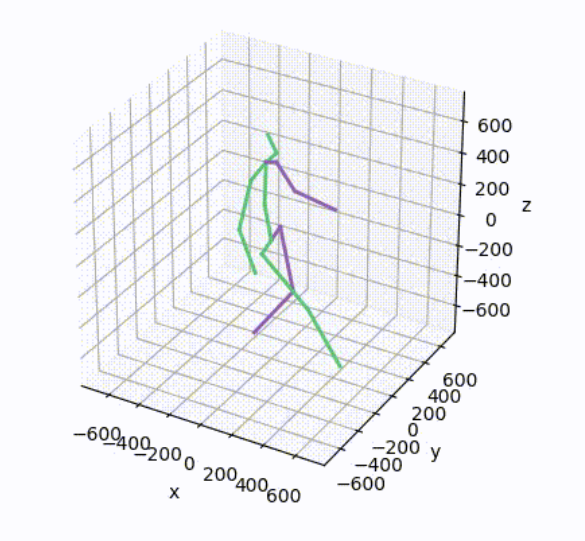
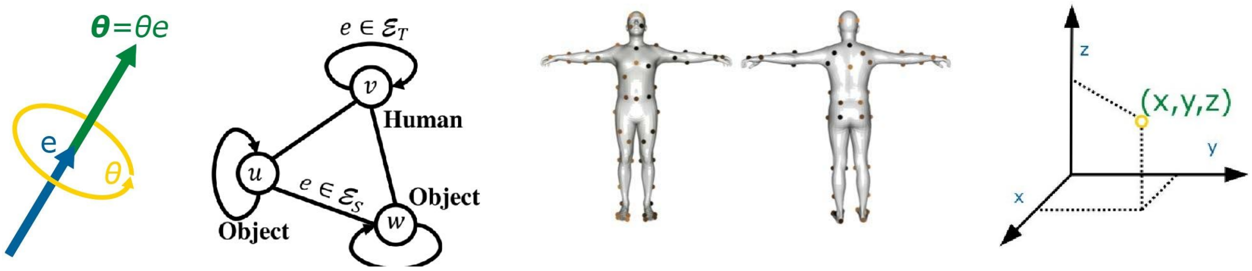


Figure 1: representation of human movement

II. HUMAN BODY REPRESENTATION

Some terms linked to the 3D body are used in the literature on human motion prediction and are common to related subjects. For instance, the complete joint that conforms to one human body or position throughout each time step is referred to as a human body or pose. Each pose is made up of a number of interconnected joints that together build a model that resembles the human skeleton. The word "joint" designates a particular joint's position or angle within the human skeleton. Because it depends on the motion sequence dataset utilised, there is no predefined number of joints in a posture.

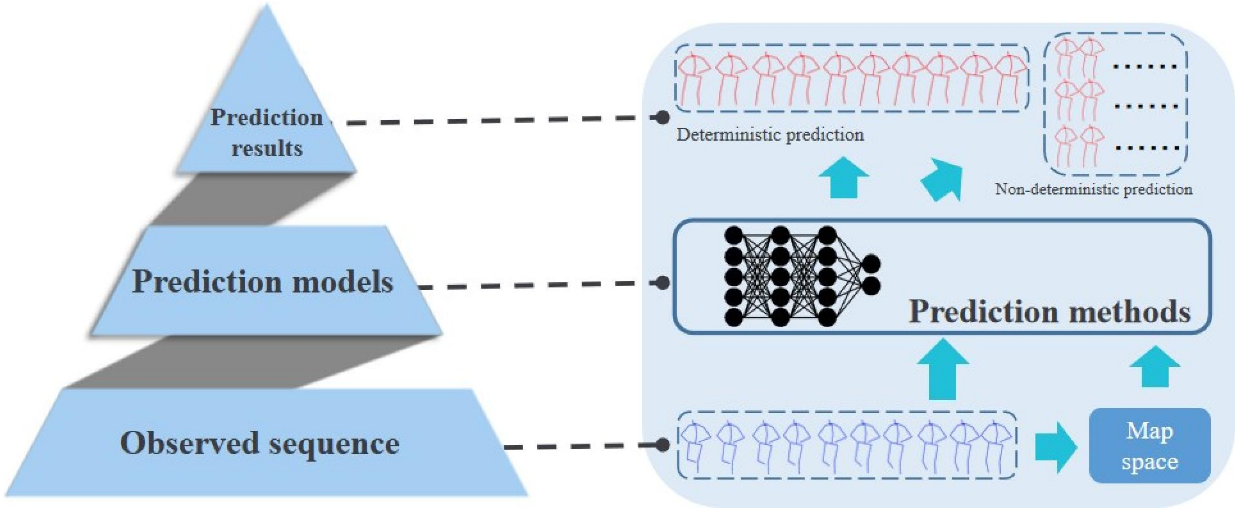
The joint representation employed in many human-motion prediction networks comes in a few different forms. An angle- or location-based joint representation is the norm for poses.



(a).Axis Angle (b).ST Graph (c).Marker Location (d).3D co-ordinates

Figure 2:Some of the joint types utilised in the literature are explained visually.

The axis angle representation of the joints, or more particularly, the exponential map, is referred to as the angle-based/joint angle (Figure 2(a)). The location of the joint is defined by the location-based joint representation in 3D Cartesian coordinates (Figure 2(d)).



III. DATASET

They are able to employ an action sequence dataset that has some sort of 3D body annotation because the deep human motion prediction algorithm is self-supervised. The majority of mo-cap datasets are employed because they meet the criteria needed to train a deep human motion prediction network. In this part, popular datasets are discussed together with their specifics and recommended preprocessing settings for each dataset. All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

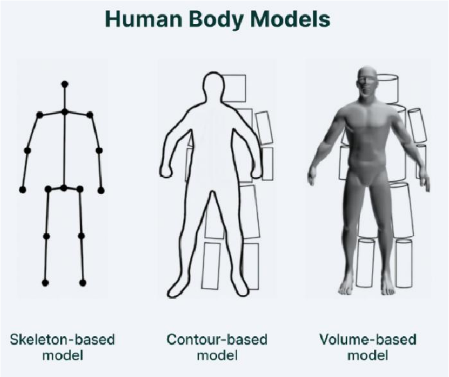
In many different domains, datasets constitute a crucial component of algorithm research. They often serve a key role in facilitating network learning and performance measurement as a common ground. As a result, more and more datasets are created to address the issues.

IV. HUMAN POSE ESTIMATION

A method for recognising and categorising the joints in the human body is called Human Pose Estimation (HPE).In essence, it's a technique for recording a set of coordinates for each joint (arm, head, torso, etc.), which is referred to as a key point that can characterise a person's position. A pair is the relationship between these points. Not all points can pair up since there must be a meaningful connection between them. The initial goal of HPE is to create a skeleton-like representation of the human body, which will subsequently be further processed for task-specific applications numerals. For example, see heading "III. Page Style" of this document. The two level-1 headings which must not be numbered are "Acknowledgment" and "References".

There are three types of approaches to model the human body :

1. Skeleton-Based Model
2. Contour-Based Model
3. Volume-Based Model



4.1 Human Pose Estimation using Deep Neural Networks

New difficulties arose as HPE's research and development programme began to grow. The multi-person pose estimation was one of them. DNNs are quite good at predicting a single human posture, however they struggle to estimate many human poses for the following reasons:

1. Several persons in various positions can be seen in one photograph.
2. Computational complexity rises with population size as a result of increased interactivity.
3. A rise in real-time inference time is frequently caused by an increase in computing complexity.

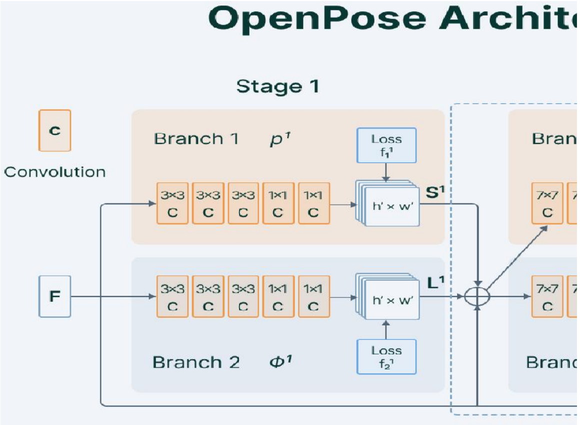
The researchers presented two strategies to address these issues:

1. Top-Down: Locate the people in the photograph or video, estimate their body parts, and then determine their poses.
2. Bottom -Up: Estimate the human body components in the image from the bottom up, then determine the position.

Let's look at deep learning models utilised for multi-human posture estimation right now.

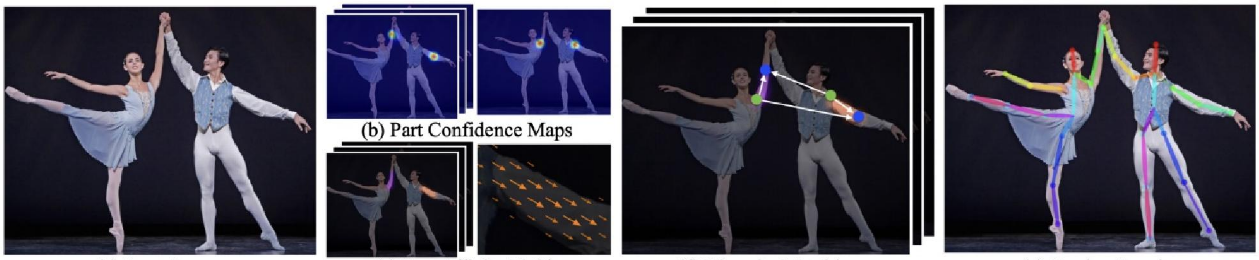
A. Open Pose

Zhe Cao et al. proposed OpenPose in 2019.The network initially recognises the body parts or key points in the image using a bottom-up approach, after which it maps the appropriate key points to form pairs.CNN is also used by OpenPose as its primary architecture. To extract patterns and representations from the input, it uses a VGG-19 convolutional network. Convolutional networks have two branches from which to take their output, the VGG-19.While the second branch predicts Component Affinity Fields (PAFs), which provide a degree of association between body parts, the first network predicts a collection of the confidence maps for each body part. Pruning the weaker links in the bipartite graphs is also helpful.



The architecture of OpenPose, a multi-stage CNN, is seen in the figure up top

According to the number of humans present in the input, the predictions from the two branches are essentially concatenated with the traits for the following stage to create a human skeleton. CNNs are employed in successive stages to improve the prediction.



(a) Input Image

(b) Part Confidence Maps

(c) Part Affinity Fields

(d) Bipartite Matching

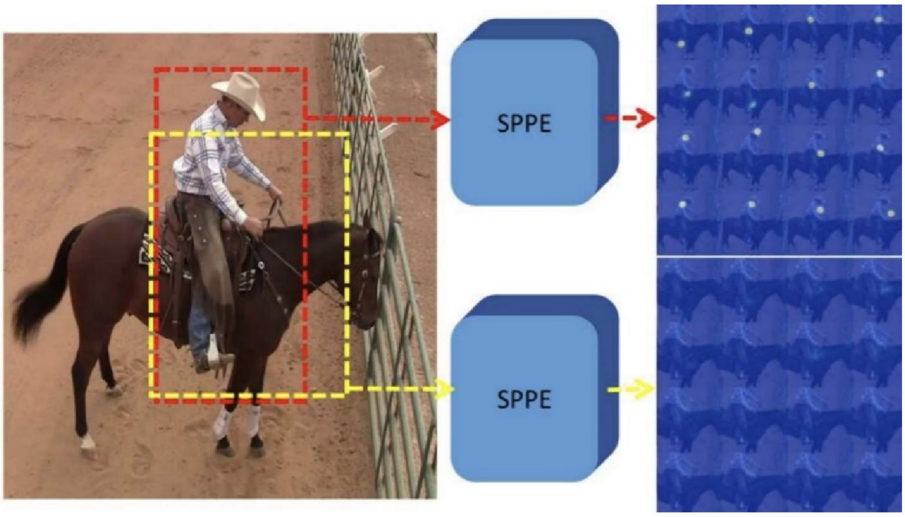
(e) Parsing Results

DOI: 10.48175/IJARSCT-7516

The OpenPose pipeline is shown in its entirety in the graphic above.

AlphaPose (RMPE)

HPE is implemented using a top-down strategy using Regional Multi-person Pose Estimation (RMPE) or AlphaPose. The top-down approach to HPE is quite difficult since it generates a lot of localization error and prediction inaccuracy.

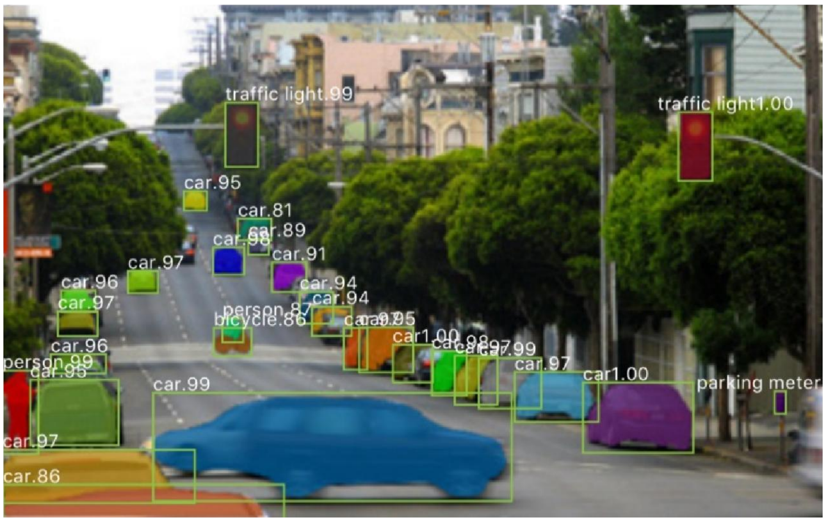


Deep Cut

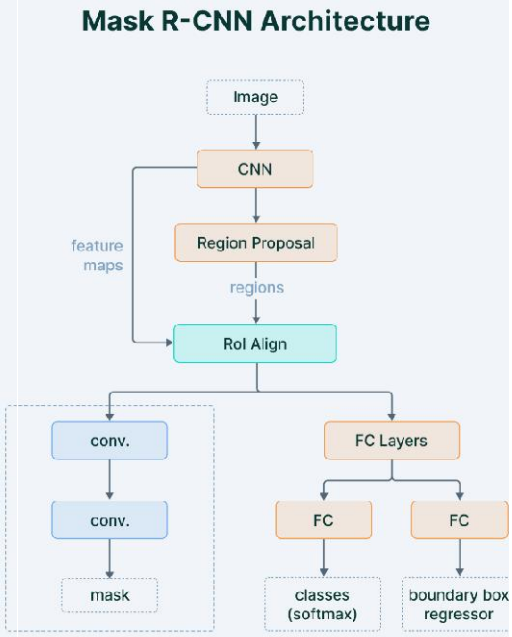
The goal of Leonid Pishchulin et al DeepCut .'s was to jointly solve the tasks of detection and pose estimation at the same time. In order to estimate human pose, a bottom-up method is used.The goal was to identify all potential body parts in the image, name them with terms like "head," "hands," "legs," and so on, and then separate the body parts that belonged to each person.

Mask R-CNN

A widely popular approach for segmentation is mask R-CNN. By drawing a bounding box around an object and making a segmentation mask, the model can localise and categorise that object at the same time.



For problems involving human pose estimation, the fundamental design is easily extensible.In order to extract features and representation from the input, Fast R-CNN makes use of CNN.Through a Region Proposal Network, the extracted features are then used to suggest potential locations for the object's presence (RPN).The extracted features are normalised using a layer called RoIAlign so that they are all of the same size because the bounding box can have different sizes, as seen in the image above.The network's parallel branches receive the extracted characteristics and use them to fine-tune the suggested region of interest (RoI) in order to produce segmentation masks and bounding boxes.

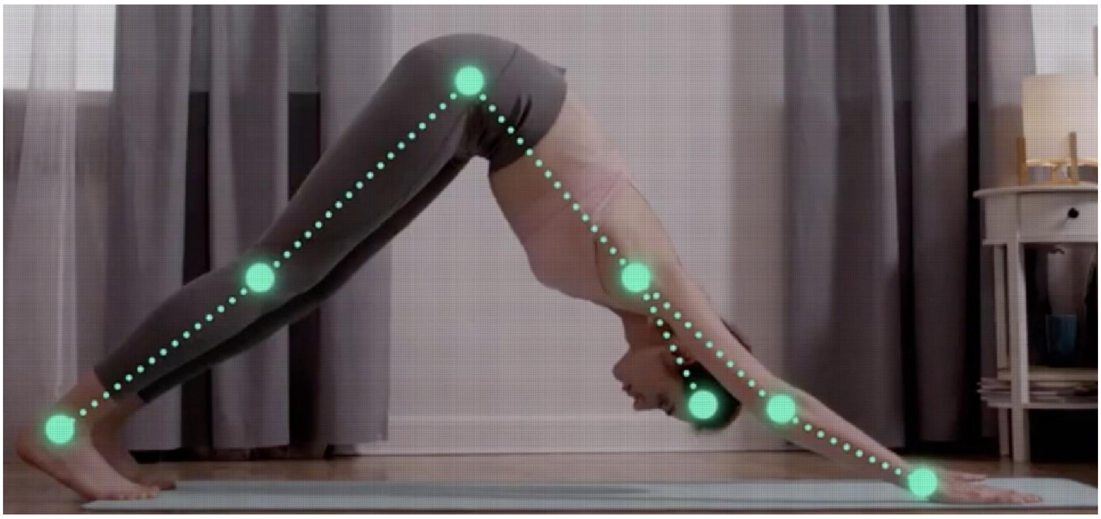


The mask segmentation output produced by the network can be used to identify humans in the input when it comes to human pose estimation. Mask segmentation, in this case person detection allows for very accurate object detection, making it possible to estimate the human position with relative ease. The person detection stage is carried out concurrently with the part detection stage in this method, which is similar to the top-down approach. In other words, the keypoint and person detection stages operate apart from one another

V. 7 HUMAN POSE ESTIMATION APPLICATIONS

7.1 AI-Powered Personal Trainers

These days, maintaining our physical health has become essential to living a fulfilling life, and working with a qualified trainer can help us get to the fitness level we want. It's not surprising that the industry has been oversaturated with apps that use AI to improve fitness.



For instance, the AI-powered yoga app Zenia leverages HPE to direct you toward adopting the ideal posture while doing yoga. Your stance is detected using the camera, and it calculates how accurate it is; if it is accurate, the anticipated pose will be shown in green, like in the image above. The red colour will take the place of the green one if the stance is incorrect.In addition to yoga, HPE has shown use in other types of physical activity. For instance, it is now frequently utilised in weight lifting, where it may instruct app users on how to conduct a proper weight-lift by looking for common errors and provide advice on how to correct them to avoid injury.

7.2 Robotics

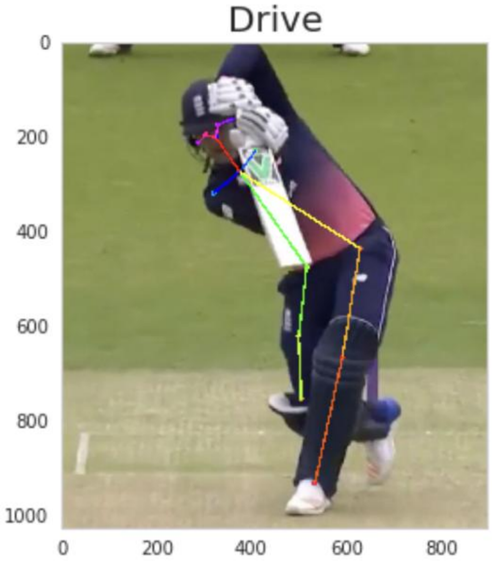
One of the development fields with the quickest growth has been robotics. Deep learning techniques can help, as programming a robot to follow a routine can be laborious and time-consuming. Robots can be successfully trained using methods like reinforcement learning, which employ a simulated environment to reach the accuracy level necessary to carry out a specific task.

7.3 Augmented Reality And Capture Motion:

CGI is yet another intriguing HPE application. The film industry, in particular, spends a lot of money on computer-generated visuals to produce special effects, enigmatic creatures, fantastical landscapes, and much more. Because it takes so much work such as wearing specialised clothing and masks to capture motion, producing surface-level effects in the approximated position, processing power, and significant time commitments on top of that—CGI is expensive.HPE can build a 3d rendering of a 2d input by automatically extracting important points, which may then be used to add effects, animations, and other things.

7.4 Detection of Athelete Pose

Nowadays, data analysis plays a significant role in practically all sports. Pose recognition can assist players in honing their technique and getting better outcomes. Pose detection also enables analysis and understanding of the strengths and weaknesses of the adversary, which is crucial for professional athletes and their coaches.



7.5 Motion Trackinng for Gamanaling

Pose estimation has another intriguing usage in in-game applications, where users can add poses to the gaming environment by using HPE's motion capturing technology. The objective is to develop an interactive game environment. For instance, Microsoft's Kinect tracks player movements and employs 3D pose estimation (using data from IR sensors) to render character actions digitally into the gaming world.

7.6 Infant Motion Analysis

The analysis of newborn movements can also be done using HPE. This is highly useful for examining the baby's behaviour as it develops, especially for determining the pace of its physical growth. When cerebral palsy, movement difficulties, or traumatic injuries are to blame, infants may be born with major health problems affecting their muscles, joints, and neurological system. Motion analysis can be used to pinpoint which muscles or joints need to be adjusted. Pose estimate can identify minute irregularities in the infant's movement, which the doctors can examine and treat appropriately. Additionally, HPE can be used as a tool for advice on how to help children develop their physical skills so they can reach their full potential.

DOI: 10.48175/IJARSCT-7516

7.7 Evaluation Metrices for Human Pose Estimation Model

To effectively learn the distribution during training and to perform well during inference, deep learning algorithms need the right evaluation metrics. Metrics used for evaluation depend on the current tasks.

The four evaluation metrics needed by HPE are covered in this section.

A. Percentage of Correct Parts (PCP)

PCP is used to gauge accurate limb detection. The limb is considered detected if the distance between the two anticipated joint positions and the genuine limb joint locations is almost less than half of the limb length. But occasionally it penalises shorter limbs, like a lower arm.

B. Percentage of Detected Joints (PDJ)

A new metric was suggested as a solution to the PCP problem. The percentage of identified joints is a measurement of the distance between the expected and the actual joint within a particular fraction of the torso diameter (PDJ).Since all joints' detection criteria for PCP are based on the same distance threshold, PDJ aids in localization precision, which mitigates its disadvantage.

C. Percentage of Correct Key-Points (PCK)

PCK is used as an accuracy metric to determine whether the projected keypoint and the real joint are close enough to one another. The bounding box's topic, which is scaled according to the PCK, is typically taken into consideration.

The threshold could be:

- PCKh@0.5 is when the threshold = 50% of the head bone link
- PCK@0.2 = Distance between predicted and true joint < 0.2 * torso diameter
- Sometimes 150 mm is taken as the threshold.
- It alleviates the shorter limb problem since shorter limbs have smaller torsos and head bone links.
- PCK is used for 2D and 3D (PCK3D)

D. Object Keypoint Similarity (OKS) based mAP

OKS is commonly used in the COCO keypoint challenge as an evaluation metric.

It is defined as:

$$\frac{\sum_i \exp\left(-d_i^2/2s^2k_i^2\right)\delta\left(v_i > 0\right)}{\sum_i \delta\left(v_i > 0\right)}$$

Where, di is the euclidean distance between the ground truth and predicted keypoint
s is the square root of the object segment area
k is the per-keypoint constant that controls fall off.
vi is considered to be a visibility flag that can be 0, 1 or 2 for not labeled, labeled but not visible and visible and labeled respectively.

Because OKS is used to calculate the distance (0-1), it shows how close a predicted keypoint is to the true keypoint.

VI. CONCLUSION

We review the most current developments in human motion prediction in this article. We noticed that there has been a considerable increase in interest in this subject each year an advancement over past works. Human motion prediction is still a difficult topic to solve, and most studies have struggled to come up with reliable outcomes for long-term prediction while excelling at short-term prediction. The availability of datasets is also a drawback because it is difficult to create new motion sequence datasets, and motion prediction techniques can only infer the distribution of action already in the dataset. We are still a long way from using motion prediction techniques in real-world applications due to these problems, error accumulation, and the tendency for long-term predictions to converge to the mean posture. However, we noticed a rise.

DOI: 10.48175/IJARSCT-7516

REFERENCES

[1]. Matthew Marchellus, (Graduate Student Member, IEEE) and In Kyu Park, (Senior Member, IEEE), .”Deep Learning for 3D Human Motion Prediction: State-of-the-Art and Future Trends”, Date of Publication: March 30, 2022.

[2]. KediLyu, Haipeng Chen, Zhenguang Liu, Beiqi Zhang, Ruili Wang, .”3D Human Motion Prediction : A Survey”.

[3]. Director: Francesc Moreno-Noguer, Codirector: AlejandroJos´e Hern´andez Ruiz, Advisor: Ren´eAlqu´ezar Mancho,

[4]. ”A Human Shape-Motion Predictor with Deep Learning”.

[5]. https://www.v7labs.com/blog/human-pose-estimation-guide#

DOI: 10.48175/IJARSCT-7516