IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

REVIEW OF LITERATURE SURVEY ON DIFFERENT HUMAN POSE ESTIMATION AND POSE COMPARISON TECHNIQUES

[1] Deep Gojariya, [2]Vatsal Shah, [3] Viraj Vaghasia,

[1,2,3]B. E Students, Department of Information Technology, D.J Sanghvi College of Engineering, Mumbai, India.

Abstract**:** Computer vision is a very good application of deep learning and is used to solve many problems in the vision category. One of its applications is Human pose estimation. Using pose estimation various applications can be made such as yoga training, physiotherapy training, also pose estimation can be used in sports analytics. Due to its wide range of application, it becomes very much important to learn how the things actually work and how the poses are estimated from a moving video or a still image. This paper consists of the literature-study related to how human pose estimation works and by which techniques one can perform the pose estimation along with that some pose comparison methods are also discussed in the paper which are required for developing the above-mentioned applications.

Index Terms: **Pose Estimation, Key-points, CNN (Convolutional Neural Network), FPS (Frames per Second), Computer Vision, Dynamic Time Warping, PAF (Part affinity fields).**

I. INTRODUCTION

Pose Estimation is a computer vision task that infers the pose of a person or an object in an image or a video. We can also think of pose estimation as the problem of determining the position orientation of a camera relative to a given person or an object. This is typically done by identifying, locating, and tracking a number of key-points on a given person. These key-points represent major joints like elbows, knees, shoulders, wrists, etc [1]. With pose estimation we're able to track a person in real-world space at an incredibly granular level. This powerful capability opens up a wide range of possible applications such as Augmented reality, animation, gaming, robotics, etc.

II. LITERATURE SURVEY

2.1 POSE ESTIMATION TECHNIQUES

Pose estimation can be performed using the various deep learning based pose estimation algorithms, they are mainly divided into two categories top-down and bottom-up pose estimation algorithms or approaches.

The top-down approach of human pose estimation is a very naive and traditional method. Given an image or a video of people it first detects where a person is present in that image and then draws a bounding box around it using object detection. After obtaining the bounding box it is fed to a pose estimator which then extracts the body key-points from that bounding box. This approach is very simple but has some drawbacks like the runtime is directly proportional to the number of people and high computational cost.

The bottom-up approach is exactly opposite to that of the top-down approach yet so powerful. It first draws the key-points on the image and then tries to map it with different people in that image using part affinity maps. This method is not only fast but also more accurate as compared to the top-down approach. All modern day pose estimation algorithms are inspired by this approach. We have studied some major bottom-up algorithms like DeepPose, Convolutional Pose Machines, Openpose, Posenet and Blazepose.

2.1.1 DEEPPOSE

DeepPose was the first major development research that applied deep learning to human pose estimation. In this approach pose estimation is formulated as a CNN-based regression problem towards body joints. They also use a cascade of such regressors to refine the pose estimates and get better estimates. The model consists of AlexNet as its baseline CNN model which has (7+1 extra final layer) that outputs 2k joint coordinates - (x_i , y_i) for i € {1,2,3,4,…..,k}, where k is the number of joints and this model was trained using L2 loss [2].

2.1.2 CPM (CONVOLUTION POSE MACHINES)

Convolutional Pose Machines uses the concept of pose machines. A pose machine consists of an image feature computation module followed by a prediction module. CPM are completely differentiable and their multi-stage architecture can be trained end-to-end. A CPM can consist of more than 2 stages and the number of stages is a hyperparameter. Stage 1 is generally fixed and stages > 2 are just repetitions of Stage 2 which takes heatmaps and image evidence as an input which it refines through subsequent stages [3].

2.1.3 OPENPOSE

OpenPose, developed by researchers at the Carnegie Mellon University can be considered as the state-of-the-art approach for real-time human pose estimation. It follows an architecture in which first an image is passed through a state of art CNN like VGG-16 or VGG-19 which will give multidimensional tensors as its output which are also referred as feature maps which are then passed through two branches of the Stage-1. The task of branch 1 is to output confidence maps for different body parts and the task of branch 2 is to output part affinity maps (PAF's). These PAF's are the direction vectors representing the directions of different body limbs. After Stage-1 the confidence maps, PAF's and the feature maps obtained from the CNN network are combined together and given as an input to Stage-2. This stage is the replica of stage-1 and is there to provide more refined output. In the architecture of Openpose we can have more than one stage-2 if we have a good amount of computing resources which will give us more correct predictions in the form of human pose [4].

2.1.4 POSENET

Another algorithm is Posenet, it can also be termed as a lighter version of Openpose because instead of VGG-16 or VGG-19 it uses a CNN model from the MobileNet family which is developed for achieving faster fps while live detection also it was mainly introduced for deploying deep learning applications on mobile phones and therefore it is lighter. Moving on to the architecture of Posenet it does not have a multi-stage architecture like that of Openpose, it has only one stage which calculates the confidence maps and offset vectors which are none other than PAF's and combines them to get the exact location of body key-points in the pose [5].

2.1.5 BLAZEPOSE

Blazepose is a lightweight CNN architecture for human pose estimation that was developed by Google that can compute (x,y,z) coordinates of 33 body key-points. Blazepose consists of two machine learning models: a Detector and an Estimator. The pose estimation is done with a two-step detector tracker ML pipeline. Using detector pose region-of-interest (ROI) is first detected. The tracker then predicts 33 pose critical points from the ROI. For video use cases, the detector is run only on the first frame. For subsequent frames we derive the ROI from the previous frame's pose key-points [6].

2.2 POSE COMPARISION TECHNIQUES

Pose estimation when combined with pose comparison can give rise to many applications. Pose comparison is a technique of comparing two different poses on the basis of how similar they are or how dissimilar they are. Pose comparison can be easily performed if we have the body key-points coordinates of both the poses. There are several techniques to perform comparison of poses. We studied some of them like superimposing poses, cosine similarity and dynamic time warping.

2.2.1 SUPER IMPOSING POSES

One of the ways to compare the poses is to superimpose one pose skeleton on another. This technique has some properties like a) lines map to lines, b) parallel lines remain parallel, c) Origin does not necessarily map to origin, d) body ratio is preserved. Also, this method will not work very well in cases where the two poses have a lot of dissimilarity in their body structure e.g.: pose 1 is of a short person and pose 2 is of a relatively tall person [7].

2.2.2 COSINE SIMILARITY

Second way for comparing poses is a method which uses cosine similarity. Previously we were comparing the length of the body parts but instead in this method comparison of angles between joints is done. The angles are independent of physical length and hence it would give us good results. For calculating joint angles the cosine triangle rule is used. But this method also has a little drawback which is that two different poses can also have the same joint angle. In order to overcome that drawback along with the joint angle we also check the coordinate position of the limbs which are connected to that joint. If for both the poses we get the same coordinates then we can say that both poses are matching [7].

2.2.3 DYNAMIC TIME WARPING

Dynamic time warping is a fast and an efficient algorithm for measuring similarity between two sequences of videos with different length. Similarity is measured by aligning two sequences and computing distance between them at each phase. It can handle sequences with different scale and translation and also it has less effect of noise and therefore it enhances the functionality of the applications that use it [8].

This was basically a brief overview of the techniques used for human pose estimation and pose comparison.

III. COMPARITIVE STUDY

We have compared the different bottom-up techniques for pose estimation on the basis of a few parameters like type of architecture, baseline CNN model, average accuracy and FPS achieved.

The algorithms/techniques compared in this study are DeepPose, Convolutional Pose Machines, Openpose and Posenet. All of these algorithms have a multi-stage architecture which have multiple stages and can have different baseline CNN models like Alexnet for DeepPose, custom CNN for CPM, a state-of-art CNN like VGG-16 or VGG-19 for Openpose and a relatively lighter CNN model like Mobilenet for Posenet. Due to the model configurations and its multi-stage architecture these techniques have different average accuracy. Openpose has the best recorded accuracy among the four techniques then comes CPM following CPM are Posenet and DeepPose respectively. Another factor of comparison can be the fps achieved while testing the techniques and it also depends on the hardware of the system and as we know that Posenet has Mobilenet CNN and therefore it achieves the highest fps then comes DeepPose due to its simpler architecture the fps achieved by CPM and Openpose were relatively less as compared to the other two techniques since Openpose requires heavy computational resources [2][3][4][5][6].

Table 1: Comparison of bottom-up techniques

Parameter	**DeepPose**	**CPM**	**OpenPose**	**PoseNet**	**BlazePose**
Multistage	Yes	Yes	Yes	No	Yes
Baseline CNN Model	AlexNet	Custom CNN	VGG-16/VGG 19	MobileNet	Single Shot Detector
Average Accuracy	0.61 [2]	0. 66 [3]	0.79 [4]	0.75 [5]	0.67 [6]
FPS	Good	Moderate	Bad	Very Good	Excellent

IV. CONCLUSION AND FUTURE SCOPE

In this paper we have reviewed the different pose estimation and pose comparison techniques. The paper focuses on two modules: pose estimation techniques and pose comparison methods. In the pose estimation module, we have discussed five different techniques i.e., DeepPose, CPM, Openpose, Posenet and Blazepose along with their working and in the next module we discussed the different methods used for performing pose comparison. After exploring all the techniques and methods we compared the pose estimation techniques on a few parameters and our attempt will be to develop an application which uses pose estimation and comparison which will teach children some basic dance moves, hence promoting their overall growth to a new level.

V. REFERENCES

[1] Derrick Mwiti, "A 2019 Guide to Human Pose Estimation," August 5,2019. https://heartbeat.comet.ml/a-2019-guide-to-human-pose-estimation-c10b79b64b73

[2] A. Toshev and C. Szegedy, "DeepPose: Human Pose Estimation via Deep Neural Networks," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1653-1660, doi: 10.1109/CVPR.2014.214.

[3] S.-E. Wei, V. Ramakrishna, and T. K. and Yaser Sheikh, "Convolutional pose machines. " In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[4] Z. Cao, G. Hidalgo, T. Simon, S. -E. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172-186, 1 Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[5] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson and Kevin Murphy, "PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model," arXiv:1803.08225v1 [cs.CV] 22 Mar 2018.

[6] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang and Matthias Grundmann, "BlazePose: On-device Real-time Body Pose tracking," arXiv:2006.10204v1 [cs.CV] 17 June 2020.

[7] Pradnya Krishnanath Borkar , Marilyn Mathew Pulinthitha , Mrs. Ashwini Pansare, 2019, Match Pose – A System for Comparing Poses, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 08, Issue 10 (October 2019).

[8] Stan Salvador, and Philip Chan. “FastDTW: Toward accurate dynamic time warping in linear time and space.” Intelligent Data Analysis 11.5 (2007): 561-580.