Human Pose Estimation

Flavio Amurrio-Moya, G00593001

Octobe 2021

1 Proposal

I will be planning on using Humam Pose Estimation to gather data on different type of human activities. Using this data, I would then build a Machine Learning Model and/or Deep Neural Network with the human pose data which would allow me to classify body positions. One possible application is to have a model that would detect when a student has their hand raised. This would turn on a light on the instructor's desk. This could be done for both in person and online courses. Another possible application is to detect human positions when playing games such as charades and have those positions be described to a visually impaired person.

2 Survey

2.1 Introduction

In recent years, the application of Human Pose Estimation has progressed exponentially in terms of speed and accuracy. From being able to detect simple 2D poses of one person at the time to real-time multi-person 3D Human Pose detection in videos, this rapid evolution is mainly thanks to the advances of Deep Neural Networks, which help with the detection of body part sections and joints within the human body. Human Pose Estimation originally began as a means of basic object detection before its broader application into healthcare, entertainment, and videography in general. This document will seek to explain Human Pose Estimation's history, it's conception and progression according to the latest research, as well as its shortcomings in current applications.

The origin of Pose Estimation was first hinted at by Martin A. Fischler and Robert A Elsachelager [5] in 1973 in the form of matching Pictorial Structures, where, given a description of an object, one could find the object in a picture. One of the applications mentioned on [5] was to perform some variance of facial detection. During this time, there was a strong need for object detection which prompted hastened research in Pictoral Structures. As technology advanced, more intricate and complex renditions of the human anatomy were required. This would later evolve beyond the scope of mere physical appearance to human behavior and movement as well, which will be further explained as we delve into the more technical aspect in the below survey.

2.2 Survey

To achieve a realistic depiction of the human form, Human Pose Estimation requires, in broad terms, two main steps - preprocessing and body parts parsing [8]. It also requires training deep learning models to predict the poses. We will further elaborate the method preprocessing in detail below.

Before the popular models and approaches to solving this problem can be explained, we must discuss that pose estimation falls under two different methodologies, starting with the top-down approach. In the top-down approach, human detection is done using CNNs in two steps. First, much like the object detection algorithms, the framework will first detect bounding boxes around the humans. Later it involves estimating the pose inside of the boxes. Estimating poses requires finding important coordinates on a human body. These would include key points, such as knees, elbows, etc. For example, in the much widely used datasets, such as the COCO Dataset [7], 17 coordinates are used. In other research papers, and datasets, different numbers of keypoints could be used. These coordinates are paired with an extra field, which indicates whether the coordinate is visible or not. There are often occlusions to the body parts and it is the neural network's job to infer the points regardless, and thus the training data must contain the points.

In the bottom-up, CNNs are also used to solve the detection problem. However, the approach will first estimate all the key points in an image and then go on to classify which key points are for a person [1]. This helps in both "performance and efficiency"because top-down will estimate poses per person and is a two step process. There are many open source models that resolve this problem, and we will discuss the most recognized and state of the art model OpenPose [1]. The methods described here fall under the bottom-up, as do most of the state of the art approaches such as Regional Multi-Person Pose Estimation (AlphaPose) [4] . These models have been optimised to that they could be used in browser and/or mobile without the use of a GPU.

2.3 Prepocessing

Preprocessing can be generously defined as the work that must be done in order to best optiomized the data that will be used for the Machine Learning Model. In this discipline in particular, it involves the initial division of the human body into several parts [9], taken with the exact measurements as expressed by a camera's calibration. The camera, in this instance, requires varying levels of perspective for it's beginning depiction, before it can be compiled through a process known as Data calibration [10]. A popular body capturing device used to gather data is Microsoft's Kinect - most likely due to its dominating presence in the market and low cost [3].

Another proponent of preprocessing involves body localization [8]. Body localization, as the name suggests, involves using a human detection model to recognize and locate a humanoid figure. Perhaps a robust example of this could be asking an Machine learning model which could distinguish a human body while it stands in front of a painted mosaic. This method then lends itself into the usage of Human detection, which involves a similar concept in the recognition of a person's location and space in a given image. Much like other vision algorithms such as object classification, segmentation, and object detection, deep learning has improved upon traditional methods for pose estimation.

Human detection is done using a CNN in two steps. First, much like the object detection algorithms, the framework will first detect bounding boxes around the humans. Later it involes estimating the pose inside of the boxes. Estimating poses require finding important coordinates on a human body. These would include key joints, such as knees, elbows, and wrists.

The final method of preprocessing involves Background Subtraction. Background Subtraction involves removal of data not needed. This can be achived by the usage of dense regions [6] where lighting and context change can be utilized to distinguish between what is in the foreground and what is in the background.

Note that the above processes are components of preprocessing, but are not inherent steps required in any particular order (at least according to the resources gathered). It is an educated estimation that while each method contains its strengths and weaknesses, a combination of the methods summarized above would lead to the greatest dataset needed to create and image. When attempting to replicate the human form, it is important to note that each process requires the same need for feature extraction [8].

The second, and most admittedly most recognized portion of HPE is body parts parsing [8]. This is the process of identifying each section of a person body. There are four categories for this process which include 2D and 3D single and multi-person parsing. Each one of these categories have specific methods for feature extraction, appereance model and structure models. The compilation of these models and tooling all work to create a holistic image.

Out of the four categories of body parts parsing, two approaches by Chen/Yuille [2] and Thompson [12] are perhaps the most popular, efficient and accurate. This can be surmised by the results of the subsequent benchmark performed by [8]. Results were strikingly similar and equally positive.

2.4 Conclusion and Final Remarks

Human Pose Estimation, like with all disciplines in technology, is a constantly evolving process - and along with each of it's accomplishments, lies the existence of controversy and failure. With respect to AI (and it's inferred infallibility when given the right dataset), we then must defer to controversies that are imposed mostly by human error, bias, and methodology. That is to say, that most of the controversy surrounding Human Pose Estimation is caused by the application of it's use, rather than by the technology itself. One striking topic that has arose since the emergence of HPE is fitness and the tools we use to improve it. The question then arises as to how we should cater to a more diverse group of users. As Maksym Tatriants [11] cite's, “fitness is on trend,”and HPE has entered the playground as a helpful tool in analyzing a user's posture during regular exercise. While the idea of this application is beneficial, it has been found to have several shortcomings in real world application.

Take Maksym Tatariants, author of ”Challenges of Human Pose Estimation in AI-Powered Fitness Apps” [11], and his experience practicing a simple squat using a fitness AI model as an example. In his experience, Tatariants remarks that before the exercise session can even begin, there are already shortcomings in its attempt to guide the user towards the correct posture. “Men and women’s bodies are physiologically different,” he explains. “If the model was trained only on men’s images, it will return accurate results for only male users but not for females.” This also brings into question the height and weight of dataset the AI was given versus the same measurements of the user performing the exercise. In this day and age, what is considered average height and weight? How broad of a range in this matter can be determined to be safe before the exercise can be performed correctly? These are popular media questions that, while they do may neccessarily contribute to the advancement of Human Pose Estimation, may very well effect it’s application and use in the future.

In spite of this, Human Pose Estimation and it’s achievements in all fields, not just fitness, will continue to push it forward. As the public becomes gradually more comfortable with the assistance of AI-rendered HPE in their everyday lives, the more unique and diverse datasets will become to provide a human assisted application that will benefit everyone.

References

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2021.

[2] Xianjie Chen and Alan Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. 2014.

[3] Jose Antonio Diego-Mas and Jorge Alcaide-Marzal. Using kinect™ sensor in observational methods for assessing postures at work. *Applied ergonomics*, 45(4):976–985, 2014.

[4] Haoshu Fang, Shuqin Xie, and Cewu Lu. RMPE: regional multi-person pose estimation. *CoRR*, abs/1612.00137, 2016.

[5] M.A Fischler and R.A Elschlager. The representation and matching of pictorial structures. *IEEE transactions on computers*, C-22(1):67–92, 1973.

[6] Constant Guillot, Maxime Taron, Patrick Sayd, Quoc Cuong Pham, Christophe Tilmant, and Jean-Marc Lavest. Background subtraction adapted to ptz cameras by keypoint density estimation. In *Proceedings of the British Machine Vision Conference*, pages 34.1–34.10. BMVA Press, 2010. doi:10.5244/C.24.34.

[7] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild, 2018.

[8] Zhao Liu, Jianke Zhu, Jiajun Bu, and Chun Chen. A survey of human pose estimation: The body parts parsing based methods. *Journal of visual communication and image representation*, 32:10–19, 2015.

[9] 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, 2018.

[10] C Stoll, N Hasler, J Gall, H Seidel, and C Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *2011 International Conference on Computer Vision*, pages 951–958. IEEE, 2011.

[11] Maksym Tatariants. Challenges of human pose estimation in ai-powered fitness apps, Oct 2020.

[12] Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. 2014.