Human Pose Estimation Using AI/Machine Learning Algorithms

Muhammed Sahal (*Author)*
Dept. of Computer Science and Engineering
Mangalam College of Engineering
Ettumanoor, India

Melvin M Abhraham (*Author)*
Dept. of Computer Science and Engineering
Mangalam College of Engineering
Ettumanoor, India

Nafcy N (*Author)*
Dept. of Computer Science and Engineering
Mangalam College of Engineering
Ettumanoor, India

Lijin Laiju (*Author*)
Dept. of Computer Science and Engineering
Mangalam College of Engineering
Ettumanoor, India

Ms.Sruthy K Joseph (Author)
Assistant Professor
Dept. of Computer Science and Engineering
Mangalam College of Engineering
Ettumanoor, India

Abstract **- Human pose estimation localizes body points to accurately identify individual poses given an image. This step is a prerequisite for many tasks in computer vision, including human recognition. This article contains an overview of the Human pose estimation techniques using machine learning and also proposed an AI-based system which can work as a personal fitness advisor. It is based on an algorithm that looks at your exercise chart in real time and tells you what's right and what's wrong! The methods used in human estimation are briefly described before listing some of the applications and problems encountered in estimation. Next, focus on briefly discussing the research that has had a major impact on human prediction, and examine each new model, motivation, architecture, process (policy work) and its advantages and disadvantages, the data used, and the methods used for evaluation. metrics for the model. This review serves as a foundation for novices and guides researchers to discover new trends by examining methods and architectural flaws in current research.**

Keywords— ***sklearn library, convolution neural network (CNN), machine learning.***

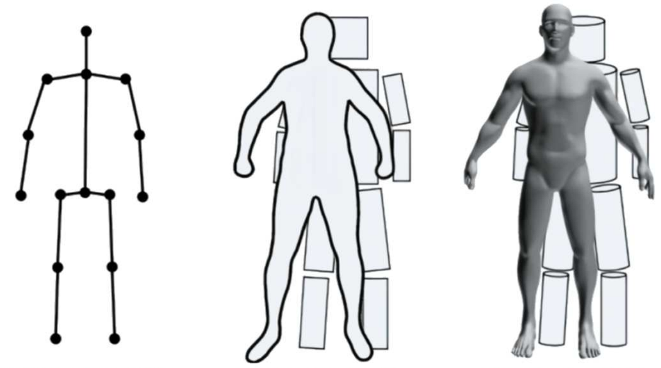
I. INTRODUCTION

Human Pose Estimation (HPE) is a technique used to identify and classify the joints in the human body. HPE is a computer vision technique to capture a set of coordinates for each joint (head, neck, arm, limps, torso, etc.,) and these coordinates are referred to as a **keypoint** that can be used to describe a pose of a person. The connection between these key points is termed as a **pair**.

Exposure detection is an active area of research in computer vision. Exposure estimation is a computer vision method used to track the movement of a person or object. Finding keywords for a product usually does this. Based on these key points, we can compare various actions and form and draw predictions. HPE is often used in augmented reality (AR and VR), animation, games, healthcare, Online coaching/training, surveillance and robotics.

The aim of Human Pose Estimation is to form a skeleton-like representation of a human body. There are three models of Human pose estimation and are used to represent the human body using computer vision.

1. Skeleton-based model or Kinematic model
2. Contour-based model or Planar model
3. Volume-based model or Volumetric model


1) Skeleton-based model (Kinematic model)
2) Contour-based model (Planar model)
3) Volume- based model (Volumetric model)

Convolution Neural Network (CNN)

The basis of most of the Human pose estimation techniques with machine learning uses Convolutional Neural Network or Visual Transformers. Vision Transformers will be explained later in this article.

Today, there are many other models that make predictions. Some of the methods used for human pose estimation are given below:

A Convolutional Neural Network (CNN) is a type of artificial neural network used in machine learning (and Deep Learning) and commonly used in Computer Vision. Computer vision is a field of Artificial Intelligence that enables a computer to understand and interpret the image or visual data. CNNs are designed based on the mathematical operation called convolution in place of general matrix multiplication in at least one of their layers.

Main three state-of-the-art models (SOTA models) for running real-time pose estimation:

1) MoveNet (detects 17 key points)
2) BlazePose (detects 33 key points)
3) PoseNet (able of detecting multiple poses)

There are many other models that make predictions. Some of the methods used for human pose estimation are given below:

1) Open pose
2) Deep Pose
3) Dense pose
4) Deep cut

Prefer one model over another depending on the application. In addition, factors such as working time, model size, ease of use will also be many reasons for model selection. Therefore, it is best to know your needs from the beginning and choose models accordingly

II. SYSTEM STUDY

A. Existing System

Current methods allow people to perform physical exercises with or without the assistance of a strength and conditioning trainer. This can damage the system and cost more than we need. Weaknesses of Existing Systems When we analyze existing systems, we find many things that are not right, downside;

1. Getting a physical trainer is most expensive
2. Time management is difficult

B. Proposed System

With the proposed system, the deficiencies of the existing system can be eliminated. We are trying to automate GST training with the help of machine learning and Python web application. The app can identify where they are doing the exercise and adjust the user's position by giving on-screen instructions.

C. Advantages of proposed system

- No need a physical trainer
- Can do workout any time with help of application
- Cost effective

The proposed system uses ViTPose, a Visual Transformer yet effective vision transformer used for human pose estimation. [17].

III. LIBRARY DESCRIPTION

Sklearn library

Scikit-learn (Sklearn) is the most useful Python library for machine learning. It provides a selection of efficient tools for machine learning such as classification, regression, clustering and dimensionality reduction through a python based interface. Sklearn library, which is largely written in Python, is built upon commonly used python libraries such as NumPy, SciPy and Matplotlib, which are used for scientific and numeric applications.

IV. METHODOLOGY

We will use Blaze Pose to capture the human pose and extract important details. This model can be easily implemented with a handy library called Media Pipeline. Media Pipe Media Pipe is an open source, crossplatform framework for building vario us machine learning pipelines. It can be used to make decision s such as face detection, multihand tracking, hair segmentatio n, detection and tracking. Exposure detection is an active area of research in computer vision.

You can find hundreds of case studies and many test models to solve the discovery problem. The reason why many machine l earning enthusiasts are interested in prediction is its wide appli cation and effectiveness. In this article, we'll cover detection a nd prediction applications using machine learning and some us eful Python libraries.

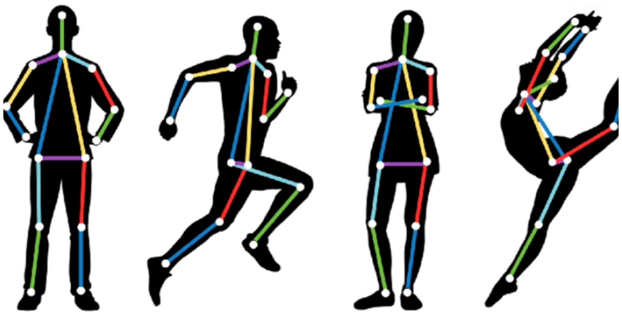


Figure 1: Pose estimation

Human Pose Estimation is a popular AI solution used to determine the position and orientation of a person or an object.

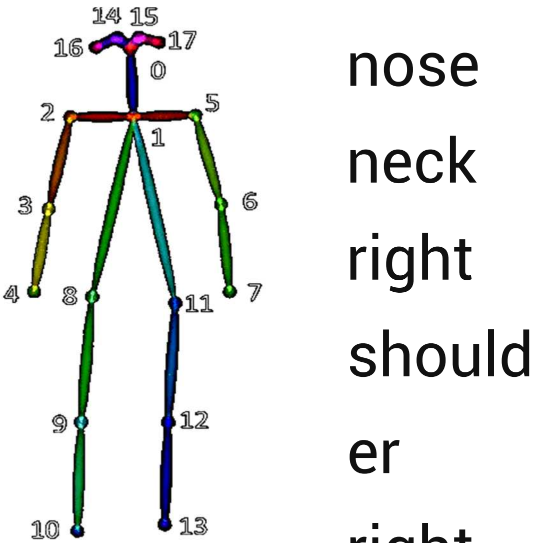
Exposure estimation is a computer vision method used to track the movement of a person or object. This is usually done by finding the location of key points for a particular item. Based on these key points, we can compare various actions and form and draw predictions. Human pose estimation is often used in augmented reality, animation, games, AI-powered personal trainers and robotics.

There are many models that make predictions today; some projection methods include: 1. Open exposure 2. Clear exposure 3. Flame exposure 4. Deep Exposure 5.Intense stance 6. Deep cut.

Prefer one model over another depending on the application. In addition, factors such as working time, model size, ease of use will also be many reasons for model selection. Therefore, it is best to know your needs from the beginning and choose models accordingly. For this article, we will use Blaze Pose to capture the human pose and extract important details. This model can be easily implemented with a handy library called Media Pipeline.

MediaPipe-Media Pipeline is an open source,cross platform framework for building various machine learning pipelines. It can be used to make decision models such as face detection, multihand tracking, hair segmentation, object detection and tracking.

OpenPose keypoints

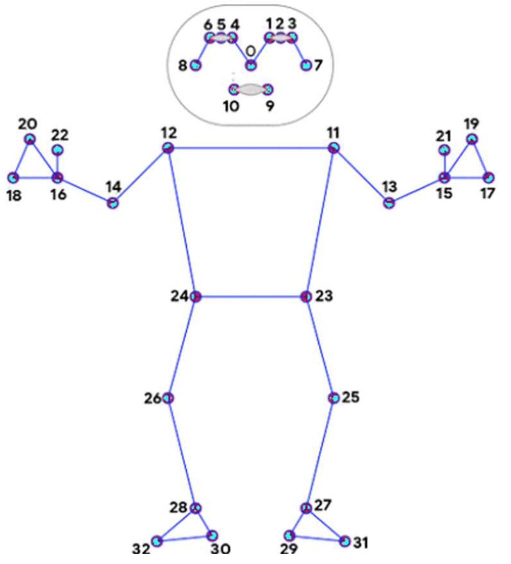


Blaze Pose Detector – BlazePose is a newer pose estimation model that can run on a smartphone. BlazePose (Full Body) is a machine learning model developed by Google that can compute (x,y,z) coordinates of 33 skeleton keypoints. It is a pose detection model and can be used in fitness applications.

Where most of the pose detection relies on COCO topology consisting of 17 key points, the blaze pose detector predicts 33 human key points including torso, arms, leg, and face. The inclusion of more key points is necessary for succeeding applications of domain-specific pose estimation models, like for hands, face, and feet. Each key point is predicted with three degrees of freedom along with the visibility score. The blaze pose is a sub-millisecond model and can be used for real-time applications with an accuracy better than most of the existing models. The model is available in two versions Blaze pose lite and Blaze pose fully to provide a balance between speed and accuracy.

Blaze pose offers several applications including fitness and yoga trackers. These applications can be implemented by using an additional classifier like the one we are going to build in this article itself.

BlazePose keypoints



Landmarks

0. Nose
1. Left_eye_inner
2. Left_eye
3. Left_eye_outer
4. Right_eye_inner
5. Right_eye
6. Right_eye_outer
7. Left_ear
8. Right_ear
9. Left_mouth
10. Right_mouth
11. Left_shoulder
12. Right_shoulder
13. Left_elbow
14. Right_elbow
15. Left_wrist
16. Right_wrist
17. Left_pinky
18. Right_pinky
19. Left_index
20. Right_index
21. Left_thumb
22. Right_thumb
23. Left_hip
24. Right_hip
25. Left_knee
26. Right_knee
27. Left_ankle
28. Right_ankle
29. Left_heel
30. Right_heel
31. Left_foot_index
32. Right_foot_index

Landmarks [16]

2D vs 3D pose estimation - Human Pose estimation can be done either in 2D or in 3D. 2D pose estimation predicts the key points from the image or video images through pixel values. Whereas 3D pose estimation refers to predicting the three-dimensional spatial arrangement of the key points as its output.

Graph Convolutional Networks (GCN)

Graph Convolutional Networks (GCN) can be used for estimating 3D human pose from a sequence of 2D human poses from images or videos. GCN is a method for semi-supervised learning on graph-structured data. GCN initially generates a spatial-temporal graph on consecutive 2D poses and then applies a GCN based local-to-global network to generate 3D poses.

Multi-person human pose estimation
The two approaches used for multi-person human pose estimation:
1) Top-down approaches and
2) Bottom-up approaches

Top-down approach
In the top-down approach, each person is first detected from the input image or video image and then the poses are estimated separately.
Top-down approach can be divided into two stages:
1) Person detector
2) Single-person pose estimator

Bottom-up approach
In the bottom-up approaches, it will first estimate all the human parts from the input image or video images and then associate the body parts to different persons.
Bottom-down approach can be divided into two stages:
1) Joint detector
2) Joint Partitioner

Vision transformers
Vision Transformers (ViT) is a model that uses self-attention mechanisms to process images. It is used in many vision tasks such as image recognition. The Vision Transformer architecture consists of a series of **transformer blocks**. Each transformer block consists of two sub-layers:
1) a multi-head self-attention layer (MHSA)
2) a feed-forward layer (FFN)

The proposed ViT system for Human Pose Estimation has two types of lightweight decoders to process the features extracted from the backbone network and localize the keypoints. ViTPose uses vision transformers as backbones to extract features for a human instance and a classic decoder for human pose estimation.

IV. ARCHITECTURE

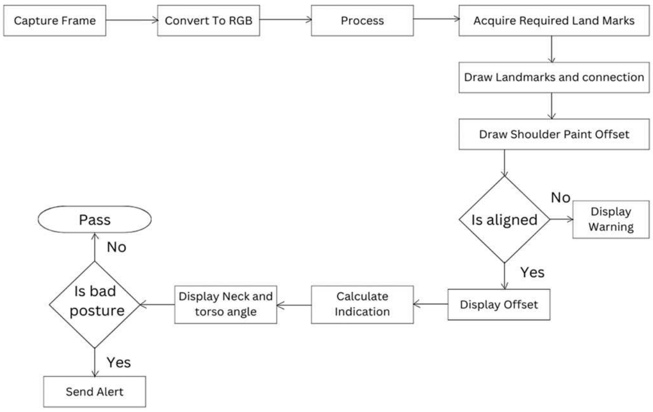
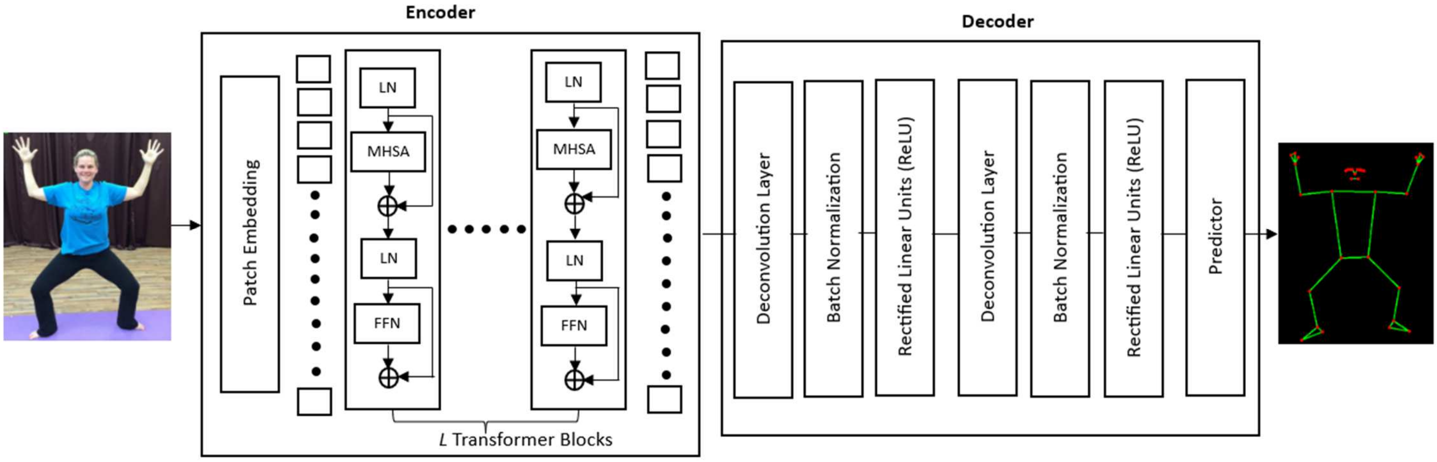


Figure 2: Human pose estimation workflow diagram

The system proposed an additional method where they implemented the cascade of such regressors in order to get more precise and consistent results. The proposed Deep Neural Network can model the given data in a holistic fashion, i.e. the network has the capability to model hidden poses, which was not true for the classical approach.

Patch embedding layer embeds the images into tokens then the embedded tokens are processed by several transformers.

At Transformer Encoder, Vision transformer uses a pure transformer to images without any convolution layers. It splits the image into patches and applies a transformer on patch embeddings. Patch embeddings are generated by applying a simple linear transformation to the flattened pixel values of the patch.



Block diagram of Visual Transformer for Human Pose Estimation

The system then a decoder. It is composed of two deconvolution blocks, each of which contains one deconvolution layer followed by batch normalization and ReLU. A convolution layer with kernel size 1 × 1 is taken to get the localization heatmaps for the keypoints,

FEASIBILITY STUDY

A feasibility study is undertaken to determine the possibility of either improving the existing system or developing a completely new system. It helps to obtain an overview of the problem and to get a rough assessment of whether a feasible solution exists. This is essential to avoid committing large resources to a project and needs for feasibility study.

ECONOMIC FEASIBILITY

- The system can be developed technically and the installed system must be a good investment for the organization.
- It can reduce the workload of user or admin than the existing system
- It should provide many benefits to the customer.
- Hence it is economically feasible.

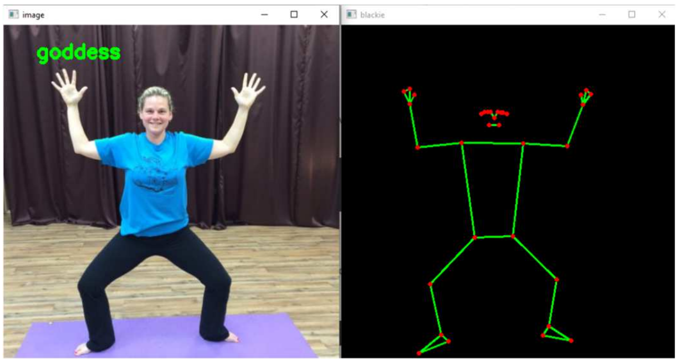
TECHNICAL FEASIBILITY

- Due to the existing system, they can only store less data. In the new system we can handle large amounts of data.
- Whenever we want, we can expand the system based on the user requirement.
- New system provides some features like security, accuracy, and reliability.
- Hence the system is technically feasible.

SOCIAL FEASIBILITY

- Computer installation has something to do with changing the job status, turnover etc.
- The introductory must take effort to educate the candidate and train the staff for the conducting business.
- The people may not be educated well. The proposed system can be designed so that they can understand the correct system.
- So, the system requires not much effort to train and educate people, the system is that much socially feasible

V. RESULT



From the above images, you can observe that the model has correctly classified the pose. You can also see the pose detected by the blaze pose model on the right side. In the first image, if you observe closely, some of the key points aren't visible, still, the pose is classified correctly. This could be possible because of the visibility of the key points attribute given by the blaze pose model.

VI .CONCLUSION

Human pose estimation is a computer vision technique used to identify and classify the joints in the human body. Pose detection is an active area of research in the field of machine learning and offers several real-life applications. In this article, we tried to introduce Human Pose Estimation techniques using machine learning and work on one such application and methods for human pose detection. We learned about human pose detection and several models that can be used for pose estimation and detection methods. We selected the blaze pose model for our purpose and learned about its pros and cons over other models. In the end, we built a classifier to classify yoga poses using the support vector classifier from the sklearn library. The proposed system uses ViTPose, a vision transformer to detect human pose while exercising/workouts and alert a message for the bad pose. We also built our own dataset for this purpose which could further be extended easily using more images.

VII .REFERENCES

[1] W. Gong et al., "Human pose estimation from monocular images: A comprehensive survey," Sensors, vol. 16, no. 12, p. 1966, 2016.

[2] H. B. Zhang, Q. Lei, B. N. Zhong, J. X. Du, and J. L. Peng, "A Survey on Human Pose Estimation, Intelligent Automation & Soft Computing," Intelligent Automation & Soft Computing, vol. 22, no. 3, pp. 483-489, 2016/07/02 2016, doi: 10.1080/10798587.2015.1095419.

[3] S. Kreiss, L. Bertoni, and A. Alahi, "PifPaf: Composite Fields for Human Pose Estimation," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11969- 11978, 2019.

[4] S. C. Babu, "A 2019 guide to Human Pose Estimation with Deep Learning," 2019.

[5] X. Chen and A. L. Yuille, "Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations," in NIPS, 2014.

[6] D. Mwiti, "A 2019 Guide to Human Pose Estimation," 2019.

[7] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 20-25 June 2009 2009, pp. 1014-1021, doi: 10.1109/CVPR.2009.5206754.

[8] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[9] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3D pose estimation and tracking by detection," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 13-18 June 2010 2010, pp. 623-630, doi: 10.1109/CVPR.2010.5540156.

[10] S. Johnson and M. Everingham, "Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation," in BMVC, 2010.

[11] Yufei Xu, Jing Zhang, Qiming Zhang, Dacheng Tao2, "ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation"

[12] https://developers.google.com/ml-kit/vision/pose-detection

[13] L. Pishchulin, M. Andriluka, P. V. Gehler, and B. Schiele, "Poselet Conditioned Pictorial Structures," 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 588-595, 2013.

[14] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in CVPR 2011, 20-25 June 2011 2011, pp. 1385-1392, doi: 10.1109/CVPR.2011.5995741.

[15] Y. Yang and D. Ramanan, "Articulated Human Detection with Flexible Mixtures of Parts," IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 35, no. 12, pp. 2878–2890, 2013, doi: 10.1109/TPAMI.2012.261.

[16] F. Wang and Y. Li, "Beyond Physical Connections: Tree Models in Human Pose Estimation," in 2013 IEEE Conference on Computer Vision and Pattern Recognition, 23-28 June 2013 2013, pp. 596-603, doi: 10.1109/CVPR.2013.83.

[17] M. Sun and S. Savarese, "Articulated part-based model for joint object detection and pose estimation," in 2011 International Conference on Computer Vision, 6-13 Nov. 2011 2011, pp. 723- 730, doi: 10.1109/ICCV.2011.6126309.

[18] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2D Articulated Human Pose Estimation and Retrieval in (Almost) Unconstrained Still Images," International Journal of Computer Vision, vol. 99, no. 2, pp. 190-214, 2012/09/01 2012, doi: 10.1007/s11263-012-0524-9.