# iPOSE – Real Time Movement Tracking Application

**Tejal Deshpande[1], Amit Marathe[2], Ronak Singh Rajput[3], Shubham Maurya[4], Ninad Tawade[5]**

*[1]Assistant Professor, Dept. of Electronics and Telecommunication Engineering, Xavier Institute of Engineering, Maharashtra, India*

*[2,3,4,5]Student, Dept. of Electronics and Telecommunication Engineering, Xavier Institute of Engineering, Maharashtra, India*

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract -** *Real-time analysis of posture detection has been quite challenging and the existing solutions lack accuracy with limited use. In this paper, we present a comparison mechanism for two or more individuals' bodily actions by detecting their physical movements that have proliferated accuracy. The proposed system should be applicable for automated learning of exercise, sports, dance, yoga, etc. The user would have to provide the system with a video of an individual performing some action that the user wishes to recreate. The system shall perform a comparison between the actions of the user and the expert to point out errors. For the learning process to be more effective we present a technique to modify the orientation of the individual in the video to increase the depth at which the user can go for learning the action.*

***Key Words*: Computer-vision, Assisted learning, Pose estimation, Tracking application, Convolutional neural networks.**

## 1. INTRODUCTION

Two-dimensional human posture estimation is a visual recognition task dealing with the autonomous localization of anatomical human joints or "key points" in RGB images and videos [1]. It is widely considered a fundamental problem in computer vision due to its many downstream applications, including action recognition and human tracking. In particular, it is a precursor to 3D human pose estimation that serves as a potential option for invasive marker-based motion capture.

Identifying the human body parts often involves many challenges especially when socially engaged individuals are involved. It also depends on the kind of activity being performed. Identification is easier for a single person performing as compared to a group of 10. The limbs of the people performing a particular action may overlap with the other, making the association of parts quite rigorous. This challenge intensifies with the involvement of multiple people in the frame. Also, the runtime complexity tends to increase with the increase in the number of people. A top-down approach is the common way for detection and association but suffers from early commitment. Moreover, for each individual detection, this approach runs a single-person pose estimator. Ergo, the more the number of people, the greater the runtime and computational cost. Conversely, bottom-up approaches are efficient as they provide early commitment and have the potential to decouple runtime complexity from the number of individuals. We focused on increasing the accuracy with which the model detects the limbs of a person and overall performance boost.

Once the individuals are recognised, the system we propose requires us to be flexible with the orientation of the subjects in the frame for the comparison with the user's key points for the feedback aspect. The following paper discusses the mechanism for changing the orientation and comparison of key points for automated learning of actions like yoga, gym exercises, sports maneuvers, etc.

## 2. LITERATURE REVIEW

Interest points or key points detection is a prime building block for several computer vision tasks, such as SLAM - simultaneous localization and mapping, camera calibration, and SfM - structure from motion. Keypoint detection has a long history predating deep learning, and many great algorithms in wide industry applications (such as ORB, SIFT, and FAST) are based on hand-crafted features. As in many other computer vision tasks, people have been exploring deep learning to outperform hand-crafted algorithms. In this paper, we will review some recent advances in this field [2].

Deep learning has dominated state-of-the-art semantic keypoint detection. Mask R CNN (ICCV 2017) and PifPaf (CVPR 2019) are two representative techniques for detecting semantic key points. Both processes are supervised learning and need extensive and expensive human annotation. This makes the application of interest point detection challenging because interest points are semantically ill-defined; hence a human annotator cannot reliably identify the same set of interest points. Therefore, it is impossible to formulate interest point detection as a supervised learning problem.

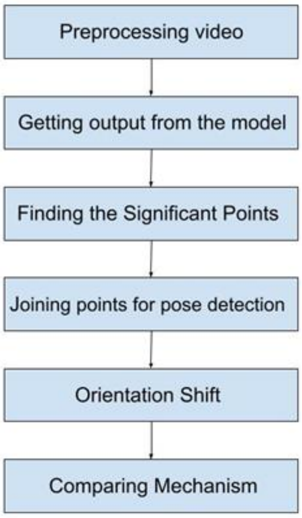
The general notion of prediction, multiple models predict the key points, and this machine performs regression[3]. Learning from the past model, the wrong could be corrected over time. So, for the next one, the general idea is the same from the past but now we are using convolutional neural networks. Hence better suited. Primarily, stacked AE on top of one another to perform regression at the end of the layer [4]. Here, the Affinity field is used to perform regression on the human pose. There is an implementation of this method and overall, it is up to the mark [5]. They incorporated the pose information into segmentation since, in a lot of videos, when a human being is covered by another, the performance degrades.

## 3. METHODOLOGY

In the following section the approach followed in developing the system is mentioned

### 3.1. Workflow of the Application

The following section comprises the steps of processing the video, feeding it to the model and post processing done on its output.

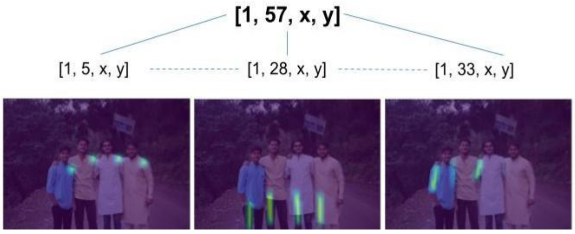

Preprocessing video
Getting output from the model
Finding the Significant Points
Joining points for pose detection
Orientation Shift
Comparing Mechanism

#### 3.1.1. Preprocessing

Extracting the frames from the video is the initial step. Reference videos with high frame rates can be scaled down to the native frame rate of the camera's video (currently 24 fps) by sampling. The model used in the application adjusts the aspect ratio of the images by itself so no further processing is required for the frames.

#### 3.1.2. Model Output

The Model outputs a [1, 57, x, y] dimensional tensor. The x and y are the dimensions of the image. The second dimension represents the 57 outputs per image which constitute the points and part affinity fields that are used to detect the pose in further sections.



#### 3.1.3. Detecting the Points

The model uses two techniques for detecting the significant points for constructing the skeleton of the individual: confidence maps and part affinity fields (PAFs). The model mentioned in the paper produces an output that is an aggregation of the two. The first half, predicts a heatmap over the image where it thinks the points lie. The areas on the image where there is a high probability of the occurrence of these points are expressed with bright colours. Part affinity fields are accountable for generating a vector field over the part of the image where the model detects the existence of a joining link between the significant points. These too are expressed as heatmaps with an elongated shape. These vectors also propose the direction, to prevent confusion in the presence of multiple subjects. The challenge occurs when there are many people in the frame since it would result in several sets of the same key points and figuring out the connections is not so easy. A unit vector along the direction of the predicted elongated shapes helps in identifying correct sets of points that belong to a particular individual. These two concurrently predict the significant points and the links that encode part-to-part association forming the skeleton using CV2's line function.

#### 3.1.4. Orientation Shift

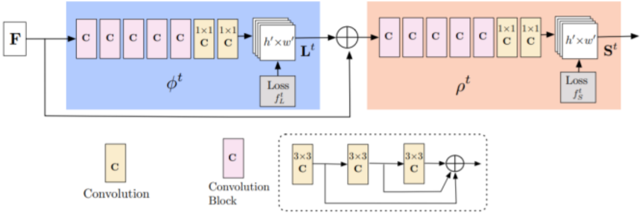
Once we have the coordinates of the points of an individual, a vertical axis is chosen by the mean values of the 'X' coordinates and must pass through the midpoint of the line that emerges from the point marked as the neck and bisects the line joining the points marked as knees. This is the axis of rotation; we would be able to rotate the points around it to recreate the figure and track its movements in any orientation. The challenge is to figure out in which direction every point is going to move for the rotation. That direction depends on the distance of the point on the z-axis, how far in or out it is from the axis. Since we cannot measure depth, we use the standard ratios of the lengths of the lines. Any extension than usual shall give us an idea of the perspective. Closer body parts will be enlarged and the ones farther away will be smaller to scale. The points are moved and their projection is given by the cosine of the angle of rotation which is decided by the user. The scaling of the connector lines is done based on their relative positions to the original and new axis.

### 3.1.5. Comparison

The system in place for comparing the movements of the reference and the user notes the coordinates of both points for comparison. The metric for judging the correctness of the user's actions is based on three attributes, the relative scale of the lines, their slopes, and the inner product of the unit vectors of lines if the reference and the user. The degree of freedom for what classifies as an error can be adjusted by the user. The exact body part that lags in motion or is out of sync with the reference is highlighted to indicate an error.

### 3.2. Model Architecture

The model that we have chosen was proposed by [6][7][8]. It's a multimodal approach, two multi-staged Convolutional Neural Networks that are responsible for detecting the points and the lines joining them. The first stage is a classic CNN with increased depth that refines the prediction over time. It includes several convolutional kernels of size 7x7 with an architecture resembling the DenseNET. The second stage also has similar architecture. The outputs from both are ultimately concatenated and passed on as a final output.



## 4. CONCLUSION

This paper has presented an approach for adapting existing pose-detection models to assist people in the movement and fitness-related activities. The mechanism for shifting the orientation of the person in the reference video gives a lot of control to the user. The application can be further designed to save and share workout routines for future use. The capabilities of the model could be further extended to incorporate the detection of fingers which would refine the experience even further. Pose detection and a system to correct errors together can translate sign language with the addition of a model to translate. It could aid in teaching it to others as well.

## REFERENCES

[1] William McNally Kanav Vats Alexander Wong and John McPhee, "EvoPose2D: pushing the boundaries of 2d human pose estimation using neuroevolution," Systems Design Engineering, University of Waterloo.

[2] Patrick Langechuan Liu 2020, Self-supervised Keypoint Learning — A Review, accessed 22 January 2021, https://towardsdatascience.com/

[3] Yunji Kim, Seonghyeon Nam, In Cho and Seon Joo Kim, "Unsupervised keypoint learning for guiding class-conditional video prediction," Yonsei University.

[4] Daniel DeTone, Tomasz Malisiewicz and Andrew Rabinovich, "SuperPoint: self-supervised interest point detection and description," Cornell University.

[5] Shangze Wu, Christian Rupprecht and Andrea Vedaldi, "Unsupervised learning of probably symmetric deformable 3d objects from images in the wild," Visual Geometry Group, University of Oxford.

[6] Zhe Cao, Tomas Simon, Shih-En Wei and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," The Robotics Institute, Carnegie Mellon University.

[7] Tomas Simon, Hanbyul Joo, Iain Matthews and Yaser Sheikh, "Hand keypoint detection in single images using multiview bootstrapping,".

[8] Shih-En Wei, Varun Ramakrishna, Takeo Kanade and Yaser Sheikh, "Convolutional pose machines,".