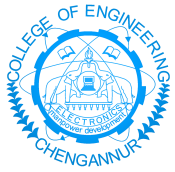
Single Person Pose Estimation and Classification

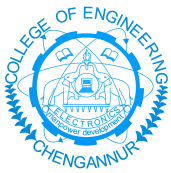
03CS6902 Mini Project

CHN20CSIP03 Praveena K M
praveenamurali1004@gmail.com
M. Tech. Computer Science & Engineering (Image Processing)



Department of Computer Engineering
College of Engineering Chengannur
Alappuzha 689121
Phone: +91.479.2165706
http://www.ceconline.edu
hod.cse@ceconline.edu

College of Engineering Chengannur
Department of Computer Engineering



C E R T I F I C A T E

This is to certify that, this report titled ***Efficient Pose:Single Person Pose Estimation and Classification*** is a bonafide record of the work done by

CHN20CSIP03 Praveena K M

Second Semester M. Tech. Computer Science & Engineering (Image Processing)

student, for the course work in **03CS6902 Mini Project**, under our guidance and supervision, in partial fulfillment of the requirements for the award of the degree, M. Tech. Computer Science & Engineering (Image Processing) of **APJ Abdul Kalam Technological University**.

Guide

Radhu Krishna
Asst. Professor
in Computer Engineering

Coordinator

Ahammed Siraj K K
Associate Professor
in Computer Engineering

Head of the Department

October 6, 2021

Dr. Smitha Dharan
Professor
in Computer Engineering

Permission to Use

In presenting this mini project dissertation at College of Engineering Chengannur(CEC) in partial fulfillment of the requirements for a Postgraduate degree from APJ Abdul Kalam Technological University, I agree that the libraries of CEC may make it freely available for inspection through any form of media. I further agree that permission for copying of this dissertation in any manner, in whole or in part, for scholarly purposes may be granted by the Head of the Department of Computer Engineering. It is understood that any copying or publication or use of this dissertation or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to CEC in any scholarly use which may be made of any material in this mini project dissertation.

Praveena K M

Statement of Authenticity

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at College of Engineering Chengannur(CEC) or any other educational institution, except where due acknowledgement is made in the report. Any contribution made to my work by others, with whom I have worked at CEC or elsewhere, is explicitly acknowledged in the report. I also declare that the intellectual content of this report is the product of my own work done as per the **Problem Statement** and **Proposed Solution** sections of the mini project dissertation report. I have explicitly stated the major references of my work. I have also listed all the documents referred, to the best of my knowledge.

Praveena K M

Acknowledgements

Primarily, I thank Lord Almighty for his eternal support through out my project work.

I express my sincere thanks to **Dr. Jacob Thomas V**, Principal, College of Engineering Chengannur for extending all the facilities required for doing my seminar. My heartfelt words of gratitude to **Dr. Smitha Dharan**, Professor and Head of Department of Computer Engineering, for providing constant support.

Now I express my gratitude to my miniproject co-ordinator **Mr. Ahammed Siraj K K**, Associate Professor in Computer Engineering and my project guide **Ms.Radhu Krishna**, Assistant Professor in Computer Engineering who played a great role for valuable suggestions and expert guidance.

Praveena K M

Abstract

This project is an attempt to implement an approach for human activity recognition and classification using a person's pose skeleton in images. This implementation is divided into two parts; a single person poses estimation and activity classification using the extracted poses. Pose Estimation consists of the recognition of 18 body key points and joints locations. Using these pose information ,the images are classified into different action classes such as standing ,sitting using a binary classifier CNN.

Contents

Chapter 1

Introduction

The goal of a Human Activity Recognition (HAR) system is to predict the label of a person's action from an image or video.Understanding human behavior in images gives useful information for a large number of computer vision problems and has many applications like scene recognition and pose estimation.One of the popular vision based Human Activity Recognition systems uses pose information. Poses have had remarkable success in human activity recognition, Poses provide useful information about human behavior. Pose estimation is a computer vision task that infers the pose of a person or object in an mage or video.It is a technique used to estimate how a person is physically positioned, such as standing, sitting, or lying down.There are two approaches: a bottom-up approach, and a top-down approach. With a bottom-up approach, the model detects every instance of a particular keypoint (e.g. all left hands) in a given image and then attempts to assemble groups of keypoints into skeletons for distinct objects. A top-down approach is the inverse – the network first uses an object detector to draw a box around each instance of an object, and then estimates the keypoints within each cropped region.

1.1 Proposed Project

This project aims to implement a human activity recognition system by extracting the pose information of a single person in the image.

1.1.1 Problem Statement

This project aims to localizing human skeletal keypoints of a person from an image ,and then the classification of the poses into action classes such as standing,sitting using the extracted pose key points .

1.1.2 Proposed Solution

The proposed solution for this pose detection and classification consists of two sequential tasks.The system takes, as input, a color image of size w $\times$ h and produces the 2D locations of anatomical keypoints for the person in the image using openpose and then the identification of salient actions through the observation of the extracted pose key points with the help of KNN classification algorithm .

Chapter 2

Report of Preparatory Work

2.1 Literature Survey Report

1. **Densepose: Dense human pose estimation in the wild, R. Alp Guler, N. Neverova, and I. Kokkinos, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018 Source: densepose.org**[7]
   The DensePose refers to Facebook's real-time approach for mapping all individual pixels of 2D RGB images into a 3D surface-based model of the human body.Dense human pose estimation aims at mapping all human pixels of an RGB image to the 3D surface of the human body. They propose DensePose-RCNN, a variant of Mask-RCNN, to densely regress part-specific UV coordinates within every human region at multiple frames per second.The strategy of the research project is to find dense correspondence by dividing the surface into many parts. The system determines for every pixel:which surface part it belongs to,where on the 2D parameterization of the part it corresponds to.The architecture consists of using a fully convolutional network (FCN) that combines classification and regression tasks. In a first step, it classify a pixel as belonging to either background or one among several region parts which provide a coarse estimate of surface coordinates. This amounts to a labeling task that is trained using a standard cross-entropy loss. In a second step, a regression system indicates the exact coordinates of the pixel within the part. We break it into multiple independent pieces and parameterize each piece using a local two-dimensional coordinate system, that identifies the position of any node on this surface part.

2. **RMPE: Regional Multi-person Pose Estimation Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, Cewu Lu,Published in: 2017 IEEE International Conference on Computer Vision (ICCV)**[8]
   Alpha pose /RMPE is a popular top-down method of Pose Estimation.Proposed a regional multi-person pose estimation (RMPE) framework for estimation in inaccurate human bounding boxes. The framework consists of three components: a Symmetric Spatial Transformer Network (SSTN), Parametric Pose NonMaximum-Suppression (NMS), and a Pose-Guided Proposals Generator (PGPG). This framework achieves a 76.7 mAP on the MPII (multi-person) dataset. The authors posit that top-down methods are usually dependent on the accuracy of the person detector, as pose estimation is performed on the region where the person is located. Hence, errors in localization and duplicate bounding box predictions can cause the pose extraction algorithm to perform sub-optimally.

3. **Alexnet** AlexNet[1] is a Classic type of Convolutional Neural Network, and it came into existence after the 2012 ImageNet challenge.The input to AlexNet is an RGB image of size 256×256. This means all images in the training set and all test images need to be of size 256×256.If the input image is not 256×256, it needs to be converted to 256×256 before using it for training the network. To achieve this, the smaller dimension is resized to 256 and then the resulting image is cropped to obtain a 256×256 image.AlexNet is the name of a convolutional neural network which has had a large impact on the field of machine learning, specifically in the application of deep learning to machine vision. It famously won the 2012 ImageNet LSVRC-2012 competition by a large margin . The network had a very similar architecture as LeNet by Yann LeCun et al but was deeper, with more filters per layer, and with stacked convolutional layers. It consisted of 11×11, 5×5,3×3, convolutions, max pooling, dropout, data augmentation, ReLU activations, SGD with momentum. It attached ReLU activations after every convolutional and fully-connected layer. AlexNet was trained for 6 days simultaneously on two Nvidia Geforce GTX 580 GPUs which is the reason for why their network is split into two pipelines.

4. **DeepPose: Human Pose Estimation via Deep Neural Networks (CVPR, 2014),Alexander Toshev and Christian Szegedy**

   This paper proposes using deep neural networks(DNNs) to tackle pose estimation task.Presented a cascade of such DNN regressors which results in high precision pose estimates. The DNN is able to capture the content of all the joints and doesn't require the use of graphical models. The convolution layer and fully-connected layer are the only layers that have learnable parameters. They both contain linear transformations followed by a rectified linear unit. The network takes an input image of size 220 × 220 and the learning rate is set to 0.0005.

5. **Pose classification using support vector machines ,In proceedings of Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference onVolume: 6 February 2000 Source:IEEE Xplore Conference: DOI:10.1109/IJCNN.2000.859415**
   In this work,they presented a software architecture for the automatic recognition of human arm poses. Our research has been carried on in the robotics framework. A mobile robot that has to find its path to the goal in a partially structured environment can be trained by a human operator to follow particular routes in order to perform its task quickly. The system is able to recognize and classify some different poses of the operator's arms as direction commands like "turn-left", "turn-right", "go-straight", and so on. A binary image of the operator silhouette is obtained from the gray-level input. Next, a slice centered on the silhouette itself is processed in order to compute the eigenvalues vector of the pixels covariance matrix. Finally, a support vector machine is trained to classify different poses using the eigenvalues array

2.2 System Study Report

Human pose estimation has been studied extensively over the past years. As compared to other computer vision problems, human pose estimation is different as it has to localize and assemble

human body parts on the basis of an already defined structure of the human body. The goal of a Human Activity Recognition (HAR) system is to predict the label of a person's action from an image or video. This interesting topic is inspired by many useful real-world applications, such as simulation, visual surveillance, understanding human behavior, etc. Action recognition through videos is a well-known and established research problem. In contrast, image-based action recognition is a comparably, less explored problem, but it has gained the community's attention in recent years. Because motion activities cannot be estimated through the still image, recognition of actions from images remains a tedious and challenging problem. It requires a lot of work as the methods that have been applied to video-based systems cannot be applicable in this. However, the approach is not the only problem faced in this task. There are many other challenges too, especially the changes in clothing and body shape that affect the appearance of the body parts, various illumination effects, estimation of the pose is difficult if the person is not facing the camera, definition, and diversity activities, etc Application of pose estimation in fitness and sports can help prevent injuries and improve the performance of people's workouts.

Chapter 3

Project Design

The proposed system for activity recognition and classification consists of two sequential tasks, pose estimation from images, and then the classification of the activities using extracted pose key points as input with the help of classification algorithms.

3.1 Human pose estimation Using OPENPOSE

Human pose estimation refers to the process of inferring poses in an image. Essentially, it entails predicting the positions of a person's joints in an image or video. This problem is also sometimes referred to as the localization of human joints.A Human Pose Skeleton represents the orientation of a person in a graphical format. Essentially, it is a set of coordinates that can be connected to describe the pose of the person.Each co-ordinate in the skeleton is known as a part (or a joint, or a keypoint). A valid connection between two parts is known as a pair (or a limb).Here the OpenPose [1] framework is used for estimating the pose from the input image.
The OpenPose network first extracts features from an image using the first few layers .The features are then fed into two parallel branches of convolutional layers. The first branch predicts a set of 18 confidence maps, with each map representing a particular part of the human pose skeleton. The second branch predicts a set of 38 Part Affinity Fields (PAFs) which represents the degree of association between parts.Successive stages are used to refine the predictions made by each branch. Using the part confidence maps, bipartite graphs are formed between pairs of parts (as shown in the above image).Using the PAF values, weaker links in the bipartite graphs are pruned. Through the above steps, human pose skeletons can be estimated and assigned to every person in the image.

3.1.1 Part Affinity Field Maps (L)

It contains two-dimensional vectors that encode the body part's positions and orientations in an image. It encrypts your data in the form of a double link be-tween body parts.

$$L = (L1, L2, L3 \ldots . Lc)(1)$$
$$L_c \varepsilon R^{w*h*2} \ ,c \ \ \varepsilon 1 ... C$$

, where C is the total number of limbs, R is the real number, L is the set of part affinity field maps, and w x h is the dimension of each map in the set L.

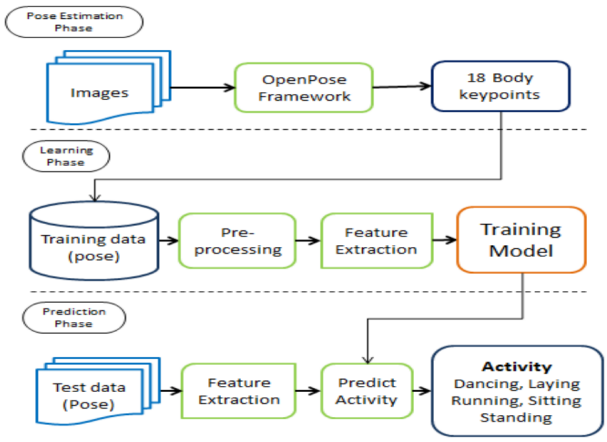


Figure 3.1.1: proposed framework

3.1.2 Confidence Map S

It is a two-dimensional representation of the belief that a particular part of the body can be placed on a specific pixel.

$$S = (S1, S2, S3. . . .Sj)(2)$$
$$L_c \varepsilon R^{w*h} \ , \ j \ \ \varepsilon 1. . . J$$

, where J is the total number of body parts, R is the real number, and S is the set of confidence maps.
The number of keypoints detected is dependent upon the dataset has been trained.The 18 different body key points are $\mathrm{R}_{Ankle}, R_{Knee}, R_{Wrist}, L_{Wrist}, R_{Shoulder}, L_{Shoulder}, L_{Ankle}, L_{ear}, R_{Ear}, R_{Elbow}, L_{Elbow}, L_{Knee}, L_{Eye} R_{Eye}, R_{Hip}, L_{Hip}, Nose, and Neck is used.$

3.2 Activity Classification

The classification algorithm takes 18 body keypoints (x-axis and y-axis values of each point) as input and label the images as upright(standing) and sitting.Here uses a supervised learning approach as our dataset contains body keypoints with an activity label.Convolutional nueral networks show a great promise in pose classification tasks, thus making it a highly desirable choice. They can be trained on keypoints of joint locations of the human skeleton or can be trained directly on the images.MobileNet is a CNN architecture model for Image Classification and Mobile Vision.The core layer of MobileNet is depthwise separable filters, named as Depthwise Separable Convolution. The network structure is another factor to boost the performance.To convert pose landmarks to a feature vector, here uses pairwise distances between predefined lists of pose joints, such as distances between wrist and shoulder, ankle and hip, and two wrists.. The CNN takes as input a full image and outputs a vector of numbers representing the probabilities of each of the activity labels for

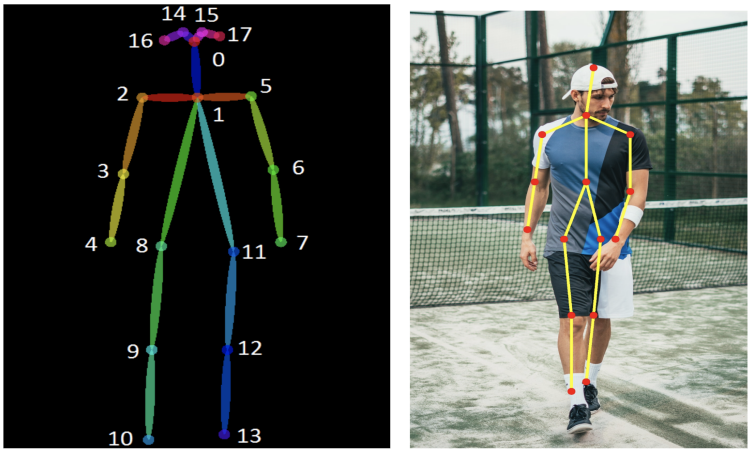


Figure 3.1.2: 18 keypoint skeltal representation

either sititng or standing activity categories, depending on the ground truth lables passed in and the size of the final fully connected output layer.

3.3 Hardware & Software Requirements

Operating System : Any Operating System
Supporting softwares libraries : Python,opencv, Tensorflow
Processor : Intel Core i5 7th Gen 2.50GHz
RAM : 8GB
Monitor : Any colour monitor
Dataset :COCO Dataset, imgNet

Chapter 4

Implementation

4.1 Human Pose Estimation

Openpose framework is used to obtain the skeltal representation of the person in the image. The number of keypoints detected is dependent upon the dataset has been trained.The 18 different body key points are $R_{Ankle}, R_{Knee}, R_{Wrist}, L_{Wrist}, R_{Shoulder}, L_{Shoulder}, L_{Ankle}, L_{ear}, R_{Ear}, R_{Elbow}, L_{Elbow}, L_{Knee}, L_{Eye} R_{Eye}, R_{Hip}, L_{Hip}$, Nose and Neck.

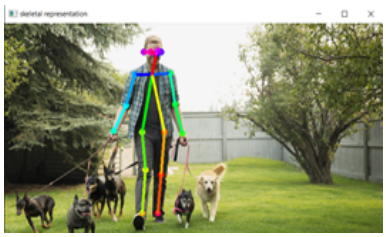


Figure 4.1.1: keypoints detected skeletal representation of pose

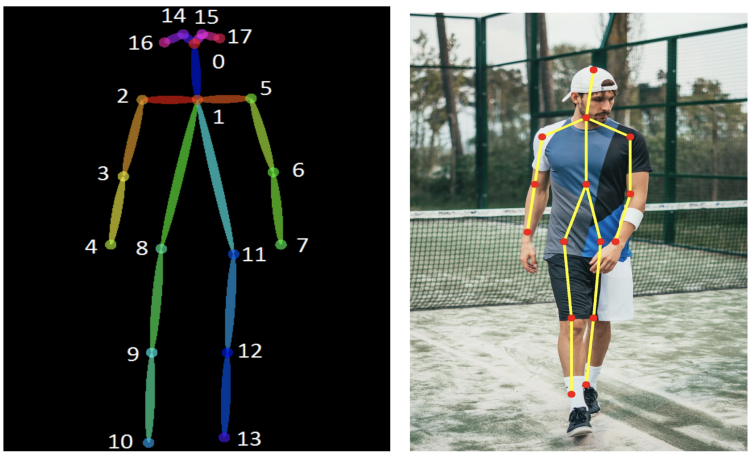


Figure 4.1.2: skeltal representation 2

4.2 Activity Classification

The mobilenet binary classifier takes 18 body keypoints (x-axis and y-axis values of each point) as input and label the images as upright(standing) and sitting.

Figure 4.2.1: class label -sitting

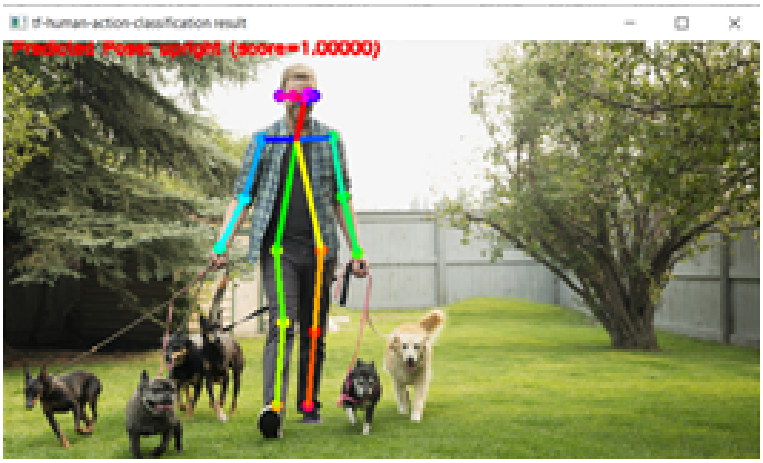


Figure 4.2.2: class label-standing 2

Chapter 5

Results & Conclusions

5.1 Results

I have implemented tensorflow openpose library for human body keypoint extraction and then mobilenet binary classifier for labelling the images as either standing or sitting. The obtained results are shown in the following figures.

Figure 5.1.1: input image 1

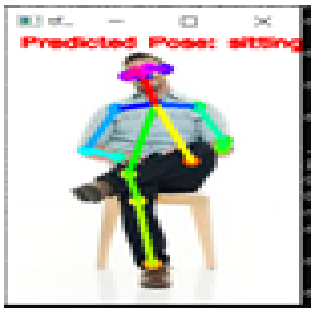


Figure 5.1.2: output image 1 with class label -sitting

Figure 5.1.3: input image2

Image 2 is labelled as standing with detected keypoint representation.

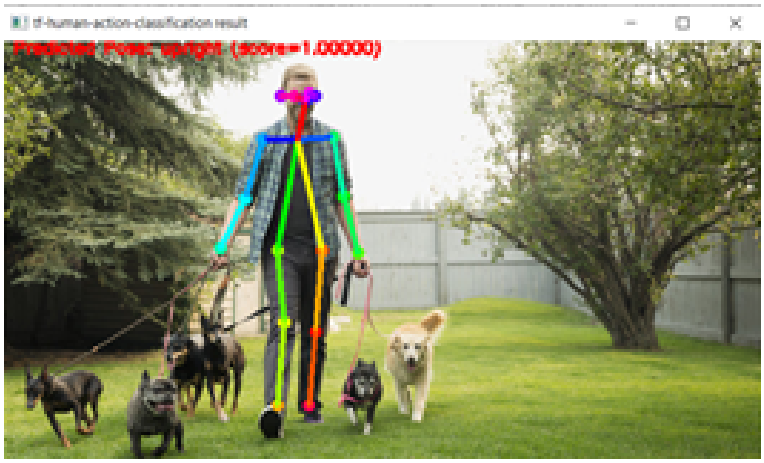


Figure 5.1.4: output image 2 with class label-standing

Figure 5.1.5: input image3

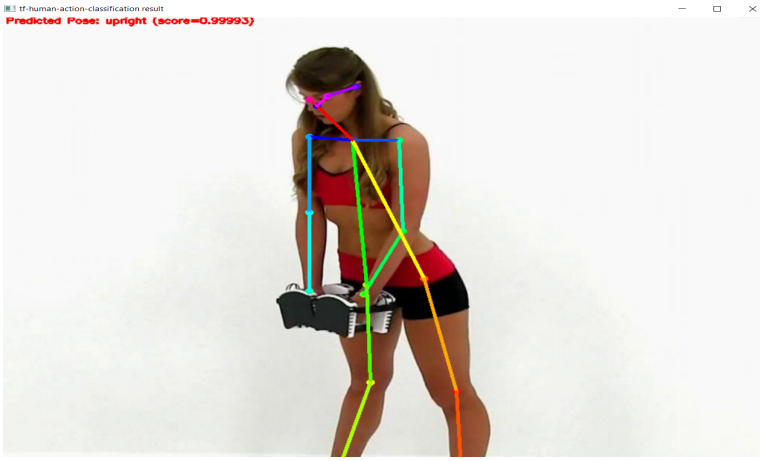


Figure 5.1.6: output image 3 with class label-standing

5.2 Conclusion

Human pose estimation has been studied extensively over the past years. As compared to other computer vision problems, human pose estimation is different as it has to localize and assemble human body parts on the basis of an already defined structure of the human body. Application of pose estimation in fitness and sports can help prevent injuries and improve the performance of people's workouts.In the proposed approach for human activity recognition from still images by extracting the skeletal coordinate information (pose) using OpenPose API and then further utilizing this pose information to classify activity with the help of a mobilenet binary classifier.In practice, there are a lot of different activities that humans use to perform in everyday life.Here only label the pose just as standing and sitting and most of the cases it yields 100 percent accuracy.The definition and diversity of activities also make it more complicated for machines to understand. Some more activities can be added to extend the scope and usefulness of the work in the future.

References

[1] Zhe Cao; Gines Hidalgo; Tomas Simon; Shih-En Wei; Yaser Sheikh. Openpose: Re-altime multi-person 2d pose estimation using part affinity fields.IEEE Transactionson Pattern Analysis and Machine Intelligence, pages 128–137, july 2019.

[2] A Novel Image Classification Approach via Dense-MobileNet Models,Research Article — Open Access Volume 2020 —Article ID 7602384 — https://doi.org/10.1155/2020/7602384Deep Learning in Mobile Information Systems 2020

[3] Heri Ramampiaro Espen AF Ihlen Daniel Groos. Efficientpose: Scalable single-person pose estimation.Springer journal of Applied Intelligence, 51:2518–2533,6,November 2020.

[4] Gehler P Schiele B Andriluka M, Pishchulin L. 2d human pose estimation: newbenchmark and state of the art analysis. In IEEE Conference on computer visionand pattern recognition (CVPR), 2014.

[5] https://medium.com/dsaid-govtech/human-pose-estimation-and-human-action-recognition-experimenting-for-public-good-dabde16521b3

[6] DeepPose: Human Pose Estimation via Deep Neural Networks (CVPR, 2014),Alexander Toshev and Christian Szegedy

[7] Densepose: Dense human pose estimation in the wild, R. Alp Guler, N. Neverova, and I. Kokkinos, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018 Source: densepose.org

[8] RMPE: Regional Multi-person Pose Estimation Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, Cewu Lu,Published in: 2017 IEEE International Conference on Computer Vision (ICCV)