International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Person Classification based on Spatio-Temporal Features from Kinect Skeletal Structure using Machine Learning Techniques.

Vishnu Nair[1], Sachin Nair[2], Jitin Nambiar[3], Abhilash Nair[4], Prof. Prakash Bhise[5]

[1-4]BE student, Department of Computer Engineering, Pillai College of Engineering, Navi Mumbai, India-410206
[5]Assistant Professor, Department of Computer Engineering, Pillai College of Engineering, Navi Mumbai, India-410206

Abstract - *There has been a recent demand for person classification based on their behavioral pattern. This is used for many security and analysis purposes. Gait recognition is one of the most intriguing and systematic studies of human and animal locomotion. The kinect skeletal structure or human pose key points helps in the formation of external skeletal structure. The features are then analyzed using machine learning technique to classify them based on the pose and behavioral pattern. Considering an environment, the behavior or activity of a person can be classified as suspicious or normal. Thus the model will classify activities or actions performed by a person or a group as suspicious and normal activities considering an environment.*

Key Words: **Person Classification, Gait recognition, kinect skeletal structure, human pose key points, machine learning**

1. INTRODUCTION

The human identification and classification area has consequently gained popularity worldwide in recent times. To identify a person standard biometric identification and recognition has been used commonly. Fingerprints, facial patterns, iris scans are common biometric standards. Systems which facilitate automatic identification and recognition of subjects are gaining increasing popularity. Such systems play a vital role in the field of surveillance and human identification.

Gait analysis refers to the scientific study of body movements that are responsible for locomotion in human beings. Gait parameters are known to be reliable indicators of neuromuscular and skeletal health [1]. Gait features focuses on limb movements during a walk and hence the characteristics of non-contact, long distance, cross-view recognition and hard to disguise, all these gaps are filled and has gained immense popularity in long distance identification in the public security industry.

The spatio-temporal feature descriptors involve distance (spatial) and time (temporal) such as the walking speed, gait cycle etc. In order to find changing motions of human actions, spatio features such as height, length of body arms and legs need to be considered. Gait features of a human walking sequence are represented in the form of a vector.

1.1 Fundamentals

Until recent years there was only little research in the field of pose estimation, which was mainly due to lack of effective datasets. Some challenging datasets have been released in recent years which have now made a far better progress in the research of pose estimation. The Gait based classification system uses one of the important techniques in pose estimation that is detecting key points hence pose estimation is also referred to as key point's detection. Key points are nothing but the major parts or joints of the body (eg. shoulder, knee, ankle etc.)

1.2 Objective

The main objective of person classification based on spatio-temporal features is to classify an individual based on the actions and thereby minimizing the cooperation required in order to identify a person such as in biometrics. And to analyze various gait features and study their use in pose estimation and classification based on pose and actions.

2. RELATED WORKS

There has been a lot of research going on the gait analysis as it has now become very popular of motion and walking pattern analysis, biometric recognition and in other applications. In the field of action recognition there exist many works which include methods such as using model based approach suggested in [2] and the use

of deep learning neural networks for accurate and better Kinect based gait recognition [3]. AlphaPose, OpenPose are two open sources pose estimators which are based on COCO and MPII dataset and also possess a good mean Average Precision rate on both the datasets.

2.1 Human Activity recognition (HAR)

Cho Nilar Phyo et al. [4] proposed the method of using Deep learning and Machine Learning models for Human Activity Recognition through motions of skeletal joints. They have considered the development of a productive information based HAR. With the help of experimental results conducted on two famous public datasets of human daily activities, the proposed system is known to outperform other state-of-the-art methods on both datasets.

2.2 Pose Estimation

Pose estimation is the technique of identifying human figures from images and videos. It uses the technique of estimating key body points also known as localization of the human joint. Various researches have been made on pose estimation to view the human gait pose from various videos and images. M Andriluka et al. in [14] developed a new benchmark for human 2D pose estimation. Similarly, Zhe Cao Tomas et al. in [15] suggested a real time multi person pose estimation using affinity fields.

2.3 OpenPose based multi-person gait recognition

OpenPose is a multi-person keypoint detection library which is used to get the gait features of a human. The complexity of human anatomy which provides 255 degrees of freedom with 230 joints makes the pose estimation a really difficult problem of research. A Viswakumar et al. in [7], proposed a cost effective, marker less approach to human gait analysis. They proposed the use of 2D Pose estimation to find knee flexion.

3. EXISTING WORK

There has been a lot of research going on gait analysis as it has now become very popular for motion and walking pattern analysis, biometric recognition and in other applications. In the field of action recognition there exist many works which include methods such as Cluster segmentation approach [9] for human detection and tracking in which the feature extraction is done using HOG (Histogram of Gradient) and SVM for classification. Various other methods were also discussed in [10] [11] where neural networks are used to recognize and classify pose and actions.

4. METHODLOGY

The Gait based classification system detects the human pose using OpenPose. The detected point's also known as key points are the joints detected on a human body.

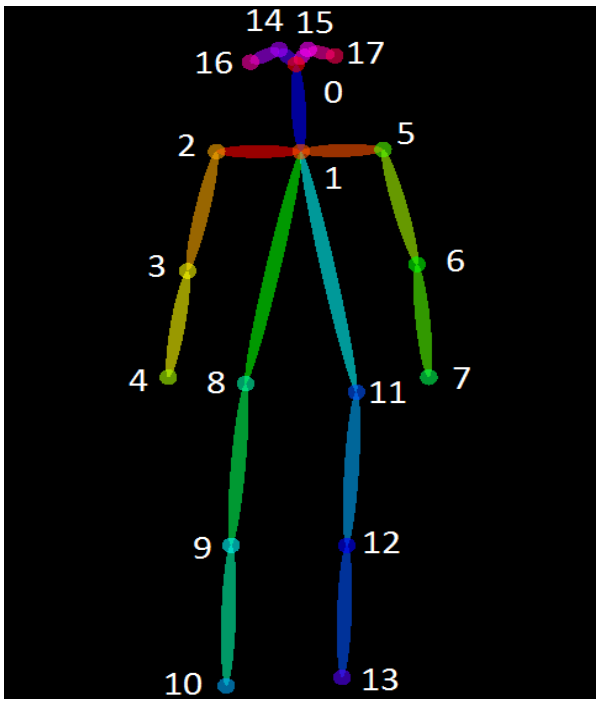

Fig -1: Human Pose Key Points

Fig -1 represents the key points in a human pose which are labeled by numbers. It is not always that all the key points mentioned are present in a human pose. If we consider a video of a person dancing, the key points detected will vary in every frame due to the constant movement of the body and hence only some joints will be detected.

4.1 Proposed System

The model takes a video of a person doing some activity as an input and the key body points are extracted which will be helpful in identification of a person's pose from the given input video. The human joints are trained using a neural network, which is then used to classify the action based on the trained features. The classification done is based on the action performed by the person and

whether the action is normal considering the environment in which the action is performed.
For example if we consider a scenario in which a person is trying to destroy an ATM machine. In that scenario the action is considered to be suspicious.

The system is designed in such a way that we have considered a scenario of an ATM environment. We have considered some actions that a person normally performs in an ATM. The system will be classifying only those actions as normal or suspicious. Actions such as fighting or destroying the machine will be considered suspicious by the model whereas standing, walking will be normal actions

4.2 OpenPose Architecture

The approach employed for recognition of pose of a person is by using the VGG net which is a multi-layered convolutional neural network of OpenPose. The model takes input as a video and produces 2D key points for each person in the frame. VGG net uses the first 10 layers in order to create feature maps for the given input. A CNN (multi stage two branched) is used to predict the confidence map for different body points and to encode the degree of association between different parts. The confidence and associated key points are then parsed together to form a 2D Key Points for all the people in the input.

As mentioned earlier, the OpenPose estimator which is a multi-person pose estimator has two models trained on two challenging datasets namely - COCO dataset and the MPII dataset. The COCO model consists of 18 key points and the MPII model consists of 15 key points. The top layer of VGG net creates a set of feature confidence maps ***C****, which can be mathematically written as,*

$$\boldsymbol{C} = (\boldsymbol{C_1}, \boldsymbol{C_2}, \boldsymbol{C_3}, \ldots\ldots, \boldsymbol{C_j})$$
$$\text{Where,}$$
$$\boldsymbol{C_j} \varepsilon R^{wxh}$$
$$j \varepsilon \{1 \ldots j\} \qquad \ldots(1)$$

j, the total number of key points, depends on the dataset which is used to train OpenPose. If we consider the MPII dataset it has 15 key points. j will be 16 (No. of key points + 1 background) whereas for the COCO dataset, it has 18 key points, hence j will be 19.

Consider a MPII dataset with set $\boldsymbol{C}$ *as* $\boldsymbol{C_1}, \boldsymbol{C_2}, \ldots, \boldsymbol{C_{16}}$*, and suppose C1 corresponds to a confidence map of key point id "0" which refers to the body part "nose".*

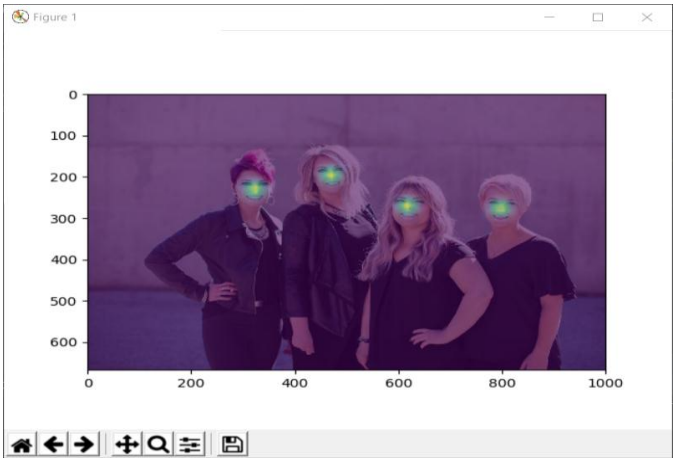

Fig -2: Confidence map for nose key point

Fig -2 represents the confident map corresponding to the key point "nose". Similarly confidence maps of other key points can also be generated. Using those confident maps, the key point can be parsed to obtain 2D key points of the person.

Table -1: Confidence Map Representation

0	0	0	0	0
0	0	**0.9**	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

In table 3.1 the cell with score 0.9 represents the confident map score for key point nose in the input.

The multistage CNN which is the bottom layer in the VGG network produces a set of key point affinity maps based on the confident maps generated in above layers. The part affinity maps ***P*** *can be mathematically represented as follows,*

$$\boldsymbol{P} = (\boldsymbol{P_1}, \boldsymbol{P_2}, \ldots., \boldsymbol{P_l})$$
$$\text{Where,}$$
$$\boldsymbol{P_l} \in R^{w \times h \times 2}$$
$$l = \{1,2,\ldots l\}, \qquad \ldots(2)$$

Here we can consider limbs as body part pairs, and totally depends on the dataset which OpenPose is trained with. For COCO, the body pairs can be considered as (1,2), (1,3), (2,3) etc. Here the set P, has each element of size w x h, where each cell is a 2D vector representing the direction of the key point pairs.

The input is first analyzed by a pre-trained CNN such as the first 10 layers of VGG network to produce feature maps ***F****.*

Stage 1: The network produces a set of confidence maps C, and a set of part affinity fields ***P****. Symbol* $\boldsymbol{\rho}$ *is a function variable of the CNN with input* ***F*** *to produce confident maps* ***C****. Similarly symbol* $\boldsymbol{\phi}$ *is a function variable of CNN with input F to produce part affinity fields* ***P****. Annotation "1" at the top represents the stage 1 of the CNN.*

$$C^1 = \rho^1(F) \quad \ldots(3)$$

$$P^1 = \phi^1(F) \quad \ldots(4)$$

Stage n: The predictions from both branches in the previous stage, along with the original image features F, are concatenated and used to produce more refined predictions.

4.3 Algorithm

The algorithm workflow is as follows:

1. Get Joints using OpenPose.
2. Track each person, using Euclidean distance between the joints of two skeletons.
3. Extract features of body, and normalized joint positions.

The input to the system is a video stream, either coming from a camera or a video file. Then the OpenPose algorithm [12] is adopted to detect the human skeleton (joint positions) from each frame. Next, a sliding window of size N aggregates the skeleton data of the first N frames. These skeleton data are preprocessed and used for feature extraction, which are then fed into a classifier to obtain the final recognition result.

A feature of a person or any object can be considered as a distinctive attribute which defines the object or person by itself. In simpler words, it is something which is unique to that object. Once the preprocessing step is done, the joints or key points are good to use for further process. Some of the features which are computed from the input are , direct concatenation of joints of all the N frames, average height of the skeletons, the next position , all the joints position, length of limbs, and joint angles computed from the joint positions.

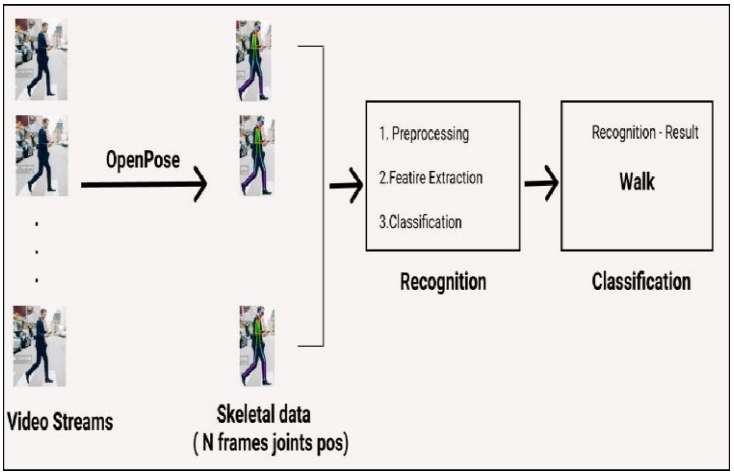

Fig -3: *Workflow of the system*

The overall workflow of the action recognition algorithm is shown in Fig.-3

4.4 Detecting Human Skeletal Data

The OpenPose algorithm [12] is adopted to detect human skeletons from the input. The main aim of OpenPose is using Convolutional Neural Network to produce two heatmaps, one for predicting joint positions (confident maps), and the other for associating the joints into human skeletons (pose affinity field maps). In short, the input to OpenPose is an image, and the output is the skeletons of all the humans this algorithm detects.

As mentioned earlier OpenPose estimator [12] has a CNN (multi stage two branched) is used to predict the confidence map for different body points and to encode the degree of association between different parts. The confidence and associated key points are then parsed together to form a 2D Key Points for all the people in the input.

Each skeleton has 18 joints, including head, neck, arms and legs, and it's key points as shown in Fig. 4 .Each joint position is represented in the image coordinate with coordinate values of x and y, so there are a total of 36 values of each skeleton.

Fig -4: *Multi person pose and its keypoints*

5. CLASSIFICATION

The total training data is split into two sets: 75% for training, and 25% for testing. The implementation of these methods is from the Python library "sklearn" and "keras-tensorflow. We have used tuned parameters of the classifier in order to get the better.

5.1 Datasets

We have used the COCO dataset, which is large-scale object detection, segmentation, and captioning dataset. COCO has several features [13]. The other important dataset to be mentioned is the MPII Human Pose dataset is a state of the art benchmark for evaluation of articulated human pose estimation [14]. Table 2 represents the information regarding datasets for human pose estimation.

Table -2: Datasets

Datasets	Classes/Categories	Items	Features
1.COCO Dataset[13]	80+	330K Images	200K + Labelled Images. 250000 People with Key Points.
2.MPII Dataset[14]	400+ Human Activities	25K + Images	Labeled images with 40K+ People with annotated body joints

It is observed that though COCO Dataset is slower than MPII Dataset but the results are comparatively better.

The COCO output format has 18 key Points whereas the MPII-Human Pose dataset has 15 key points. The output format of those key points is mentioned below.

COCO Output Format:

Nose – 0, Neck – 1, Right Shoulder – 2, Right Elbow – 3, Right Wrist –4, Left Shoulder – 5, Left Elbow – 6, Left Wrist – 7, Right Hip – 8, Right Knee – 9, Right Ankle –10, Left Hip – 11, Left Knee – 12, Left Ankle – 13, Right Eye – 14, Left Eye – 15, Right Ear – 16, Left Ear – 17, Background – 18

5.2 Performance analysis

A classification model is trained using CNN with classes such as Fighting, Destroying, and Walking etc.

Main focus is given on activities related to an environment such as an ATM wherein activities such as destroying or fighting can be considered as suspicious. The model has been trained with images of the above mentioned activities with over 350+ images of various activities.

Video classification is actually more than just simple image classification —with video we can typically make the assumption that subsequent frames in a video are correlated with respect to their contents.

Using videos an action can be classified as follows:

1. Go through every frame from the input video file...

2. For every frame, pass the frame through the neural network.

3. Consider each frame individually and independently of each other for classification.

4. Choose the label with the largest probability.

5. Label the frame and save the output frame.

Accuracy of the model is determined by the ratio of the number of frames correctly recognized to the total number of frames of the input. The training loss and accuracy on the trained dataset is shown in the Fig - 5. In which the graph shows the model has obtained accuracy close to 97% when trained on above mentioned dataset images of 3 activities - Destroying. Fighting, and Walking.

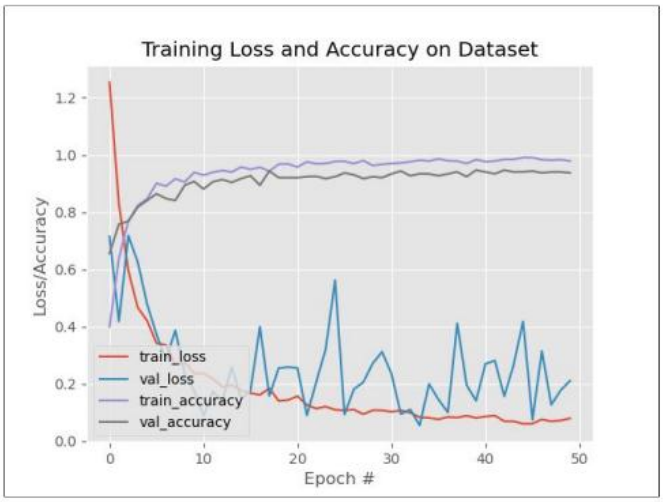

Fig -5: *Accuracy/Loss training plot*

6. RESULT AND OUTPUT

A GUI is built which has options to open a video input, get the pose from the input and classify the input. Moreover, the confidence map and key points can also be viewed. The result from the classification model is the activity recognition i.e. action performed by the person and further the action is classified as suspicious or not based on the environment in which the person is performing that particular action.

The GUI has several other options such as view confident map (show Cmap) and to view key points (show Key points) and to view only pose.

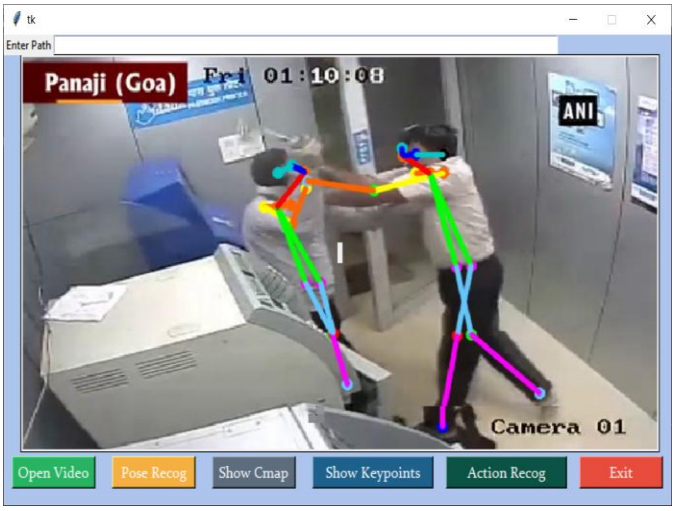

Fig-6: GUI-Action Recognition

7. Software and hardware requirements

7.1 Software requirements

Python version 3 or more with the following specifications and packages:

Tensorflow-gpu version 1.13 or more, sklearn ,OpenCv 3.4 to work with video and image inputs\ and output, Keras , Tkinter for GUI development ,numpy, pandas etc. for computational purposes

7.2 Hardware requirements

Operating system: Windows 10 / Linux / Mac

Processor: Pentium i5 and above or NVDIA Graphic Support (for tensorflow support)

RAM: 8GB

GPU support is recommended.

8. CONCLUSION

The study of gait analysis and its applications is presented in this report. Many different aspects of gait analysis and human action recognition were studied and implemented using various techniques and data available. The pose estimation was based on the OpenPose multi-person system [12] on the COCO dataset [13]. The classification model was developed by training images having people performed various activities considering a particular environment. The recognition accuracy was up to 97% on the training set composed of more than 1000 samples. The developed model achieved stable and good recognition performance on several video inputs which were tested.

9. ACKNOWLEDGEMENT

We would first like to extend our sincere thanks to our Principal Sir Dr. Sandeep Joshi to learn and explore our technical knowledge and evaluate ourselves. We are deeply thankful to the HOD of the Computer Engineering department, Dr. Sharvari Govilkar madam for giving us this opportunity to present this project. We sincerely thank our project guide, Prof. Prakash Bhise for all the valuable guidance and encouragement in carrying out this project. We would also like to thank Prof. Rupali Nikhare for helping us in group formation and updating us with instructions to be followed. We also take this opportunity to thank all our professors and friends who have directly or indirectly helped us in making this project. We also acknowledge the authors of MPII and COCO human pose datasets which made 2D human pose estimation in the wild possible.

REFERENCES

[1] *A. Viswakumar, V. Rajagopalan, T. Ray and C. Parimi, "Human Gait Analysis Using OpenPose," 2019 Fifth International Conference on Image Information Processing (ICIIP), Shimla, India, 2019, pp. 310-314, doi: 10.1109/ICIIP47207.2019.8985781.*

[2] *Gupta, Jay & Singh, Nishant & Dixit, Pushkar & Semwal, Vijay & Dubey, Shiv. "Human Activity Recognition Using Gait Pattern". 10.4018/ijcvip.2013070103.*

[3] *A. S. M. H. Bari and M. L. Gavrilova, "Artificial Neural Network Based Gait Recognition Using Kinect Sensor," in IEEE Access, vol. 7, pp. 162708-162722, 2019, doi: 10.1109/ACCESS.2019.2952065.*

[4] *C. N. Phyo, T. T. Zin and P. Tin, "Deep Learning for Recognizing Human Activities Using Motions of Skeletal Joints," in IEEE Transactions on Consumer Electronics, vol. 65, no. 2, pp. 243-252, May 2019, doi: 10.1109/TCE.2019.2908986.*

[5] *M. Ye, C. Yang, V. Stankovic, L. Stankovic and S. Cheng, "Distinct Feature Extraction for Video-Based Gait Phase Classification," in IEEE Transactions on Multimedia, vol. 22, no. 5, pp. 1113-1125, May 2020, doi: 10.1109/TMM.2019.2942479*

[6] *Cholwich Nattee, Nirattaya Khamsemanan, " A Deep Neural Network Approach for Model-based Gait Recognition", in Thai Journal of mathematics , vol 17, no 1, 2019.*

[7] *A. Viswakumar, V. Rajagopalan, T. Ray and C. Parimi, "Human Gait Analysis Using OpenPose," 2019 Fifth International Conference on Image Information Processing (ICIIP), Shimla, India, 2019, pp. 310-314, doi: 10.1109/ICIIP47207.2019.8985781*

[8] *M. Qi, "Gait based human identification in surveillance videos," 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Guilin, 2017, pp. 2317-2322, doi: 10.1109/FSKD.2017.8393133.*

[9] *K. Seemanthini and S. S. Manjunath, "Human Detection and Tracking using HOG for Action Recognition," Procedia Comput. Sci., vol. 132, no. Iccids, pp. 1317–1326, 2018.*

[10] *Huang, Yi and Lai, Shang-Hong and Tai, Shao-Heng, "Human Action Recognition Based on Temporal Pose CNN and Multi-Dimensional Fusion", The European Conference on Computer Vision (ECCV) Workshops,September,2018*

[11] *C. Li, X. Min, S. Sun, W. Lin and Z. Tang, "DeepGait: A learning deep convolutional representation for view-invariant gait recognition using joint Bayesian", App. Sci., vol. 7, no. 3, pp. 15, 2017.*

[12] *G. Hidalgo, Z. Cao, T. Simon, S.-E. Wei, H. Joo, Y. Sheikh, "OpenPose Library",https://github.com/ CMU-Perceptual-Computing -Lab/openpose .*

[13] *Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollár, "Microsoft COCO: Common Objects in Context" May 1,2014.*

[14] *M. Andriluka, L. Pishchulin, P. Gehler and B. Schiele, "2D Human Pose Estimation: New Benchmark and State of the Art Analysis," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 3686-3693, doi: 10.1109/CVPR.2014.471.*

[15] *Zhe Cao Tomas Simon Shih-En Wei Yaser Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", arXiv:1611.08050v2 [cs.CV] 14 Apr 2017.*

[16] *Working of Xbox Kinect- Jameco Electronics.https://www.jameco.com//xboxkinect.html.*

[17] *Imed Bouchrika. Gait Analysis and Recognition for Automated Visual Surveillance. School of Electronics and Computer Science, University of Southampton, 2008.6.*