Dicle University
Journal of Engineering
https://dergipark.org.tr/tr/pub/dumf
duje.dicle.edu.tr

Research Article

# Human posture prediction by Deep Learning

**Hediye Nupelda KANPAK[1*], M. Ali ARSERİM[2]**

[1] Dicle University, Electrical-Electronics Engineering Department, h.nkanpak@gmail.com, Orcid No: 0000-0001-5806-7126
[2] Dicle University, Electrical-Electronics Engineering Department, marserim@dicle.edu.tr, Orcid No 0000-0002-9913-5946

ARTICLE INFO

*Article history*:

Received 10 December 2021
Received in revised form 29 December 2021
Accepted 30 December 2021
Available online 31 December 2021

*Keywords*:

*Human pose estimation, motion detection, human body recognition and tracking*

*Doi: 10.24012/dumf.1051429*

* Corresponding author

ABSTRACT

İnterpreting the human posture in human videos and pictures constitutes the most basic structure of human posture prediction. A system is created that decides what the movement is and what purpose it is made by evaluating pictures and videos. In this way, a structure has been created that determines and classifies human movements as an automatic system. A mechanism of motional meaning contained in the created system has been recognized in such away that the pattern is expressed. It is intended to take advantage of these components by taking instant information. A result was obtained by primarily inferring instant still images and eliminating time intervals that do not contain information range. A classification was made according to their accuracy. Based on the location coordinates of the images and videos, it was tried to determine how people might react in the neck stage. Thanks to the analysis performed through the joints with optical flow calculation, motion information was obtained and classifications and analyses expressing the power of motion were created. Motion information on the region determined in the image is determined by the detection of joints, revealing the power generated by movement. The created histograms provide ease of classification of motion. Based on the reliability of the descriptions, which include the concept of the time in a sequential way with the detection of joints, it was desired to create a sliding classification mechanism within the framework of these joints. As a result of this study, it was aimed to obtain a functional structure that can recognize and understand the autonomous movement of stationary or moving beings. An efficient structure has been created in terms of providing a useful and facilitating mechanism by solving the problems in estimation.

## Introduction

Human posture prediction is a very important step to understand the actions of people in the video or pictures discussed. The main action is to determine the locations of the joints that make up the skeleton[1]. Determining, perceiving and interpreting the human movements in the human appearance is important in terms of making sense of the images discussed in the videos. The multifaceted content of human movements has made these studies more important. The perception of human movements, recognition of mimics, editing of videos, easy search of the desired content in videos, and the common class created by guessing provide great convenience. This class contains large changes within the data[2]. Depending on these changes, it has been tried to create a forecasting environment in a challenging area. The discovery of smart solutions for video, which is increasing today, increases the need for human posture predictions[3].

It has been seen that a system that can analyze human movements can help in obtaining the necessary information[4]. Contrary to popular belief, estimation is more difficult than expected. It differs from person to person, although it is within certain periods, including the natural behaviors that people do without realizing it.

The positive or negative effects of factors such as light, environment and clothing also greatly affect the estimation. There are great differences even between movements belonging to the same person in the same class. Perceiving and interpreting these movements and classifying them is difficult and complex, contrary to what it seems. Generally, a system is formed by grouping people from top to bottom. Structures that can be estimated are used in applications where various video and camera images are available online and offline, as well as pictures[5].

Monitoring vehicles from cameras in public transportation such as subway, detection of criminals or suspects from security cameras on streets and avenues, detection of thieves from cameras at home or in companies can be given examples to online applications. On the other hand, detection of the movements of the athletes in the close-up areas from the cameras in various competitions, detection of the movements of the reporters and people in the news, looking at the details of the performances of the actors who took part in the concerts and various shows can be accepted as offline applications[6].

Situations expressing the concept of motion are called primitive. Primitive actions constitute actions. Movement activities occur when actions form a meaningful and interpretive whole.

İn human motion detection studies, extracting images from videos and classifying the corresponding movements are generally fundamental. It is based on how the system works as a whole by extracting features from the images and then creating a classification obtained from these features. In classification, both internal and external variations are among the factors that directly affect the correct conclusion in estimation. The most difficult and restrictive studies in motion recognition are camera movement and dynamics with unwanted movements[7].

Disappearance of certain or all parts of the human body in the image, the change of camera angle and the focus area are the most basic factors that can make estimation difficult.

Situations such as timing, motion start and end can make a difference in every situation, as well as adversely affect the generalization situation in motion recognition, make classification difficult. The temporal dimension is the biggest problem, even if everything else is well and there are no problems[8]. It is aimed to increase the success percentage by creating a label of the movement class and making widespread classification in performing the actions in the definition category at low, medium and high levels[9]. Figure 1 shows the skeleton prediction model in different postures.

Figure 1. Guessing pose figures [37]

In this study, it is aimed to establish a systematic detection of key points in the mechanism created for body estimation, to find a wide range of joints, to detect these points, to determine how to eliminate the errors seen and to predict instantaneous movements in a systematic way. In the mechanism, which was carried out on a piece-based basis, a focal point window was created in each joint frame, ensuring that factors such as background influence at a minimum level and the resulting error rate in estimation was kept lower compared to other methods. With the help of deep learning , the system , which was maintained as 7 layers, was started. The biggest advantage in these methods is that by focusing only on the content of the image, the desired image was obtained and that image is estimated. This neural network was used to detect each joint and joint regressor. Graphic models were used for the formulation system. Human posture positions was strengthened when a high-resolution mechanism was created with the detection of interactions between joints by graphic modeling. These approaches have been found to perform well in appearance and in estimating moments of difference. Evaluation of the datasets, with the help of the python software used in the study and the opencv library, results based on human posture estimations were obtained. The football matches, which was in the domain of the data set, were based on the goal moments of Cristiano Ronaldo. The time intervals at the time of the goal were included in these images. Estimating through these intervals yielded highly successful results. The efficiency of the estimation system was measured by using snapshots according to the right, left and rear views. In Python, it is aimed to make a prediction mechanism of data sets in accordance with the prediction created by the deep learning model. With this method, it was deemed appropriate to use a deep learning system model to identify key points in the body and connect the joints. When the body parts are used with window openings, the skeleton mechanism has been made independent of the background.

The following sections were described in the article, respectively: In the Introduction section, the study in the article was explained. Material and Methods used in the study were explained in the Material and Methods section. In the Method part, it was determined how the estimation mechanism will be finalized section by section. In the Results and Discussion section, the results obtained in the study were given. In the conclusion part, the data on the accuracy of the study was shared and the efficiency of the study is shown.

**Material and Methods**

In this study, snapshots in videos and pictures available to the public are used. Data set is consisted of images belong to football matches and human prediction is made in the python software. When the images were examined, the movements of the people were made sense. While estimating, joint points were numbered first. Then, these joints were combined with their key points and the prediction results were finalized as skeletons.

**Method**

The current methodology consists of several steps. The estimation mechanism is created by determining the silhouettes with Euclidean extensions and forming them gradually.

Determining the edge points and detecting the joints by utilizing the optical flow properties is the most effective method in general.

A motional signification mechanism in the created system is recognized in such a way that the pattern is expressed. It is aimed to take advantage of these components by taking instant information. A result is obtained by eliminating the time intervals that do not contain information by firstly extracting the snapshots. A classification was made according to the accuracy margins. It has been tried to determine how people can react in the

next stage, based on the location coordinates in the pictures and videos[10].

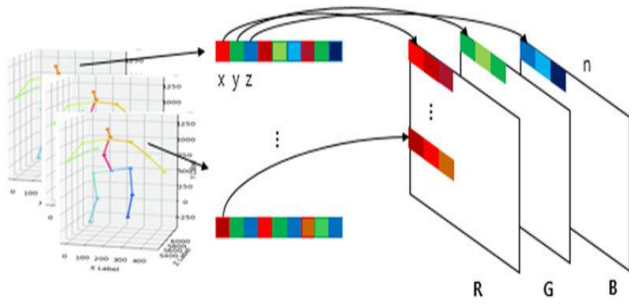


Figure 2. Joint data conversion to RGB image [6]

Before specifying the location coordinates, image analysis was done with RGB. The schematic representation of this analysis is given in figure 2.

When the matrix coordinates shown in Figure 3 are determined, it is possible to form part inferences. With simultaneous joins of pairwise matches with keypoint tags, all joints are grouped at once. The created part is called pose[11]. After the key points are detected, the connections of the joints are transferred to the multi-layer sensors. Common configurations are learned, sample data sets are created according to these configurations.

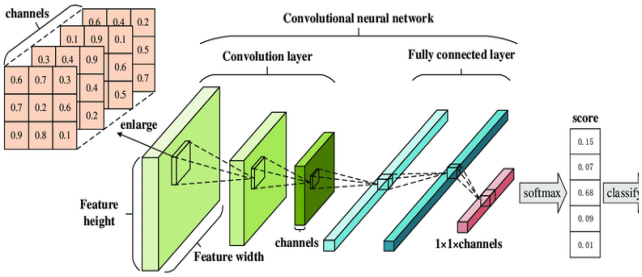


Figure 3. Convolutional neural network (CNN) structure [32]

In convolutional neural networks, an adequate understanding of the image is provided with multiple sensors. Reducing the number of parameters and at the same time minimizing the learning time improves the estimation process. The decrease in the number of data is important in terms of estimation speed.

Key points in the human body consist of 135 points in real time for the body, face, hands, feet. The number of key points in the study used was 17 as shown in Figure 4. These key points consist of nose, right and left eyes, ears, shoulders, elbows, wrists, hips, knees, ankles. In Figure 4, the skeletal mechanism is formed as a result of the combined structure of the edge pairs between the key points[12].

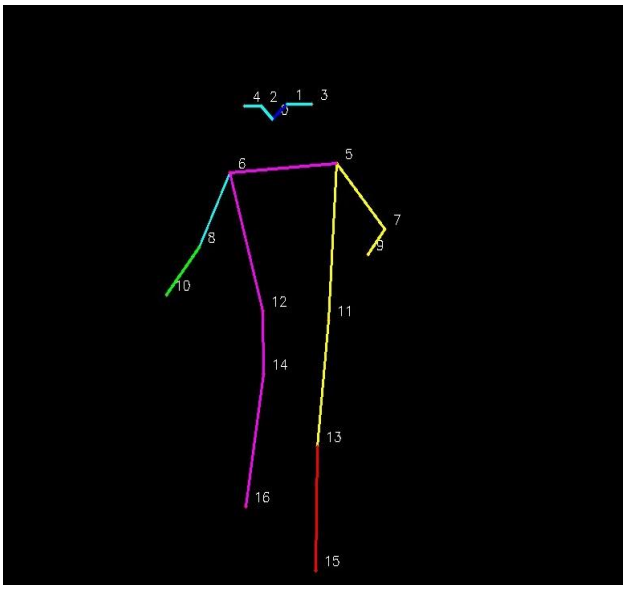

Figure 4. Key points identification and estimation methodology [35]

Human joint detection based on images is an example of symbolizing flexible modeling parts. Mixture modeling created to model joints is created with a flexible structure. Capturing is performed by establishing a relationship between the positions in the parts. After determining the relationship between the joints, templates are created[13].

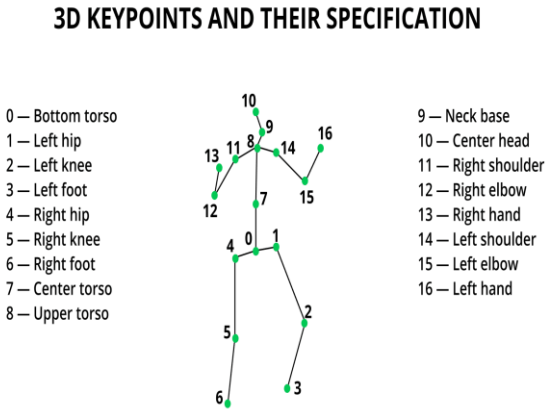


Figure 5. 3D Key Points [36]

In Figure 5, an efficient optimization joint is formed by afforestation. Everything about joints that appear local is based on spatial relations and the relations of formation of these joints.

**i) Receiving video footage**

The data containing the instant video recognition system is as in Figure 6. For this system, it measures the quality of the motion class by estimating the data of the new sequences in the data set containing the images. Categorizing the data determines the quality of the prediction by determining the class of motion prediction made.

Figure 6. Snapshot recognition dataset in football dataset

**ii) Gesture recognition based on conditions**

The first stage of the study is to remove the background outside the focused image in the images taken from the video inputs and to collect the necessary information to bring the area of interest to the foreground.

Then, the motion class is determined by filtering with the optical flow mechanism. Estimation is made by finding the key points in the gradient pattern joints, which are ordered from light to dark. Class is formed due to the differences in the human posture, the way the movements are made, the movement of the camera, the lighting, and the differences in the perspective effect. The abundance of information from its diversity provides a great advantage for this classification. Once accepted as input to an image, the direction of encoder motion is outlined by inferring convolution block features as a contraction. An architectural result is created thanks to the regressors connected to the coordinates. First of all, instead of coding the key points directly, it first creates a pose frame in that region[14]. The flow chart of the study is shown in figure 7.

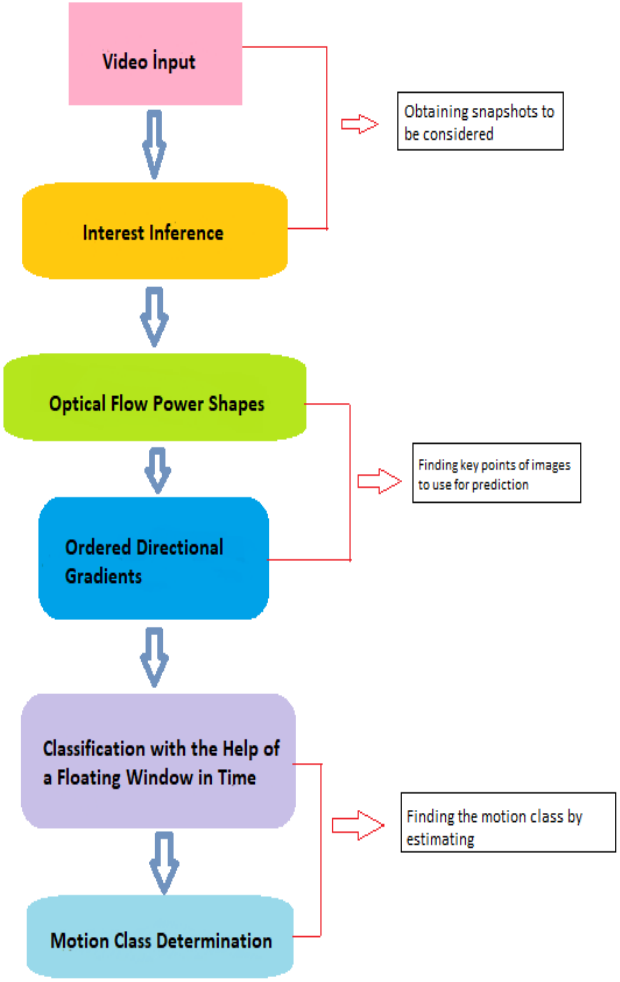


Figure 7. Estimation system structure diagram

**iii) Area of Interest Inference**

The portion of the video snapshot containing the human body silhouette is of interest to the estimation. The main factor in the inference focuses on the human body, regardless of the background. Converting color data to gray is the next step.

$$gray\ level\ value = [0.299\ 0.587\ 0.114] + [RGB]\text{^}T \qquad (1)$$

The region is accepted by combining the key points with the analysis of the coordinates of a part of the human body as a result of background subtraction. Background extraction, ambient light, brightness, external factors are minimized.

The biggest factor in taking the bounding boxes as a basis in the pose mechanism is that focusing on the human body to be estimated will leave the difficulties behind and make the estimation easier. In this system, background difficulties seem to have a minimal effect. Area of interest extraction with temporal windows is shown in Figure 8.


Figure 8. Extraction of interest in the data as a result of the analysis

**iv) Optical Flow Power Shapes**

The main purpose of the optical flow is to determine the strength of the joint parts in motion by performing the filtering process, instead of only detecting the direction of movement of the human body. Thus, the filtering process led to the disappearance of the information necessary for the prediction and ensured the protection of the necessary data. The pattern is created by combining the edge information of the moving images. When feature extraction is provided using the pattern, it provides an environment for the histogram that examines the transition features. It is obtained that the brightness in the image does not change depending on the temporal flow and does not change with time for each image[15].

Using the Hom-Schunk method, it is obtained (2) that the brightness in the image does not change depending on the temporal flow, and for each image, depending on the

(x, y, z) coordinates at time t.

$$(\frac{dI}{dt}) = 0 \tag{2}$$

The chain rule is applied after applying optical flow to the snapshot in the data set as in (3).

$$\left(\frac{dI}{dx}\right) * \left(\frac{dx}{dt}\right) + \left(\frac{dI}{dy}\right) * \left(\frac{dy}{dt}\right) + (\frac{dI}{dt}) = 0 \tag{3}$$

(u, v) shows the change in (x, y) coordinates in these diagrams.(4)

$$\left(\frac{dx}{dt}\right) = u, \left(\frac{dy}{dt}\right) = v \tag{4}$$

By arranging these inferences, the expression 5 is obtained.

$$I(x) * u + I(y) * v + I(t) = 0 \tag{5}$$

The inferences obtained above are not used alone on the properties of optical flow. Considering that the neighboring pixels of the base pixels will have similar flow movements, (x, y) coordinate components are included in the fluency constraint to form the solution of the system.

$$\left(\frac{d^2u}{dx^2}\right) + \left(\frac{d^2u}{dy^2}\right) = \Delta^{2u} \ , \left(\frac{d^2v}{dx^2}\right) + \left(\frac{d^2v}{dy^2}\right) = \Delta^2 v \tag{6}$$

In order to obtain optical flux values, minimum brightness and fluency constraints are required. It is required to have zero error for normal time. However, a brightness error occurs. To solve this, equation 7 is used in the background.

$$\varepsilon^2 = \iint \left(\alpha^2 \varepsilon^2(c) + \varepsilon^2(b)\right) d(x) d(y) \tag{7}$$

Optical flow powers were calculated by calculating the change in any direction in the coordinate plane by finding the amount of change with the method used. The edge features are determined more clearly, and the clear view of the human body posture is obtained without being dependent on the environment (8).

$$\varepsilon^2 = (u^2 + v^2)^{\frac{1}{2}} \tag{8}$$

**v) Ordered Directional Gradients**

After the optical flow power is determined, the part containing the motion information consists of these gradients. Extraction of the edge features of the human body in the image is obtained. Inferences were made with the HOG method. The HOG method finds out by recognizing the human body.

**vi) Motion Class Determination**

If HOG is used, when n windows are arranged, a successful analysis containing time and motion information in 81xn size is created. The HOG descriptor is very important in terms of the movements of the 81xn size descriptors, which are obtained by successful and highly ordered direction gradients, in the n-dimensional time window in 81xn size descriptors lined up one after the other. The system includes the concepts of time and motion[16]. Based on the speed concept of the real-time system, k-NN neighborhood is used. When using nearest neighbors , identifiers are generated and the nearest neighbors to the joints are identified to find key points [17]. When it captures similar features, it finds the closest option. The distance (m) between the data to be trained for the nearest neighbor (p) and the items to be classified (q) is calculated and the target is found.

$$((p1 - q1)^2 + (p2 - q2)^2 + \cdots + (pm - qm)^2)\text{^}(1/2) =$$

$$((\mathrm{m})\Sigma(\mathbf{i} = \mathbf{1})(pi - qi)^2)/(1/2) \quad (9)$$

The 81xn size identifiers show continuity in the motion detection part depending on the tags. With RGB and HOG ease of use motion detection, posture estimation is created[18].

**Results and Discussion**

The algorithms used in this study helped to solve the human pose estimation problem. The data set was collected from the videos of football matches, and consisted of Cristiano Ronaldo's image when he scored a goal. Tree modeling method was used and the area of interest is inferred. After the area of interest extraction, the parts called part poses were determined and the coordinates were determined. As a result of these stages, the joints are determined. Classification of joints by combining them revealed clearer results. Thus, limb points are provided. Each key point has a combination of right, left, up and down. The resulting system is the skeletal mechanism. When the systematic prediction is taken as 17 key points in the exposure mechanism, it was seen that the data in direct video images showed the best accuracy. The N-solid kinematic skeleton model has brought the forecasting quality to the highest level with the pipe system. Classification has been simplified. Ambient light, brightness in the snapshot and external factors are minimized. The concept of time was kept within the framework of importance and allowed the emergence of a successful mechanism. Cristiano Ronaldo's goal moments in the data set were examined from many angles, and it was determined in which positions he scored with the key point detection. In order to increase the accuracy in data sets, the accuracy level has reached a successful level thanks to the RGB camera when the common information is labeled in detail and matched. It has been observed that the RGB camera does not decrease the estimation accuracy. It has been observed that more accurate results are obtained when ear and eye detection is performed in data sets with specific movements. As the angle of the camera changed, the accuracy of the estimation result decreased with the decrease in key points. It has gained speed thanks to convolutional neural networks and the prediction mechanism has gained functionality with a minimum data set.

**Conclusion**

The Openpose system is aimed to realize a real-time system for 17 key points, with a total of 135 key points in the whole body in the images. Openpose has open-sourced the work in the library repository, enabling pipelines between command line interfaces Python API and Unity plugins and joints. Communities made easy with nvidia, cuda, opencl, cpu support. All joint positions were found for separation. It was determined which joints would come together for the body parts. Peak information was included for the (x, y) position. Quan Hua's deep learning model was used as a database for this posture prediction[31].

Figure 9. Human Exposure Estimation Results

| Real/Estimation | Front View | Right View | Left View | Back View |
|---|---|---|---|---|
| 0 | 92.7 | 78.4 | 77.3 | 79.7 |
| 1 | 90.6 | 72.5 | 73.6 | 73.6 |
| 2 | 91.3 | 73.7 | 74.8 | 75.3 |
| 3 | 89.5 | 79.5 | 78.5 | 78.5 |
| 4 | 87.6 | 77.4 | 79.5 | 74.6 |
| 5 | 90.2 | 73.5 | 91.7 | 91.2 |
| 6 | 89.3 | 90.4 | 70.3 | 88.3 |
| 7 | 90.5 | 79.4 | 93.4 | 91.6 |
| 8 | 89.6 | 91.2 | 90.6 | 87.5 |
| 9 | 90.9 | 70.2 | 92.8 | 86.8 |
| 10 | 89.9 | 89.9 | 85.4 | 89.5 |
| 11 | 90.3 | 75.4 | 92.4 | 91.3 |
| 12 | 89.6 | 90.1 | 71.1 | 88.2 |
| 13 | 90.7 | 74.1 | 90.6 | 88.9 |
| 14 | 89.8 | 90.8 | 72.0 | 87.8 |
| 15 | 91.1 | 81.2 | 93.2 | 92.3 |
| 16 | 90.1 | 91.2 | 70.7 | 89.2 |

Figure 10. Estimation Accuracy Rates

The table shown in Figure 10 contains the accuracy results of the estimation mechanism. Considering the key points in the data used, four positions was taken as a basis: front, back, right and left. By identifying 0-4 key points, other parts of the body was passed. Left body parts show less accuracy in right view, while right body parts show less accuracy in left view. In the rear view model, the key points covering the eyes and nose are less accurate.

**Ethics committee approval and conflict of interest statement**

There is no need to obtain permission from the ethics committee for the article prepared. There is no conflict of interest with any person institution in the article prepared.

## Authors' Contributions

Nupelda Kanpak and M. Ali Arserim : Estimation mechanism formation, Carrying out the estimation study of the data in the data set, writing the article

## References


[1] Parekh, P., & Patel, A. (2021). Deep Learning-Based 2D and 3D Human Pose Estimation: A Survey. In *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security* (pp. 541-556). Springer, Singapore.

[2] Souvenir, R., & Babbs, J. (2008, June). Learning the viewpoint manifold for action recognition. In *2008 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-7). IEEE.

[3] Yang, Y., & Ramanan, D. (2012). Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, *35*(12), 2878-2890.

[4] Wang, Y., Huang, K., & Tan, T. (2007, June). Human activity recognition based on r transform. In *2007 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-8). IEEE.

[5] Ramanan, D. (2006, December). Learning to parse images of articulated bodies. In *Nips* (Vol. 1, No. 6, p. 7).

[6] Lee, J., & Ahn, B. (2020). Real-time human action recognition with a low-cost RGB camera and mobile robot platform. *Sensors*, *20*(10), 2886.

[7] Tran, D., & Forsyth, D. (2010, September). Improved human parsing with a full relational model. In *European Conference on Computer Vision* (pp. 227-240). Springer, Berlin, Heidelberg.

[8] Weinland, D., Ronfard, R., & Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Computer vision and image understanding*, *104*(2-3), 249-257.

[9] Chang, M. C., Qi, H., Wang, X., Cheng, H., & Lyu, S. (2015). Fast Online Upper Body Pose Estimation from Video. In *BMVC* (pp. 104-1).

[10] Eichner, M., Ferrari, V., & Zurich, S. (2009, September). Better appearance models for pictorial structures. In *Bmvc* (Vol. 2, p. 5).

[11] Yang, Y., & Ramanan, D. (2012). Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, *35*(12), 2878-2890.

[12] https://mobidev.biz/blog/human-pose-estimation-ai-personal-fitness-coach

[13] Toshev, A., & Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1653-1660).

[14] https://nanonets.com/imagerecognition?&utm_source=nanonets.com%2Fblog%2F&utm_medium=blog&utm_content=How%20to%20Classify%20Fashion%20Images%20easily%20using%20ConvNets

[15] Sethi, S., Kathuria, M., & Kaushik, T. (2021). Face mask detection using deep learning: An approach to reduce risk of Coronavirus spread. *Journal of Biomedical Informatics*, *120*, 103848.

[16] Rosebrock, A. (2020). Covid-19: Face mask detector with opencv, keras/tensorflow, and deep learning. *Link: https://www. pyimagesearch. com/2020/05/04/covid-19-face-mask-detector-withopencv-keras-tensorflow-and-deeplearning*.

[17] https://github.com/prajnasb/observations

[18] Newell, A., Yang, K., & Deng, J. (2016, October). Stacked hourglass networks for human pose estimation. In *European conference on computer vision* (pp. 483-499). Springer, Cham.

[19] Quan Hua, "Human Pose Estimation in OpenCv" Link: *human-pose-estimation-opencv/LICENSE at master · quanhua92/human-pose-estimation-opencv · GitHub*

[20] Pons-Moll12, G., Taylor13, J., Shotton, J., Hertzmann14, A., & Fitzgibbon, A. (2013). Metric regression forests for human pose estimation. BMVC.

[21] Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., & Schiele, B. (2018). Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5167-5176).

[22] Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., & Schiele, B. (2018). Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5167-5176).

[23] Wang, J., & Payandeh, S. (2017). Hand motion and posture recognition in a network of calibrated cameras. *Advances in Multimedia*, *2017*.

[24] Remelli, E., Han, S., Honari, S., Fua, P., & Wang, R. (2020). Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6040-6049).

[25] Zhao, M., Li, T., Abu Alsheikh, M., Tian, Y., Zhao, H., Torralba, A., & Katabi, D. (2018). Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7356-7365).

[26] Sarafianos, N., Boteanu, B., Ionescu, B., & Kakadiaris, I. A. (2016). 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, *152*, 1-20.

[27] Rogez, G., & Schmid, C. (2016). Mocap-guided data augmentation for 3d pose estimation in the wild. *arXiv preprint arXiv:1607.02046*.

[28] Chen, W., Wang, H., Li, Y., Su, H., Wang, Z., Tu, C., ... & Chen, B. (2016, October). Synthesizing training images for boosting human 3d pose estimation. In *2016 Fourth International Conference on 3D Vision (3DV)* (pp. 479-488). IEEE.

[29] Fabbri, M., Lanzi, F., Calderara, S., Alletto, S., & Cucchiara, R. (2020). Compressed volumetric heatmaps for multi-person 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7204-7213).

[30] Liu, R. (2019). *Attention Based Temporal Convolutional Neural Network for Real-Time 3D Human Pose Reconstruction*. University of Dayton.

[31] Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2019). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence*, *43*(1), 172-186.

[32] Kang, X., Song, B., & Sun, F. (2019). A deep similarity metric method based on incomplete data for traffic anomaly detection in IoT. *Applied Sciences*, *9*(1), 135.

[33] Sengupta, A., Budvytis, I., & Cipolla, R. (2020). Synthetic training for accurate 3d human pose and shape estimation in the wild. *arXiv preprint arXiv:2009.10013*.

[34] Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., & Tian, Q. (2019). Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6569-6578).

[35] Sovit Ranjan Rath, et, al. "Human Pose Detection using PyTorch Keypoint RCNN." Machine Learning and Deep Learning , 2020

[36] Sarafianos, N., Boteanu, B., Ionescu, B., & Kakadiaris, I. A. (2016). 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, *152*, 1-20.