Human Activity Recognition Using Pose Estimation and Machine Learning Algorithm

Abhay Gupta, Kuldeep Gupta, Kshama Gupta and Kapil Gupta

National Institute of Technology, Kurukshetra Haryana, India

Abstract

Human Activity Recognition is becoming a popular field of research in the last two decades. 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. There are various methods present for activity recognition; every technique has its advantages and disadvantages. Despite being a lot of research work, recognizing activity is still a complex and challenging task. In this work, we proposed an approach for human activity recognition and classification using a person's pose skeleton in images. This work is divided into two parts; a single person poses estimation and activity classification using pose. Pose Estimation consists of the recognition of 18 body key points and joints locations. We have used the OpenPose library for pose estimation work. And the activity classification task is performed by using multiple logistic regression. We have also shown a comparison between various other regression and classification algorithm's accuracy on our dataset. We have prepared our dataset, divided it into two parts, one is used to train the model, and another is used to validate our proposed model's performance.

Keywords 1

Human Activity Recognition, Pose Estimation, Body Keypoints, Logistic Regression, OpenPose

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. 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.

Activity recognition through smartphones and wearable sensors is very common; there are various benchmarks available. But these systems rely on collecting data from sensors installed on the devices and user needs to wear these devices that are uncomfortable in practical. Vision-based systems are a better alternative for this kind of problem due to the fact that the user doesn't need to carry or wear any device. Instead, tools like a camera are

ISIC'21: International Semantic Intelligence Conference, February 25–27, 2021, New Delhi, India
EMAIL: abhaygupta190@gmail.com (A. Gupta);
ORCID: 0000-0001-8529-8085 (A. Gupta);


installed in the surrounding environment to capture data [1]. One of the popular vision-based HAR systems uses pose information. Poses have had remarkable success in human activity recognition, and researchers are widely using them in this problem these days. Poses provide useful information about human behavior. The concept is beneficial in various tasks such as HAR, content extraction, semantic understanding, etc. It uses a convolutional neural network (CNN) because they are much efficient in dealing with images. They are similar to traditional neural networks in that they consist of neurons with biases and learnable weights. In this study, we proposed a pose-based HAR system that overcomes the issues we discussed in the smartphone and wearable sensor approach. We extract human pose (18 body keypoints location in the two-dimensional plane) from images using OpenPose library, which internally uses CNNs. Finally, activity is classified using the pose information through a supervised machine learning algorithm.

The rest of the paper is structured as in section 2 literature survey of some selected research papers in the area is mentioned. Section 3 contains the methodology and architecture of the proposed approach. A brief description of dataset and evaluation metrics (like precision, recall, and f1-score) used in this work is given in sections 4 and 5, respectively. Section 6 contains experiments and results of various classification algorithms applied in this work. Section 7 finally concludes the work with some future direction.

2. Related Work

Research has recently begun to recognize the behavior of humans from the images. Compared to the video-based action classification, the number of research papers and journals are less. We have stated some techniques used for HAR. Four types of approaches address classifications of actions, including image structure-based methods, pose-based systems, model-based approaches, and example-based methods. The pose-based method trains each pose using an annotated 3D image [2]. The model-based method uses a known parametric body model to match posture variables [3]. The example-based model uses classical machine learning algorithms to find actions in some image properties [4]. In the method based on image structure, the posture's representation is considered as features to the classification of the action [5].

[6] detected daily living activities by preprocessing the data collected from the Microsoft Kinect motion-sensing device for minimizing the error produced by the system and subject. [7] proposed a new approach to activity recognition by simultaneously extracting features from objects used to perform the activity and human posture. [8] applied openpose and Kalman filter to track the target body, and then a one-dimensional full CNN is used for the classification of activity.

Moreover, a single person activity can also be recognized by using smartphone sensors and wearable sensors; the smartphone-based approach uses sensors that are inbuilt in the device, such as accelerometer and gyroscope, to identify activity, whereas the wearable sensor-based approach requires the sensors to be attached on the subject body to collect action information. [9] used several machine learning algorithms (SVM, KNN, and Bagging) and collected data from smartphones' accelerometers and gyroscope sensors, and detected six different activities. [10] recognize human activity using an accelerometer and gyroscope sensor, which is mounted on humans, and used various machine learning algorithms such as KNN, Random Forest, Naïve Bayes, and detecting three different activities. [11] collect data from the smartphone and smartwatch and used a five-fold cross-validation technique to detect five upper limb motions. [12] used wearable and smartphone-embedded sensors for detecting six dynamic and six static activities using a machine learning algorithm. [13] applied Deep learning and convolutional neural network to recognize the body's actions on data retrieved from smartphone sensors.

3. Proposed Approach

Our approach to 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 such as logistic regression, support vector machine, decision

tree, etc. Figure 1 shows the architecture of the proposed approach.

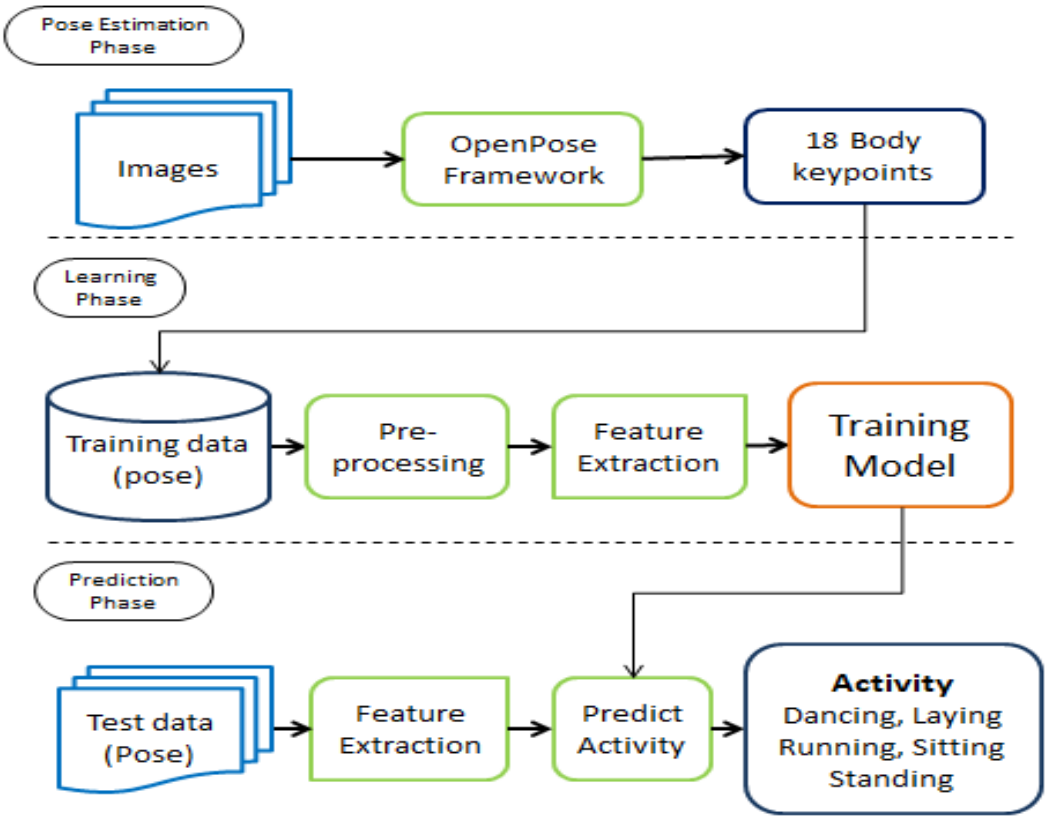

Figure 1: Proposed Architecture

3.1. Human Pose Estimation

Human Pose Estimation is the task of extracting the body's skeletal key points and joints locations corresponding to the human body parts. It uses all those key points and joints to associate the two-dimensional structure of the human body. In this work, we have used the OpenPose framework for estimating the pose from the input image.

In OpenPose, the image is sent over the CNN output network to get the features from input. The feature map is then processed in the multi-stage CNN sequential layers to generate (PAF) Part Affinity Fields and Confidence Map. The Partial Affinity Fields and Confidence map generated above are passes through a bipartite graph matching algorithm to capture human posture in the image. Figure 2 shows the OpenPose pipeline.

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.

$$L = (L1, L2, L3 \ldots. Lc) \quad (1)$$

$$Lc \in Rw*h*2$$

$c \in \{1 \ldots C\}$, 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.

3.1.2. Confidence Map

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

$$S = (S1, S2, S3 \ldots. Sj) \quad (2)$$

$$Sj \in Rw*h$$

$j \in \{1 \ldots J\}$, 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 through OpenPose dependent upon the dataset has been trained.

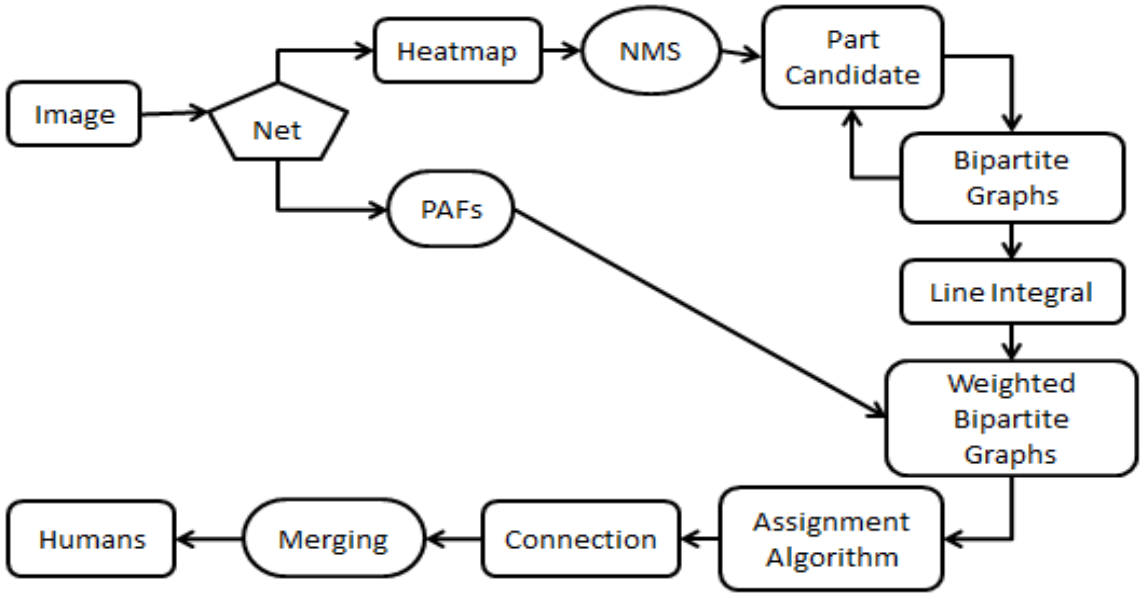

Figure 2: OpenPose Pipeline

In this work, the COCO dataset having 18 different body key points (see Figure 3) 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.

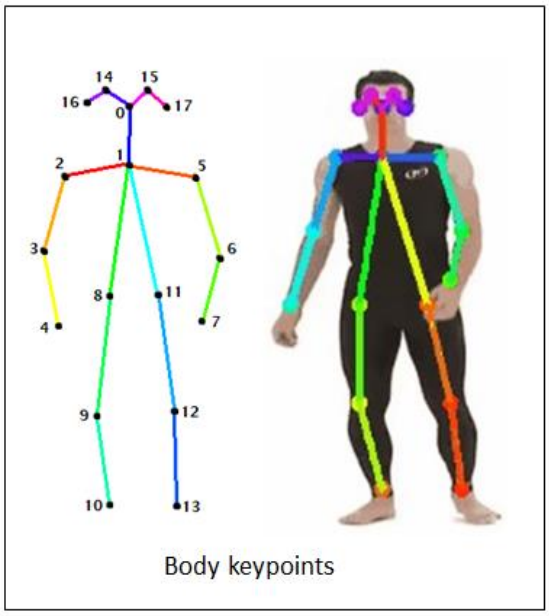

Figure 3: OpenPose Keypoints

3.2. Activity Classification

We formulate the activity classification problem as a multiclass classification problem, which can be modeled using various machine learning regression and classification algorithms. The classification algorithm takes 18 body keypoints (x-axis and y-axis values of each point) as input for our model's training and testing. We used a supervised learning approach as our dataset contains body keypoints with an activity label. Among all the algorithms, we use multiple logistic regression, and random forest provides significantly greater accuracy.

4. Dataset

OpenPose uses the COCO keypoints detection dataset for the pose estimation task, which contains more than 200K images labeled with keypoints [14]. We have collected images from Google for classification purposes, and some photos are clicked by a smartphone camera. We prepared our dataset on approximately 1000 images (five activity categories, namely, sitting, standing, running, dancing, and laying). Each activity category has more than 170 images. We divided our dataset into training and testing in the ratio 90:10. Data collected from different sources contain unequal width and height images, while our model requires the same width and height. We have resized all images to fixed-size 432x368 pixels, and then key points are extracted from them.

5. Evaluation Metrics

For performance evaluation, Recall, Precision, and F1-score are used in this

experiment. We have also shown the Confusion Matrix of some classifiers.

5.1. Precision

Precision(P) is the ratio of the number of true positives (Tp) to the sum of false positives (Fp) and true positives. It can also be defined as how many images classified into this class belong to this class.

$$P = Tp/(Tp+Fp) \quad (3)$$

5.2. Recall

Recall(R) is the ratio of the number of true positives (Tp) to the sum of false negatives (Fn) and true positives. It can also be defined as how many images that belong to this class are classified into this class.

$$R = Tp/(Tp+Fn) \quad (4)$$

5.3. F1-Score

F1-Score is calculated as the harmonic mean of recall and precision. Eqs.5 calculates it.

$$F1\text{-}Score = 2\ (P*R)/(P+R) \quad (5)$$

5.4. Confusion Matrix

It is a two-dimensional matrix used to measure the overall performance of the machine learning classification algorithm. In the matrix, each row is associated with the predicted activity class, and each column is associated with the actual activity class. The matrix compares the target activity with the activity predicted by the model. This gives a better idea of what types of errors our classifier has made.

6. Experiments and Result

The following five activities are considered for pose estimation and activity recognition and classification: sitting, standing, dancing, laying, and running. The experiments are conducted in Scikit Learn (0.23.1) and Python (3.6.6) in Windows 10 Operating System with Intel i5 Processor 3.40 GHz with 8 GB RAM and using five classification algorithms for activity classification. These algorithms are described below with their confusion matrix. The performance results are provided in Table 1, which shows the recall, precision, and f1-score of various classifiers used in the proposed approach.

6.1. Classification Algorithm

6.1.1. Logistic Regression

This algorithm is based on supervised learning, and it is used in classification problems. In this work, multiple logistic regression is used for classifying activities, and 'sag' is used as a solver because it solves only L2 regularization with primal formulation or no regularization and Uses dummy variables to represent the categorical outcome.

Table 1
Performance Evaluation on Different Classifier

Algorithms	Precision (%)	Recall (%)	F1-Measure (%)
Logistic	80.72	81.47	80.95
KNN	77.90	77.89	77.12
SVM	80.43	81.14	80.46
Decision Tree	74.49	75.80	73.50
Random Forest	80.75	80.34	79.43

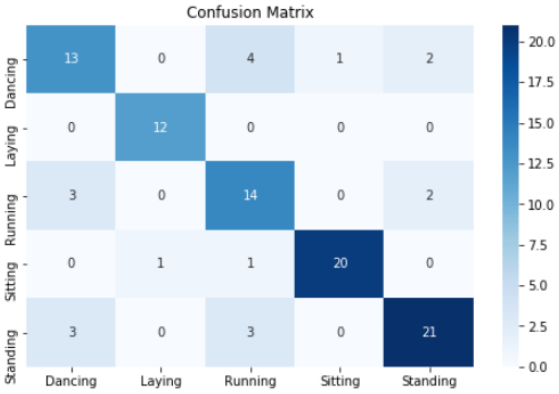

Figure 4:Confusion Matrix (Logistic Regression)

6.1.2. K-Nearest Neighbors

K-nearest neighbor (KNN) is a supervised machine learning algorithm used for classification, and it's a non-parametric, lazy algorithm. Despite this simplicity, we got very competitive results that are one reason for using this algorithm in our work. We used different values for k and got the highest accuracy in 5. The distance function(d) used in this algorithm is given in Eqs.6 and for the confusion matrix (see Figure5).

$$d(p,q) = \sqrt{\sum (qi - pi)2} \qquad (6)$$

where p, q are vectors containing keypoints of two different images and i=1 … n.

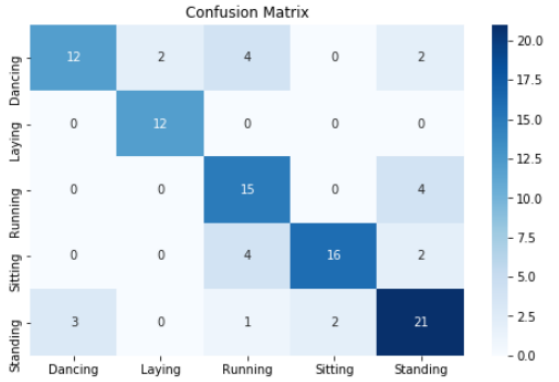

Figure 5: Confusion Matrix (KNN)

6.1.3. Support Vector Machine

It also comes under supervised learning algorithms and is mainly used in classification and regression problems. We plotted all available data as points in two-dimensional space. The classification is done by finding a hyperplane that provides similarly different outputs between the two classes. The confusion matrix is provided in Figure6.

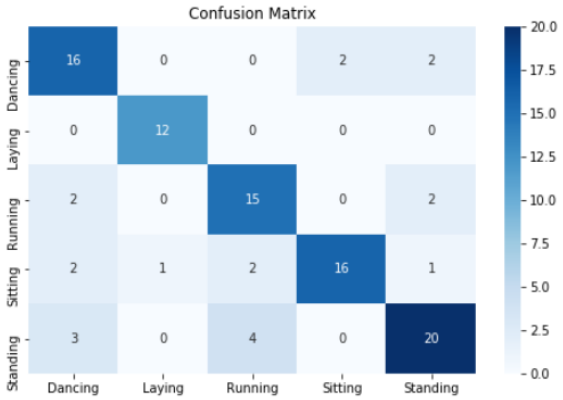

Figure 6: Confusion Matrix (SVM)

6.1.4. Decision Tree

The decision tree comes under supervised learning. It is the most powerful and accepted tool for prediction and classification. This algorithm uses learning to predict a target pose's activity and make decisions from previously trained data. Predictions for activities are made from the root of the tree. The record attribute value is compared to the root attribute value. The confusion matrix is given in Figure7.

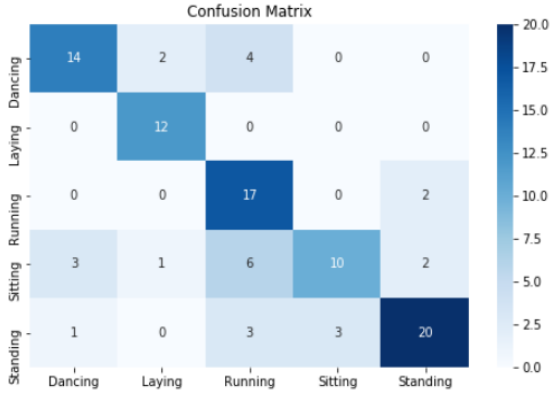

Figure 7: Confusion Matrix (Decision Tree)

6.1.5. Random Forest

Random decision forest is a supervised learning algorithm, and it is an ensemble learning method for classification and regression. It is also one of the most used and popular algorithms because it gives better results without tuning hyper-parameter. It creates multiple decision trees and selects the best solution using voting. We use a random forest because it predicts activity with good accuracy and runs efficiently even for big

datasets. The confusion matrix is shown in Figure8.

7. Conclusion

In this study, we proposed our 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 supervised machine learning algorithm. We prepared our dataset for this work, which contains five different activities, viz. sitting, standing, laying, dancing, running. We have used five algorithms (Logistic Regression, SVM, KNN, Random Forest, and Decision Tree) to find better results for our model. From our experiment results, we observed that Multiple Logistic Regression, SVM, and Random Forest are showing the highest accuracy of 80.72%, 80.43%, and 80.75%, respectively, and the other two algorithms KNN and Decision Tree, are underperforming. We have shown accuracies of some recent researches on HAR in table 2.

Although much research has already been done to a certain extent to deal with the activity recognition problem, more convincing actions must be taken. In practice, there are a lot of different activities that humans use to perform in everyday life. Detecting all of them isn't an easy task because it requires a very large dataset to train the model. Although the dataset is not the only problem, 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. Besides adding activities, we can apply some data preprocessing techniques for handling missing keypoints of the body. We can also experiment with some other machine learning algorithms that can provide better results.

Table 2
Comparative Study

S.No.	Authors and Year	Dataset	Activities	Model Used	Accuracy (%)
1.	Nandy et al., 2019 [12]	Accelerometer and heart rate sensor	Walking, climbing stairs, sitting, running	Multilayer Perceptron	77.0
				Linear Regression	53.92
				Gaussian Naïve Bayes	73.73
				Decision Tree	93.54
2.	Ghazal et al., 2018 [15]	Images from the internet	Sitting on the chair or ground	Decision-making algorithm with feedforward CNN	95.2
3.	Gatt et al., 2019 [16]	COCO keypoints	Abnormal activity such as fall detection	Used pre-trained models of PoseNet and OpenPose	93

Figure 8: Confusion Matrix (Random Forest)

8. References

[1] A. Gupta, K. Gupta, K. Gupta and K. Gupta, "A Survey on Human Activity Recognition and Classification," 2020 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2020, pp. 0915-0919, doi: 10.1109/ICCSP48568.2020.9182416.

[2] G. Sharma, F. Jurie and C. Schmid, "Expanded Parts Model for Human Attribute and Action Recognition in Still Images," 2013 IEEE Conference on

Computer Vision and Pattern Recognition, Portland, OR, 2013, pp. 652-659, doi: 10.1109/CVPR.2013.90.

[3] J A. Gupta, A. Kembhavi and L. S. Davis, "Observing Human Object Interactions: Using Spatial and Functional Compatibility for Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 10, pp. 1775-1789, Oct. 2009, doi: 10.1109/TPAMI.2009.83.

[4] Yang Wang, Hao Jiang, M. S. Drew, Ze-Nian Li and G. Mori, "Unsupervised Discovery of Action Classes," 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 2006, pp. 1654-1661, doi: 10.1109/CVPR.2006.321.

[5] J. Shotton et al., "Realtime human pose recognition in parts from single depth images," CVPR 2011, Providence, RI, 2011, pp. 1297-1304, doi: 10.1109/CVPR.2011.5995316.

[6] B. M. V. Guerra, S. Ramat, R. Gandolfi, G. Beltrami and M. Schmid, "Skeleton data preprocessing for human pose recognition using Neural Network*," 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 2020, pp. 4265-4268, doi: 10.1109/EMBC44109.2020.9175588.

[7] B. Reily, Q. Zhu, C. Reardon and H. Zhang, "Simultaneous Learning from Human Pose and Object Cues for Real-Time Activity Recognition," 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 2020, pp. 8006-8012, doi: 10.1109/ICRA40945.2020.9196632.

[8] H. Yan, B. Hu, G. Chen and E. Zhengyuan, "Real-Time Continuous Human Rehabilitation Action Recognition using OpenPose and FCN," 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Shenzhen, China, 2020, pp. 239-242, doi: 10.1109/AEMCSE50948.2020.00058.

[9] E. Bulbul, A. Cetin and I. A. Dogru, "Human Activity Recognition Using Smartphones," 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technolo-gies (ISMSIT), Ankara, 2018, pp. 1-6, doi: 10.1109/ISMSIT.2018.8567275.

[10] R. Liu, T. Chen and L. Huang, "Research on human activity recognition based on active learning," 2010 International Conference on Machine Learning and Cybernetics, Qingdao, 2010, pp. 285-290, doi: 10.1109/ICMLC.2010.5581050.

[11] K. -S. Lee, S. Chae and H. -S. Park, "Optimal Time-Window Derivation for Human-Activity Recognition Based on Convolutional Neural Networks of Repeated Rehabilitation Motions," 2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR), Toronto, ON, Canada, 2019, pp. 583-586, doi: 10.1109/ICORR.2019.8779475..

[12] A. Nandy, J. Saha, C. Chowdhury and K. P. D. Singh, "Detailed Human Activity Recognition using Wearable Sensor and Smartphones," 2019 International Conference on Opto-Electronics and Applied Optics (Optronix), Kolkata, India, 2019, pp. 1-6, doi: 10.1109/OPTRONIX.2019.8862427..

[13] R. Saini and V. Maan, "Human Activity and Gesture Recognition: A Review," 2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3), Lakshmangarh, Sikar, India, 2020, pp. 1-2, doi: 10.1109/ICONC345789.2020.9117535.

[14] Z. Cao, T. Simon, S. Wei and Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 1302-1310, doi: 10.1109/CVPR.2017.143.