www.cadanda.com
Computer-Aided Design™
AND APPLICATIONS

Human Joint Profile Extraction using Deep Learning Approaches

Miri WeissCohen[1], Andrea Vitali[2], Daniele Regazzoni[2]

[1] Braude College of Engineering, miri@braude.ac.il
[2]University of Bergamo, andrea.vitali1@unibg.it, daniele.regazzoni@unibg.it

Corresponding author: Miri Weiss Cohen, miri@braude.ac.il

Abstract. Digital human modeling and gait analysis are essential for improving hip replacement surgery (HRS). In this study, Convolution Neural Networks (CNN) are used as a machine learning method to extract the most accurate stick-model from videos captured on a simple camera to represent gait and body components. We developed and tested multiple approaches to create an equitable skeleton model from an image. This process consists of two main parts: defining the joint locations using a CNN network in different architectures, and defining the connections into the final skeletons. A CNN has been trained, validated, and tested using the OpenPose software, which combines two different networks that have been tested on three data-sets for learning and evaluation. The results were satisfactory, but MobileNetV1 was evaluated for optimization of OpenPose computations and definitions. Several hyper-parameters were investigated to provide better representations. As a result of utilizing OpenPose methodology in conjunction with heavily optimized network design and post-processing code, and implementing MobileNet, the proposed solution has provided improved accuracy ratios.

Keywords: 2D Joint profile, skeleton feature extraction, Deep Learning, CNN
DOI: https://doi.org/10.14733/cadaps.2023.704-715

1 INTRODUCTION

Osteoarthritis of the hip joint (coxarthrosis) is the most common disease of the hip joint in adults. Despite advances in research on osteoarthritis, no known treatment can stop its progression. It is universally recognized that prosthetic hip replacement surgery (HRS), also known as Total Hip Arthroplasty (THA), is the most effective treatment for hip osteoarthritis. Numerous studies have evaluated the therapeutic success of total hip arthroplasty (THA) using clinical controls, functional evaluations, and radiographic controls in accordance with the established protocols [22]. In particular, the recovery in the immediate postoperative period is an interesting research topic in relation to the ability of the patient to resume normal walking activity quickly after surgery. Furthermore, the short-term recovery for these patients is also important for hospital management.

Fast-track recovery paths have been developed to reduce the time spent in the hospital in recent years [20]. Currently, healthcare centers lack the knowledge of quantitative methods for assessing patients' gait patterns, and there is evidently a need for methods to determine objectively how patients are performing. For the problem at hand, there are no diffused methods for comparing the pre- and postoperative conditions.

In traditional observational approaches, data are collected exclusively through qualitative measures (such as VAS scores and Barthel scores), and no objective information on the patient's walking improvement is available [5, 10]. The current state of objective assessment methods is inadequate for obtaining quantitative and measurable parameters for comparing preoperative and postoperative conditions in the same patient. The combination of objective data and medical information may provide an indication of the level of functional recovery after THA, especially during the immediate postoperative period [19].

Multidisciplinary approaches are necessary to determine which gait parameters are appropriate, when they must be considered, and how to interpret the results. In order to determine the ranking of the different surgical procedures, a gait analysis is conducted based on the acquired data [17, 11], for providing a feedback to surgeons and healthcare providers. It is very crucial for the patient's short recovery to evaluate the performance of his gait and essential for assessing the success of hip arthroplasty [21].

Digital human modeling and gait analysis are key factors in moving towards a better surgical approach in healthcare regarding hip replacement surgery (HRS). Moreover, to gain an understanding about the patient's recovery process as well as to evaluate the gait of the person before and after surgery, a skeleton' if accurate can be used. As with gait analysis of patients with THA, pose detection of humans requires greater accuracy for use as a medical tool for assessing patients' recovery.

In order to perform the objective analysis and to find a set of key parameters to identify the difference between pre and post-operative gait performance, motion capture (mocap) systems are an excellent solution. The mocap system can create a virtual avatar that follows a person's movements. Using mocap systems in the medical field opens up new possibilities for tracking body movements, such as gait, posture, and any other gesture, in a specific context at hand.

Marker-based mocap systems are used in research studies where the gap between preoperative and postoperative mocap acquisitions is greater than six months [9]. However, marker-based solutions are costly and require a dedicated staff and environment to be used correctly [15]. Another mocap solution is the marker-less mocap system, whereby a device is placed around the zone where the patient will walk rather than being worn along the body while the patient walks. Marker-less mocap systems use two types of sensors: RGB and RGB-D. The Microsoft Kinect is the most widely used RGB-D sensor to design semi-automated methods for motion capture without wearing markers. Marker-less mocap systems are generally more affordable and easier to use than marker-based solutions. As for gait analysis, marker-less mocap systems have been widely adopted in recent years, and the measured accuracy of these systems has been evaluated well enough for the purposes of assessing human movements during several rehabilitation processes, such as shoulder, knee and hip [8, 23].

In spite the fact that a marker-free solution is considered a good tool for human motion analysis, several research studies have identified significant limitations associated with non-frontal tracking and body occlusions [16]. In addition, RGB-D sensors may not be readily available for daily use in hospitals; for example, the KinectV2 device has been discontinued since 2017.

In this study, we propose to monitor a patient's movement during assessment using a simple video camera. Creating performance scores and metrics requires elucidating the data, which depends on the context and the performance being evaluated. A machine learning method of Convolutional Neural Networks (CNN) is used for determining the most accurate skeletons from video taken with a camera. By analyzing gait using simple camera technology, gait analysis can be added to clinical evaluations and provide a quantitative measure of recovery within a short period of time following surgery.

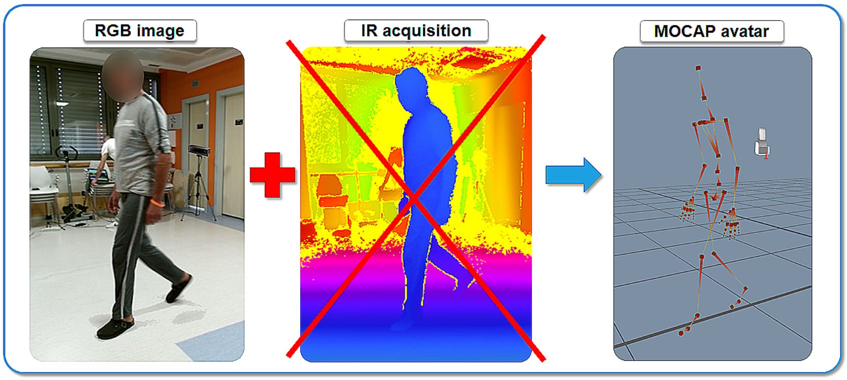

Figure 1: Proposed gait evaluation process

2 SKELETON FEATURE EXTRACTION

Using a visual input, a human body representation (such as a skeleton pose) is built based on the location of human body parts. Pose estimation is much more accurate using human body modeling. Visual input data can be used to extract feature points and keypoints using it. Model-based approaches are often used to describe and infer human body poses, as well as render 2D and 3D poses. In previous studies [6] a Kinect sensor was used to extract a skeleton model in the form of a compact representation of major points on the body such as the head, shoulders, elbows, hands, hips, knees, and feet. Despite the Kinect's ability to estimate skeletons, it has been limited in its use for clinical monitoring in spite of its built-in solution.

Several approaches have been taken to retrieve skeleton models from an image [1]. By training a CNN network, we first track the joint location, and then we transform this information into the final skeleton [11]. The logic is straightforward, where each connection sharing the same part detection candidate, is assembled together. Essentially, if two connections share the same part, they are merged. Finally, a set of human sets is constructed, with each human containing an index, relative coordinates, and a score. Recent reviews of pose estimation methods are found in [24, 3, 18] Initially, the results were not promising and the predictions were inaccurate and certainly not accurate for medical purposes. This problem can be partially resolved by extracting a skeleton using OpenPose software [2]. Two CNN networks are trained in parallel, one for tracking the joints and the other for tracking the limbs. The limbs encode the degree of association between parts, which is absent from the previous existing models . Adding this missing component greatly enhances accuracy. To prevent further medical complications and to accurately assess patients' situations in the medical field, human motion recognition needs to be performed at an extremely high level of precision avoiding any errors or miscalculations. With deep learning technologies such as CNNs, similar tasks such as human motion recognition and classification can be accomplished [25]. A variety of approaches were employed to reconstruct skeleton models from images. First, we tracked joint location by training a CNN network, then we transformed these detected connections into the final skeleton. We followed a straightforward procedure. Connections with the same part detection candidates are grouped together. For example, if two connections have the same part, they are merged into one. The final step is to construct a set of human sets, with each human representing a collection of components, each component containing its index, coordinates, and score. For medical purposes, which require extreme precision, the results and predictions were not accurate. See Figure 2 for an example of very poor results.

OpenPose [13] uses the first few layers of an image to extract features from it. The extracted features are then input into two separate divisions of a convolutional network in parallel. Based on the first division, a set of 18 confidence maps is predicted for each of the 18 parts of the human pose skeleton. During the next branch

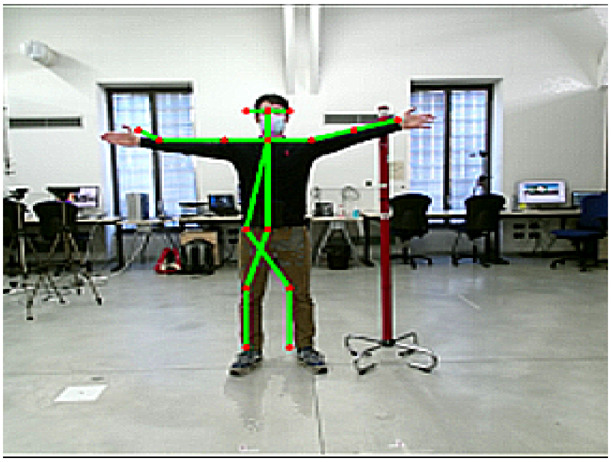

Figure 2: Basic Open-Pose results

of analysis, 38 Part Affinity Fields (PAFs) are predicted, which help us to determine the degree of association between the different parts. Later in the process, the branches are used to clean up the predictions, and based on the confidence maps, bipartite graphs are generated between pairs of parts. Furthermore, PAF values are used to prune weak links from bipartite graphs. Following all of the above steps, human pose skeletons can be calculated and assigned to each individual in the picture.

In Figure 3, the two-branch, multi-stage architecture of CNNs is depicted. Two-dimensional confidence maps (top branch) show the locations of body parts. There are total 19 confidence maps, 18 of which are body parts, and one is the background. A set of part affinity field maps (bottom branch) encodes the 2D vectors for a given limb. 38 PAF maps contain the x and y directions for a given limb (total of 19 limbs). The image will be processed through a convolutional network in order to generate a set of feature maps. In an iterative prediction architecture, the feature maps will be refined at successive stages. At each stage, part affinity fields and detection confidence maps are generated and refined. Each heat-map shows the confidence level for a

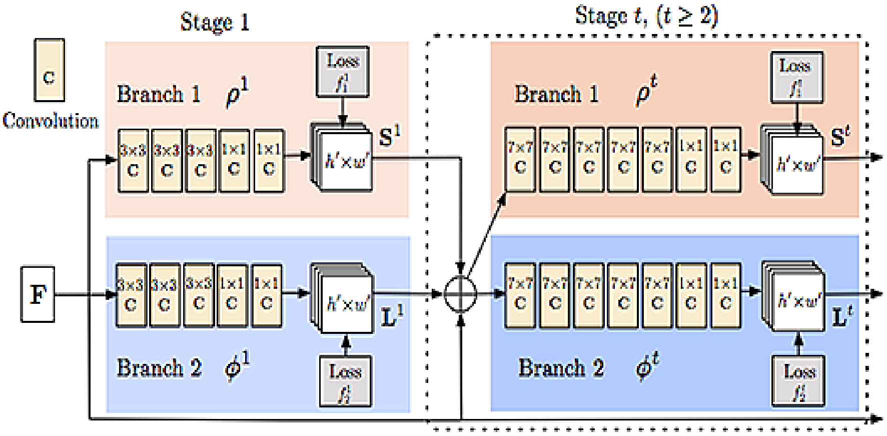

Figure 3: Two-branch, multi-stage architecture of CNNs [2]

specific part of the image. In order to transform confidence into certainty, a local maximum greater than a predetermined threshold needs to be applied. The next step is to connect the pairs, which is a well-known problem in graph theory. Calculated by connecting the candidate part locations on a line segment, in which possible associations are assigned confidence scores based on the strength of the association. Thereafter, the line integral is approximated by sampling and summing uniformly spaced values. In the final stage of the process, the predictions from the two branches are concatenated with the image features.

By using OpenPose, we have compared the data-sets to determine which data-set would be most applicable to the case at hand. To evaluate the quality of the solution, we will use Figure 4 as an example.

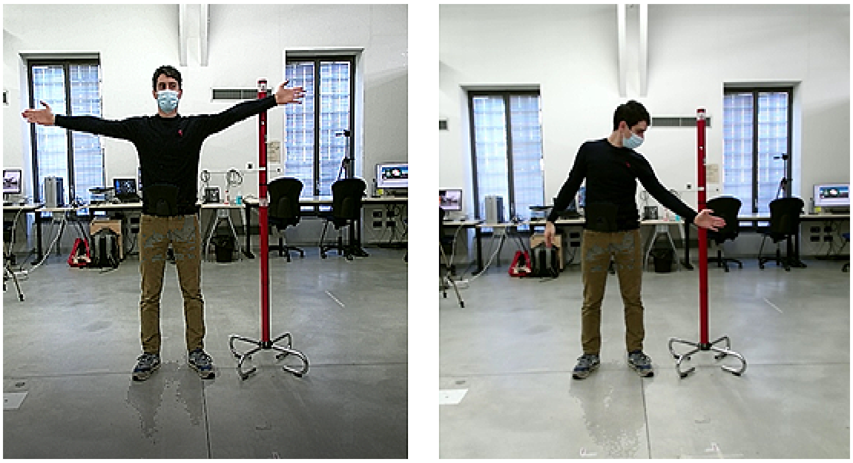

Figure 4: Examples of test cases

MPII data-set training We used the MPII Human Pose data-set [14] in order to examine the accuracy of articulated human pose estimation. This data-set contains 25K images of over 40K people, with their 17 joints annotated. The images were systematically collected based on a taxonomy of everyday human activities. According to Figure 5, errors in the definition of joints have a significant impact on the resulting skeleton. It was not possible to construct a correct skeleton with the OpenPose because some joints were defined outside the head in the image.

COCO data-set training One of the most popular tools for training deep learning algorithms is the COCO database, which is an open source object recognition database. The database contains 250000 images of people with 17 joints [12, 4]. Figure 6 depicts the results of training the OpenPose on the data-set. The results can clearly be seen to be better than when training on MDII; however, looking at the right figure, one can see that the joints in this pose are still outside the image. Such errors remain challenging to correct.

Body-25 data-set training The COCO data-set and foot keypoints are combined in Body-25 to produce skeletons with 25 keypoints. In Figure 7 it is clearly seen that a demi-skeleton has been found within an object (the right hand side of the image) that is unrelated to the pose.

3 PROPOSED APPROACH

The main challenge of this study is two-fold, first, tracking movements and identifying joints from video data, which makes the use of affordable devices a valuable tool. Furthermore, to develop a software solution to determine a patient's gait by defining a skeleton in an accurate and appropriate manner.

In this work 30 training videos were used, with patients being recorded. Due to the lack of accurate data in the input videos, each video has been divided into single frames to improve accuracy and key points detection

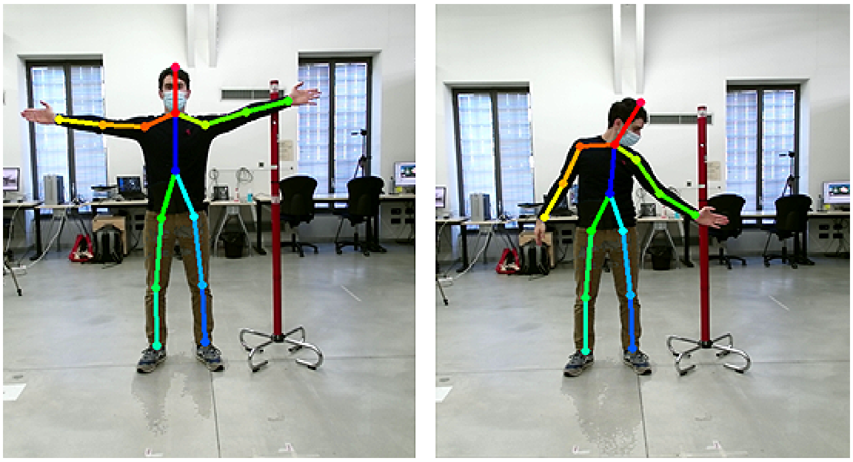

Figure 5: MPII results

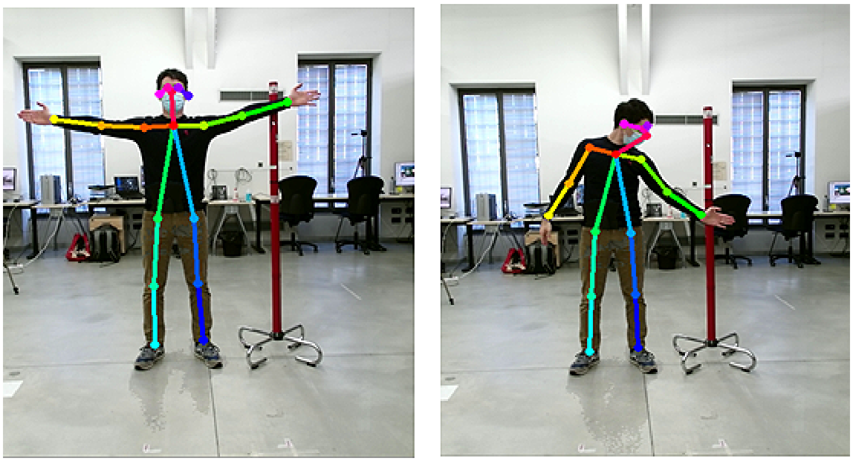

Figure 6: COCO results

in the human figure. It was imperative that each frame of the video be encapsulated by itself to achieve the highest level of (x, y) coordination, so dividing the process was essential. Moreover, the data must also be coordinated as precisely as possible since they are used for medical purposes, such as estimating an operation's success or tracking a patient's healing process after surgery.

In order to correct the faults found in using OpenPose, we propose to use MobileNetV1 to optimize OpenPose results [7]. It was chosen because of its computational efficiency and the ability to run the network on a CPU rather than a GPU. Using depth-wise separable convolutions, MobileNets are based on a lightweight architecture for building deep neural networks, which uses an inverted residual structure for shortcut connections between the thin bottleneck layers. Layers of intermediate expansion are lightweight depth-wise convolutions that serve to keep the representation intact Table 1 describes a partial description of MobileNetV1 and its ConV layers.

These sequential models can be built layer by layer for the vast majority of problems, but they cannot share layers or have multiple inputs and outputs. However, this sequential model is the most suitable for our problem. It is evident that the sequential model described above facilitates layer-by-layer creation of models, but it is

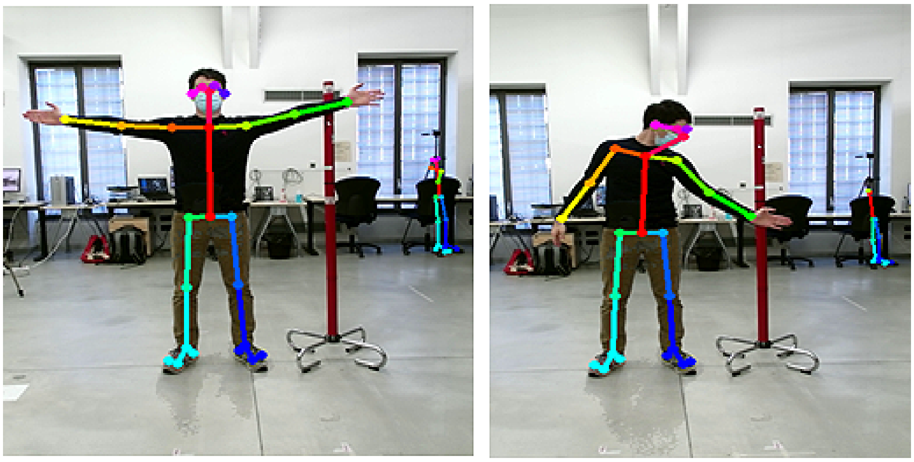

Figure 7: Body25 results

limited in that it is not able to handle models with more than one input or output.

The image is first passed through a convolutional network to generate a set of feature maps. The feature maps are refined iterative, and the predictions become more accurate as more stages are added. In the multistage CNN used in this work, two maps are predicted: A confidence map (heat-map) and a Part Affinity Field (PAF)map. The heat-maps provide the level of confidence of each part on an image, which is then tested for the local maximum that exceeds a predefined threshold. As an example: Heat map indexes to find each $limb : limb_type = 0$ indicates the nose, as well as PAF indices containing the $xandy$ coordinates of the limb. Figure 8 depicts partial heat map results. It can be seen that the left elbow, the hip, ankles etc. are visualized. The line integral of the corresponding Part Affinity Fields, can be calculated along the line segment connecting the candidate part locations. According to the scores assigned to each possible connectivity, we can estimate the confidence we have in their relationship by sampling and summing values evenly spaced out. Figure 9 illustrates the results of converting confidence into certainty. Since there are several detected

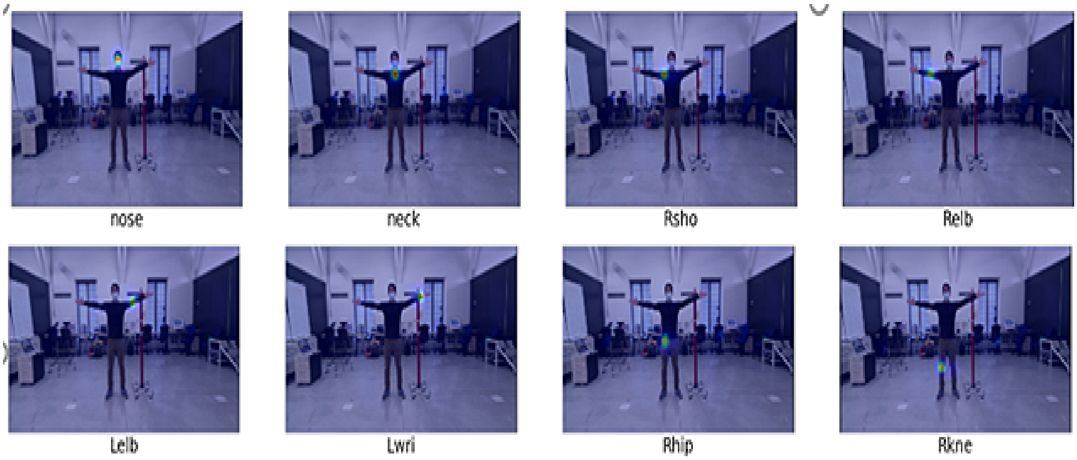

Figure 8: Partial Heatmap peak information and Part Affinity Field (PAF) results

Table 1: Set of parameter values for ILS/D algorithm calibration

Sequential(i)	Cov/layer
$Seq - 0 : 0$	Conv2d(3, 32, kernel$_size = (3,3), stride = (2,2), padding = (1,1), bias = False$)
$Seq - 0 : 1$	BatchNorm2d(32, eps=1e-05, momentum=0.1, affine=True, track$_running_stats = True$)
$Seq - 0 : 2$	ReLU(inplace=True)
$Seq - 1 : 0$	BatchNorm2d(32, eps=1e-05, momentum=0.1, affine=True, track$_running_stats = True$)
$Seq - 1 : 1$	BatchNorm2d(32, eps=1e-05, momentum=0.1, affine=True, track$_running_stats = True$)
$Seq - 1 : 2$	ReLU(inplace=True)
$Seq - 1 : 3$	Conv2d(32, 64, kernel$_size = (1,1), stride = (1,1), bias = False$)
$Seq - 1 : 4$	BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True, track$_running_stats = True$)
$Seq - 1 : 5$	ReLU(inplace=True)

candidates for each part definition (each containing the x, y, coordinate, a score, and an ID), the heatmap is pre-processed using the Gaussian filter. The maximum peaks are found, normalized, and a score is calculated. A unique ID is assigned to each detection. These detected connections are merged into the final skeletons

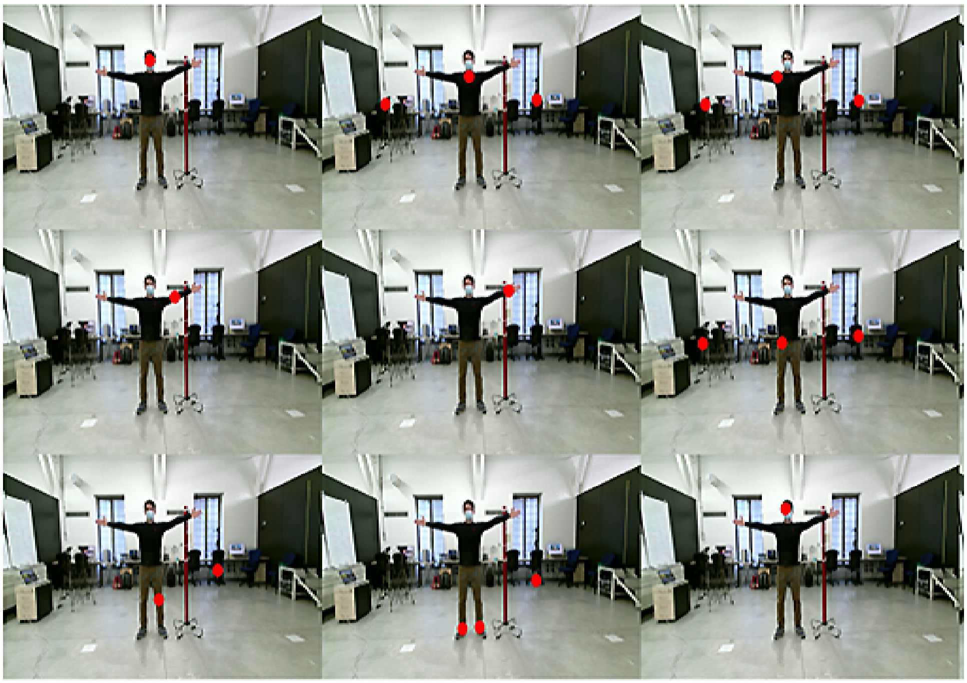

Figure 9: Results of converting confidence into certainty

because connections that share a part detection candidate are grouped together. The final step is to transform detected connections into final skeletons.

4 TEST CASES AND RESULTS

The videos were divided by five seconds between each frame in order to clearly distinguish between poses and the skeleton of the model. A range of hyperparameters was used to train, validate, and test the system for the purpose of finding the most effective ones. Through simulations done in the research process, we have found that the following hyper-parameters yielded the best results:

- Scale search function - scaled images can be generated to simulate, allowing a greater level of accuracy to be achieved while sacrificing inference time. Scale-search =[0.5, 1.0, 1.5, 2.0],
- Boxsize is the height of the baseline image that should be scaled, box-size =368,
- Openpose model has Stride = 8 for VGG backbone to obtain heat and PAF feature maps,
- PadValue: The padding of the input image which saves the image size by padding through the activation maps. PadValue = 128
- Heatmap-Avg: 19 parts: 18 parts and 1 background
- PAF-Avg: 38 channels representing 19 connections

Figure 10 illustrates four different cases of skeleton extraction using the COCO database and the MobileNetV1 model. The validity of the method is demonstrated by the estimation of a single person pose.

5 CONCLUSIONS

In this paper, we studied the problem of human pose estimation network, which is suitable for real-time performance on edge devices. A solution based on the OpenPose method with a heavily optimized network design and post-processing code. Using a dilated MobileNetv1 feature extractor with depth wise separable convolutions and a lightweight refinement stage with residual connections ave proved to provide much more accurate results as depicted in the examples.

REFERENCES

[1] Aubry, S.; Laraba, S.; Tilmanne, J.; Dutoit, T.: Action recognition based on 2d skeletons extracted from rgb videos. In MATEC Web of Conferences, vol. 277, 02034. EDP Sciences, 2019.

[2] Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition, 7291–7299, 2017.

[3] Chen, Y.; Tian, Y.; He, M.: Monocular human pose estimation: A survey of deep learning-based methods. Computer Vision and Image Understanding, 192, 102897, 2020.

[4] COCO-DataSet: https://cocodataset.org/.

[5] Gerhardt, D.M.; Mors, T.G.; Hannink, G.; Susante, J.L.V.: Resurfacing hip arthroplasty better preserves a normal gait pattern at increasing walking speeds compared to total hip arthroplasty. Acta Orthopaedica, 90, 231–236, 2019. ISSN 17453682. http://doi.org/10.1080/17453674.2019.1594096/SUPPL_FILE/IORT_A_1594096_SM0244.PDF.

[6] Ghorbel, E.; Papadopoulos, K.; Baptista, R.; Pathak, H.; Demisse, G.; Aouada, D.; Ottersten, B.: A view-invariant framework for fast skeleton-based action recognition using a single rgb camera. In 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Prague, 25-27 February 2018, 2019.

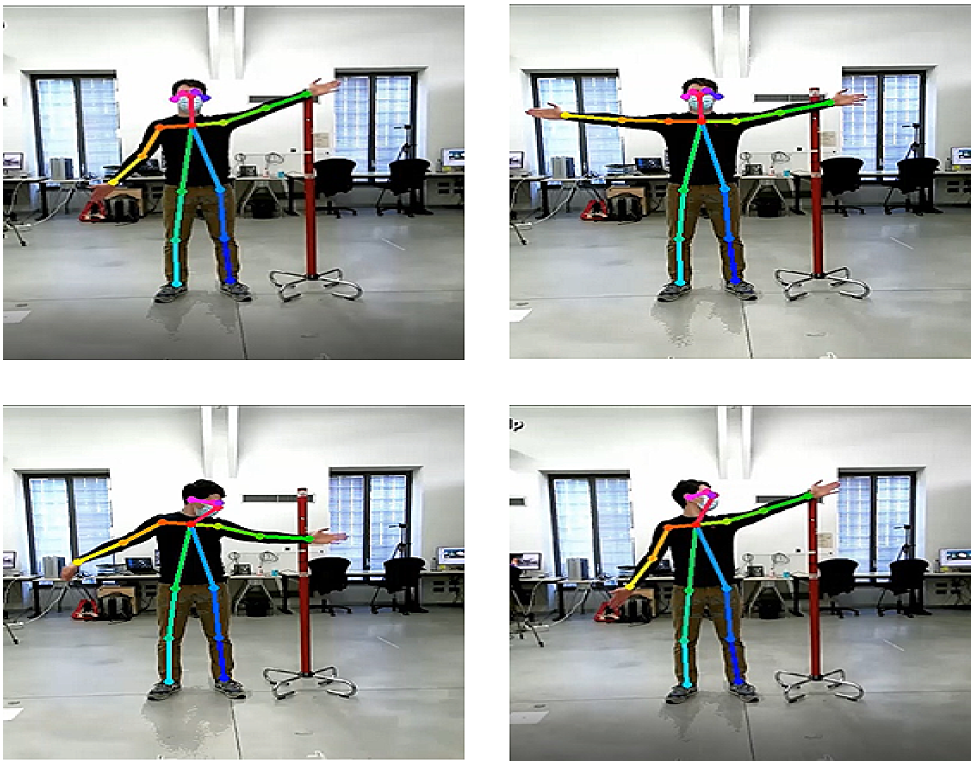

Figure 10: Results of different examples of skeleton extractions using the COCO database and the MobileNetV1 model

[7] Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[8] Hu, G.; Wang, W.; Chen, B.; Zhi, H.; Li, Y.; Shen, Y.; Wang, K.: Concurrent validity of evaluating knee kinematics using kinect system during rehabilitation exercise. Medicine in Novel Technology and Devices, 11, 100068, 2021. ISSN 2590-0935. http://doi.org/10.1016/J.MEDNTD.2021.100068.

[9] Huang, C.H.; Foucher, K.C.: Step length asymmetry and its associations with mechanical energy exchange, function, and fatigue after total hip replacement. Journal of orthopaedic research : official publication of the Orthopaedic Research Society, 37, 1563–1570, 2019. ISSN 1554-527X. http://doi.org/10.1002/JOR.24296.

[10] Kimura, A.; Mitsukura, Y.; Oya, A.; Matsumoto, M.; Nakamura, M.; Kanaji, A.; Miyamoto, T.: Objective characterization of hip pain levels during walking by combining quantitative electroencephalography with machine learning. Scientific Reports 2021 11:1, 11, 1–10, 2021. ISSN 2045-2322. http://doi.org/10.1038/s41598-021-82696-1.

[11] Laraba, S.; Brahimi, M.; Tilmanne, J.; Dutoit, T.: 3d skeleton-based action recognition by representing motion capture sequences as 2d-rgb images. Computer Animation and Virtual Worlds, 28(3-4), e1782, 2017.

[12] Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L.: Microsoft coco: Common objects in context. In European conference on computer vision, 740–755. Springer, 2014.

[13] Martınez, G.H.: Openpose: Whole-body pose estimation, 2019.

[14] MPII-DataSet: http://human-pose.mpi-inf.mpg.de/.

[15] Otte, K.; Kayser, B.; Mansow-Model, S.; Verrel, J.; Paul, F.; Brandt, A.U.; Schmitz-Hbsch, T.: Accuracy and reliability of the kinect version 2 for clinical measurement of motor function. PloS one, 11, 2016. ISSN 1932-6203. http://doi.org/10.1371/JOURNAL.PONE.0166532.

[16] Plantard, P.; Auvinet, E.; Pierres, A.S.L.; Multon, F.: Pose estimation with a kinect for ergonomic studies: Evaluation of the accuracy using a virtual mannequin. Sensors 2015, Vol. 15, Pages 1785-1803, 15, 1785–1803, 2015. ISSN 1424-8220. http://doi.org/10.3390/S150101785.

[17] Regazzoni, D.; Vitali, A.; Colombo Zefinetti, F.; Rizzi, C.: Gait analysis in the assessment of patients undergoing a total hip replacement. In ASME International Mechanical Engineering Congress and Exposition, vol. 83518, V014T14A003. American Society of Mechanical Engineers, 2019.

[18] Singh, J.P.; Jain, S.; Arora, S.; Singh, U.P.: Vision-based gait recognition: A survey. IEEE Access, 6, 70497–70527, 2018.

[19] Temporiti, F.; Zanotti, G.; Furone, R.; Molinari, S.; Zago, M.; Loppini, M.; Galli, M.; Grappiolo, G.; Gatti, R.: Gait analysis in patients after bilateral versus unilateral total hip arthroplasty. Gait Posture, 72, 46–50, 2019. ISSN 0966-6362. http://doi.org/10.1016/J.GAITPOST.2019.05.026.

[20] Vainieri, M.; Panero, C.; Coletta, L.: Waiting times in emergency departments: A resource allocation or an efficiency issue? BMC Health Services Research, 20, 1–10, 2020. ISSN 14726963. http://doi.org/10.1186/S12913-020-05417-W/TABLES/3.

[21] Varin, D.; Lamontagne, M.; Beaulé, P.E.: Does the anterior approach for tha provide closer-to-normal lower-limb motion? The Journal of arthroplasty, 28(8), 1401–1407, 2013.

[22] Wacha, H.; Domsel, G.; Herrmann, E.: Long-term follow-up of 1217 consecutive short-stem total hip arthroplasty (tha): a retrospective single-center experience. European Journal of Trauma and Emergency Surgery, 44(3), 457–469, 2018.

[23] Wochatz, M.; Tilgner, N.; Mueller, S.; Rabe, S.; Eichler, S.; John, M.; Vller, H.; Mayer, F.: Reliability and validity of the kinect v2 for the assessment of lower extremity rehabilitation exercises. Gait Posture, 70, 330–335, 2019. ISSN 0966-6362. http://doi.org/10.1016/J.GAITPOST.2019.03.020.

[24] Zheng, C.; Wu, W.; Yang, T.; Zhu, S.; Chen, C.; Liu, R.; Shen, J.; Kehtarnavaz, N.; Shah, M.: Deep learning-based human pose estimation: A survey. arXiv preprint arXiv:2012.13392, 2020.

[25] Zoppo, G.; Marrone, F.; Pittarello, M.; Farina, M.; Uberti, A.; Demarchi, D.; Secco, J.; Corinto, F.; Ricci, E.: Ai technology for remote clinical assessment and monitoring. Journal of wound care, 29(12), 692–706, 2020.