Title:
Human Joint Profile Extraction using Deep Learning Approaches

Authors:
Miri Weiss Cohen, miri@braude.ac.il, Braude College of Engineering
Andrea Vitalli, andrea.vitali1@unibg.it, University of Bergamo
Daniele Regazzoni, daniele.regazzoni@unibg.it, University of Bergamo

Keywords:
2D Joint profile, skeleton feature extraction, Convolution Neural Networks (CNN)

DOI: 10.14733/cadconfP.2022.2022.215-219

Introduction:
Coxarthrosis (degenerative osteoarthritis) is the most common form of osteoarthritis of the hip joint. Despite advances in understanding osteoarthritis, no known therapy can prevent its progression and the most common surgical procedure is prosthetic hip replacement, also called Total Hip Arthroplasty (THA). Many reasons exist for healthcare systems and surgeons to search for a better surgical approach and accuracy. By doing so, patients will have a safer, more effective, and faster path to healing [1].

The purpose of this study is to provide a short-term assessment procedure for patients' gait ability after THA. 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 [2, 3]. To provide feedback to surgeons and healthcare providers, a patient's short-term recovery performance is essential for assessing the success of hip arthroplasty.

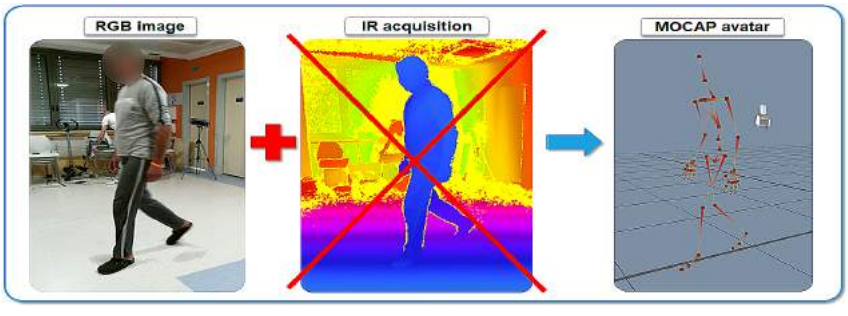

Fig. 1: proposed approach: Gait evaluation process

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. Medical motion tracking is opening up new possibilities for clinical and home applications because

it allows detecting and tracking human body movements, such as gait or any other gesture or posture, in a specific context however, it is very expensive and difficult to operate[4]. 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.

Main Idea

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. 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 previous studies [5], 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. In this work 30 training videos were used, with patients being recorded. In order to improve accuracy and key point detection, we divided the videos into single frames for better assessment. 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.

Several approaches have been taken to retrieve skeleton models from an image[6]. By training a CNN network, we first track the joint location, and then we transform this information into the final skeleton. 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. 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[7]. 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. Training, validation and testing was done using OpenPose on three different data-sets:

- MPII Human Pose dataset for evaluating articulated human pose estimation. Around 25K images of over 40K people were included in the dataset, with their joints annotated. Using a taxonomy of everyday human activities, the images were systematically collected. 17 joints were analyzed [8].
- COCO - This open source object recognition database is one of the most popular tools for training deep learning algorithms. The dataset contains 250000 people images and includes 17 joints [9].
- BODY-25 COCO dataset and MPII dataset combined. There are 25 joints in this dataset.

The image is first passed through a convolutional network to generate a set of feature maps. The feature maps are refined iteratively, 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 [3]. 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. Figure 2 depicts partial heat map results. The line integral of the corresponding Part Affinity Fields, can be calculated

Fig. 2: Heatmap peak information retrieval partial results

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 3 depicts partial heat map results. To optimize the

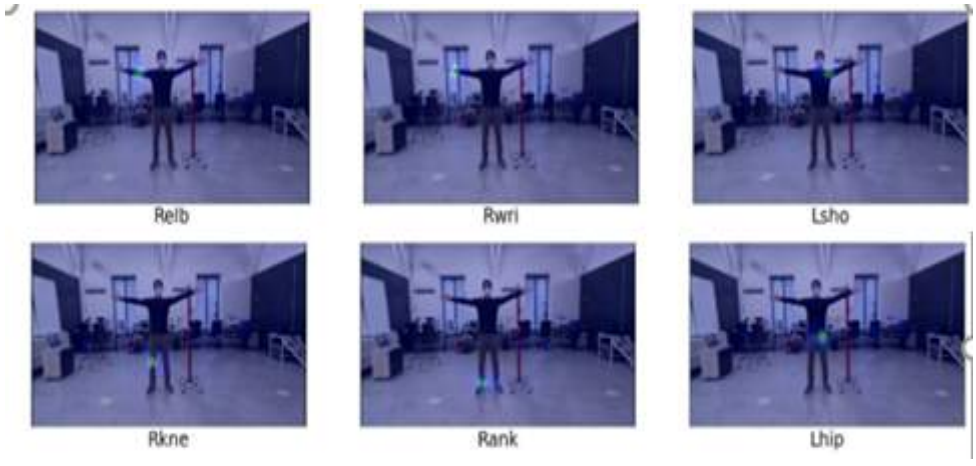

Fig. 3: PAF indices examples containing the x and y coordinates for the limb and Neck

results of the trained model described in the previous stage, we propose to retrain the output results on a different CNN architecture model named MobileNetV1. This CNN architecture employs an inverted residual structure in which shortcut connections are made between the thin bottleneck layers. Using lightweight depth-wise convolutions, the intermediate expansion layer solves this non-linearity. Table 1

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)

describes a partial description of MobileNetV1 and its ConV layers. The Sequential model described, allows layer-by-layer creation of models, but is limited in that it does not allow creation of models with multiple inputs or outputs.

Results

Figure 4 depicts four different examples of skeleton extractions using the COCO database and the Mo-

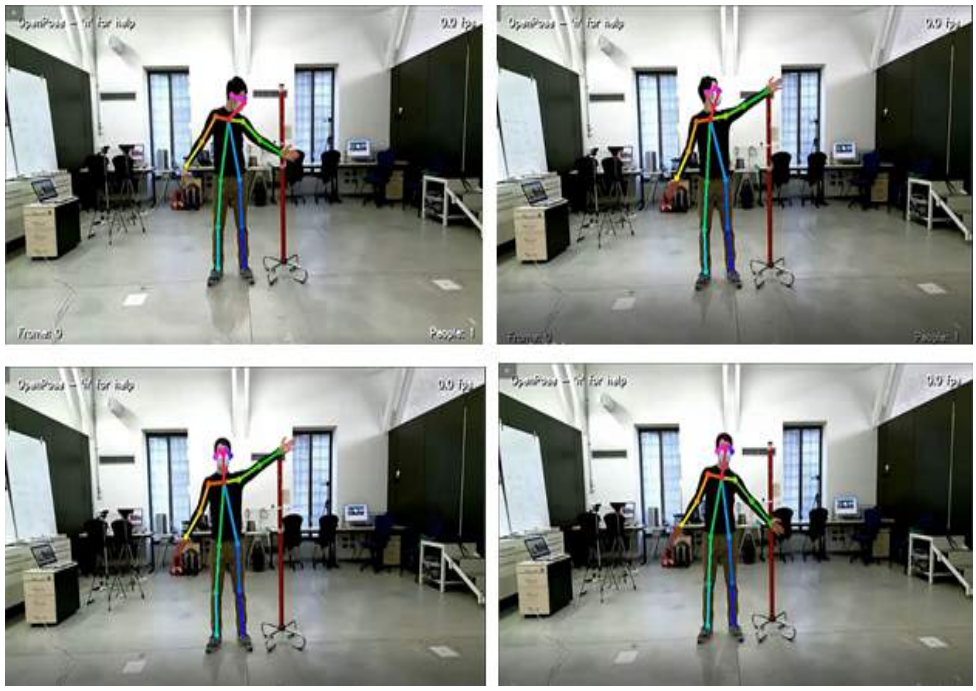

Fig. 4: Results of different examples of skeleton extractions using the COCO database and the MobileNetV1 model.

bileNetV1 model. During this procedure, a single person pose is estimated, demonstrating its reliability. The system was trained, validated and tested for various hyperparameters, ultimately resulting in the best ones. In order to clearly distinguish between the poses and the skeleton provided by the model, the videos were divided by five seconds between each frame.
The following hyper-parameters were used: Scale search functions: generate a series of scaled images to simulate, allowing for greater accuracy while sacrificing inference time. scale-search =[0.5, 1.0, 1.5, 2.0], Boxsize is the baseline image height that needs to be scaled, box-size =368, stride: openpose model has stride = 8 for VGG backbone to obtain heat and PAF feature map, padValue: the input image is desired to be divisible by 8. padValue = 128, heatmapavg: 19 channels: 18 parts and 1 background, pafavg: 38 channels representing 19 connections with x, y.

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 heavily optimized network design and post-processing code was proposed. Using a dilated MobileNetv1 feature extractor with depthwise separable convolutions and a lightweight refinement stage with residual connections, we increased accuracy versus complexity by more than six times.

References:

[1] Regazzoni, D.,; Vitali, A.; Colombo Zefinetti, F.; Rizzi, C.: Gait Analysis in the Assessment of Patients Undergoing a Total Hip Replacement. ASME International Mechanical Engineering Congress and Exposition (Vol. 83518, p. V014T14A003). American Society of Mechanical Engineers, 2019.

[2] 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, p. 02034. EDP Sciences. 2019.

[3] 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 pp.7291-7299, 2017.

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

[5] 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.

[6] Ijjina, E.P.: Classification of human actions using pose-based features and stacked auto encoder, Pattern Recognition Letters, 83, pp.268-277, 2016.

[7] Zavala-Mondragon, L.A.; Lamichhane, B.; Zhang, L.; Haan, G.D.: CNN-SkelPose: a CNN-based skeleton estimation algorithm for clinical applications, Journal of Ambient Intelligence and Humanized Computing, 11(6), pp.2369-2380, 2020.

[8] MPII Human Pose Dataset: http://human-pose.mpi-inf.mpg.de/

[9] Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollr, P.; Zitnick, C.L.:. Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham, 2014.