A Pilot Study for Investigating Gait Signatures in Multi-Scenario Applications

[1*] Sumit Hazra
MIBM Lab, Dept. of CSE
National Institute Of Technology
Rourkela, India
sumhaz15@gmail.com

[2] Priyankar Roy
Dept. of Information Technology
IIEST Shibpur
Howrah, India
priyankarroy1711@gmail.com

[3] Anup Nandy
MIBM Lab, Dept. of CSE
National Institute Of Technology
Rourkela, India
nandy.anup@gmail.com

[4] Rafał Scherer
Częstochowa University of Technology
Częstochowa, Poland
rafal.scherer@pcz.pl

***Abstract*—Human pose estimation in a gait sequence is an essential step for solving human identification problems, medical diagnosis, monitoring, and rehabilitation. In this paper, a low-cost Kinect V2.0 sensor is used for investigating motion signatures obtained from normal healthy adults. The purpose of this study is to determine the accuracy and reliability of observational assessments of spatio-temporal features. A novel approach for human detection and tracking is proposed, which involves gait feature learning principles from depth and RGB video. In the first step, a human object from the depth image is extracted using the proposed semi-dynamic object tracking algorithm, and a stick model is generated using body aspect ratios to extract hip angles. In the second step, the gait energy image (GEI) representation is utilized for training a 2D Convolutional Neural Network (AlexNet) for automatic feature extraction. A key point detection algorithm is proposed for estimating knee, hip, and ankle joints from RGB gait videos. The reliability analysis of motion signatures is performed using various statistical methods to ensure feature learning for multi-scenario applications. The statistical results are promising for evaluating the methods which influence the inter-record differences among motion signatures.**

***Index Terms*—Region Of Interest, Gait Energy Image, Convolutional Neural Network, Gait, ANOVA**

I. Introduction

Human walking is a complex process that requires much balance and coordination. It is generated, maintained, and guided by the neuro-muscular system. Human motion is analyzed in terms of gait. It is a subject of extensive research. The systematic movement of limbs resulting in bipedal locomotion called gait is considered to be a significant biometric trait with applications on a large scale. Gait biometrics can be obtained without the person's attention, unlike DNA, fingerprints, and other biometric traits. The data acquisition for human gait analysis is performed opting for a vision-based approach using Kinect v2.0 sensors. Gait identification is significant in surveillance security and medical implications. Computer vision-based techniques are used for tracking, detecting, and identifying human due to their unobtrusive and non-invasive properties. Such biometric methods obtain promising results despite the fact that the camera and the subject are distant enough from each other. The video-based gait analysis is computationally intensive for handling segmentation, tracking, and silhouette extraction algorithms. The sensor-based technique for gait data acquisition can bring wrong recognition results due to factors like displacement of sensors during walking, errors in measurement of sensor readings, and discomfort in wearing sensor-based bodysuit while walking. In [1], a sensor-based data acquisition suit utilizing force sensors (USL06-H5-500N-C) was built to measure ground reaction forces during the walking phases. The spatial-temporal representation of the entire gait sequence was first proposed in [2], which is known as GEI. This GEI concept was further extended to solve gait identification problems, which we also use in our research as inputs for a Convolutional Neural Network (CNN). Independent Component Analysis technique for choosing amongst the best features for classification with the matching pursuit algorithm for feature extraction was proposed in [3]. A feature selection technique based on the statistical learning theory for Gait Energy Images after the application of cross-validation techniques was proposed in [4].

The contribution of our work is to create a usable solution for people detection and tracking for clinical environment gait analysis of their gait patterns from the extracted data. This involves gait spatio-temporal feature extraction from the depth and RGB video acquired via Kinect V2.0 sensors. Features are extracted from the depth video using two approaches: Firstly, a stick model is generated using the body aspect ratio of humans. The joint angles are then extracted from the model, which are used as features for gait. Secondly, features are extracted automatically after the application of the Convolution Neural Network on the GEIs of the subjects obtained via the frames extracted from a single gait cycle using auto-correlation. The classification of gait recognition is done using CNN, which gave features as well for reliability analysis. From the RGB video, on the extracted frames, a human key point (pose) detection algorithm is used to locate the centroids at various body joint positions to extract the body joint angles. Then, the features obtained from both the techniques are compared, analyzed, and reliability tested using Hypothesis Testing (Z-test as sample size$>$30), Fisher's Discriminant Ratio (FDR), and Analysis Of Variance (ANOVA). The results are found to be promising. The performance is demonstrated with proper graphical representations to provide a clear understanding of the error rates.

The paper is organized as follows: Section II presents related works, Section III talks about the proposed methodology,

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Section IV projects the results and discussion, and Section V puts forth the conclusion and the future work.

II. Related Works

For the last few years, it is noticed that an increase in activity has been happening in the gait analysis research area. Faster processing power and improvement in the mobility of today's computers act as an aid to this research. Recollecting some of the previous works focussed on people detection [5], [6], background extraction technique is given much emphasis. These are applicable for any kind of images such as color or depth and produce satiating results subject to certain specific requirements. The advent of RGB-D cameras, such as the Asus Xtion or Microsoft Kinect, has been beneficial for computer vision research because of the availability of both color and depth images on the same device. Simultaneously the cost incurred is also very less when compared to other 3D or thermal cameras, such as Laser Range Finders or Long-Wave Infrared (LWIR) cameras. Some of the works such as [7], [8] and [9], present the utilization of these new type of cameras for human identification, detection, and tracking. Basically, there are two main approaches to dealing with human detection. The first being mostly on machine learning methods, such as AdaBoost [10]–[12] or multi-class Support Vector Machines(SVM) that use features like Histogram of Oriented Gradients (HOG) [13] or Local Surface Normals (LSN) [14] to classify objects as human or non-human. Another technique that is used is template matching, employable in systems such as [5] or [8]. The spatial pose model for the estimation of body parts, joints, etc. can be classified into two categories. One is based on tree-structured graphical models [15], [16] which encode the relationship between the parts which are adjacent following a kinematic chain. The other is a family of non-tree models [17]–[21], that add additional edges to the tree structure to capture certain features such as symmetry, occlusion and relationships which are of long-range. To obtain reliable observations of body parts locally, Convolutional Neural Networks (CNNs) have been used, which improved the accuracy on body pose estimation [22]–[34]. In [35], CNN was used to capture and understand the global spatial dependencies by proposing and implementing networks with quite large receptive fields.

In this work, we extract spatio-temporal features of gait, both from RGB and depth video. We apply various machine and deep learning techniques, the most notable being a 2D CNN (AlexNet). After this, various statistical techniques are applied to check and analyze which features are more reliable than the other and can be adopted in the future to carry out work in the domain of gait analysis.

III. Proposed Methodology

The workflow diagram for this research work is presented in Fig. 1. Two main approaches are explicitly explained in this paper. Spatio-temporal gait features are obtained from color depth and RGB video using Kinect v2.0 sensors. The features include the pattern of the knee and hip angles with a gait speed of 5 km/h. The video is taken at 29-30 fps, and each frame is separately analyzed to extract the features. The subjects are instructed to walk on a treadmill at 5 km/h, and the Kinect camera is placed at a distance of 2.5 meters from the treadmill. A step is taken to ensure that no object is found between capturing probe and subject. The frames extracted from video undergo further processing. Finally, the features are tested for reliability analysis using various statistical techniques, as discussed in Section IV. The features extracted are the hip angles from color depth and RGB video and knee joint angles from RGB video. Normally humans have a fixed gait pattern, and thus more or less same plot for hip and knee angles are obtained whereas someone with deformity or trouble walking will show deviation from what we obtain from the gait cycle of normal humans.

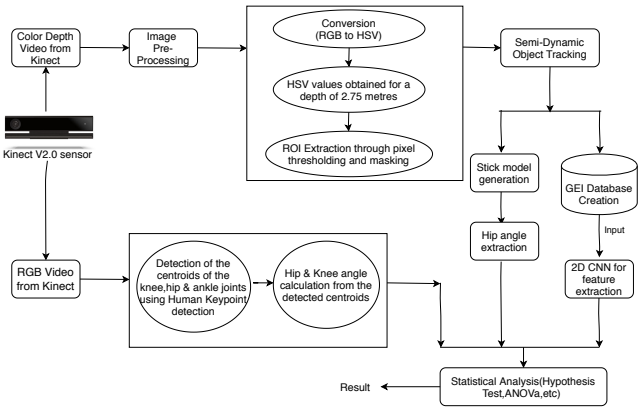


Fig. 1: Workflow of the proposed method.

A. Data Acquisition for Depth Video

The Kinect v2.0 sensor is placed at a distance of 2.5 meters from the treadmill and the subject. The color depth video is recorded with a Windows-based system of 8 GB RAM and Intel i5 core processor when the subject is moving on the treadmill at a speed of 5 km/h. A screen recorder is used because Kinect provides no utility software to record the depth video at .mp4, .avi, or any such commonly readable formats. The recorded video is stored with .xef, which is accessed using Kinect Studio v2.0. Gait features are extracted from the depth video using two approaches: Firstly, a stick model is generated using the body aspect ratio of subjects, and then the extracted joint angles are used as features. Secondly, the features are extracted automatically after applying Convolution Neural Network model on GEIs. These GEI images are the frames extracted from a single gait cycle measured using the auto-correlation technique.

B. Obtaining Regions Of Interest

The first stage of our workflow is the detection of regions of interest in the images, known as ROIs. They work as masks indicating the regions of the image for both color and depth, that generally, a human object populates.

1) Image Segmentation through Thresholding: Processing of the color depth image and the extraction of the ROI is done via human detection and tracking using the Kinect camera. Usually, for the extraction of ROIs from depth images, histogram analysis is performed, and then the ROI is extracted as described in [36]. However, in this case, it cannot be performed because of the three different color streams in R, G and B. A color depth image assigns different colors depending on the subject's distance from the camera irrespective of the variation in the lighting condition, which is used for the extraction of ROI. As the subject is placed at a distance of 2.5 meters, its color will lie in a specific range (yellow to green). Thus, thresholding and binary masks are used to get the human as desired in our research work, which is a novelty in our case. All work is performed using the HSV format to ease the task of thresholding.

2) Dynamic Object Tracking via Rectangular Bounding Box: The procedure discussed in Section III-B1 also resulted in getting the area of the floor between the human on the treadmill and the Kinect camera apart from the required object. As we are only concerned with the human subject whose gait features are to be taken into consideration, the human is specifically extracted. This is done considering a legitimate assumption. The assumption being that if the subject is standing with his back upright, the distance from the feet to the head will be the longest patch of pixels with almost the same depth (< 30 % variation in HSV). Then, a semi-dynamic (as it is capable of expanding or contracting along its width) bounding box is fit around the human subject to be able to extract it quickly.

The main ROI is the human on the treadmill. All the objects inside the box are considered as the ROI. As per our assumption, it has the same color with minimum deviation, considering the 3D nature of the human body. So the longest distance is calculated, and then its midpoint is assumed to be the body's vertical center. From there, the width is found out by traversing left and right as finding the off pixels. The semi-dynamic nature is added by finding the maximum width on pixels, and if it turned out to be more than the initial width of the bounding box, its width is increased and vice-versa. The proposed semi-dynamic object tracking algorithm is given in Algorithm 1. In the proposed algorithm, we need to account for every patch of pixels, so we need to parse the entire image matrix. After the bounding box is created, we need to operate only within the region of interest. Therefore, the complexity of the overall algorithm is O(height of image * width of the image). Hence, the bounding box that is generated for every frame is efficient as it is quite adept in adjusting its width and contains only the test subject. Fig. 2 shows an extracted image with a rectangular bounding box around the subject on the treadmill along with the stick model generation, which is explained explicitly in Section III-C1.

C. Feature Extraction Technique from Depth Image

The extracted bounding box will contain ROI, from where we extract the features via two techniques explained in the following sections.

Algorithm 1 Pseudo-Code for Dynamic Object Tracking

Input: $im = input\ image$
$im\_hsv = image\ in\ HSV\ form.$
$on\_list =\ column\ containing\ the\ highest\ no.\ of\ 'On'\ pixels$
$P_0 = the\ top\ most\ pixel\ of\ the\ calculated\ roi$
$P_{end} = the\ bottom\ most\ pixel\ of\ the\ calculated\ roi$
Output:
$b_{semi\_box} = a\ semi\ dynamic\ bounding\ box\ around\ the\ roi$
Procedure:

$read\_im(im)$
$im_{gray} = rgbToGray(im)$
$im\_er =\ Erode(image)$
$hsv =\ convertToHSV(im\_er)$
$im_{thres} =\ extractROI(hsv)$
$im_{thres\_new} =\ modify\_threshold(im_{thres})$
$res =\ mask(im_{thres\_new})$
for i in range (v_len(res)) **do**
  for j in range(h_len(res) **do**
    $s = \sum_{j,i=0,0}^{h\_len(res),v\_len(res)}(P_{ji})\ ,\ where\ P_{ji} > 0$
    $on\_list.append(s)$
for i in range(v_len(res)) **do**
  if $(intensity(i)\ < 0.3 * intensity(i-1))$ **then**
    $length =\ length\ +\ 1$
  else
  $list.append(P_0, Pend)$
$patch\_max =\ len_{patch\_large\_30(on\_list.append(list))}$
$mid =\ P_0\ +\ (patch\_max/2)$
$width =\ traverse_{(p<mid_{<=30\%}\ \&\ p>mid_{<=30\%})}$
$b_{semi\_box} =\ box((P_{end}(v\_len), left), (P_0(v\_len), right))$
$return\ b_{semi\_box}$

1) Feature Extraction Technique via the Proposed Stick Model Generation: The region of interest obtained in the previous step is further processed to obtain the required spatio-temporal feature of gait. The image is decomposed into three independent regions, namely a head node, body torso, and the leg region. It is done in accordance with the anatomical ratio of the human body [37]. The respective heights of the head, body torso and the leg region are calculated utilizing the body

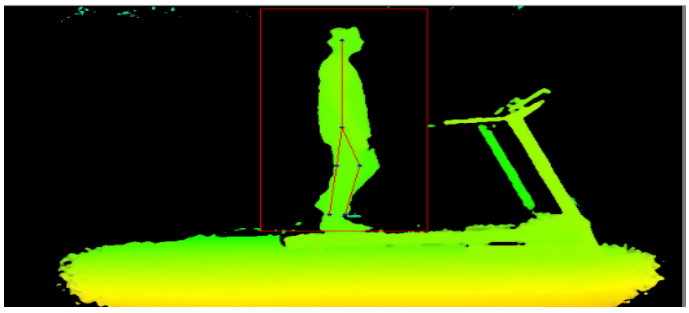


Fig. 2: Extracted image with bounding box and skeletal frame.

segment ratios. The human gait signature is extracted from each independent region of the human body. We are concerned with the lower portion of the body starting from the hip. Thus, the boundary coordinates are extracted for the head node, body torso, and lower limb. The feature metrics are deposited as a 1-D distance signal. The Euclidean distance metric is used to calculate the distance between the boundary coordinates and centroid of each region of the human body. The hip angles are the features extracted from the depth image captured from the depth video. The hip angle calculation from the stick model, as shown in Fig. 2 is explicitly explained diagrammatically (Fig. 3) with mathematical formulations (Equations (1), (2), (3) and (4)) in two cases.

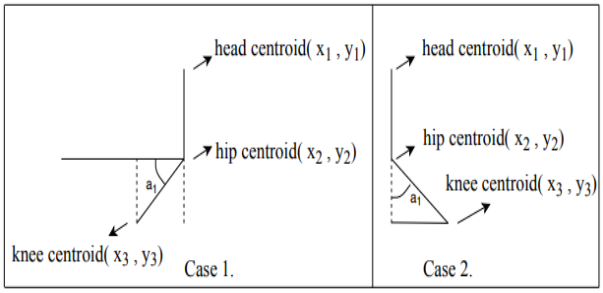


Fig. 3: Hip angle calculations in Cases 1 and 2.

- Case 1: *When hip centroid is behind the knee centroid*

$$H_\theta = 90^\circ + a_1, \tag{1}$$

$$a_1 = \tan^{-1}\frac{(y_3 - y_2)}{(x_3 - x_2)}, \tag{2}$$

where a_1 is an angle as shown in Figure 3. Case 1, H_θ is the hip angle.

- Case 2: *When knee centroid is ahead of the hip centroid*

$$a_1 = \tan^{-1}\frac{(x_3 - x_2)}{(y_3 - y_2)}, \tag{3}$$

$$H_\theta = 180^\circ + a_1, \tag{4}$$

where a_1 & H_θ have their usual meanings as in Case 1.

2) Feature Extraction Technique via 2D CNN (Convolution Neural Network): Feature Extraction is an integral part of all the classical pattern classification problems. The gait signal possesses spatio-temporal information of human movements. So, to represent human walking properties, GEI is utilized for feature selection. Various factors, such as walking speed, view angles, and many others, increase the intraclass variance. Thus, it is an utter requirement to extract a unique gait feature to remove any kind of ambiguity in subject identification, thus minimizing within-class variance. A GEI gait frame is obtained by calculating the average of the gait frames comprising one complete gait cycle. As it involves averaging the sequence of frames, the computational cost is reduced to a significant extent. Further, utilizing the correlation between the obtained images, we calculate the gait cycle.

- *Calculation Of Gait Period*: For a complete human gait cycle, the stance and the swing period gets repeated. A gait period is defined as the number of frames comprising the full gait cycle. The complete movement from one heel strike to the next heel strike of a leg is known as a gait cycle. For the calculation of the gait period, we follow a correlation-based approach. On analyzing the frames of the depth video, we get the sequence of silhouette frames of a person's walking. In this, the correlation coefficient of the first frame is calculated with all the remaining subsequent frames. The calculation of the correlation coefficient is as demonstrated with the following mathematical formulation

$$Corr = \frac{\sum_{h=1}^{M}(SS_h - \overline{SS})(NS_h - \overline{NS})}{\sqrt{\sum_{h=1}^{M}(SS_h - \overline{SS})^2}\sqrt{\sum_{h=1}^{M}(NS_h - \overline{NS})^2}} \tag{5}$$

where $Corr$ is a correlation coefficient, SS is the first frame of gait sequence, NS is the next frame of gait sequence, and M is the total number of frames in gait sequence. The correlation coefficient is calculated for all the obtained frames, following which the maxima locations are found out. The gait period is the number of frames between two successive maxima. The same process is repeated for all the ten subjects for a speed of 5 km/h at our lab. The mean of all the gait periods is taken as the length of the gait cycle, considering only a particular speed. Background subtraction technique helps in obtaining only the object with the help of the semi-dynamic bounding box in a gray scale. Hence, background subtraction is done for the depth image as well. After that, auto-correlation is performed. Thus, Fig. 4 is the plot of the correlation coefficient versus the image index(no. of frames). The peaks appearing in the curve depicts the locations where the correlation coefficients possess maximum values. Gait period is calculated as the distance between two corresponding maxima (shown by red dots in Fig. 4) The determination of the peak is solely based on the knowledge of the simple convex property in which a maximum is defined as a location at which a point has its value greater than all the M preceding values and the M following values. The window has a length M, which depends upon the speed of the walking pattern.

- *Gait Energy Image (GEI)*: GEI is the space and time normalized average of the extracted silhouette frames in a gait period. It retains the maximum amount of spatio-temporal gait information for a subject that is helpful for the feature extraction process. Fig. 5 represents GEI for a subject walking at a speed of 5 km/h on the treadmill. The mathematical expression for GEI is as follows

$$GEI(a,b) = \frac{\sum_{j=1}^{k}(GP_t(a,b))}{k}, \tag{6}$$

where $GP_t(a,b)$ is t^{th} frame in a gait period k, (a,b) are the corresponding pixel coordinates in the images.
It is understood from Fig. 5 that the main trunk of the

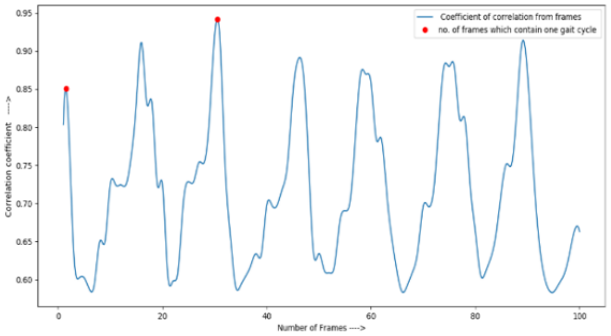


Fig. 4: Plot of the correlation coefficient vs. the number of frames at 5 km/h gait speed for a subject (depth video).

body is always constant. Thus, no significant information can be extracted from that part. So, without loss of any generality, each and every frame is substractable. The obtained GEI is then fed into 2D CNN (AlexNet) for automatic feature extraction.

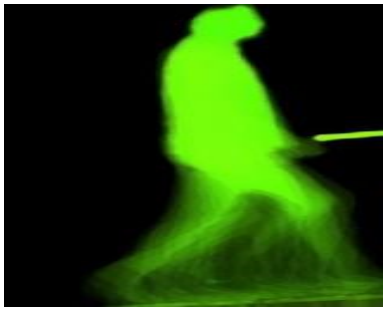


Fig. 5: Sample GEI image of a person walking at 5 km/h.

AlexNet was the winning entry in ILSVRC 2012. It solves the problem of image classification where the input is an image of one of 10 different classes (GEI of different subjects), and the output is a vector of 10 numbers. The ith element of the output vector is interpreted as the probability that the input image belongs to the ith class. Therefore, the sum of all elements of the output vector is 1. The input to 2D CNN is an RGB image of size 224×224. Thus, the dimensions of all the images in the training and test sets are changed to 224×224.

Our 2D CNN consists of 5 convolutional layers and three fully connected layers. Multiple convolutional kernels (a.k.a filters) extract interesting features in an image. In a single convolutional layer, there are usually many kernels of the same size. For example, the first conv. layer of AlexNet contains 96 kernels of size $11 \times 11 \times 3$. For our work, the width and height of the kernel are usually the same, and the depth is the same as the number of channels. The first two convolutional layers are followed by the overlapping max-pooling layers that we describe next. The third, fourth, and fifth convolutional layers are connected directly. The fifth convolutional layer is followed by an overlapping max-pooling layer, the output of which goes into a series of two fully connected layers. The second fully connected layer feeds into a softmax classifier with 10 class labels. The diagram of the 2D CNN architecture used in our research is given in Fig. 7. The GEIs are given as inputs to the 2D CNN. We use our dataset with 50 GEIs for prediction and cross-validation. Our GEI cross-validation dataset is shown in Fig. 6. The first max-pooling layer of the 2D CNN gives us a total of 96 features. Features of multiple subjects obtained from various layers of the 2D CNN undergo various statistical tests for further analysis.

D. Data Acquisition for RGB Video

The setup is similar to the one done for the depth video. The only difference being, this time, an RGB video is recorded using the Kinect camera and a screen recorder. Again a sequence of silhouette RGB frames are extracted from the video, but this time automatically by the key point detection algorithm via the OpenPose Framework. The centroids of the joints are detected frame by frame. The keypoints are detected using a pretrained Caffe model trained on the COCO dataset using the Openpose Framework developed by CMU-Perceptual-Computing-Lab [38]. The OpenPose Framework helps in 2D pose estimation. It is explained explicitly in the following section.

E. Feature Extraction Technique from RGB Image

Pose estimation is one of the most common problems in computer vision. It basically involves the detection of the keypoint locations of the object, which encompasses the position and the orientation of the objects as well. So in our work, we are concerned with the detection and the localization of the major joints of the body, namely the hip, knee, and ankle, respectively. Thus, using the OpenPose Framework

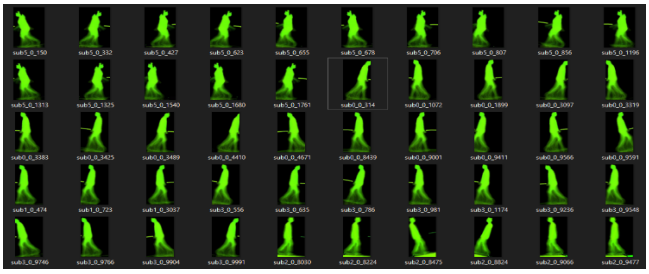


Fig. 6: GEI cross-validation dataset.

model [38], the detected keypoints for our work is numbered as shown later in Table 1.

Introduction of a novel feature

A new point is created for the hip centroid by taking the average of the points 11 and 8. It is indexed as 19 (Table 1). We only focus on points 19, 12, 13, 9, 10 in our research. These are used to calculate the required spatio-temporal features of gait. The first ten layers use a VGGNet feature map of the image. Also, the confidence and affinity maps produced are parsed by greedy inference to produce the 2D keypoints for all people in the image. Here the aforementioned pre-trained Caffe model ($pose\_iter\_440000. - caffemodel$) is

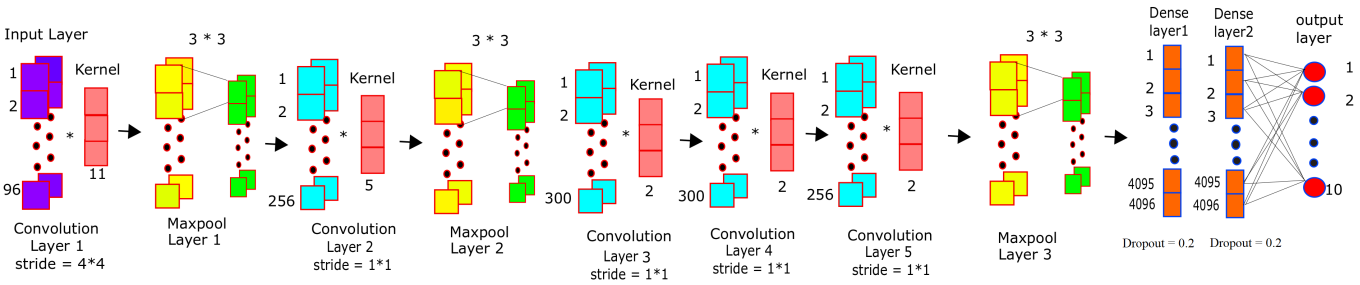


Fig. 7: Proposed 2D CNN architecture.

passed through the inbuilt OpenCV framework, which employs its own deep neural network techniques and the keypoints are predicted. For our work, we take a frame rate of 29 to 30 fps (frames per second) for video capture. This is again similar to the procedure followed for the depth videos initially. Background subtraction is performed for RGB video as well. Following which we perform auto-correlation (Fig. 8) to find the number of frames in a gait cycle. As a gait period (cycle) in our case consists of approximately 30 frames, we take the fps as justified. Due to which the prediction is sometimes inaccurate as the camera is not able to provide sharp images due to this low framerate. However, the inaccuracies are dealt with using median filters. As per our analysis of the features extracted from the individual frames (a total of approximately 30 frames) of the gait cycle, the sharp increase in the plots occurs as the movement of the hip and knees is not uniform for every frame. The swing of the left leg varies for each frame, thus providing us with non-evenly spaced discrete points for the joint angle plots (Fig. 11). The method of averaging the neighborhood pixels does suppress the noise, which is isolated and out-of-range. In the bid to remove the noise, certain other changes such as line features, sharp edges, so forth also get blurred. All of them which are affected correspond to high spatial frequencies. Thus, to come to the rescue, the median filter is used. It acts as an effective method that distinguishes the out-of-range isolated noise to a particular extent from a specific image.

With the help of the Key-Point Detection Algorithm [38] the detected human body keypoints on the extracted RGB is shown in Fig. 9a. The corresponding stick model obtained via the detected keypoints is further depicted in Fig. 9b. Thus, from the skeletal frame (Fig. 9b), the joint angles of the hip and knee joints are obtained. The hip angle is obtained using the same formulations as mentioned in Eq. (1)-(4) respectively. The knee joint angles are obtained, as discussed below in Fig. 10 and in Eq. (7)-(12) respectively. Fig. 11 demonstrates the knee angles obtained from the RGB images along with two essential spatio-temporal features of gait, namely toe-off and heel strike. Toe-off is defined as the point when the toe of reference swings in the air, and it marks the beginning of the

TABLE I: Table showing the detected keypoints along with the proposed point as per the COCO Output Format

Body Joints	Keypoints	Body Joints	Keypoints
Nose	0	Right Ankle	10
Neck	1	Left Hip	11
Right Shoulder	2	Left Knee	12
Right Elbow	3	Left Ankle	13
Right Wrist	4	Right Eye	14
Left Shoulder	5	Left Eye	15
Left Elbow	6	Right Ear	16
Left Wrist	7	Left Ear	17
Right Hip	8	Background	18
Right Knee	9	**Center Hip**	**19**

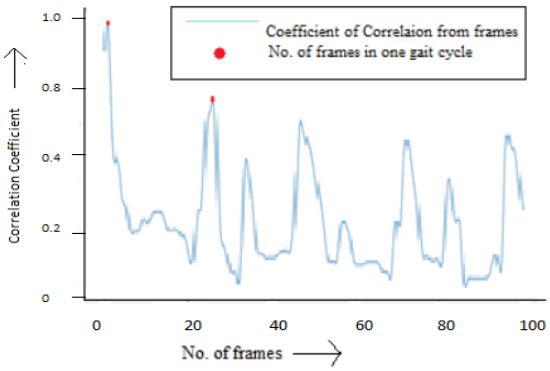
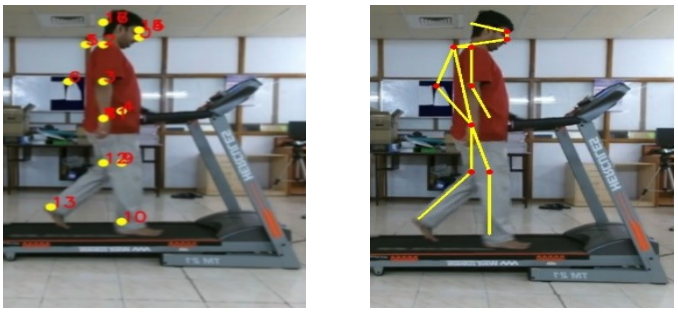


Fig. 8: Plot of the correlation coefficient vs. the number of frames at 5 km/h gait speed for a subject (RGB video).

swing phase in a gait cycle. Heel strike is the phase when the first bone of the reference foot, i.e., the heel touches the ground and marks the beginning of the stance phase in a gait cycle. A gait cycle comprises of a stance phase and a swing phase. Multiple subjects are used, but then the data of only one subject is shown for convenience. Gait cycles are extracted from RGB video following the same procedure as from depth video. As the video is in RGB format, there is



(a) Detected keypoints (b) Skeletal Frame

Fig. 9: Proposed pose estimation model.

a bit of noise in the thresholded image, but the extraction of the gait cycle through correlation of frames works properly, giving good results. As we can see by comparing the graphs Fig. 4 and Fig. 8, we can accurately determine the gait period.

For knee angle calculation:

- Case 1:*When knee is ahead of the centroid normal to the hip (ankle is behind the knee).*

$$a_1 = \tan^{-1}\frac{(y_2 - y_1)}{(x_2 - x_1)}, \tag{7}$$

$$a_2 = \tan^{-1}\frac{(y_3 - y_2)}{(x_2 - x_3)}, \tag{8}$$

$$K_\theta = a_1 + a_2, \tag{9}$$

where a_1 & a_2 are angles as shown in Fig. 10, Case 1, K_θ is the knee angle.

- Case 2: *When knee is ahead of the centroid normal to the hip (ankle centroid ahead of knee).*

$$a_1 = \tan^{-1}\frac{(y_2 - x_1)}{(x_2 - x_1)}, \tag{10}$$

$$a_2 = 90^\circ + \tan^{-1}\frac{(x_3 - x_2)}{(y_3 - y_2)}. \tag{11}$$

Thus

$$K_\theta = a_1 + a_2. \tag{12}$$

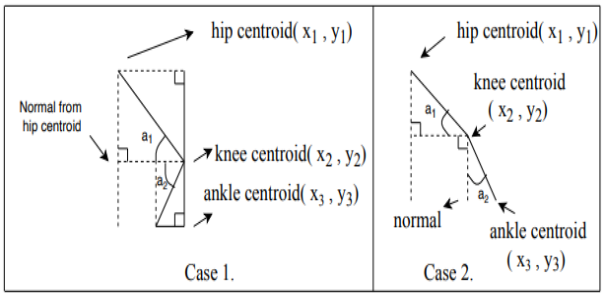


Fig. 10: Knee angle calculation in Cases 1 and 2.

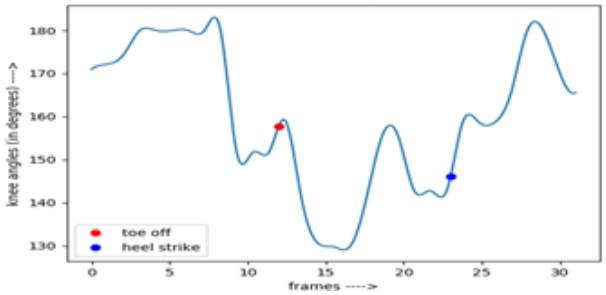


Fig. 11: Knee angles for a single gait cycle (RGB Video).

IV. Result Analysis and Discussion

The data collected via Kinect v2.0 sensors undergo a lot of processes and sequential steps involving deep learning, machine learning, and statistical tests, respectively. The data of all ten subjects are adequately analyzed, and the results are hereby discussed. The proposed 2D CNN model that we use consists of 9 layers and gives us a total of approximately 96 features for each layer. The same has been portrayed in the previous sections. Table II gives an overview of the different filters tried in the different layers and hence the models. The table gives the accuracies attained by the respective trials, out of which we choose the maximum accuracy for feature reliability for gait analysis. Fig. 12 shows the accuracy versus epoch curve for the training and the validation accuracies for the CNN model used.

TABLE II: Accuracy for various models tested

Filters in	Model No. 1	**Model No. 2**	Model No. 3
1st convolution layer	50	**96**	96
2nd convolution layer	265	**256**	256
3rd convolution layer	384	**300**	300
4th convolution layer	384	**300**	300
5th convolution layer	256	**256**	200
1st dense layer	4096	**4096**	3500
2nd dense layer	4096	**4096**	4096
Accuracy(in %)	84.44	**93.33**	93.2

As is visible from Fig. 12, the training versus validation graph converges well. Thus, the accuracy obtained is above 93 %, as is seen from Table II. Hence, we can infer that the features extracted from the depth video after the application of the 2D CNN model are informative ones. Hypothesis testing, FDR, and the Anderson-Darling test are the statistical tests performed on the obtained data. They are then analyzed to find out which data is most reliable for future works. The outcomes are discussed in this section and depicted in Table IV. The features used for testing are the knee and hip angles obtained from RGB video, the hip angles obtained from color depth video and the features extracted from the nine layers of the CNN via the feature maps.

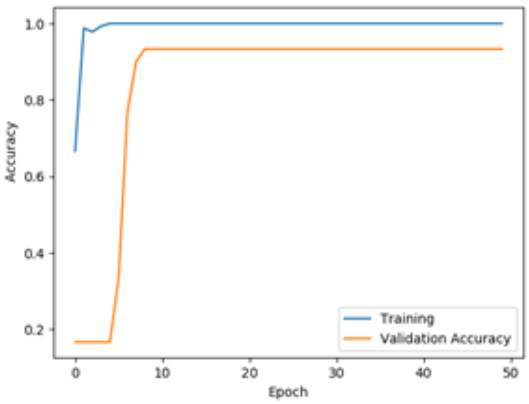


Fig. 12: Accuracy versus epoch curve for 50 epochs.

A better fit is obtained by producing lower AD values. The final distribution of the data is chosen to compare with the lowest AD values between distributions. P-value: The highest p-value (e.g.>0.05) indicates that the feature points follow a particular distribution. For our extracted features, the back knee angle from RGB video is not considered as most informative because it does not follow the normal distribution, but it does seem to stick close to the line. The front knee angles obtained from the RGB videos are quite informative. On the contrary, the hip angles obtained from the depth video provide more information when compared to those obtained from the RGB video, as the plot seems to fit the line better in its case. The CNN features are not much prominent to follow the normal distribution, therefore being the least informative. The Anderson Darling test results are obtained for all the extracted gait features. It is shown diagrammatically for some of the features, as in Fig. 13 and in Fig. 14 respectively.

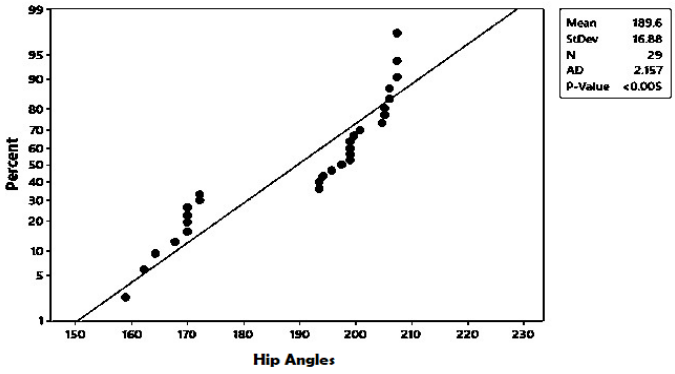


Fig. 13: The Anderson-Darling test plot for hip angles obtained from the depth video.

Further details are understood by evaluating the ANOVA results, the Z scores, and the FDR values, respectively. As mentioned earlier, ANOVA depends both on the f-value and the p-value obtained. The f-value tells us how significant our results are, but the null hypothesis (samples come from a

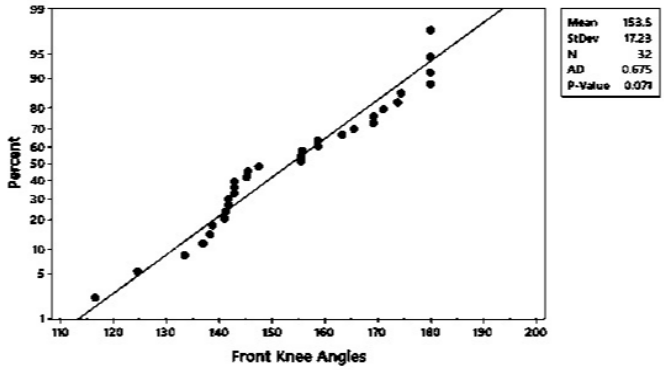


Fig. 14: The Anderson-Darling test plot for front knee angles (RGB video).

population with the same mean) can only be rejected when the p-value is also below the critical level. Over here, the standard critical value is 5 %. Thus, if the p-value is less than 0.05, the null hypothesis is safely rejected. Table III gives the complete ANOVA results.

The back knee angles obtained from the RGB video (Table III) show a p-value of 0.29, which is a lot more than 0.05 critical value. So, we can safely assume that the data from multiple samples come from the same population mean. Similar results are obtained for all the other features, but the f-value of the hip angle from depth video is a way too much. It signifies a low variability relative to the variability within each group, which may be caused due to the inherent noise, which remains after filtering. Therefore, despite the low f-value, the sample comes from populations of more or less same mean and hence the angles are more or less accurate.

The Anderson-Darling test finds out that, for working with a particular feature, the proposed way will provide informative results or not. However, the ANOVA and the z-test tell us that given the results are informative or non-informative, how much generalized they will be for different subjects (sample populations). All the features seem to have crossed the critical value for the ANOVA test, thus proving that all come from more or less same sample populations and show less variability when subjects (sample populations) are changed. Thus, clarifying that the hip and the knee angles will follow similar patterns though the values may differ corresponding to different subjects. A similar trend is observed by analyzing the results of the z-test and FDR as all are having their p-statistic above the critical value of 0.05. The variance within the dataset is more than between two datasets (subject's data), thus giving an FDR value less than 1.0.

V. Conclusion and Future Work

We use Kinect v2.0 sensors to capture the color depth and RGB video of subjects, which are then analyzed to extract features. From the RGB video, we extract the knee and hip angles. From their plots, we understand how the knee and the hip angles change in a gait cycle, which is an essential feature for clinical gait analysis. All these are done using the

TABLE III: Overall ANOVA summary for all the extracted gait features.

	Data Summary					ANOVA Summary					
Features	Groups	N	Mean	Std. Dev.	Std. Error	Source	Degrees Of Freedom(DF)	Sum Of Squares(SS)	Mean Square(MS)	F-Stat	P-Value
Hip Angle (Depth)	Group 1	31	189.7895	12.9728	2.33	Between Groups	1	6.581	6.581	0.0295	0.8643
	Group 2	31	190.4501	16.685	2.9967	Within Groups	60	13400.4829	223.3414		
						Total	61	13407.064			
CNN	Group 1	50176	126.3459	54.1858	0.2419	Between Groups	1	6149.6712	6149.6712	2.0271	0.1545
	Group 2	50176	126.841	55.9854	0.2498	Within Groups	100350	304433975.2141	3033.7217		
						Total	100351	304440124.8853			
Front Knee Angle(RGB)	Group 1	32	153.46	17.2323	3.0463	Between Groups	1	568.7653	568.7653	1.8365	0.1803
	Group 2	32	147.4978	17.9568	3.1743	Within Groups	62	19201.3637	309.6994		
						Total	63	19770.129			
Back Knee Angle(RGB)	Group 1	32	160.5337	17.1664	3.0346	Between Groups	1	293.3958	293.3958	1.1031	0.2977
	Group 2	32	156.2515	15.403	2.7229	Within Groups	62	16490.0686	265.9688		
						Total	63	16783.4644			
Hip Angle (RGB)	Group 1	32	204.6874	41.5251	7.3407	Between Groups	1	4057.2823	4057.2823	2.2769	0.1364
	Group 2	32	220.6116	42.8895	7.5819	Within Groups	62	110479.1373	1781.9216		
						Total	63	114536.4197			

TABLE IV: Overall statistical analysis values

Tests	Back Knee Angle	Front Knee Angle	Hip Angle (RGB Video)	Features from CNN	Hip angle (Depth Video)
Hypothesis test (Z-score, p-value)	(1.05, 0.29)	(1.35, 0.18)	(-1.50, 0.13)	(1.4, 0.15)	(0.174, 0.861)
Fisher Discriminant Ratio(FDR)	0.03	0.05	0.07	0.000043	0.00143
Anderson Darling test(A-score , p-value)	(0.914, 0.018)	(0.671, 0.071)	(3.604, <0.005)	(2700, <0.005)	(2.157, <0.005)

OpenPose framework with our augmented novelty. Color depth video is also used for feature extraction. The obtained GEIs are fed to CNN for feature extraction. These extracted features from the color depth and RGB videos are utilized for statistical analysis. The knee angles obtained are as expected. The curve may see some steep edges, but that occurs because the data is extracted from every individual frame, and the leg movement is not the same in every frame. As a result, the graphs show a little bit of irregularity. Also, the used data are interpolated, which helped us in finding out the intermediate points to form a smooth cubic graph. As observed from ANOVA and the Z-tests, the data obtained from each process is more or less consistent, and there is not much variation with the change of subjects. Considering these, the inferences drawn about the various features extracted are as described hereby. The back knee angles proved to be the most promising feature of all the features obtained from RGB video. From the Anderson Darling plots of the front knee angles (Fig. 14), it is seen that they lie somewhat on the straight line and hence are informative. The hip angles from RGB videos are never an optimal choice. The hip angles obtained from the depth videos (Fig. 13) are better as the points seem to fit the line more than for RGB videos. The CNN features obtained from the GEI (Fig. 5) provide sufficient but less information for gait analysis as compared to the other techniques. A problem that we encounter during the calculation of the angles from the depth image is that we are not able to find the elevation of the centroids when they move as the movement of the joints is not always translational. Thus, we are not able to find the knee angles from the depth image, which needs a future analysis. Detection of foot joints and using them for ankle detection from both the color depth and RGB images will also be a part of the future work.

ACKNOWLEDGMENT

We are extremely thankful to Science and Engineering Research Board (SERB), DST, Govt. of India to support this research work. The Kinect v2.0 sensors used in our research experiment are purchased from the project funded by SERB with FILE NO: ECR/2017/000408. We would also like to extend our sincere thanks to the students of Department of Computer Science and Engineering, NIT Rourkela for their uninterrupted co-operation and participation catering to the

data collection. The presentation of this work has also been funded by the project financed under the program of the Polish Minister of Science and Higher Education under the name "Regional Initiative of Excellence" in the years 2019-2022, project number 020/RID/2018/19.

REFERENCES

[1] T. Liu, Y. Inoue, and K. Shibata, "A wearable ground reaction force sensor system and its application to the measurement of extrinsic gait variability," *Sensors*, vol. 10, no. 11, pp. 10240–10255, 2010.

[2] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 2, pp. 316–322, 2005.

[3] F. Dadashi, B. N. Araabi, and H. Soltanian-Zadeh, "Gait recognition using wavelet packet silhouette representation and transductive support vector machines," in *2009 2nd International Congress on Image and Signal Processing*, pp. 1–5, IEEE, 2009.

[4] K. Bashir, T. Xiang, and S. Gong, "Feature selection on gait energy image for human identification," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 985–988, IEEE, 2008.

[5] I. Haritaoglu, D. Harwood, and L. S. Davis, "W/sup 4: real-time surveillance of people and their activities," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 809–830, 2000.

[6] G. L. Foresti, L. Marcenaro, and C. S. Regazzoni, "Automatic detection and indexing of video-event shots for surveillance applications," *IEEE transactions on multimedia*, vol. 4, no. 4, pp. 459–471, 2002.

[7] F. Guan, L. Li, S. S. Ge, and A. P. Loh, "Robust human detection and identification by using stereo and thermal images in human robot interaction," *International Journal of Information Acquisition*, vol. 4, no. 02, pp. 161–183, 2007.

[8] L. Xia, C.-C. Chen, and J. K. Aggarwal, "Human detection using depth information by kinect," in *CVPR 2011 workshops*, pp. 15–22, IEEE, 2011.

[9] S. N. Krishnamurthy, "Human detection and extraction using kinect depth images," *Bournemouth University*, 2011.

[10] S. Ikemura and H. Fujiyoshi, "Real-time human detection using relational depth similarity features," in *Asian Conference on Computer Vision*, pp. 25–38, Springer, 2010.

[11] J. W. Davis and M. A. Keck, "A two-stage template approach to person detection in thermal imagery," in *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)-Volume 1*, vol. 1, pp. 364–369, IEEE, 2005.

[12] M. Correa, G. Hermosilla, R. Verschae, and J. Ruiz-del Solar, "Human detection and identification by robots using thermal and visual information in domestic environments," *Journal of Intelligent & Robotic Systems*, vol. 66, no. 1-2, pp. 223–243, 2012.

[13] L. Spinello and K. O. Arras, "People detection in rgb-d data," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3838–3843, IEEE, 2011.

[14] F. Hegger, N. Hochgeschwender, G. K. Kraetzschmar, and P. G. Ploeger, "People detection in 3d point clouds using local surface normals," in *Robot Soccer World Cup*, pp. 154–165, Springer, 2012.

[15] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International journal of computer vision*, vol. 61, no. 1, pp. 55–79, 2005.

[16] D. Ramanan, D. A. Forsyth, and A. Zisserman, "Strike a pose: Tracking people by finding stylized poses," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 271–278, IEEE, 2005.

[17] Y. Wang and G. Mori, "Multiple tree models for occlusion and spatial constraints in human pose estimation," in *European Conference on Computer Vision*, pp. 710–724, Springer, 2008.

[18] L. Sigal and M. J. Black, "Measure locally, reason globally: Occlusion-sensitive articulated pose estimation," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 2041–2048, IEEE, 2006.

[19] X. Lan and D. P. Huttenlocher, "Beyond trees: Common-factor models for 2d human pose recovery," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 1, pp. 470–477, IEEE, 2005.

[20] L. Karlinsky and S. Ullman, "Using linking features in learning non-parametric part models," in *European Conference on Computer Vision*, pp. 326–339, Springer, 2012.

[21] M. Dantone, J. Gall, C. Leistner, and L. Van Gool, "Human pose estimation using body parts dependent joint regressors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3041–3048, 2013.

[22] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*, pp. 483–499, Springer, 2016.

[23] W. Ouyang, X. Chu, and X. Wang, "Multi-source deep learning for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2329–2336, 2014.

[24] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 648–656, 2015.

[25] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Advances in neural information processing systems*, pp. 1799–1807, 2014.

[26] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *Advances in neural information processing systems*, pp. 1736–1744, 2014.

[27] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1653–1660, 2014.

[28] V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 468–475, IEEE, 2017.

[29] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *European Conference on Computer Vision*, pp. 717–732, Springer, 2016.

[30] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1831–1840, 2017.

[31] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1281–1290, 2017.

[32] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial posenet: A structure-aware convolutional network for human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1212–1221, 2017.

[33] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 190–206, 2018.

[34] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 713–728, 2018.

[35] T. Pfister, J. Charles, and A. Zisserman, "Flowing convnets for human pose estimation in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1913–1921, 2015.

[36] L. Ferreira, A. Neves, A. Pereira, E. Pedrosa, and J. Cunha, "Human detection and tracking using a kinect camera for an autonomous service robot," *Advances in Aritificial Intelligence-Local Proceedings, EPIA*, pp. 276–288, 2013.

[37] D. A. Winter, *Biomechanics and motor control of human movement*. John Wiley & Sons, 2009.

[38] C. R. Nair, "A voting algorithm for dynamic object identification and pose estimation," 2019.