Journal of Computer Science

Original Research Paper

Computer Users Sitting Posture Classification Using Distinct Feature Points and Small Scale Convolutional Neural Network for Humana Computer Intelligent Interactive System During COVID-19

[1, 2, 3]Jheanel Estrada and [2]Larry Vea

[1]School of Information Technology, Mapua University, Makati, Philippines
[2]College of Information Technology Education, Technological Institute of the Philippines, Manila, Philippines
[3]Energy Research Institute @ Nanyang Technological University, Singapore

Article history
Received: 09-11-2022
Revised: 04-01-2023
Accepted: 18-01-2023

Corresponding Author:
Jheanel Estrada
Energy Research Institute @ Nanyang Technological University, Singapore
Email: jheanelestrada29@gmail.com

Abstract: The sudden change in our workplace practices from face-to-face work to work from home setup due to the pandemic has brought positive and negative impacts on our overall health. In literature, the use of deep learning and specialized cameras in the estimation of the human pose is popular even if there is a need for high computational resources and complex models. For this purpose, this study developed an intelligent and interactive system utilizing a human estimation model with the use of distinct keypoint such as thoracic, thoraco lumbar, and lumbar points in the spine. An objective type of a dataset captured in a work from home environment with the knowledge and guidance of Licensed Physical Therapists to assess proper and improper sitting posture was developed. The study developed and implemented a small-scale convolutional network and low-cost smartphone camera to recognize body key points. Once all the feature points' locations were extracted, additional features such as cosine similarity and point distances were calculated. Next, feature selection and optimization were utilized to classify proper and improper sitting postures. As a result, the study developed (2) datasets and (2) models with an accuracy of 85.18 and 92.07% and kappa of 0.691 and 0.838 respectively.

Keywords: CNN, Human Pose Estimation, HCI, Model Development, Work from Home

Introduction

Human pose estimation is one of the challenging tasks of Computer Vision which aims in determining the position by detecting the pixel location of different body parts/joints of a person in each image or video (Kreiss *et al*., 2019; Babu, 2019). Human pose estimation is usually performed using image observations in either 2D or 3D (Chen and Yuille, 2014; Mwiti, 2019). These estimations obtain the pose of the detected person which consists of joints and selected body points. Several approaches were proposed in the literature from the classical use of morphological operators to articulated human pose estimation using a convolutional neural network (Andriluka *et al*., 2009; 2010; Johnson and Everingham, 2010; Pishchulin *et al*., 2013). These methods face challenges such as inaccuracy in determining the point's location and finding the correlation between variables (Yang and Ramanan, 2011; 2012).

Human pose estimation progressed significantly due to the existence of deep learning and widely and publicly available datasets. This can be seen in the applications in the fields of animation and human monitoring (Lin *et al*., 2014; Andriluka *et al*., 2014). Then this can be applied to video surveillance and monitoring, assistance systems used for daily living, and driver systems (Toshev and Szegedy, 2014; Krizhevsky *et al*., 2017).

However, as this field emerges, challenges such as joints that are barely visible, multiple people in a frame, and spatial differences because of clothing, lighting, backgrounds, and complex positions were being looked into. Since the cost of 3D camera sensing has been decreasing over the last few years and the emergence of machine learning, these would bring new and innovative solutions and approaches to these problems.

The COVID-19 pandemic has greatly disrupted many working routines, switching the traditional and typical office based setups to a new normal on short notice. Many studies show that this new normal setup will give considerable negative impacts on employees' overall well-being and productivity (Girshick *et al.*, 2014). Nevertheless, many studies also show positive impacts of Work From Home (WFH) setting (Girshick, 2015; Simonyan and Zisserman, 2014; He *et al*., 2016; Yao *et al*., 2012). With this, human pose estimation systems could be used to assess sitting postures for those employees in a work from home setup.

Direct measurement and/or the use of intrusive device methods such as Inertial Measurement Units (IMUs) and surface Electromyography (sEMG) sensors have been used to assess MSDs risk factors. However, these methods are usually used for monitoring workers' body movements of a few muscles, such that, they are difficult to acquire the ground reaction force data of the whole body (Sasikumar, 2018; Chandna *et al*., 2010). In addition, these methods require sensors to be attached to the workers' skin (Killough *et al*., 1995; Riihimäki, 1995), which makes them feel uncomfortable and inconvenient while performing a given task.

Vision based methods and/or the use of less or non-intrusive devices have been used to assess risk factors for MSDs in relation to their posture. For example, motion tracking systems have been used because of their precision and non-invasiveness. Despite its ability to identify awkward postures, this is limited by the fact that a direct line of sight is required to register the movements.

The popularity of deep learning and convolutional neural network can give way to developing new models to recognize proper and improper sitting posture using distinct key points, low cost and small-scale.

Gathering the entire mentioned gap from the previous studies such as:

a. The existence of deep learning and Convolutional Neural Networks (CNN) running in high performing machines is costly, therefore the development of small-scale CNN is needed
b. When physical and social distancing is a challenge during this time of COVID-19, and physical markers were not feasible therefore the need for virtual markers is more appropriate
c. Though the cost of sensors is slowly decreasing, the use of less expensive and less intrusive devices is needed
d. Determining the relevant body points that contribute to a more accurate classification of sitting posture

Problem, Gap, and Opportunity

With the research mentioned above, human posture recognition using non-contact and non-intrusive methods offers a deeper analysis of image processing techniques. It is one of the challenging tasks due to the variation of human appearances, changes in the background and illumination, additional noise in the frame, and diverse characteristics and amount of data generated. Aside from these, generating a high configuration of recognition of human body parts, occlusion, almost alike similar parts of the body, variations of colors due to clothing, and all other various factors make this task one of the hardest tasks in computer vision. To satisfy this task, some have developed models that used RGB images through background subtraction or geometric transformation, but this leads to high computational cost and a controlled environment (Boulay *et al*., 2003; Moeslund and Granum, 2000; Agarwal and Triggs, 2005). Others have used RGB-D sensors (e.g., kinect) to analyze depth information but this leads to operational costs and highly computational machines (Laptev and Lindeberg, 2005; Laptev *et al*., 2008). This gives way to a non-intrusive posture recognition system using virtual markers and cameras. Compared with sensor based or intrusive devices, image recognition can be done with the easy acquisition, long distance, non-contact, and non-invasive techniques. However, the studies mentioned show the use of large-scale computational devices to meet the computational requirement of convolutional neural networks and deep learning algorithms for this task. With all the research gaps, this study provides opportunities to investigate the use of virtual markers placed in the human body using a capturing tool that has a less computational cost and less powerful devices (i.e., smartphones); lastly the addition of the examination of other human body points, distances, and demographics.

Review of Related Literature and Studies

As the growth of mobile and embedded systems becomes more pervasive, the interaction between its users and computers is increasing. The developments of autonomous and adaptable systems have increased over the decades (Zaslavsky, 2002). Pervasive computing is defined as a paradigm that can produce computational activities in a manner where it can impact and support works. This includes the adaptability of the system when the environment changes. With this, the use of wireless connections has been widespread. It has been widely used to sense people, sense locations, or sense their environments through sensors and wireless capabilities (Li *et al*., 2016). Some of its wide range sensing applications include sports (Baca *et al*., 2009), gesture-based recognition systems (Abdelnasser *et al*., 2015), and posture recognition systems (Yao *et al*., 2015).

With the entire fields of human computer interaction and computer vision, many intelligent interactive systems were developed. Intelligent interactive systems is an interdisciplinary field that shows how to develop natural systems that provide guidelines customized to their user preferences (Munir and Nadeem, 2018).

In the sensor based intelligent interactive system (direct measurement/intrusive devices) various types of sensors have been used to capture relevant points that lead to the detection of posture. Early studies discover the use of one of the most common sensors such as force or pressure sensors (Huang and Ouyang, 2012; Tessendorf *et al*., 2009; Kamiya *et al*., 2008).

The study of Kamiya *et al*. (2008), used a total of (64) Flex force pressure sensors placed on a sheet, and each sensor is placed 30 mm apart from the other sensors on 8 × 8 cells. This used a sampling rate of 12.5 Hz and a set of 64 sensor values is obtained every 80 ms. During the data gathering, when a person sits down on the chair and varies his/her posture, the sum of the pressure values also changes. Therefore, there will be "stable", "unstable", and "changing" parts. This study only captures the stable part that was extracted using the high-frequency cut filter to a given time series. To do this, the study extracted the periods that have a length of more than 1 sec and a difference from the previous frame is less than 0.

However, some issues arose in this study person dependency and time dependency, and weight dependency. This was solved by applying normalization of the pressure sensor values by the distribution of the starting time. To normalize the values, the study recorded the normal posture vector which is equivalent to the ten first frames of the first stable part, and average them. A total of 10 male university students with ages ranging from 21-24 years old and weighing 57-90 kgs were the participants of the study. Then, to get the normalized posture vectors, each subject was asked to sit down back and lean deeply and sit in an upright position. The study captured (9) sitting positions and each posture was maintained for 2-3 sec and breaks in between trials were exhibited. Each subject performed five trials, and 10 frames (~1 sec) were extracted from each stable part cut automatically in each posture. Thus, 450 (= 5 trials × 9 postures × 10 frames) frames were collected for each subject. This study shows an accuracy rate of 93% with position and weight normalizations, however, it is necessary to analyze a long time series of sensor data.

On the other hand, the study of (Mu *et al*., 2010), utilizing a less or non-intrusive device (i.e., 2D camera), used the face's location and size as the features of the sitting posture. To capture the normal posture vector, the user is required to sit and maintain a correct sitting posture for at least (1) min at the start of the data gathering. Then, the face's size and location are extracted and compared using pattern matching based on Harsdorf distance measurement. Using a controlled environment with the same background over time, but still allowing a minimal change in the user's position. The Sobel operator was utilized to get the intersection of the edge image and the frame difference images. Using the face's size and location and Harsdorf distance to calculate the degree of resemblance between the pattern and the part of the image that the pattern is superimposed on. However, this requires a high computational load and to solve that, the search scope is limited by skin color analysis. YCbCr color space is used for skin color identification. The study runs in a controlled environment and therefore is limited to only one user who uses the computer and appears in the image. The user's most left and right points can be detected from the profile image. Thus, the points that have a similar color to the skin in the background can be removed. However, since this study has limited user movement, it shows the problem in tracking head tilt and rotation and if the user's clothes are close to the skin and background color. Other factors such as lighting may cause deviation and noise in the identification of the search box.

In the study of Huang and Ouyang (2012), (7) force sensors were placed under a chair cushion. Since force sensors were very sensitive that even without any action, there are small changes in the information. Using the data sets of the force sensors:

$$Si = \left\{ f_1 \text{ , } f_2 \text{ ,..., } f_n \right\}$$

where, i = 1, 2, 3,.., 7, this will be used to normalize the values using time series analysis with n is the time series equal to 10.

The model presented is as follows:

a. If the total force is equivalent to {1,1,1,1,1,0,0}, then most likely the posture is proper type 1
b. If the total force is equivalent to {1,1,1,1,0,0,0}, then most likely the posture is proper type 2
c. If the total force is equivalent to {1,1,1,0,1,0,0}, then most likely the posture is improper posture of crossing legs type 1 (a)
d. If the total force is equivalent to {1,1,0,1,1,0,0}, then most likely the posture is improper posture of crossing legs type 1 (b)
e. If the total force is equivalent to {1,1,1,0,1,0,0}, then most likely the posture is improper posture of crossing legs type 2 (a)
f. If the total force is equivalent to {1,1,0,1,1,0,0}, then most likely the posture is improper posture of crossing legs type 2 (b)
g. If the total force is equivalent to {1,1,0,1,0,0,1}, then most likely the posture is the improper posture of crooked sitting posture type (a)
h. If the total force is equivalent to {1,1,0,1,0,1,0}, then most likely the posture is the improper posture of crooked sitting posture type (b)

This study shows a simple model but able to do the task of a posture recognition system, however, since the force sensors are very sensitive, this causes a change in the obtained data. Additionally, weight and other attributes were not utilized.

Since the pressure sensor was used to measure posture and shows a great accuracy rate, some points arise on the use of pressure and force sensors as data capture tools in measuring proper posture (Table 1).

With this, accelerometer sensors have been widespread in the recognition of posture. However, in order to determine the reliability of these results, the placements of these sensors are vital. For whole body movement recognition, sensors on the waist, sternum, and lower back show optimal; some sensors were placed in the thigh and ankle to measure leg movement (walking, jogging).

In 2016, since the emergence of intrusive and direct measurement sensors, (Ma *et al*., 2016), they have developed a model to classify sitting postures through an accelerometer sensor attached to the back of the neck. (5) Sitting postures were covered; (1) proper and (4) improper were captured. (6) Subjects with no reports of spinal disorders participated and each subject keeps the (5) postures for (5) min. The features were extracted using Principal Component Analysis (PCA) from the acceleration data and classified using a support vector machine and k-means clustering for the transformed vectors. The experiment was applied to (6) individuals. As a result, SVM correctly classified rate is 95.33% and k-Means Clustering shows a rate of 89.35%. Results show that SVM and k-means provide almost the same capabilities in identifying proper sitting posture, however, SVM shows consistency in other sitting postures. Since the accelerometer sensor was placed at the back of the neck, the study provides a lower accuracy rate in the improper sitting postures such as sitting back arch and sitting cross legged caused by the limited data source.

The initial study of the proponents was conceptualized using the previous study. Estrada and Vea (2016) utilized the use of smartphones (built-in accelerometer sensors) as a direct measurement tool to capture the inclination degrees of (3) spine points. The (3) points namely: Thoracic, thoraco lumbar and lumbar were measured using the built-in accelerometer sensor of the smartphones. These were placed inside an adjustable girdle. On a go signal of the Licensed Physical Therapist, the inclination degrees were captured and stored in a CSV File. The data set has a total of (60) participants (30 female and 30 male) with no reported or known spine disorders as per the assessment of the experts. This contributes to a total of (600) instances. On a go signal of the experts standing on the lateral and anterior side of the participant, the tool will capture (9) improper and (9) proper sitting postures, and in between breaks are utilized to lessen the bias. To normalize the data, using the controlled environment and guidance of the experts, the study mapped the demographics of the participants and the readings of the spinal points. Some of the well-known classifiers such as SVM, random forest, decision tree, and neural network were compared. As a result, the decision tree classifier shows the most acceptable model which has an accuracy rate of 87% and a kappa of 0.8. The study shows promise in utilizing smartphones and their built-in accelerometer sensors; however, the weight and size of the mobile phones made the participants uncomfortable and therefore, affect their sitting positions.

Aside from direct measurements using intrusive devices (i.e., pressure sensors, accelerometer sensors) and vision-based measurement using non or less intrusive devices (i.e., 2D and/or stereo cameras), computer vision is also a widely used technique for human posture recognition systems. There were different studies published to establish an efficient way to recognize human sitting posture. Some of the research deal with the use of visible light cameras, stereo cameras, or RGB-sensor.

Estrada and Vea (2017) the proponents made another contribution, additional features were gathered such as chin, manubrium, and acromion process using physical markers. This was done to measure head and shoulder posture. Using the laptop’s built-in camera, the study developed a PC-based application utilizing MATLAB to get the coordinates of these key points. These raw data were transformed into the distance of the key points such as:

a. Chin to acromion process (left)
b. Chin to the acromion (right)
c. Chin to Manubrium
d. The difference in the Y-Axis of the left and right acromion process

A total of (600) instances (10 captures × 60 participants) were captured. These features were compiled and run into a separate model. As a result, it shows a more acceptable model which has an accuracy of 95% and a kappa of 0.9. This study shows that Body Mass Index (BMI) is a significant factor in the recognition. As a recommendation, this study suggests replacing physical markers with virtual markers.

At this point, it is important to choose the significant feature points (body key points) to recognize proper and improper sitting postures. In 2019, the study of (Kappattanavar *et al*., 2020) utilized a combination of Inertial Measurement Unit (IMU) sensors and a Kinect camera to track and locate (5) key points (i.e., 3 points in the spine, right hip, and the fifth at the sternal angle).

Utilizing (2) Kinect cameras placed to record the depth images of the upper and lower part of the body. The study has (6) subjects (5 male and 1 female) ages 27-34 years old, with weight and height in the range of 61-91 Kg and 169-180 cm respectively. To capture the raw data, the subjects were made to sit for (3) to (5) sec.

Then, the subjects were instructed to sit in four (4) different sitting postures:

a. Forward
b. Backward
c. Lean right
d. Lean left

Table 1: Key insights from the early studies (2008-2016)

Features	# of Samples	Key discovery	Gap
64 Flexiforce sensors	10 male (age 21-24) (weight 57 to 90 kg)	Data normalization using time series (start time). Each posture was maintained for 2-3 sec	Person dependency, weight dependency, and longer time analysis is needed
Face's size and location	10 sample frames	The user is required to sit and maintain a correct sitting posture for at least (1) min in the start of the data gathering. The use of distance measurement comparing the pattern from the normal posture vector and the newly captured image frame	Was not able to recognize head tilt and rotation and if the user's clothes are close to skin color and background color. Other factors such as lighting are also an important factor
7 Force sensors		Simple model that does the task of posture recognition	Force sensors are very sensitive that even without any action is exhibited, there is still a small change in the obtained information
1 Accelerometer sensor placed at the back of the neck	6 subjects	The duration of the capturing is 5 min. Raw accelerometer data was transformed into PCA data	The study provides a lower accuracy rate in the improper sitting postures such as sitting back arch and sitting cross-legged caused by limited data sources
3 Smartphones with built-in accelerometer placed in 3 Spinal points	60 subjects	The raw data captured by the accelerometer built-in the smartphone is the inclination degree that is being are transmitted synchronously to a web server. Aside from the inclination degree, additional data were captured such as age, gender, height, weight, and wrist size which contributes to the user's Body Mass Index (BMI)	Intrusive devices such as these provide discomfort to the users due to its size and weight

Each posture was captured three times (3 ×) and in between takes, the subject is instructed to sit straight. The data was collected in 100 Hz and for approximately 6 min per subject. During the pre-processing, the synchronization of all the sensors was done by measuring the peak of data caused by jumping. To normalize the raw data, mean and standard deviation were extracted. For feature selection, the Recursive Feature Elimination (RFE) with cross validation method with multinomial Logistic Regression (LR) classifier and linear kernel Support Vector Machine (SVM) were used separately. This study suggests that the accuracy of the sitting posture classification task also depends on the location and the optimal number of IMU sensors, an aspect that has been insufficiently addressed in extant research. Moreover, the placement of a single sensor at the thoracic region seems to be a more favorable position when using only one sensor as compared with the other four locations in this study. However, the spine angles cannot be measured using a single sensor, since at least two are needed for this. This gives way to a careful evaluation to assess optimal sensor placement with respect to position and number.

Due to the popularity of neural networks and computer vision, in 2022, the study of (Katayama *et al*., 2022) includes a compact size lidar that scans the environment and generates point cloud data. Then a neural network is designed to extract features and recognize sitting posture from the captured point cloud. The proposed neural network maps the point cloud into pseudo images and then these will be used to train the posture classification part of the network. This is the pilot study that uses 3D point cloud data of LIDAR devices for posture recognition systems. Light Detection and Ranging (LIDAR) is a device to estimate the distance to the surrounding surfaces by measuring the time for a laser pulse to travel from the sensor to a surface and get reflected (Jolly *et al*., 2022). In the study presented, the 3D point cloud data was transformed into a 2D pseudo image using a feature encoder, then a pillar feature network discretized the point cloud into an evenly spaced grid in the x-y pillar. Afterward, (6) dimensions were created for the points. This has been done by adding the distance of the point from the 3D pillar center in each dimension. Thereafter, a stacked pillar tensor is created, enabling extracting a set of features that can be scattered back to a 2D pseudo image. The extracted pseudo image is then fed to the posture recognition network responsible for classifying the posture in the image. The network consists of a 2D-CNN followed by a flattened layer and a SoftMax layer. The SoftMax layer is the output layer consisting of several neurons corresponding to the considered posture. The detected posture is one with the highest SoftMax probability. This initial study showed promise in the posture recognition systems but existing LIDARs in the market are expensive, non-portable, and large. Also, gathering the point cloud in real-time gathering the *x*, *y*, and *z* coordinate information connected to a Raspberry *Pi* 4, and running the neural network is computationally expensive. Therefore, this needs to be transformed into a small-scale neural network running on less computationally expensive devices.

Another study that made significant contributions, (Jolly *et al*., 2022) converts 2D images into 3D using (3) major components (a) Recognizing the face; (b) Converting the image into a 3D model, and (c) Detecting the posture. In the first step of the study, in recognizing the face, a convolutional neural network calculates the distance of vector coordinates. Next, convert a 2D image into a 3D model, using a webcam and Three.js (a general-purpose 3D library) that is used to build and demonstrate animated 3D objects in a web browser using WebGL. Afterward, when the 3D model is generated, the study uses different key points such as:

a. Eyes: Right eye, left eye
b. Shoulders: Left shoulder, right shoulder
c. Ears: Left ear, right ear
d. Eyebrows: Left eyebrow, right eyebrow
e. Nose and other facial points: Mouth

Table 2: Key insights from the early studies (2017-2022)

Features	# of Samples	Key discovery	Gap
7 Feature points	60 subjects (30 male and 30 female)	The use of head and shoulder posture (Chin, Manubrium, left and right shoulder)	Consideration of deep learning and CNN (Estrada and Vea (2017)
3 Points in the spine, 1 neck and 1 hip	6 subjects	Combination of IMU sensors and Kinect cameras	Consideration of more areas/points with respect to their position (Kappattanavar *et al*., 2020)
Pointcloud data (3D images)		A pilot study utilizing LIDAR sensors to capture 3D images then convert them into 2D with the use of convolutional neural network	LIDARs in the market are expensive, non-portable and large. Capturing the 3D point cloud and transformation is computationally expensive. Therefore, this needs to be transformed into a small-scale neural network running in less computational expensive devices (Ahmad *et al*., 2017)
10 Feature points		Converting 2D into 3D images and applying posenet.js and ResNet50 model for the detection of these points	A small-scale model that can run an efficient and accurate model detection of more feature points is recommended (Ma *et al*., 2016)

To detect these points, this study used Posenet.js and ResNet50. Once the 3D input frame passed the model, the algorithm returns a final model incorporating all the key points detected.

To normalize and calibrate all the values, the user is asked to place their face inside the box displayed on the screen. Then the user can move continuously, and the system starts to record the *x*, *y*, and *z* coordinates. The study was able to recognize proper and improper sitting posture based on the threshold value set to 0.2 and 0.3. In the future, considering more user posture data and the number of key points that could detect accurately will help develop a more advanced model. Also, once the model has been trained, convert it into smaller fragments so that they can be easily processed and delivered using the content management system of the web browser. To summarize the studies for the past few years, Table 2.

As a summary of all the mentioned related literature and studies in Appendix A.1, the study was able to identify relevant gaps and useful methodology and implementations. Many studies (Estrada and Vea, 2017; Kappattanavar *et al*., 2020; Ahmad *et al*., 2017) used direct types of measurement (intrusive devices) that vary from the use of pressure sensors, accelerometer sensors, and notch sensors. Due to the availability of sensors, the captured feature points were limited from 1 up to 3 only, therefore it may result in to increase accuracy rate due to a smaller number of samples. However, due to a smaller number of attributes, they recommended exploring other relevant attributes, especially in the upper extremity points. Some studies used vision-based approach (non or less intrusive) (Ma *et al*., 2016; Wang *et al*., 2019; Bhatlawande and Girgaonkar, 2022; Katayama *et al*., 2022; Jolly *et al*., 2022) varies from RGB, RGB-D to Point cloud data. These studies used high powered devices such as Jetson Nano which is good for image processing due to its capability for high computational load. Because of this, the mentioned studies increased the number of attributes/features that range from 5-10. This includes some upper extremity points such as the eyes, shoulder, ear, mouth, and spine. These studies need high computing devices, therefore, making them costly and inaccessible to some.

Materials and Methods

Data Gathering-Participants

This covers males and females with a given height and wrist size. Wrist size together with the participant's height sums up to get their body frame. Each body frame has (5) participants. A total of (60) computer literate individuals will be the respondents of this study. Upon data gathering procedures, all the participants signed an ethical clearance, and an initial examination will be done by the licensed Physical Therapists (PTs) to ensure that all participants are healthy and do not have any existing posture problems such as kyphosis and like (Table 3).

Basic exploratory analysis was utilized in the total number of participants. The age of the participants ranges from 18-44 years old (mean = 26), table height ranges from 17-31 inches (mean = 28), chair height ranges from 16.5 to 29 inches (mean = 17), and a distance of 20 inches from computer monitor to the participant across all samples.

Keypoints

The virtual markers were attached to the upper extremity points. It covers the wrist, arm, shoulder, neck, head, and back of the participant. The same keypoints were used for both the left and right sides. Keypoints 1-4 were on the right side of the participant while keypoints 5-8 were on the left side. Keypoints 9-11 were found at the back of the participants and captured in the lateral view. Chin and nose were added to measure the distances of some keypoints and will be used later for the feature extraction. The keypoints were stated in Table 4.

Data Gathering Setup

The setup for the data gathering is as follows:

1. The left camera (smartphone) was positioned at a 90° angle perpendicular to the knee of the participant at a given height of 2.5 feet with a distance of 3 feet

2. The right camera (smartphone) was positioned in the same manner as the left camera
3. A front camera was also provided so the physical therapists (experts) will be able to monitor and communicate with the participants

    a. The distance of the monitor from the user is 20 inches
    b. The height of the monitor varies based on the preferences of the user

4. Before the data gathering took place, these (2) cameras were connected to a single network and sends the video (frames) on a real time basis to a specific web address
5. On the side of the PT, the web address was given to them and accessible using a web browser
6. Once the setup was done, a zoom meeting link will be given to the experts
7. Additionally, the technical team will measure the table height and chair height

After the live streaming, due to a very big amount of data, only the 2 cameras (left and right) will be recorded and saved into the server.

Data Capturing Tool

1. The study is composed of (2) portable smartphone cameras that were placed at the lateral (1 left and 1 right) of the participant and (1) web camera for monitoring purposes

The smartphone's minimum specifications for capturing videos are:

    a. 1280 × 720 resolutions
    b. 30 fps capturing capability

2. Using the web camera, a zoom meeting was facilitated and essential for the Licensed Physical Therapists (PTs) to monitor the data gathering procedures and check the sitting posture every ten 10 sec
3. While the (2) smartphone cameras placed in the lateral are essential for live streaming purposes so that the experts will be able to see the movements of the participants in real time. The smartphones were connected to a single server and accessible thru a web browser. This is done so the experts will be able to access and see the videos
4. Only the (2) smartphones recording will be used for feature extraction in the later stages. An informed consent form was signed by the participants to ensure that confidential data (audio) will not be recorded

The setup for the data gathering is as follows (Fig. 1):

1. The left camera (smartphone) was positioned at a 90° angle perpendicular to the knee of the participant at a given height of 2.5 feet with a distance of 3 feet
2. The right camera (smartphone) was positioned in the same manner as the left camera
3. A front camera was also provided so the participants could see and talk to the PT (Physical Therapists)

    a. The distance of the monitor from the user is 20 inches
    b. The height of the monitor varies based on the preferences of the user

4. Before the data gathering took place, these (2) cameras were connected to a single network and sends the video (frames) on a real time basis to a specific web address
5. On the side of the PT, the web address was given to them and accessible using a web browser
6. Once the setup was done, a zoom meeting link will be given to the experts
7. Additionally, the technical team will measure the table height and chair height
8. After the live streaming, due to a very big amount of data, only the 2 cameras (left and right) will be recorded and saved into the server

Table 3: Body frame categories

	Height	Wrist size	Category	No. of participants
Female	5' 2” or less tall	<5.5"	Small	5
		5.5" to 5.75"	Medium	5
		>5.75"	Large	5
Female	5' 2” to 5' 5” tall	<6"	Small	5
		6" to 6.25 "	Medium	5
		>6.25"	Large	5
Male	5'5” or less tall	<6"	Small	5
		6" to 6.25 "	Medium	5
		>6.25"	Large	5
Male	taller than 5' 5”	<6.25"	Small	5
		6.25" to 6.5"	Medium	5
		>6.5"	Large	5

Table 4: List of keypoints

	Keypoints	Remarks
1	Sternocleidomastoid process (right)	Neck right
2	Brachioradialis (right)	Elbow right
3	Deltoids (right)	Shoulder right
4	*Trapezius muscle* (right)	Upper back right
5	Sternocleidomastoid process (left)	Neck left
6	Brachioradialis (left)	Elbow left
7	Deltoids (left)	Shoulder left
8	*Trapezius muscle* (left)	Upper back left
9	Thoracic	Upper mid back
10	Thoraco-lumbar	Middle back
11	Lumbar	Lower back
12	Mentalis	Chin
13	Nose	Nose

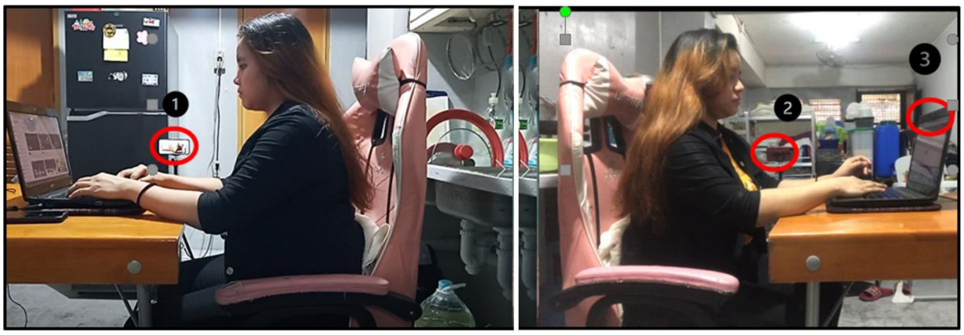


Fig. 1. Data gathering setup

Keypoints Extraction Tool

Mediapipe was used to extract keypoints. It was used to capture This keypoints extraction tool is intended for capturing and tracking high fidelity body pose. This runs in small-scale devices such as desktops/laptops and mobile phones. Other machine learning solutions require powerful desktop environments to generate accurate results.

The pipeline of mediapipe consists of pose detection and tracking each keypoint. The advantage of this framework is its fast detection and low latency. Therefore, to achieve the best and fast performance of detection and tracking, head as the most visible part of the frame will be detected and tracked. This will help in the calculation and detection of the subject person within the frame. Then, this adds (2) additional virtual points that determines the center of the human body, its rotation and scale.

Figure 6, a circle was created to predicts the center of the human body using the pose detector. Afterwards, the midpoints of a person's hips in the frame, the radius of a circle to map the whole person in the frame, and the incline angle of the line connecting the shoulder and hip midpoints will be predicted.

This study by running the pose detector in the first frame of the input video that will localize the region of interest (person) and draw a bounding box. The pose tracker will then predicts all 33 key points (Fig. 2) and run through all the subsequent frames using the previous frame's ROI. This only calls the detection model when it fails to reach the target confidence score (which means fails to track the person).

Distinct Feature Points Recognition

Aside from the recognized feature points, this study used distinct feature points namely chin, thoracic, thoraco-lumbar, and lumbar (spinal points). The distinct key points were computed using the midpoint formula shown in the equation below:

$$midpoint = \left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2} \right) \quad (1)$$

To locate the chin, the midpoint of the left or right shoulder and nose is computed. To locate the thoracic, the midpoint of the left and right shoulder is computed. To locate the lumbar, the midpoint of the left and right hips is computed, and lastly, to locate the thoraco lumbar, the midpoint of the thoracic and lumbar is computed.

As a result, these points will calculate the location of each distinct point. After this, the study will calculate the distances of relevant feature points using the distance formula as shown below:

$$f\left(\left(x_1, y_1\right), \left(x_2, y_2\right)\right) = \sqrt{\left(\left(x_1 - x_2\right)^2 + \left(y_1 - y_2\right)\right)^2} \quad (2)$$

The distances of the following feature points will be calculated (Fig. 3 and Appendix B):

a. The difference between thoracic, thoraco lumbar, and lumbar (spinal points) to Y-Axis
b. Shoulder left (*Trapezius muscle*) and nose distance
c. Shoulder right (*Trapezius muscle*) and nose distance
d. Angle computed from shoulder, elbow, and wrist
e. Angle computed from the shoulder, thoraco lumbar, and lumbar
f. Nose and thoracic distance

To get the angle mentioned in D and E follow the procedure mentioned Table 6:

1. First, find the distance between point A (shoulder left) and point b (elbow left) using the distance formula
2. Second, find the distance between point b (elbow left) and point c (wrist left)
3. Then, find the distance between point c (wrist left) and point a (shoulder left)
4. Lastly, apply the cosine rule:

$$elbow angle = \cos^{-1}\left[\frac{\left(a^2+b^2-c^2\right)}{2ab}\right] \quad (3)$$

With all these raw and computed feature points, the final list of features is listed Table 5.

Video Annotation

Once the video is recorded, annotation is needed to extract relevant features. The annotation process was done as follows:

1. (3) Experts (licensed physical therapists) served as annotators
2. This will be done in a blind annotation process (each annotator will work independently)
3. To check the postures, a guideline was set and stated in Table 7
4. The posture checking is done every (10) sec and subdivided into (7) categories

    a. From the first (1st) to (the 10th) second, the annotator will check for the dominant posture
    b. Dominant posture is determined by a greater number of seconds (e.g., (6) sec of proper and (4) sec of improper, then the dominant posture is proper, otherwise improper)

5. The overall class as seen in Fig. 4 is computed on a dominant posture based on (7) categories (the dominant posture among the (7) categories, if proper > = 4 then proper, otherwise improper)
6. Lastly, the csv file is composed of (2) parts as seen in Fig. 4

    a. Part 1 is predefined attributes based on interviews and measurements conducted before the actual data gathering
    b. Part 2 is the annotation from the experts

Data Preparation

Once the video annotation was done, there will be three (3) CSV files as the output from the (3) annotators. The data preparation is done as follows:

1. Part 1 of the CSV file is the same as the (3) CSV files
2. Part 2 of the CSV file will be compared among the three (3) CSV files (experts' annotation)

    a. To do this, each of the categories will be compared among the (3) CSV files (Fig. 5)
    b. If two out of three (2 out of 3) annotators agreed on a label then it will be the overall label for that instance (e.g., Row 2, PT1 = Improper, PT2 = Proper, PT3 = Proper, then the overall label for it is proper)

Table 5: List of demographic features

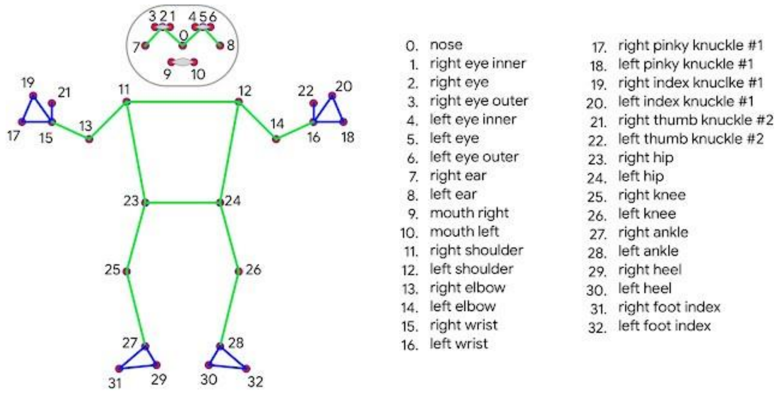
#	Feature/attribute	Description
1	Name	Name of the participants (used as an identifier)
2	Age	Integer
3	Gender	Male or female
4	Table height	Table height appropriate to the user's body frame (upper part of the table o floor)/inches
5	Chair height	Chair height appropriate to the user's body frame (knees to the floor)/inches
6	Distance	Distance between the table and chair of the user (distance from the end of the table to the end of the chair/inches
7	Category	F_small, F_medium, F_large, F1_small, F1_medium, F1_large M_small, M_medium, M_large, M1_small, M1_medium, M1_large

Table 6: List of upper extremity points

#	Feature	Description
1	Diffy axis	Thoracic, thoraco-lumbar, and lumbar with respect to Y-Axis
2	Nose to left shoulder	Nose to left shoulder (deltoids) distance
3	Nose to right shoulder	Nose to right shoulder (deltoids) distance
4	Shoulder_elbow_dist	Shoulder to elbow (brachioradialis) distance in the left or right camera
5	Elbow_wrist_dist	Elbow (brachioradialis) to wrist distance in the left or right camera
6	Wrist_shoulder_dist	Wrist to shoulder in the left or right camera
7	Shoulder_mid_dist	Shoulder to mid (thoraco-lumbar) distance in the left or right camera
8	Mid_hip_dist	Mid (thoraco-lumbar) to Hip (Lumbar) distance in the left or right camera
9	Hip_shoulder_dist	Hip (lumbar) to shoulder distance in the left or right camera
10	Nose to neck	Nose to Neck (Thoracic) distance in the right camera
11	SEW angle A	Cosine similarity of angle A (Shoulder-Elbow-Wrist)
12	SEW angle B	Cosine similarity of angle B (Shoulder-Elbow-Wrist)
13	SEW angle C	Cosine similarity of angle C (Shoulder-Elbow-Wrist)
14	Label	Proper or improper

Table 7: Posture states criteria

No.	Landmark	Remarks
1	Head	Eye level with monitor, not too forward or close to monitor
2	Neck	Neutral (not too bent backward, forward, or to the side)
3	Shoulders	Levelled (not raised or rounded)
4	Elbows	Not too flexed or extended
5	Wrist	Slightly extended or neutral. Not too flexed or extended
6	Upper back	No kyphotic or lordotic posture, or rounded or shifting to one side
7	Lower back	No kyphotic or rounded or shifting to one side



0.	nose
1.	right eye inner
2.	right eye
3.	right eye outer
4.	left eye inner
5.	left eye
6.	left eye outer
7.	right ear
8.	left ear
9.	mouth right
10.	mouth left
11.	right shoulder
12.	left shoulder
13.	right elbow
14.	left elbow
15.	right wrist
16.	left wrist
17.	right pinky knuckle #1
18.	left pinky knuckle #1
19.	right index knuclke #1
20.	left index knuckle #1
21.	right thumb knuckle #2
22.	left thumb knuckle #2
23.	right hip
24.	left hip
25.	right knee
26.	left knee
27.	right ankle
28.	left ankle
29.	right heel
30.	left heel
31.	right foot index
32.	left foot index

Fig. 2: List of keypoints

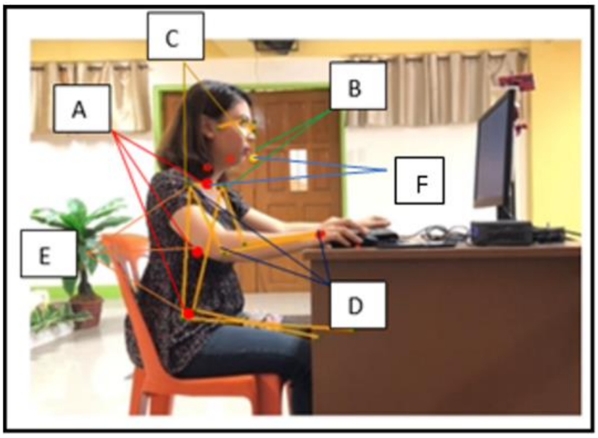
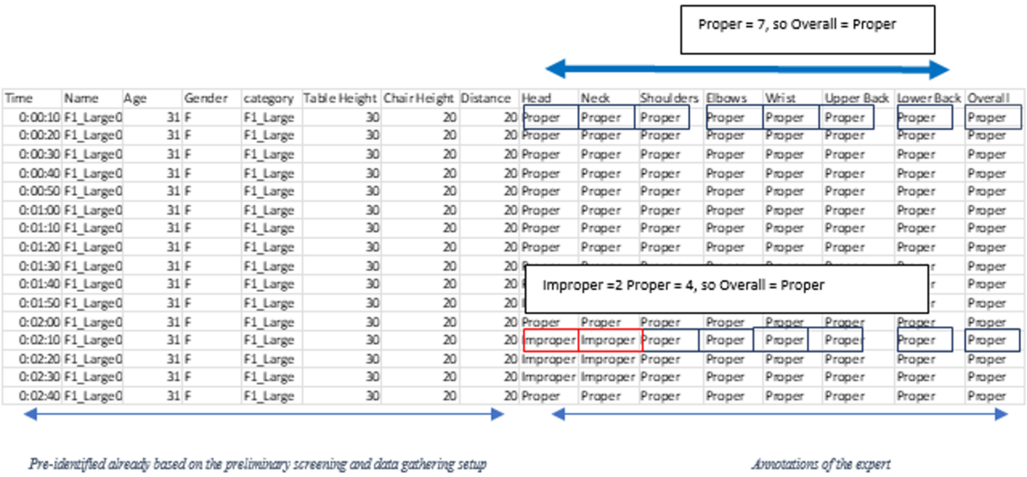


Fig. 3: Additional computed features



Proper = 7, so Overall = Proper

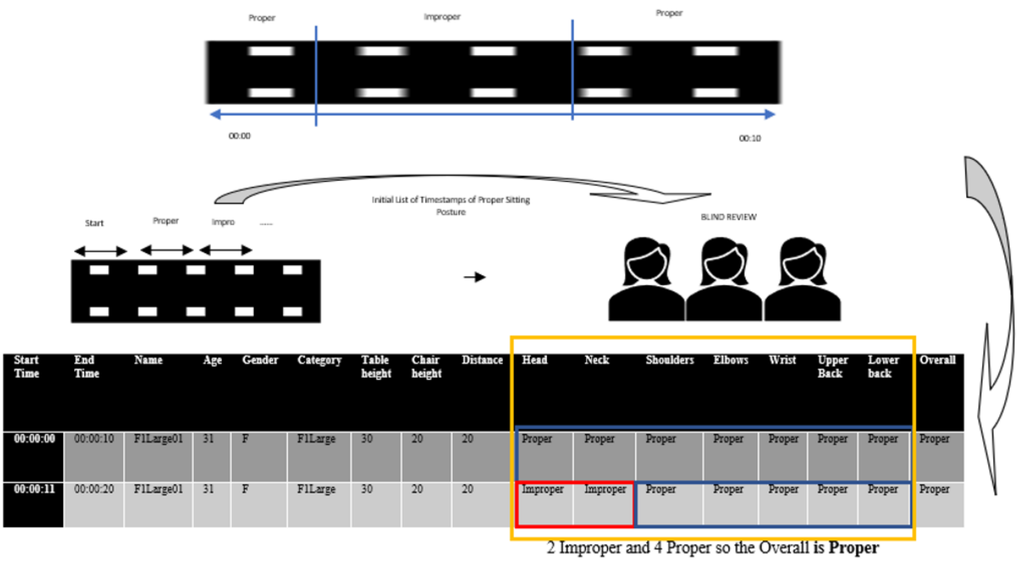
Time	Name	Age	Gender	category	Table Height	Chair Height	Distance	Head	Neck	Shoulders	Elbows	Wrist	Upper Back	Lower Back	Overall
0:00:10	F1_Large0	31	F	F1_Large	30	20	20	Proper	Proper	Proper	Proper	Proper	Proper	Proper	Proper
0:00:20	F1_Large0	31	F	F1_Large	30	20	20	Proper	Proper	Proper	Proper	Proper	Proper	Proper	Proper
0:00:30	F1_Large0	31	F	F1_Large	30	20	20	Proper	Proper	Proper	Proper	Proper	Proper	Proper	Proper
0:00:40	F1_Large0	31	F	F1_Large	30	20	20	Proper	Proper	Proper	Proper	Proper	Proper	Proper	Proper
0:00:50	F1_Large0	31	F	F1_Large	30	20	20	Proper	Proper	Proper	Proper	Proper	Proper	Proper	Proper
0:01:00	F1_Large0	31	F	F1_Large	30	20	20	Proper	Proper	Proper	Proper	Proper	Proper	Proper	Proper
0:01:10	F1_Large0	31	F	F1_Large	30	20	20	Proper	Proper	Proper	Proper	Proper	Proper	Proper	Proper
0:01:20	F1_Large0	31	F	F1_Large	30	20	20	Proper	Proper	Proper	Proper	Proper	Proper	Proper	Proper
0:01:30	F1_Large0	31	F	F1_Large	30	20	20	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Proper
0:01:40	F1_Large0	31	F	F1_Large	30	20	20	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Proper
0:01:50	F1_Large0	31	F	F1_Large	30	20	20	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Proper
0:02:00	F1_Large0	31	F	F1_Large	30	20	20	Proper	Proper	Proper	Proper	Proper	Proper	Proper	Proper
0:02:10	F1_Large0	31	F	F1_Large	30	20	20	Improper	Improper	Proper	Proper	Proper	Proper	Proper	Proper
0:02:20	F1_Large0	31	F	F1_Large	30	20	20	Improper	Improper	Proper	Proper	Proper	Proper	Proper	Proper
0:02:30	F1_Large0	31	F	F1_Large	30	20	20	Improper	Improper	Proper	Proper	Proper	Proper	Proper	Proper
0:02:40	F1_Large0	31	F	F1_Large	30	20	20	Proper	Proper	Proper	Proper	Proper	Proper	Proper	Proper

Improper =2 Proper = 4, so Overall = Proper

Pre-identified already based on the preliminary screening and data gathering setup

Annotations of the expert

Fig. 4: Video annotation CSV file



Start Time	End Time	Name	Age	Gender	Category	Table height	Chair height	Distance	Head	Neck	Shoulders	Elbows	Wrist	Upper Back	Lower back	Overall
00:00:00	00:00:10	F1Large01	31	F	F1Large	30	20	20	Proper	Proper	Proper	Proper	Proper	Proper	Proper	Proper
00:00:11	00:00:20	F1Large01	31	F	F1Large	30	20	20	Improper	Improper	Proper	Proper	Proper	Proper	Proper	Proper

2 Improper and 4 Proper so the Overall is **Proper**

Fig. 5: Experts' annotation

As seen in Eq. 4 below, a total of (60) participants were gathered, and (20) min of data gathering for each participant. Twenty minutes has 1200 sec but since the annotation will be done every (10) sec, a total of (120) instances per participant will be recorded. Moreover, multiplying it again by the total number of participants, the overall number of instances is (7, 200).

FPS (frame 1 sec)	Frame per 10 sec	Frame per 10 min per participant	Total # of participant	Total # of frames per setting posture	Total # of frames
10	100	6000	60	360000	720000
# of instances 1 min	# of instances in 10 min	# of instances per participant per posture	Total # of participant	Total # of instances	
6	60	120	60	7200	

$$60\ participants = 20\,minutes = \frac{1200}{100}\,seconds$$
$$= 120\ instances\ per\ participant \times 60 \tag{4}$$
$$= 7200\ total\,instances$$

Results

Kappa Agreement Among the Experts

During the data gathering, there were (3) experts to facilitate and annotate the video capturing and recording. To measure the agreement among the (3) experts who did the annotation, Fleiss' Kappa was calculated (Table 8). Based on the table, PT1 and PT2 have a moderate agreement for all the instances for both left and right CSV annotation files; PT2 and PT3 have a moderate agreement as well on the right CSV file.

Based on the results (Table 8), even the experts have a hard time coming to a unanimous observation and recognition of proper and improper sitting posture simply by using observation and a set of guidelines. During the development of the dataset with (3) experts, the highest kappa among them is 0.5211901 for the left camera and 0.591223 for the right camera. As a start of the development of an interactive intelligent system for posture recognition system, it shows moderate agreement among the experts and will be used for the development of the model.

Pose Estimation Results

The workflow starts with capturing a frame with the proper setup presented in this study. The next step is to convert the frame into an RGB frame and process using a background segmentation mask to the whole body. The next step is to locate the Region of Interest (RoI) within the frame. To locate the ROI of an image frame, it follows a two-step process (Fig. 6):

1. Detect face due to its high contrast features and comparably small in appearance with the use of blaze face model
2. Explicitly predict additional (2) virtual key points to describe and locate the center of a human body. From this, it will predict the midpoint of a person's hips that results in a consistent tracking

Training Set

Given the proper data gathering setup, to build a good classifier, the study was able to capture an acceptable number of samples covering (2) different positions proper and improper sitting posture. The study captured 10 frames per second for 20 min. A total of 12,000 frames were captured for each participant. The study has a total of 60 participants; therefore, the study has a total of 720, 000 frames.

Recognition of Raw Feature Points

The algorithm and pipeline of MediaPipe used for human pose classification requires a feature vector representation of each sample. To do this, pairwise distances between predefined lists of joints were computed (Fig. 7).

Performance of the Recognition

To evaluate the quality of the recognition model compared to other well-known models, mean average precision (mAP) and Percentage of Correct Keypoints (PCK) were used. Mean average precision is based on the following sub-metrics:

a. Confusion matrix heavily relies on True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN)
b. Intersection Over Union indicates the intersection of the coordinates of the predicted bounding box to the actual bounding box (ground truth). Higher IoU shows that the coordinates of the predicted bounding box resemble the actual bounding box coordinates
c. Precision measures how well the performance of the true positive out of all positive predictions
d. Recall measures how well the performance of the true positive out of all the predictions

The Percentage of Correct Keypoints (PCK) is the metric used for identifying if a detected joint is considered correct if the distance between the predicted and the true joint is within a certain threshold (threshold varies). For this study, it used PCK@0.2 = distance between predicted and true joint <0.2 * torso diameter. This metric is usually used for 2D and/or 3D (PCK3D).

The performance of the model is shown in the figure below. The model has deployed (3) versions namely heavy, full, and lite. As seen in the figure below, the model outperformed the other existing solutions in mAP and PCK (Fig. 8).

Model Performances (Left and Right Camera Models)

Figure 9, shows the accuracy of the model given the optimal values for each parameter mentioned in Tables 9-10. The decision tree shows 85.18 accuracies, 0.691 kappa; rule induction shows 84.5 accuracies, and 0.679 kappa for the left camera dataset.

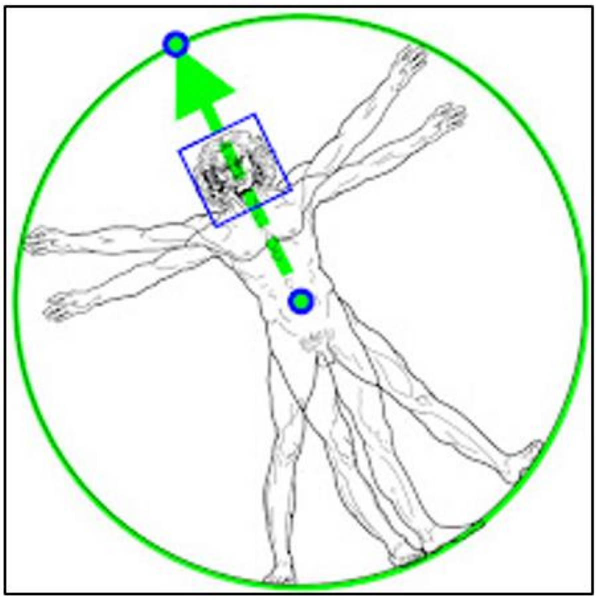


Fig. 6: Two virtual key points and the blaze face detector

The comparison of both left and right camera models were presented in Fig. 9. (2) Most appropriate models have been presented namely decision tree and rule induction. The rules of the models were presented in Tables 9-10.

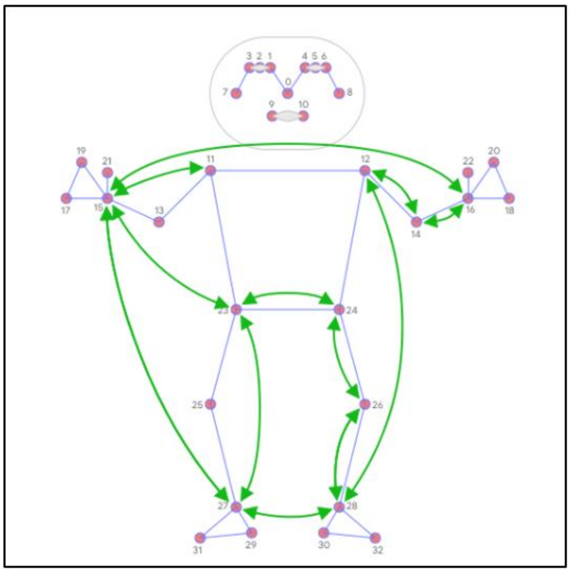


Fig. 7: Pairwise distances between predefined lists

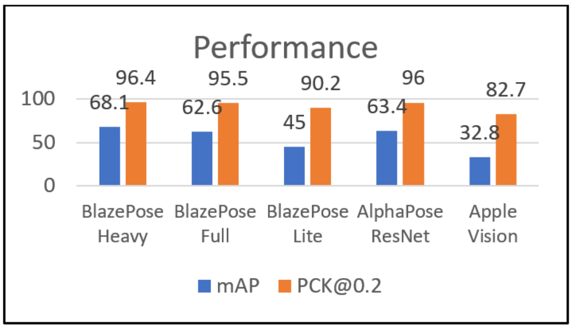


Fig. 8: Recognition performance comparison of different well-known models

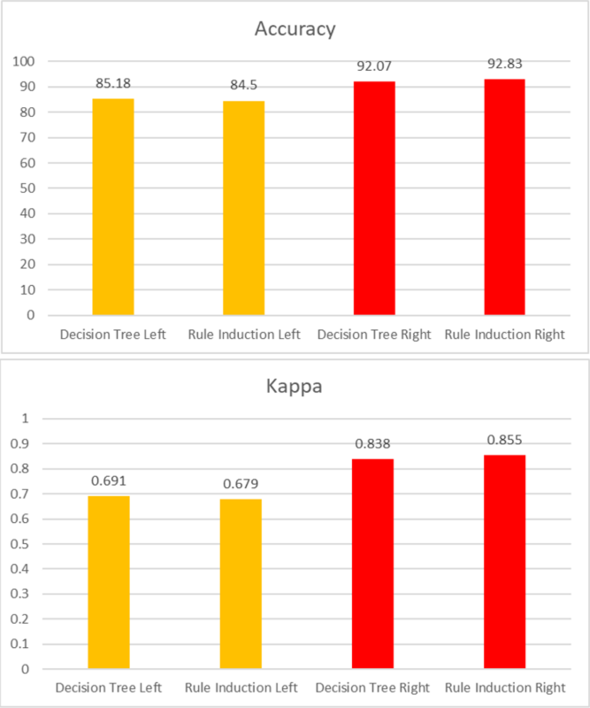


Fig. 9: Accuracy and kappa of decision tree and rule induction for right camera dataset

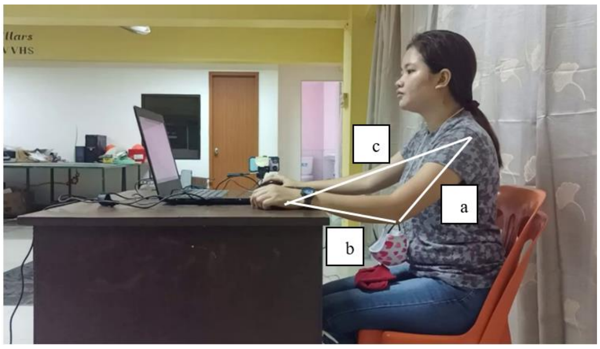


Fig. 10: Features for the computation of cosine angle (shoulder, elbow, and wrist)

Table 8: Kappa agreement among experts

	PT1 and PT2	PT2 and PT3	PT3 and PT1
Left	0.460053319	0.141669493	0.5211901
Right	0.466282672	0.466282672	0.591223
	PT1, PT2 and PT3		
Left	0.387105772		
Right	0.466282672		

Table 9: Decision tree rules for the left camera

Rules	Most likely to
If meannoseleftshoulder > 343.267 and meanshoulder_elbow_dist > 153.040 and medianelbow_wrist_dist > 203.572	Proper
If meannoseleftshoulder > 343.267 and meanshoulder_elbow_dist > 153.040 and medianelbow_wrist_dist <= 203.572	Improper
If meannoseleftshoulder > 343.267 and meanshoulder_elbow_dist <= 153.040 and medianshoulder_mid_dist > 56.117	Improper
If meannoseleftshoulder > 343.267 and meanshoulder_elbow_dist <= 153.040 and medianshoulder_mid_dist <= 56.117	Proper
IF meannoseleftshoulder <= 343.627 and Age >18.5 and stdSWEAngleA >3.293 and tableheight <=30	Improper
IF meannoseleftshoulder <= 343.627 and Age >18.5 and stdSWEAngleA <= 3.293 and meanhip_shoulder_dist <= 588.166	Improper

Table 10: Decision tree rules for the right camera

Rules	Most likely to
If medianwrist_shoulder_dist > 287.477 and stdnosetoleftshoulder <= 8.028 and meanhip_shoulder_dist <= 328.643 and Age > 19	Proper
If medianwrist_shoulder_dist > 287.477 and stdnosetoleftshoulder > 8.028 and stdelbow_wrist_dist <=24.754 and medianelbow_wrist_dist > 157.938	Improper
If medianwrist_shoulder_dist <=287.477 and tableheight <=30.500 and medianSEWAngleC <=122.449 and meanwrist_ shoulder_dist <=293.595	Improper

Discussion

Significant Attributes (Upper Extremity Points)

In the rules generated by the acceptable model, this study has found out that the significant features are the following for the left and right models respectively (Tables 11-12):

Tables 11-12, both models have almost identical attributes that are significant in the recognition of proper and improper sitting postures. Both models have a nose to left shoulder distance and elbow and wrist distance. Nose to left shoulder is significant to measure the posture of the head and shoulder and elbow and wrist distance is used to identify the typing position.

Significant Attributes (Demographics and Ergonomic Elements)

Aside from the feature points mentioned in Tables 11-12, ergonomic design is also part of the dataset which includes table height, chair height, and the distance of the user from the monitor. Table height and chair height signify a correlation to other body feature points. Specifically, table height shows a correlation to raw feature points while chair height shows a correlation to the computed feature points.

Based on the models presented in Tables 9-10, age and table height show relevance in the recognition of proper and improper sitting posture. It has been noticeable that the body frame does not have any direct relationship with the recognition.

The table height shows a correlation to cosine angles A and C consistently while chair height is highly correlated to cosine angle B. The study found out the relationship of ergonomic elements (table and chair height) to cosine angles of shoulder, elbow, and wrist.

For left camera model, the significant features are nose to left shoulder, left shoulder to the elbow, elbow to wrist, and shoulder to mid-hip (thoraco lumbar). It is also noticeable that age and table height was part of the rules in the classification (Fig. 10).

For right camera model, also shows almost the same feature points such as nose to left shoulder, left shoulder to the elbow, elbow to wrist, shoulder to wrist, and shoulder to hip (lumbar). It is also noticeable that age and table height was part of the rules in the classification. Also, even though this model particularly was captured on the right side of the participant, it only chooses the distance between the nose to the left shoulder (not the right) as part of the rules.

Therefore, the insights that we can get from these rules are the following:

a. Comparing the (2) models, the right camera models include nosetoneck and variance, which greatly impacts the accuracy of the model. With the same processes from feature selection, optimization, and model development, it shows an increase of ~7% in accuracy and ~0.2 in kappa
b. Nose, shoulder, elbow, wrist, age, and table height are present in both left and right camera models
c. Since the feature points are almost identical for both left and right camera models, but the right camera model shows an additional feature (distance between wrist and shoulder, distance between nose and neck, and shoulder elbow wrist angle C), we could utilize the right camera model for the detection
d. The distance between the nose and the left shoulder is present in both left and right camera models (the right camera model does not use the nose to the right shoulder as a determining factor for the classification)
e. Age and table height shows that they are considered important factors in the classification of sitting posture
f. The standard table height regardless of body frame is higher than 30 inches

For prototype, the application can be run using any web browser. The application has a responsive capability to fit into any device's screen. Figure 11, the user can register or log in using the given credentials. Once the user logs in, it will show the user's dashboard. It will show (2) buttons video and profile edit (Fig. 12).

Next, under the video button, this will show the following:

a. Record video this can be used in smartphones that are placed on the left and right side of the user
b. Upload video this can be used on laptops when there is already a recorded video available
c. Show list of processed videos this is a list of all the pending (uploaded or recorded video that has not been processed yet) and processed videos. This shows the position (left or right), age, table height, date created, and status

Table 11: Significant attributes (upper extremity points) in the left model

	Attribute	Description	Explanation
1	Nose left shoulder	Distance between nose to left shoulder	Monitor the movement of head (leaning forward, backward, left or right)
2	Shoulder_elbow_distance	Distance between left shoulder and left elbow	Monitor the movement of shoulder and elbow angle
3	Elbow_wrist_distance	Distance between left elbow and left wrist	Monitor the angle between elbow and wrist
4	SWE angle A	Shoulder, elbow, and wrist angle A	This will properly assess the angle of these three (3) points

Table 12: Significant attributes (upper extremity points) in the right model

	Attribute	Description	Explanation
1	Wrist_shoulder_distance	Distance between right wrist and right shoulder	Monitor the angle of shoulder and wrist
2	Nose to left shoulder	Distance between nose and left shoulder	Monitor the movement of head (leaning forward, backward, left and right)
3	Elbow_wrist_distance	Distance between right elbow and right wrist	Monitor the angle between elbow and wrist
4	SEW angle C	Shoulder, elbow, and wrist angle C	This will properly assess the angle of (3) points

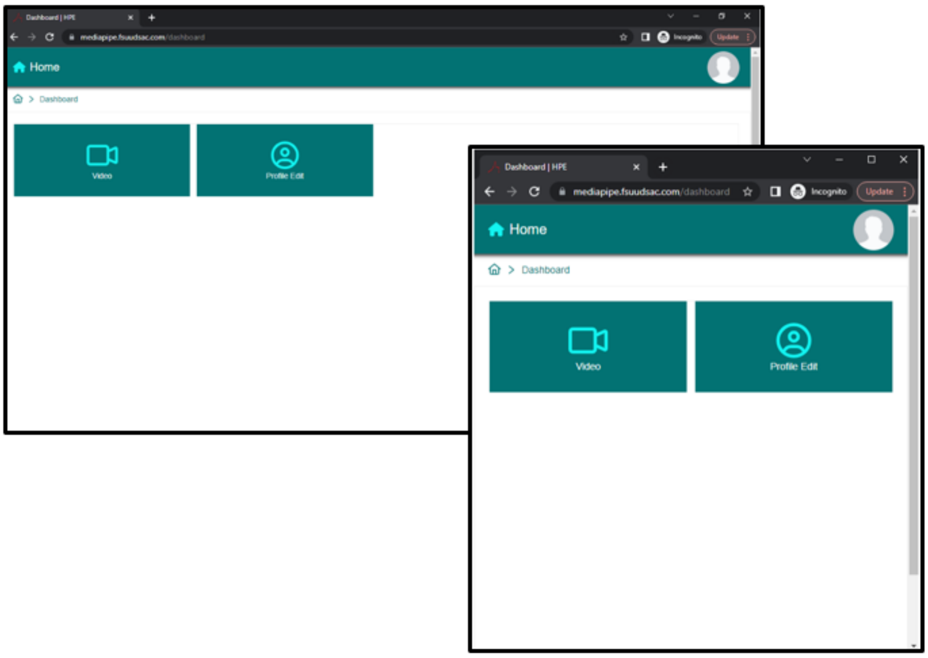
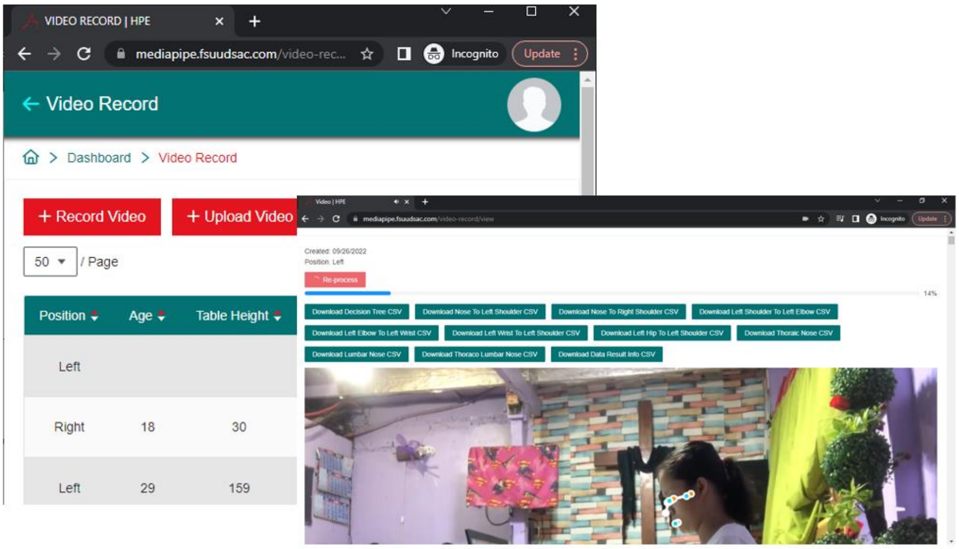


Fig. 11: Prototype dashboard


VIDEO RECORD | HPE
mediapipe.fsuudsac.com/video-rec...
Incognito
Update
Video Record
Dashboard
Video Record
+ Record Video
+ Upload Video
50
/ Page
Video | HPE
mediapipe.fsuudsac.com/video-record/view
Re-process
14%
Download Decision Tree CSV
Download Nose To Left Shoulder CSV
Download Nose To Right Shoulder CSV
Download Left Shoulder To Left Elbow CSV
Download Left Elbow To Left Wrist CSV
Download Left Wrist To Left Shoulder CSV
Download Left Hip To Left Shoulder CSV
Download Thoraic Nose CSV
Download Lumbar Nose CSV
Download Thoraco Lumbar Nose CSV
Download Data Result Info CSV

Position	Age	Table Height
Left		
Right	18	30
Left	29	159

Created: 09/26/2022
Position: Left

Fig. 12: Video processing

Figure 12 a sample processed video with the recognized points using a small scale convolutional neural network. The small-scale NN runs in both smartphones and CPU-based laptops. To process a one-minute video, the NN needs an additional 30 sec. This is a good performance considering that it only runs on CPU-based devices.

Conclusion

The study was able to recognize proper and improper sitting posture and develop an acceptable model. The conduct of the study, therefore concludes the following:

a. To develop an objective dataset, even experts had a hard time unanimously recognizing the proper sitting posture
b. Upper extremity points such as the shoulder, elbow, and wrist show great significance in recognizing proper and improper sitting posture
c. While body frame was considered a significant factor, it shows that age has a greater significance in the recognition; table height, on the other hand, is also part of the significant features
d. The left and right camera dataset were compared which shows an accuracy of 85.18 and 92.07% and kappa of 0.691 and 0.838 respectively

Acknowledgment

This study would like to thank the college of information technology education for the support they have given the authors to finish this study. This study also extends a deepest gratitude to all the participants and to the licensed physical therapists (Grace, Kath and Cristelle) for giving their time, expertise and effort to the success of this project.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Jheanel E. Estrada: Participated in design, testing and implementation. Design the data gathering and analysis. Did all the experiments relevant to the study.

Larry Vea: Designed the research plan and organized the study.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Abdelnasser, H., Youssef, M., & Harras, K. A. (2015, April). Wigest: A ubiquitous wifi-based gesture recognition system. In *2015 IEEE Conference on Computer Communications (INFOCOM)* (pp. 1472-1480). IEEE. https://doi.org/10.1109/INFOCOM.2015.7218525

Agarwal, A., & Triggs, B. (2005). Recovering 3D human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(1), 44-58. https://doi.org/10.1109/TPAMI.2006.21

Ahmad, J., Andersson, H., & Sidén, J. (2017, October). Sitting posture recognition using screen printed large area pressure sensors. In *2017 IEEE Sensors* (pp. 1-3). IEEE. https://doi.org/10.1109/ICSENS.2017.8233944

Andriluka, M., Roth, S., & Schiele, B. (2009, June). Pictorial structures revisited: People detection and articulated pose estimation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1014-1021). IEEE. https://doi.org/10.1109/CVPR.2009.5206754

Andriluka, M., Roth, S., & Schiele, B. (2010, June). Monocular 3rd pose estimation and tracking by detection. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 623-630). IEEE. https://doi.org/10.1109/CVPR.2010.5540156

Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014, June). MPII human pose dataset. In *Proc. CVPR* (pp. 3686-3693). https://doi.org/10.1109/CVPR.2014.471

Baca, A., Dabnichki, P., Heller, M., & Kornfeind, P. (2009). Ubiquitous computing in sports: A review and analysis. *Journal of Sports Sciences*, *27*(12), 1335-1346. https://doi.org/10.1080/02640410903277427

Bhatlawande, S., & Girgaonkar, I. (2022, June). Elderly Care System for Classification and Recognition of Sitting Posture. In *2022 2nd International Conference on Intelligent Technologies (CONIT)* (pp. 1-7). IEEE. https://doi.org/10.1109/CONIT55038.2022.9848298

Boulay, B., Bremond, F., & Thonnat, M. (2003). Human posture recognition in video sequence. In *IEEE International Workshop on VS-PETS, Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. https://hal.inria.fr/inria-00494249

Babu, S. C. (2019). A 2019 Guide to Human Pose Estimation with Deep Learning. [Online]. Https://Nanonets.Com/Blog/Humanpose-Estimation-2d-Guide/

Chandna, P., Deswal, S., & Pal, M. (2010). Semi-supervised learning-based prediction of musculoskeletal disorder risk. *Journal of Industrial and Systems Engineering*, *3*(4), 291-295.

Chen, X., & Yuille, A. L. (2014). Articulated pose estimation by a graphical model with image dependent pairwise relations. *Advances in Neural Information Processing Systems, 27*. https://proceedings.neurips.cc/paper/2014/hash/8b6dd7db9af49e67306feb59a8bdc52c-Abstract.html

Estrada, J. E., & Vea, L. A. (2016, May). Real-time human sitting posture detection using mobile devices. In *2016 IEEE Region 10 Symposium (TENSYMP)* (pp. 140-144). IEEE. https://doi.org/10.1109/TENCONSpring.2016.7519393

Estrada, J., & Vea, L. (2017, November). Sitting posture recognition for computer users using smartphones and a web camera. In *TENCON 2017-2017 IEEE Region 10 Conference* (pp. 1520-1525). IEEE. https://doi.org/ 10.1109/TENCON.2017.8228098

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 580-587).

Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1440-1448).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778).

Huang, Y. R., & Ouyang, X. F. (2012, October). Sitting posture detection and recognition using force sensor. In *2012 5th International Conference on Biomedical Engineering and Informatics* (pp. 1117-1121). IEEE. https://doi.org/10.1109/BMEI.2012.6513203

Johnson, S., & Everingham, M. (2010, August). Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC* (Vol. *2*, No. 4, p. 5). https://doi.org/10.5244/C.24.12

Jolly, V., Jain, R., Shah, J., & Dhage, S. (2022, January). Posture Correction and Detection using 3-D Image Classification. In *2022 International Conference for Advancement in Technology (ICONAT)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICONAT53423.2022.9725833

Kamiya, K., Kudo, M., Nonaka, H., & Toyama, J. (2008, December). Sitting posture analysis by pressure sensors. In *2008 19th International Conference on Pattern Recognition* (pp. 1-4). IEEE. https://doi.org/10.1109/ICPR.2008.4761863

Kappattanavar, A. M., Da Cruz, H. F., Arnrich, B., & Böttinger, E. (2020, November). Position Matters: Sensor Placement for Sitting Posture Classification. In *2020 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 1-6). IEEE. https://doi.org/ 10.1109/ICHI48887.2020.9374328

Katayama, H., Mizomoto, T., Rizk, H., & Yamaguchi, H. (2022, March). You work we care: Sitting posture assessment based on point cloud data. In *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)* (pp. 121-123). IEEE., https://doi.org/10.1109/PerComWorkshops53856.2022.9767292

Killough, M. K., Crumpton, L. L., Calvert, A., & Bowden, R. (1995). An investigation of using neural networks to identify the presence of carpal tunnel syndrome. In *4th Industrial Engineering Research Conference Proceedings, IIE, Norcross, GA* (pp. 659-667).

Kreiss, S., Bertoni, L., & Alahi, A. (2019). Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11977-11986).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90. https://dl.acm.org/doi/abs/10.1145/3065386

Laptev, I., & Lindeberg, T. (2005). On space-time interest points. *International Journal of Computer Vision*, *64*(2-3), 107-124. https://doi.org/10.1109/ICCV.2003.1238378

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008, June). Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-8). IEEE. https://doi.org/10.1109/CVPR.2008.4587756

Li, H., Yang, W., Wang, J., Xu, Y., & Huang, L. (2016, September). WiFinger: Talk to your smart devices with finger-grained gesture. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 250-261).

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (pp. 740-755). Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48

Ma, S., Cho, W. H., Quan, C. H., & Lee, S. (2016, October). A sitting posture recognition system based on 3 axis accelerometer. In *2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (pp. 1-3). IEEE. https://doi.org/10.1109/CIBCB.2016.7758131

Moeslund, T. B., & Granum, E. (2000, June). 3D human pose estimation using 2D-data and an alternative phase space representation. In *Workshop on Human Modeling, Analysis and Synthesis at CVPR* (Vol. *16*, p. 1).

Mu, L., Li, K., & Wu, C. (2010, April). A sitting posture surveillance system based on image processing technology. In *2010 2nd international conference on computer engineering and technology* (Vol. *1*, pp. V1-692). IEEE. https://doi.org/ 10.1109/ICCET.2010.5485381

Mwiti. D. (2019). A 2019 Guide to Human Pose Estimation. [Online]. https://heartbeat.fritz.ai/a-2019-guide-to-human-poseestimation-c10b79b64b73

Munir, S., & Nadeem, A. (2018). Intelligent Interactive Systems. *Journal of Information Engineering and Applications*.

Pishchulin, L., Andriluka, M., Gehler, P., & Schiele, B. (2013). Poselet conditioned pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 588-595).

Riihimäki, H. (1995). Hands up or back to work-Mure challenges in epidemiologie research on musculoskeletal diseases. *Scandinavian Journal of Work, Environment & Health*, 401-403. https://www.jstor.org/stable/40966435

Sasikumar, V. (2018). A model for predicting the risk of musculoskeletal disorders among computer professionals. *International Journal of Occupational Safety and Ergonomics*. https://doi.org/10.1080/10803548.2018.1480583

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. https://doi.org/10.48550/arXiv.1409.1556

Tessendorf, B., Arnrich, B., Schumm, J., Setz, C., & Troster, G. (2009, September). Unsupervised monitoring of sitting behavior. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 6197-6200). IEEE. https://doi.org/10.1109/IEMBS.2009.5334620

Toshev, A., & Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1653-1660).

Wang, H., Zhao, J., Li, J., & Wang, K. (2019, November). The sitting posture monitoring method based on notch sensor. In *2019 IEEE International Conference on Industrial Internet (ICII)* (pp. 301-302). IEEE. https://doi.org/ 10.1109/ICII.2019.00058

Yang, Y., & Ramanan, D. (2011, June). Articulated pose estimation with flexible mixtures-of-parts. In *CVPR 2011* (pp. 1385-1392). IEEE. https://doi.org/10. 1109/CVPR.2011.5995741

Yang, Y., & Ramanan, D. (2012). Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2878-2890. https://doi.org/ 10.1109/TPAMI.2012.261

Yao, A., Gall, J., & Van Gool, L. (2012). Coupled action recognition and pose estimation from multiple views. *International Journal of Computer Vision*, *100*, 16-37. https://doi.org/10.1007/s11263-012-0532-9

Yao, L., Sheng, Q., Ruan, W., Gu, T., Li, X., Falkner, N., & Yang, Z. (2015). Rf-care: Device-free posture recognition for elderly people using a passive rfid tag array. 10.4108/eai.22-7-2015.2260064

Zaslavsky, A. (2002, January). Adaptability and interfaces: key to efficient pervasive computing. In *Proc. NSF Workshop on Context-Aware Mobile Database Management* (pp. 24-25). https://cs.brown.edu/nsfmobile/wshop.html/zaslavsky.pdf

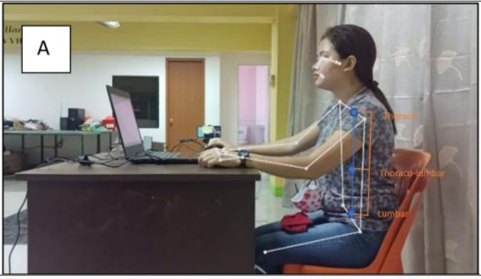
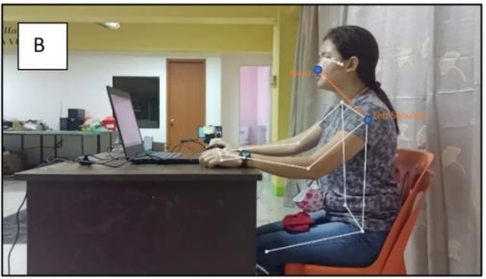
Appendix A: Summary of related literature and studies

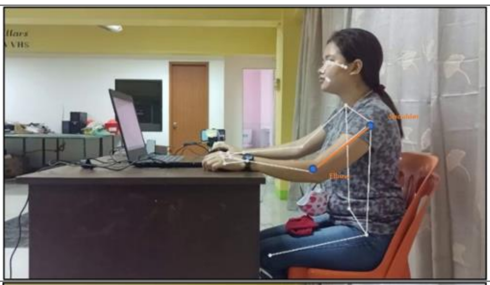
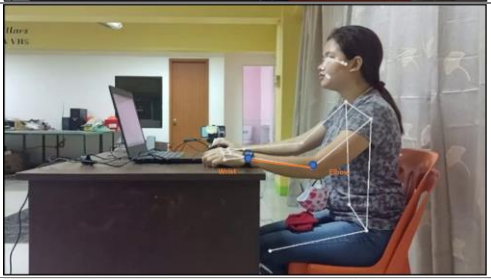
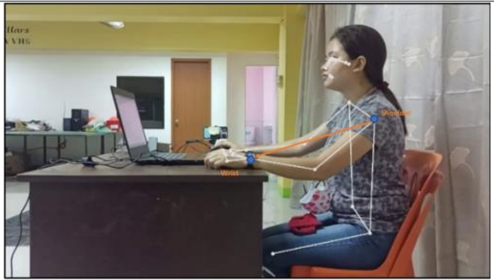
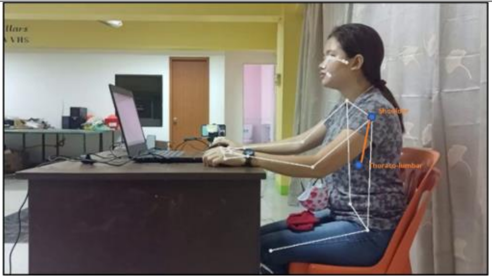
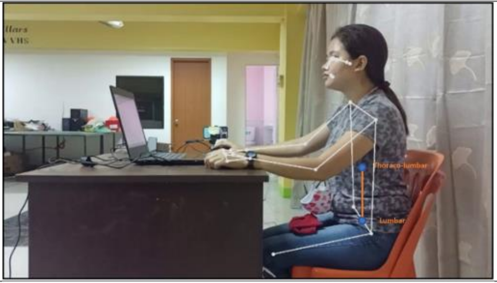
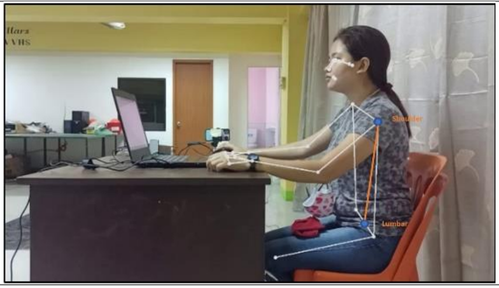
Title/Author/Year	Type of sensors	# of Subjects/samples	Placement of sensors	Results	Gap
Sitting posture analysis by pressure sensors/ 2008	Pressure sensor/direct type of measurement	10 male university students aged 21 to 24 years and weighing 57 kg to 90 kg	Chair	90.6% accuracy	One issue is person dependency and time dependency: Different persons would obviously sit in different ways. Another issue is weight dependency, the need to examine whether difference in weight has a great effect
A sitting posture surveillance system based on imag processing technology/2010	RGB/Vision-based method	<not-mentioned>	the profile features, the face's location and size using Harsdorf distance measurement		Does not cover head tilt and rotation, skin color identification problem
A Sitting posture recognition system based on 3 axis accelerometers/ /2016	3 axis accelerometer/ direct type of measurement	6 subjects	back of the subject’s neck	95.33% SVM 89.35% K-means	Consider other areas of the body that contributes to the recognition of sitting postures
Real-time human sitting posture detection using mobile devices/2016	accelerometer built in the mobile devices/ direct type of measurement	60 subjects	Thoracic, thoraco-lumbar, lumbar	96.13% decision tree	Consideration of other body points is needed
Sitting posture recognition for computer users using smartphones and a web camera/ 2017	RGB/vision-based method	60 subjects	Chin, Manubrium, left and right shoulder	95.35% decision tree	Consideration of CNN and other body points
The sitting posture monitoring method based on notch sensor/2019	Notch sensor/ direct type of measurement		lumbar and thoracic vertebrae of the human body	98-99% Simple Bayesian	Lumbar only can do the job but try to consider more areas in the upper extremity points
Position matters: Sensor placement for sitting	pressure sensors/ IMU/direct	6 subjects	12th thoracic vertebra (T), 3rd lumbar vertebra (L),	92.78% SVM 91.37% LR	Hence, these results are still subject to ulterior validation in the context of a

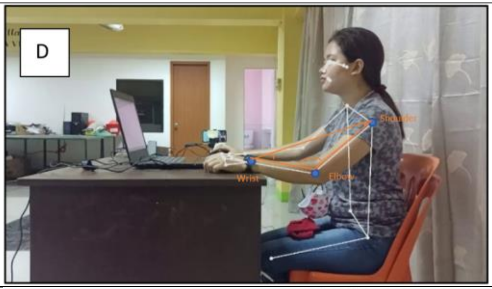
Appendix A: Continue

posture classification//2020	measurement and kinect/vision-based measurement		between 1st and 2nd vertebrae of the sacral region (S) right hip (H) sternal angle (N) Depth images of the upper and lower part of the body	80.45% HM	fully fledged study, based on the classification of real life sitting postures in the occupational settings Finally, we urge researchers in the field to conduct a careful evaluation to assess optimal sensor placement with respect to position and number, e.g., using the techniques described.
Elderly care system for classification and recognition of sitting Posture/2022	RGB/Vision-based method		<not-mentioned>	75% linear SVC 0.72 F1-score 98% polynomial SVM 76% RBF kernel SVM 75% Random forest 69% decision tree	Jetson Nano is good for image processing, but it is costly and requires a good amount of processing and calculation loads
You work we care: Sitting posture assessment based on point cloud Data/2022	Pointcloud data-compact-size LiDAR/Vision-based method		Pointcloud	87% accuracy with reduced processing time	Can only capture limited scenarios
Posture correction and detection using 3-D image classification/2022	RGB to RGBD/ vision-based method		Right eye Left eye Left shoulder Right shoulder Left ear Right ear Left eyebrow Right eyebrow Nose Mouth	98% recognition	High in the computational load
Proposed dissertation topic	When physical sensors were not available because of COVID. RGB/vision-based measurement	60 subjects	13 feature points Neck right Elbow right Shoulder right Upper back right Neck left Elbow left Shoulder left Upper back left Upper mid back Middle back Lower back Chin Nose	81.18% left model 92.07% right model	

Appendix B: List of features

SN	Feature Name	Feature description	Image representation
1	TY and TLY and LY diffyaxis	The difference of three (3) points in the spine – T (Thoracic), TL (Thoraco-Lumbar) and L (Lumbar) in Y Axis.	
2	Nose to left shoulder	The distance between Left Shoulder – TML (*Trapezius Muscle* Left) and N (Nose)	
3	Nose to right shoulder	The distance between Right Shoulder – TML (*Trapezius Muscle* Right) and N (Nose)	

4	Shoulder_elbow_dist_left	The distance between shoulder and elbow	
5	Elbow_wrist_dist_left	The distance between elbow and wrist	
6	Wrist_shoulder_dist_left	The distance between wrist and shoulder	
7	Shoulder_mid_dist_left	The distance between shoulder and mid (thoraco-lumbar)	
8	Mid_hiP_dist_left	The distance between mid (thoraco-lumbar) and left or right hip (depending on the camera model)	
9	Hip_shoulder_dist_left		
10	SEW angle A		

11 **12**	SEW angle B SEW angle C	Cosine Rule for (Brach) Brachioradialis angle (computed from left/right shoulder, elbow and wrist)	
13	Nose to neck	The distance between Nose and Thoracic Distance