Human Pose Estimation

Analysis of in-bed pose estimation

***Abstract*- Humans are able to recognise and label poses simply by visualising the location and position of different predefined human body parts. The goal of the project is to generate 2D and 3D human pose estimations and explore the challenges of in-bed pose estimation. Although human pose estimation applications have been studied broadly, yet in-bed pose estimation has been disregarded. In-bed pose estimation is of crucial interest however, it does present a series of challenges including lighting, and an absence of quality datasets. To investigate the impact of these challenges, the report discusses the way in which data can be used to estimate joints. The idea of employing a CNN model trained on datasets was explored. This paper presents an in-depth overview of human pose estimation with a focus on stationary poses. It highlights the process and challenges involved with in-bed pose estimation.**

I. INTRODUCTION

The idea of human pose estimation has been advocated from the beginnings of computer vision technology. Human pose estimation refers to the detection and positioning of specific points on the human body. Fishler and Elschlager introduced the Pictorial Structures (PS), a collection of parts arranged in a deformable configuration, which became the groundwork for human pose estimation[9]. A large variety of PS-based models have since been developed and utilised to enrich the pose estimation field. A drawback of the PS-based models is that they have resulted in typically simple binary potential and are not depending on image data.

Motion and pose analysis is required in a broad range of diagnostic hospital procedures. While active poses such as walking, ice-skating and running, allow a patient to move around freely, the stationary act of sleep-related diagnostics requires patients to be monitored for an extended period of time.

In-bed pose estimation is a branch of human pose estimation that focuses on individuals typically at rest or in a stationary position. In-bed pose estimation has shown value in fields such as; hospital patient monitoring, sleep studies and smart homes. In some instances [2], it is shown that sleeping posture can affect symptoms of multiple health issues such as impaired circulation, sleep apnea and carpal tunnel syndrome. Hence, stationary pose estimation is of crucial interest in hospital environments.

The aim of the project is to develop 2D and 3D human-joint pose estimations and explore the challenges of in-bed pose estimation.

According to sleep studies[4], the average person spends about 26 years sleeping and 7 years trying to get to sleep. Which equates to 33 years of an individual's life spent in bed. Bed bound patients can spend up to 100% of their time in bed. In-bed pose estimation is a crucial step in many human behavioural monitoring systems which are focused on prevention, prediction, and management[3] of at-rest or forced stationary conditions in both adults and children. Through improving pose estimation, it can provide assistance through computers and robotics by enabling clinicians to prevent bed sores and further injuries, detect pathological patterns and assist in further research studies on neural behaviours.

Despite many years of research, pose estimation remains a largely unsolved problem. Among the most significant challenges are: (1) lack of variability in human visual appearance, (2) fluctuation in lighting conditions, (3) bias data; lack of human physique variability, skin colour variability, no notable age variability, (4) occlusions due to medical equipment and layer of objects such as blankets, (5) complexity of the human pose, (6) ethical challenges. To this date, there is no approach to produce highly accurate results. The data quality challenges and privacy concerns have hindered the use of AI-based in-bed behaviour monitoring systems.

II. DATASET

There is a small variety of benchmarks for human pose estimation. In this work we used the limited available

datasets. The first dataset that was used was the SLP dataset which was publicly available online [6]. The dataset consists of 13,770 images, all captured from a bird's eye view. The dataset consists of 109 participants. For each participant there are 15 different poses. The images have 3 different conditions from uncover to cover1 and cover 2.

The second dataset used was the Manne2 dataset [8]which was created by the same owners of the SLP dataset. Unlike the SLP images, the Manne2 images were modelled by mannequins. The Manne2 dataset consists of 1,456 images

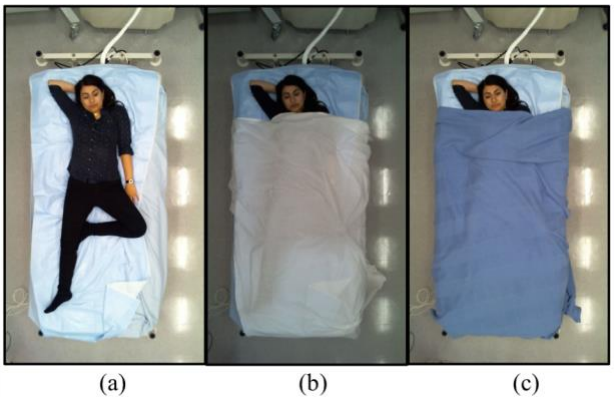

Figure1:Sample from SLP dataset

III. METHODOLOGY

The research employed the following procedures:

- Convolutional neural networks.
- Tensorflow estimation.

Tensorflow

Tensorflow is an end-to-end open source machine learning platform that is used to develop pose estimation.

Convolutional Neural Networks

CNN is a type of neural network model designed for working with two-dimensional image data. It consists of an input layer (convolutional layer), hidden layers (pooling layers) and an output layer (fully connected layer). The layers each generate several function outputs that are passed onto the next layer. Each layer extracts features then at the final layer, outputs a set of confidence values between 0 and 1.

Convolutional Neural Network (CNN) have shown to be effective in human pose estimation hence, the model used in this report was a CNN.

IV. EXPERIMENT SETUP

To produce the joint estimation, 18 body[5] parts were defined and connected. The images were then run through the code to output the estimated joint locations. The following images display two of the SLP dataset joint estimation outputs.

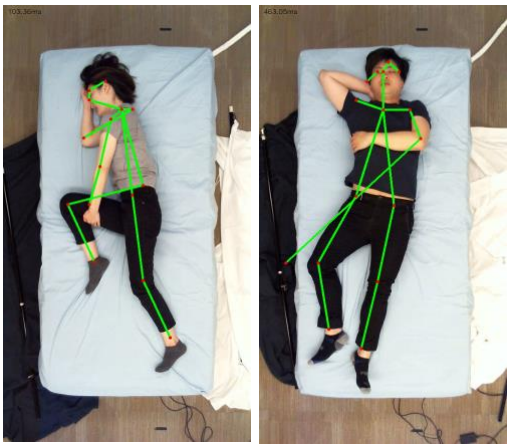

Figure 2: SLP joint estimation

The following two images are some of the outputs of the Manne2 dataset.

Figure 3: Manne2 joint estimation

In Figure 2, the joint estimation is usually successful as estimating a roughly correct pose, however, the pose is not fully snapped into place. In the first image, the predicted hip joint is higher than the actual joint. This is likely due to the inability to locate both hip bones and hence, as the code states that the leg is attached to the hip bone, the computer then connects both legs to the one hip bone. The second image however, produces some inaccuracies. There is clear confusion over joint location in the face and arm. The possible reason being that the participants' arm is draped over his chest which could be obstructing the estimation.

By looking at Figure 3, it seems that the joint estimation had difficulty locating the correct landmarks. The first image has a joined joint located between the knee and ankle with no other detected joints. This is likely due to the quality of the image. The Manne2 images are low quality images with a lack of clarity. The grainy images alongside the fact that the bodies are mannequins, is likely the reason as to why the computer is having difficulty detecting the joint locations. Unlike the first image, the second image provides more correct joint locations. However, the predicted joints on the leg appear to be overlapping. At a closer look you can see that there are two estimated joints overlapping each other on the legs meaning the model has failed to successfully recognise the other lower body joint locations. Due to the lack of recognition success, the Manne2 dataset will not be tested on the CNN model. Instead we will analyse the accuracy of the SLP data.

To build the CNN, the SLP data was split manually into separate files. The used data was split 80:20 with 80% being used for training and 20% being used for testing and validation.

The layout for our Neural Network was sourced from Rashida[7]. The following is the summary of the model's layers.

Layer	**Output Shape**	**Param #**
conv2d_2	(None, 1022, 574, 16)	160
batch_normalization	(None, 511, 287, 16)	64
max_pooling2d	(None, 511, 287, 16)	0
conv2d_3	(None, 509, 285, 32)	4640
batch_normalization_1	None, 254, 142,32)	128
max_pooling2d_1	(None, 254, 142, 32)	0
conv2d_4	(None, 252, 140, 64)	18496
batch_normalization_2	(None, 126, 70, 64)	256
max_pooling2d_2	(None, 126, 70, 64)	0
conv2d_5	(None, 124, 68 ,64)	36928
max_pooling_3	(None, 62 ,34, 64)	0
batch_normalization_3	(None, 62, 34, 64)	256
flatten	(None, 134912)	0
dense	(None, 32)	4317216
dense_1	(None, 3)	99

Table 1: Summary of layers

As shown in Table 1, the model consists of 15 layers. It implements the use of layers for techniques to assist in training the CNN model.

- Batch Normalisation: a standardisation technique used to reduce the number of epochs required for training.
- Max Pooling: a technique used to reduce the dimensions of the image.
- Flatten: converts the multi-dimensional input tensors into a single dimension so it may run through the CNN.
- Dense: feeds all outputs from previous layers into its neurons, then provides one output to the next layer.

The epochs were first set to 20 so the results could be observed. The optimiser we used was 'adam' while the

loss was set to 'categorical crossentropy' and the learning rate was set to 0.0001.To improve the overall accuracy, it was crucial to analyse where the model would begin to overfit and select an optimal number of epochs. The baseline run produced the following accuracy scores:

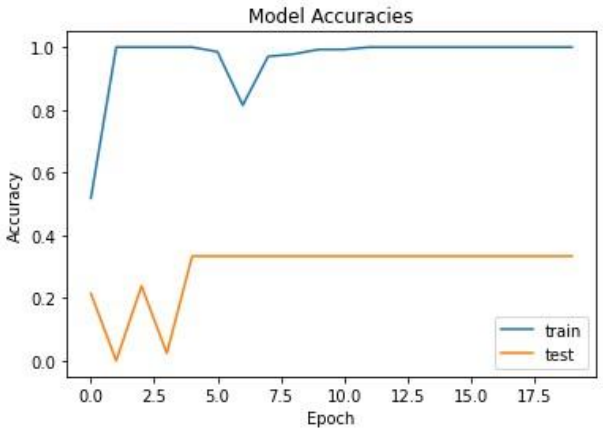

Figure 4: SLP accuracy of model with 20 epochs

By observing Figure 2, it appears that around 5 epochs the model stops fluctuating in accuracy and consistently stays around 35%. The learning rate was increased to 0.0001 and the epochs were set to 8 in order to observe any potential improvements in the accuracy. Table 2 shows the accuracy results for the epochs

Epochs	**Training Accuracy**	**Testing Accuracy**
1	0.8000	0.3333
2	0.9481	0.6667
3	0.9704	0.3333
4	1.0000	0.3333
5	1.0000	0.3333
6	1.0000	0.3333
7	1.0000	0.3333
8	1.0000	0.3333

Table 2: Accuracy table of epochs

By looking at the accuracy values, we can see that the training accuracy stays around 90% and then increases to 100% at 4 epochs. Looking at the Testing accuracy, we can see that it stays at 33% then goes up slightly at the 2nd epoch to 66% however, it immediately goes back to 33% and stays consistent.

The following figure displays the train and test accuracies.

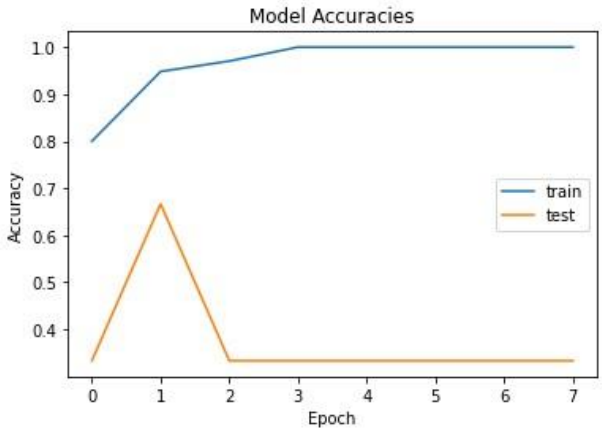

Figure 5: SLP accuracy of model with 8 epochs

V. RESULTS

Three-Dimensional (3D) human pose estimation involves estimating the articulated 3D joint locations of a human body. 3D pose estimation can be used for a variety of different human-computer interactions including pose estimation. The performance of 3D pose estimation has remained 'barely satisfactory'[1], which is likely due to the lack of depth information. Furthemore, current 2D pose estimates are inaccurate which would, as a result, produce errors in 3D models. 3D human pose estimation enables widespread applications such as autonomous driving, video surveillance, sports performance and computer interactions.

The following images display the 3D joint prediction outputs for the SLP images.

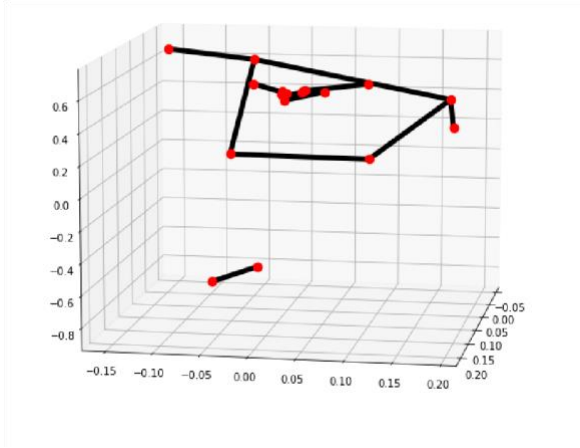

Figure 6: SLP 3D joint location 1

Figure 6 displays the 3D estimation joint locations for the Figure 2 first image. The 3D model does include the detected joints however, there is no shape or pattern to the model. It does not output an accurate bodily shape as expected.

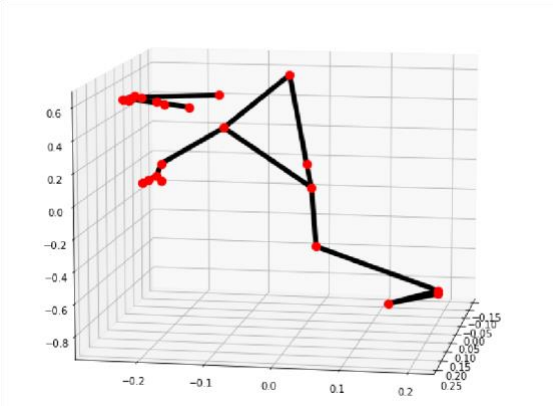

Figure 7: SLP 3D joint location 2

Figure 7 shows the 3D estimation for the Figure 2

second image. This 3D model is visually more accurate than the previous. There is a similar pattern to the estimation image. There appears to be a head, shoulder and leg region.

The reasoning for better accuracy on the second image may be due to the pose position in the image. Based on the results, a 3D joint estimation model works more effectively when the individual is lying face up.

The 3D joint prediction outputs for the Manne2 dataset are as follows.

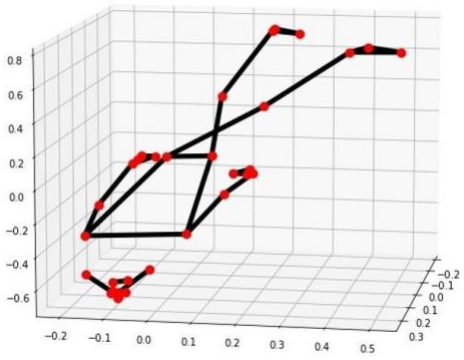

Figure 8: Manne2 3D joint location 1

Figure 8 is the 3D output for the first image in Figure 3. In comparison to the SLP 3D models, the Manne2 dataset produces outputs with a human figure compared to the lack of shape produced in the SLP images. The result is an accurate depiction of the detected joints.

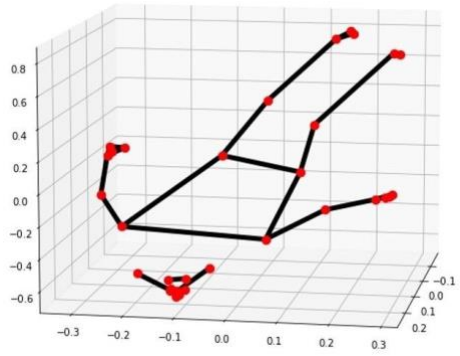

Figure 9: Manne2 3D joint location 1

Figure 9 displays the output for the second image in Figure 3. Figure 9 also outputs a good figure However, unlike Figure 8, the joints displayed on the graph are not visible on the 2D model. Hence, we can ask where the additional joints came from.

While 3D modelling is possible, there is still a long way to go in increasing its accuracy. The 3D outputs display a significant lack of accuracy in the original 2D models. If the input images are produced with more clarity and heatmap are implemented into the 3D model process, there could be an improvement in the output.

VI. CHALLENGES/DECISIONS

While humans recognise poses virtually without effort, successful human pose estimation technology is an ongoing challenge. The obstacles plaguing human pose estimation are; complex code, image clarity, lighting difficulties, occlusions, and bias data.

To improve the performance of in-bed pose estimation, most images require improved imaging conditions. A more controlled image environment would cover the following.

1) Occlusions

Occlusions in the image environment can have a negative effect on human joint recognition. Foreground and background clutter can make determining the joint more difficult. In a hospital setting, objects such as blankets, bandages, medical equipment and cords/wires can increase the difficulty to perform reliably.

2) Bias Data/Lack of Data

There is a lack of large datasets available for in-bed pose estimation. As a result of having limited data, the available images are heavily biassed. There needs to be an improvement in the lack of diversity among the images. Improving the lack of diversity could include; variety of skin colour, body shapes/sizes and age. As well as including characteristics such as bodily deformities. Currently, the datasets are comprised of middle-aged individuals with a similar body shape/size, all with (on average) medium toned skin. Lighting

The two issues with lighting is as follows:

1. Natural lighting in real-life hospital situations could fluctuate in the evenings or depending on the state of the individual.
2. The lighting conditions in the dataset images are uneven. A few individuals have overexposed lighting conditions while others vary from even to underexposed conditions. While this isn't a major concern, an improvement could be to provide all test subjects with various equal lighting conditions.

3) Image Clarity

Improving image clarity by using high quality cameras and manipulating the image to increase the contrast between the body and its surrounds.

An obstacle in producing accurate results was the lack of and difficulty of code. While there is an abundance of code provided for human pose estimation, there is a minimum supply of code for in-bed pose estimation. The code associated with the SLP dataset is one of the few tailored programs for in-bed pose estimation. Based on the report provided by the dataset[6], applying the dataset to the code produces good results. However, a major difficulty with this process was the inability to run the code due to out-dated software. Software such as OpenCV 3.1 is no longer available online hence, if you do not already have it installed then you will not be able to run this code. Therefore, the dataset provided cannot be fully optimised unless a new code is developed to cater to it.

A potential solution to the lack of available datasets on in-bed pose estimation could be to generate synthetic data using ragdoll physics. Ragdoll physics can be used to model people at rest and find statistically stable poses. It would involve generating a digital human skeletal body and dropping them onto a hospital bed in order to produce realistic positions at rest.

The human pose estimation algorithm usually requires large backbone models and high feature resolution which in turn requires fast processing GPU systems. This prohibits the majority of devices such as mobile phones and all computers with insufficient GPU's.

Furthemore, while there are some works on 3D human pose estimation, there is little to none focused on stationary poses . The vast majority of work on 3D pose estimation is captured in constrained environments with limited, bias data. The pose estimation field as a whole could benefit from unconstrained 3D human pose estimation with large, diverse datasets.

Problems outside of the computational side of in-bed pose estimation are issues surrounding ethics. Ethical questions such as what to do in situations where an individual is unable to consent to being monitored. Furthermore, if patients are uncomfortable or do not feel safe with the technology then will hospitals spend the money required to upkeep the machines?

The main aim of this project was to develop 2D and 3D human-joint pose estimations and explore the challenges of in-bed pose estimation. A convolutional Neural Network (CNN) was employed for this purpose. Employing the model yielded accuracies between 33% to 66%. Further improvements could be made with improved data and programming..

Besides the challenges discussed above, there are also many unexplored complications for the topic of pose estimation. The development of not only human pose estimation but In-bed pose estimation, would contribute heavily to the computer vision industry and the medical field in the coming years. This project presented a shallow exploration of in-bed pose estimation and the challenges that accompany applying 2D and 3D joint detection code.

VII. REFERENCES

[1] J. Wang, S. Tan, X. Zhen, S. Xu, F. Zheng, Z. He, L. Shao, *Deep 3D human pose estimation: A review*, Volume 210, 2021. https://www.sciencedirect.com/science/article/pii/S1077314221000692

[2] McCabe, S. J., Gupta, A., Tate, D. E., & Myers, J. (2011). Preferred sleep position on the side is associated with carpal tunnel syndrome. *Hand (New York, N.Y.)*, *6*(2), 132–137. https://doi.org/10.1007/s11552-010-9308-2

[3] Competitions.codalab.org. 2021. *IEEE VIP Cup 2021: SLP Human Pose Estimation*. [online] Available at: <https://competitions.codalab.org/competitions/31489>.

[4] Curtis, G., 2017. *Your Life In Numbers - The Sleep Matters Club*. [online] The Sleep Matters Club. Available at: <https://www.dreams.co.uk/sleep-matters-club/your-life -in-numbers-infographic>.

[5] GitHub. 2018. *GitHub quanhua92/human-pose-estimation-opencv: Perform Human Pose Estimation in OpenCV Using OpenPose MobileNet*. [online] Available at: <https://github.com/quanhua92/human-pose-estimation -opencv>.

[6] Liu, S., Yin, Y. and Ostadabbas, S., n.d. *SLP Dataset for Multimodal In-Bed Pose Estimation – Augmented Cognition Lab*. [online] Web.northeastern.edu. Available at: <https://web.northeastern.edu/ostadabbas/2019/06/27/multimodal-in-bed-pose-estimation/>.

[7] N.Sucky, R., 2021. *Developing a Convolutional Neural Network Model Using the Unlabeled Image Files Directly From the…*. [online] Medium. Available at: <https://towardsdatascience.com/developing-a-convolutional-neural-network-model-using-the-unlabeled-image-files-directly-from-the-124180b8f21f>.

[8] Ostaddabas, S. and Liu, S., 2017. *GitHub ostadabbas/In-Bed-Posture-Estimation: A Vision-Based System for In-Bed Posture Tracking (ICCVW2017)*. [online] GitHub. Available at: <https://github.com/ostadabbas/In-Bed-Posture-Estimat ion>.

[9] Toshev, A. and Szegedy, C., n.d. *DeepPose: Human Pose Estimation via Deep Neural Networks*. [online] Static.googleusercontent.com. Available at: <https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/42237.pdf>.