Sheffield Hallam University | Faculty of Health and Wellbeing

Ashish John Stanley
29029604
M.Sc. Sports Engineering
Dr. Marcus Dunn
Word Count:

Extending 2D pose estimation algorithms to study lower leg kinematics

ABSTRACT

Introduction

Capturing motion is an important tool for sports research to analyse physical condition, athletic performance, technical expertise and injury mechanism for both prevention, and rehabilitation (Pueo & Jimenez-Olmedo, 2017). From a biomechanical perspective, for running, motion capture can be used to study lower leg kinematics by performing gait analysis. The footwear industry uses such understanding of kinematics coupled with kinetics to gain biomechanical insights for their product development. Previous research has shown how footwear biomechanics such as fore foot bending stiffness (Roy & Stefanyshyn, 2006) and underfoot cushioning (Stefanyshyn & Fusco, 2004) can improve the performance of the athlete. Being able to track motion and gather information on kinematics of foot joints could provide enriching feedback to a product development lifecycle. It also helps to understand athletes and provide apt footwear solutions designed to improve performance.

Importance of gait analysis for studying lower extremity kinematics:

Gait analysis helps study the kinetics (forces causing motion) and kinematics (independent motion of joints or body segments) of motion. Kinematics of motion can be studied by tracking joints of interest using motion capture. Therefore, making it possible to track range of motion provided by joints in the foot during different phases of the gait cycle across various planes by motion capture. Plantarflexion (PF) and dorsiflexion (DF) are the kinematic parameters that describe range of motion of the ankle and metatarsophalangeal (MTP) joint in the sagittal plane. The MTP joint shows movement in the sagittal plane as PF and DF (Mojica & Early, 2019).

Research shows how designs of shoe have been moulded around studies regarding PF and DF during the gait cycle (Bishop, Fiolkowski, Conrad, Brunt, & Horodyski, 2006), (Bourgit, Millet, & Fuchslocher, 2008), (Ng, et al., 2014) making it important kinematic parameters for footwear development. The ankle and MTP joints show varying degrees of PF and DF during various phases of the gait cycle making it important joints to identify and track.

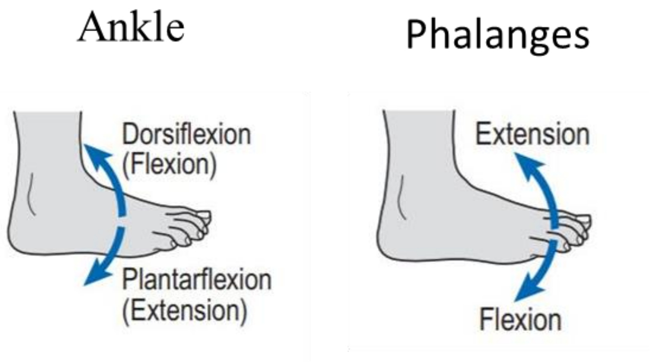


Figure 1: Range of motion parameters in the ankle and phalanges (Whittle, 2014).

Gait Cycle Phase	**Angle of ankle joint in the sagittal plane**
Foot strike to midstance	5 degrees of plantarflexion to 10 degrees dorsiflexion.
Midstance to take off	10 to 20 degrees of dorsiflexion.
Follow through	20 to 30 degrees of plantarflexion.
Forward swing	30 to 0 degrees of plantarflexion.
Foot descent	0 to 5 degrees of dorsiflexion to 5 degrees of plantarflexion.

Table 1:Normal Range-of-Motion values for ankle joint during running.

Motion Capture systems for biomechanics:

Different methods of motion capture to study kinematics can be categorised based on the technology used namely Optoelectronic measurement system (OMS), Electromagnetic Measurement System (EMS), Image Processing Systems (IPS), Ultrasonic localisation systems (ULS), Inertial Sensor Measurement systems (IMS) (van der Kruk & Reijne, 2018). Based on the goal and environment of the study, a motion capture technology that performs to requirement is chosen. OMS based systems are considered the gold standard for motion capture of lower leg kinematics (Fong & Chan, 2010), but they have their drawbacks of being expensive, sensitive to sunlight and disturbances in the set up, and are restricted to a small capture volumes (Richards, 1999).

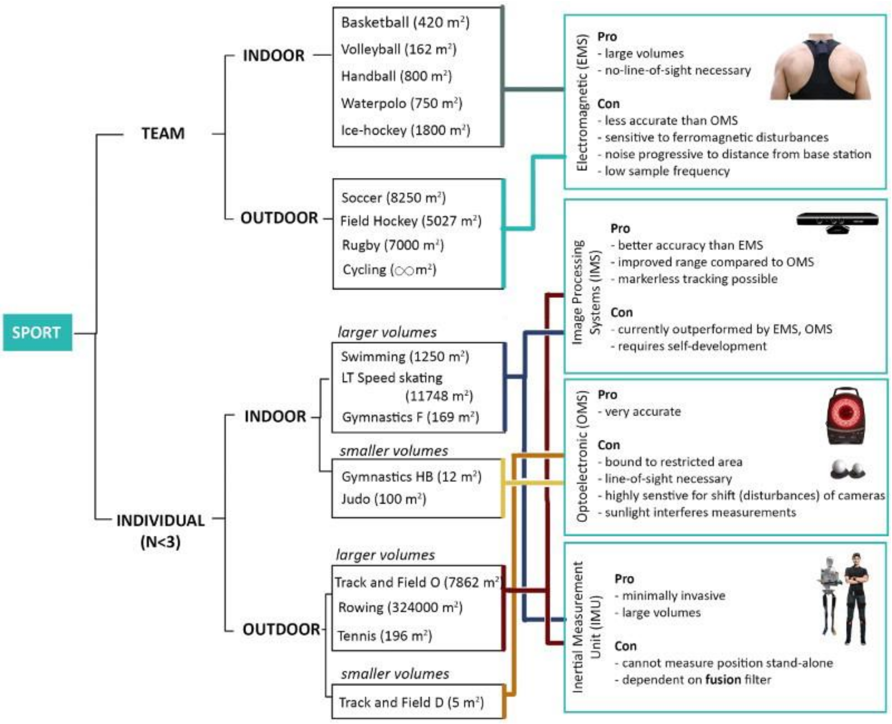

SPORT
TEAM
INDOOR
Basketball (420 m²)
Volleyball (162 m²)
Handball (800 m²)
Waterpolo (750 m²)
Ice-hockey (1800 m²)
OUTDOOR
Soccer (8250 m²)
Field Hockey (5027 m²)
Rugby (7000 m²)
Cycling (∞ m²)
INDIVIDUAL (N<3)
INDOOR
larger volumes
Swimming (1250 m²)
LT Speed skating (11748 m²)
Gymnastics F (169 m²)
smaller volumes
Gymnastics HB (12 m²)
Judo (100 m²)
OUTDOOR
larger volumes
Track and Field O (7862 m²)
Rowing (324000 m²)
Tennis (196 m²)
smaller volumes
Track and Field D (5 m²)
Electromagnetic (EMS)
Pro
- large volumes
- no-line-of-sight necessary
Con
- less accurate than OMS
- sensitive to ferromagnetic disturbances
- noise progressive to distance from base station
- low sample frequency
Image Processing Systems (IMS)
Pro
- better accuracy than EMS
- improved range compared to OMS
- markerless tracking possible
Con
- currently outperformed by EMS, OMS
- requires self-development
Optoelectronic (OMS)
Pro
- very accurate
Con
- bound to restricted area
- line-of-sight necessary
- highly senstive for shift (disturbances) of cameras
- sunlight interferes measurements
Inertial Measurement Unit (IMU)
Pro
- minimally invasive
- large volumes
Con
- cannot measure position stand-alone
- dependent on fusion filter

Figure 2: A schematic diagram giving guidance for selection of a motion capture technology based on type of sport , environment , volume of capture (van der Kruk & Reijne, 2018).

Image Processing based algorithms allow for larger range of usage in comparison to EMS systems and only depend on passive optical sensors that operate in the visible spectrum eliminating the need of attaching sensors on the athlete. They either use model based or non-model-based tracking. Non model based tracking employs simple image processing techniques that either identify a visual marker placed on the human or use differences in consecutive frames based on changes in parameters such as shape, texture, velocity (Akio & Jack, October, 1991) , colour to capture motion. Model based tracking uses a priori model which can either be a volumetric (Akita, 1984) (Rohr, 1994) or stick model (Karaulova, Hall, & Marshall, 2000) to identify the human in motion from the sequence of frames. Model based IPS are preferred over non-model based since they are more robust and can overcome difficulties in motion capture due to lighting conditions, occlusions, clothing, image quality and camera calibration.

2D Pose Estimation

A new type of learning algorithm called 2D pose estimation (keypoint detection) that can detect keypoints from images has been used to identify human keypoints or anatomical landmarks for biomechanical purposes. They work on individual frames, do not require any form of sensor/marker on the athlete or calibration for the passive optical sensor and do not need a priori models like in model based IPS.

Identifying key points (joints of interest) from consecutive frames of a video sequence could be used for gait analysis. Marker based motion capture when used as ground truth and compared with a stacked hourglass architecture of pose estimation (Newell, Yang, & Deng, 2016) reported an error of 14.72±2.96 mm (Mehrizi, et al., 2018) and when compared with Open Pose (Cao, Hidalgo, Simon, Wei, & Sheikh, 2018) reported a mean absolute error of less than 30mm (Nobuyasu, et al.). Pose estimation algorithms have proven to provide accurate results in terms of identify key points in the human body. Newell et al. work on pose estimation reported an accuracy of identifying 99% and 97% accuracy in identifying elbow and wrist joints respectively in the “Frames Labelled in Cinema” dataset (Newell, Yang, & Deng, 2016).

Usage of 2D pose estimation systems will allow for unobtrusive detection of target joints for kinematic studies in normal training and competitive environments. There will be a significant reduction in time in identifying joint location using 2D pose estimation algorithms in comparison with manual annotation process. To track motion in the sagittal plane for activities such as running and sprinting 2D coordinates suffice (Bezodis, Salo, & Trewartha, 2015).

New Balance is a world leading sports footwear and apparel manufacturer with their vision to aid athletes to achieve excellence in the performance. In their process of understanding how they can track and gain insight on performance of athletes using their products they are looking at research in markerless motion capture systems to aid studies in range of motion of different joints of the foot. This project is completed to report the feasibility of using 2D pose estimation algorithms for studying range of motion for MTP joint. As stated earlier, MTP joint shows PF and DF motion in sagittal plane (Mojica & Early, 2019) therefore , requiring only 2D coordinates to calculate kinematics range of motion angles (Bezodis, Salo, & Trewartha, 2015).

Relevant theory on pose estimation:

An artificial neural network based on supervised learning is used to implement the pose estimation problem which identifies the 2D keypoints from the RGB input image. Supervised learning is used to train the model, based on a given set of inputs and their known outputs. The model trained will be able to predict the output for new inputs.

Convolution Neural Network (CNN):

The pose estimation implementation uses convolutional neural networks (CNN) for feature extraction. CNNs learn features in the form of feature maps from an image in computer vision tasks. Two primary two layers that help CNNs function are:

Convolution Layers:

Convolutional layers that takes an input and a kernel (filter) and uses cross-correlation to give an output in the form of feature map. Given an input image *I(i,j)* and a two dimensional kernel *K(m,n)* the feature map *S* is calculated by equation 1. Convolutional layers enable CNNs to learn lower level features and translate these low-level features to richer higher-level feature deeper in the network.

$$S(i,j) = (K * I)(i,j) = \sum_{m}\sum_{n} I(i+m\,,j+n)K(m,n) \qquad (Equation\ 1)$$

While training the artificial neural network that uses CNNs for feature extraction, each kernel detects specific features and the values of the kernel are learnt in the process. CNN kernels can identify rich features and identify these features invariant of their spatial orientation in the image, making it an ideal choice for the pose estimation task.

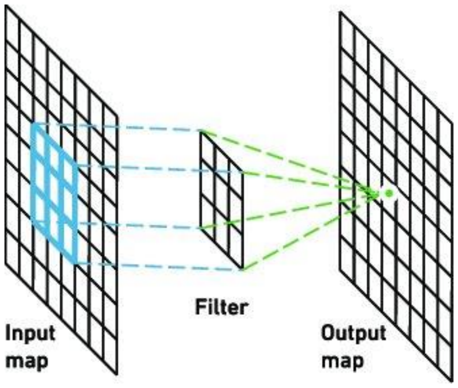


Figure 3: Image showing convolution operation of the Filter (kernel) on a section of the input with its corresponding output.

Pooling Layers:

Pooling layers are used in CNNs to reduce the dimensions of the input to a convolution layer. Reducing the dimension helps reduce the number of parameters required to be learnt thereby, reducing the computational complexity.

Feature Extraction Architectures:

ResNet18:

Resnet18 is a convolution neural network that is 18 layers deep and is a scaled down version deeper residual networks making it lightweight. It was introduced by Microsoft research (He, Zhang, Ren, & Sun, 2016) and focused on overcoming the stagnating accuracy of deep neural networks by introducing residual learning. Residual learning works on the concept of skip connections or shortcuts that jump layers in the networks. Addition of double or triple layer skips to a CNN showed improvement in accuracy of the model trained (He, Zhang, Ren, & Sun, 2016).

MobileNetv2:

MobileNetV2 is a convolution neural network introduced by Google (Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018) which overcomes stagnating accuracy and has improved memory efficiency by introducing inverted residual blocks. General residual networks reduce the input channels initially and then expand it. Inverted residual networks do the opposite by expanding the input channels initially and then reducing it.

Optimisation Algorithm:

During the training process the values of the different kernels known as the weights are being learnt to identify the features. These values are changed or updated in each iteration based on the loss or error in the model's predictions. An optimizer algorithm is used to decide how the values are updated.

Adam is an optimization algorithm known as Adaptive Movement Estimation. It works by calculating the moving averages ($m_t\,, v_t$) of the gradient (g_t) and gradient square (g_t^2) of the loss space. In equation 2 and 3 β_1 and β_2 are hyperparameters of the algorithm.

$$m_t \ = \ \beta_1 m_{t-1} + (1 - \ \beta_1)g_t \qquad (Equation\ 2)$$

$$v_t \ = \ \beta_2 v_{t-1} + (1 - \ \beta_2)g_t^2 \qquad (Equation\ 3)$$

The moving average values known as moments are corrected from bias before using it to updating the weights by equation 4 and 5.

$$m_t^{\sim} = \frac{m_t}{1-\beta_1^t} \qquad (Equation\ 4)$$

$$v_t^{\sim} = \frac{v_t}{1-\beta_2^t} \qquad (Equation\ 5)$$

The final update step for the weights (parameters) of the model are done by equation 6 where λ is the learning rate and ε is a small number used to avoid a divide by 0 arithmetic error.

$$w_t \ = \ w_{t-1} - \ \lambda\left(\frac{m_t}{(\sqrt{v_t} - \varepsilon)}\right) \qquad (Equation\ 6)$$

Methodology

In this project, the aim was to report the feasibility of identifying the joint location of the ankle, MTP and distal phalanges (toe) using 2D Pose estimation algorithms for studying range of motion of the MTP in the sagittal plane. A lightweight monocular MobilePose 2D single person pose estimation implementation (Xiu, Chen, & Fang, 2018) in PyTorch is used to predict the 2D joint locations of key points. The model is trained to predict three key points on each foot namely, ankle, MTP and toe.

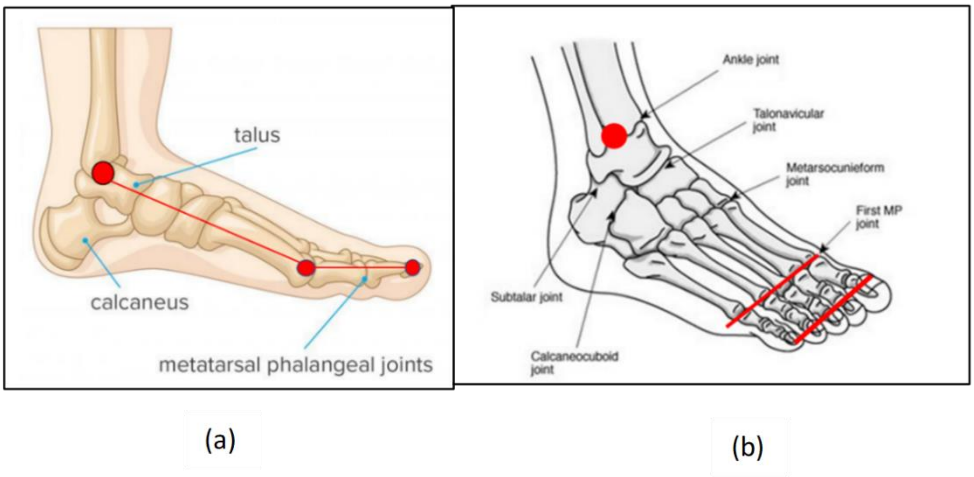


Figure 4: Images showing the joints of interests. (a) Side profile view showing the ankle, metatarsophalangeal joint and the distal phalanges(toe) highlighted. (b) Isotropic view of the joints mentioned above.

Environment Setup:

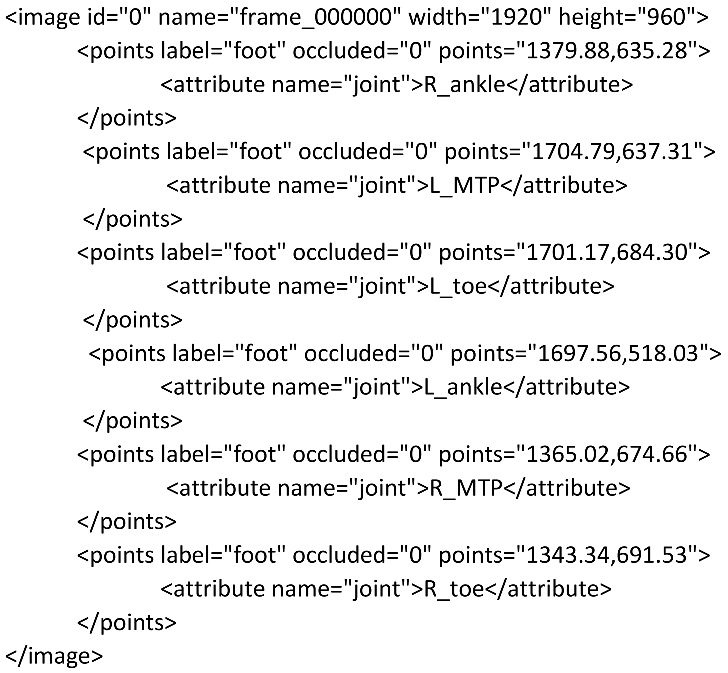
The pose estimation algorithm is implemented using Python 3.7 using PyTorch 1.0 deep learning framework. The training process is conducted on a Google Collab cloud setup using the freemium GPU (12GB NVIDIA Tesla K80) with CUDA support. Model check pointing is used to train the model to overcome time limits for the training process on Google Collab.

Pose Estimation Implementation using ResNet18 and MobileNetV2:

The pose estimation implementation generates heatmaps for each key point to be detected. The heatmaps indicates the probability of finding a key point in every pixel of the image. The heatmap is regressed in every iteration of the training process as a part of the learning process. Two convolutional neural network-based backbones are used for feature extraction namely, ResNet18 and MobileNetV2.

Data acquisition and annotation:

The dataset consists of image frames from videos collected from the internet and corresponding ground truth coordinates of the six anatomical landmarks (key points) that are manually annotated. A total of 900 frames extracted from 5 videos are manually annotated using Computer Vision Annotation Tool. The format of the coordinates of the key points are downloaded from CVAT in the XML file format in the following structure:



```
<image id="0" name="frame_000000" width="1920" height="960">
        <points label="foot" occluded="0" points="1379.88,635.28">
                <attribute name="joint">R_ankle</attribute>
        </points>
         <points label="foot" occluded="0" points="1704.79,637.31">
                 <attribute name="joint">L_MTP</attribute>
         </points>
        <points label="foot" occluded="0" points="1701.17,684.30">
                 <attribute name="joint">L_toe</attribute>
         </points>
          <points label="foot" occluded="0" points="1697.56,518.03">
                <attribute name="joint">L_ankle</attribute>
         </points>
        <points label="foot" occluded="0" points="1365.02,674.66">
                 <attribute name="joint">R_MTP</attribute>
        </points>
        <points label="foot" occluded="0" points="1343.34,691.53">
                <attribute name="joint">R_toe</attribute>
        </points>
</image>
```

Figure 5: XML format of dataset from CVAT for the 6 keypoints annotated.

Data Pre-Processing:

The XML data and image data are pre-processed before running the training process. To reduce computational complexity the images are cropped to contain only the region of interest using a bounding box and resized to 224 x 224 pixels. The XML data is filtered to exclude frames with occluded key points and converted to CSV format. Each videos CSV key point data is passed through a smooth by moving average function to reduce noise due to manual annotation. The dataset was split into a training, test and validation dataset.

Training the network:

Loss function:

The loss for each key points prediction is calculated using Euclidean distance. A regularization loss is calculated using L1 regularisation loss function. The average of the regularisation and Euclidean loss gives the overall loss. The addition of regularisation loss helps steer the training of the model towards generalisation.

Optimisation Algorithm:

The model is trained using Adam optimisation algorithm (Kingma & Ba, 2014) that has an adaptive learning rate for each network weight unlike stochastic gradient descent algorithm that maintains the same learning rate (alpha). The learning rate used for the optimiser is 1e-3. The beta values used by the optimizer to control decay rate of exponential moving average of the gradient and squared gradient are set to 0.9 and 0.999, respectively.

Batch wise training and monitoring training:

The training process uses a batch size of 32 while training to update the network's weight. For every two epochs the training testing loss is calculated and logged. The training process is monitored using a live plot of the training loss and validation loss.

Evaluation of the results:

Based on the training process a model with minimal loss before the training and testing loss converges is chosen for evaluation. The model weights are extracted from the checkpoint file and are used to generate prediction of 2D coordinates of the keypoints. The generated predictions are compared with the ground truth coordinates using the same Euclidean distance loss function used during training.

The accuracy of the model is predicted based on the following parameters:

NOTE: All results are calculated based on model predictions on the in-sample testing dataset.

Percentage of correct keypoints (PCK):

PCK is the percentage of keypoints from the testing dataset that are identified correctly. A keypoint is said to identifies correctly if it the Euclidean distance between the ground truth and the predicted point lies within a threshold. A threshold of 5 pixels is selected to report the PCK.

Area under the curve (AUC):

The area under the curve plot will provide a broader view on the performance of the model to different thresholds. The error of predicted 2D coordinates of the model are calculated for various thresholds. The PCK for various thresholds is plotted against the various thresholds for the AUC plot. A range of thresholds from 0-31 pixels is chosen for the AUC plot.

Statistical analysis on range of motion:

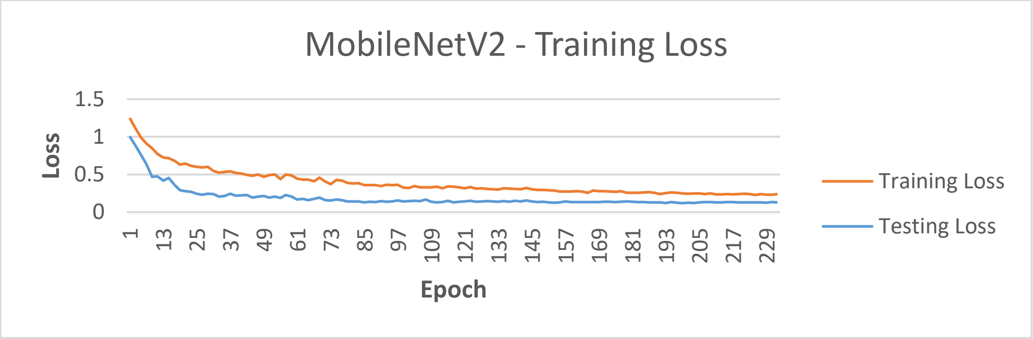
The range of motion of the MTP joint is calculated during one gait cycle for a single leg (right) using the joints predicted by the model and the manually annotated joints. The gait cycle of the leg is taken from an in-sample video (from the dataset). Based on the 2D coordinate predictions of the model the segments of the foot are extracted. The angle formed between the two segments are calculated using geometric angle calculation given the slope of the segments. Kinnovea software is used to manual annotate the joints to find the angle made by the segments.

Statistical analysis is done on the results from the two methods of calculating the angle formed by the segments. Root mean square error, inter class correlation and a Bland Altman plot (95% confidence limit) are used to analyse the results obtained from the two methods.

Results

The training loss and testing loss curve in figure 6 (a) and (b) initially show underfitting. A checkpoint model for both the architectures is chosen at a point of stability before the two loss plots start to converge. For the MobileNetV2 architecture the loss values for the epochs between 95 and 105 remain constant (loss ≈ 0.14) to the second decimal value. Similarly, for the ResNet18 model the loss values remain constant (loss ≈ 0.11) between the 91st and 99th epoch. Therefore, the 105th and 99th checkpoint model are chosen as the final model for MobileNetV2 and ResNet18, respectively. Further results are generated based on predictions of these chosen models.

(a)



(b)

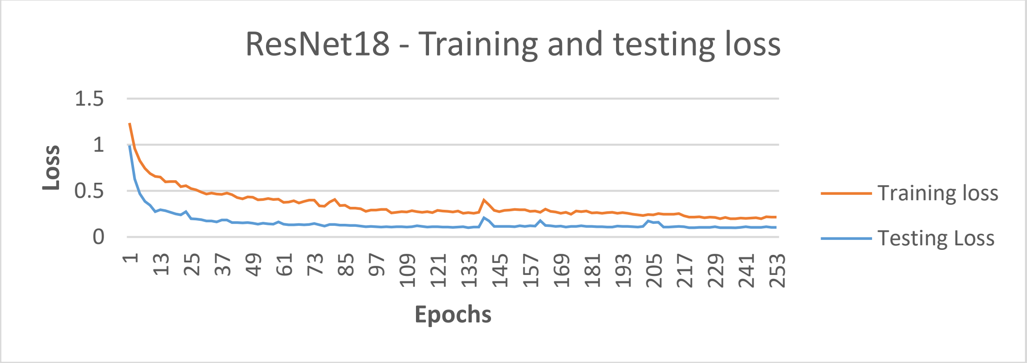


Figure 6: Training and testing loss plotted against epoch number for (a) MobileNetv2 architecture (b) Resnet18 architecture

Average Error

The ResNet18 architecture reported a smaller average error ≈ 15 pixels, in contrast to the higher average error ≈ 22 pixels for MobileNetV2. The gait analysis for the single leg is conducted with the model with the better average error, namely ResNet18.

Percentage of correct keypoints (PCK)

The overall percentage of keypoints for a 5 pixels threshold is reported to be for MobileNetv2 and for ResNet18. The following table lists the PCK for individual joints detected by both models:

Joint	**ResNet18 PCK**	**MobileNetV2 PCK**
Left Ankle	28.75%	23.75%
Left proximal interphalangeal joint	31.25%	25%
Left toe	30%	12.5%
Right ankle	17.5%	15%
Right proximal interphalangeal joint	22.5%	22.5%
Right toe	26.25%	27.5%

Table 2: Percentage of correct keypoints identified for each keypoint by the MobileNetV2 and ResNet18 architecture.

Area under the curve

The area under the curve plot of both the model does not have a steep increase for both the models, implying that the model won’t be accurate for applications with low threshold values. The plot shows the accuracy for various threshold values. The highest accuracy for the ResNet18 and MobileNetV2 model is for threshold value of 31 pixels.

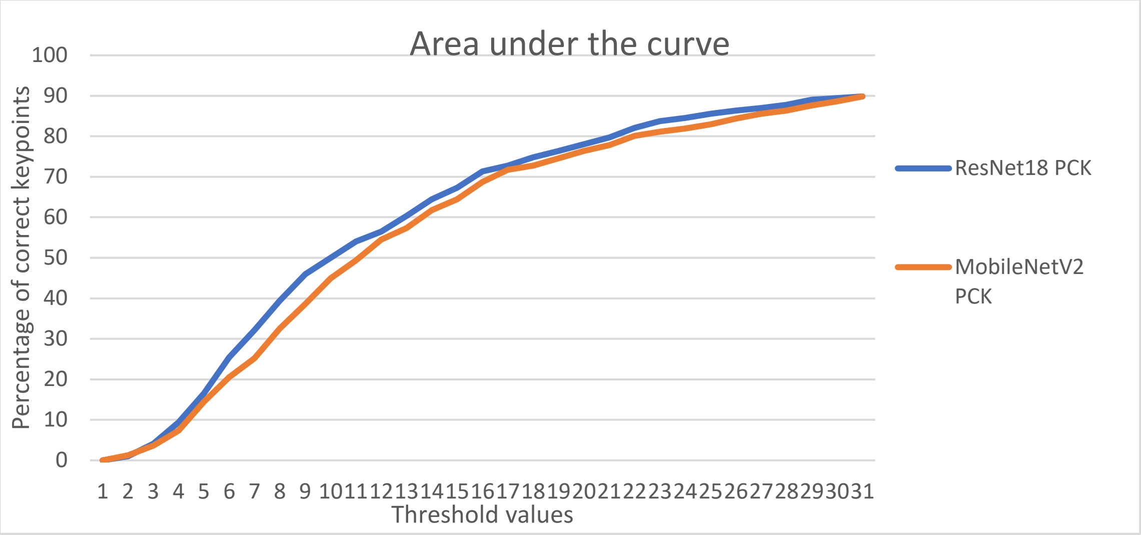


Figure 7: Area under the curve plot depicting PCK for threshold values in the range of 0 to 31 pixels for MobileNetV2 and ResNet18 architecture.

Statistical analysis of range of motion of MTP joint

The range of motion angle calculated based on the ground truth 2D coordinates and the predicted 2D coordinates of the ResNet18 model has a root mean square error of ≈ 10°.The ICC correlation between the angles measured by the two methods indicate a strong correlation (inter class correlation coefficient=0.726)

The BA plot shows high degree of variance for all the average values.

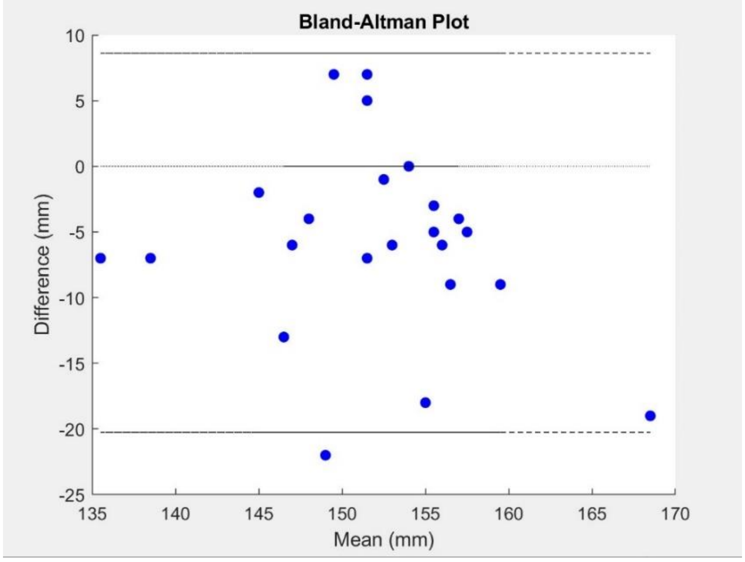


Figure 8: Bland Altman Plot showing 95% Confidence interval limits.

Discussion

The aim of the project was to report on the feasibility of locating joint locations using 2D Pose estimation algorithms to help study range of motion of the lower extremity MTP. The 2D pose estimation algorithm used two feature extraction architectures namely ResNet18 and MobileNetV2. The model with the ResNet18 feature extractor performed a better task of predicting the joint location with an average error of 15 pixels in comparison to the average error of 22 pixels by MobileNetV2 based model. Based on the average error and AUC it can be said the ResNet18 architecture does a better job at predicting the joint location.

From figure 6 it can be seen that the testing accuracy

Interpretation of Statistical Analysis:

Previous pose estimation algorithms have been trained to identify human keypoint but have not included lower extremity anatomical landmarks on the foot. In the study the keypoints chosen are with respect to a strong biomechanical vision and a first step to enable markerless motion capture to aid studies in understanding range of motion of joints in the foot. The current model is limited to identifying only 3 joints on the foot since the dataset was manually annotated which is time consuming. The foot has an overall of 33 different bones and creating a model to identify all the joints will help study all the segments and joint motions possible in the foot. Similar work (LARSSON, 2020) has been done to identify the 21 joints in the hands by 3D pose estimation using the FreiHAND dataset (Zimmermann, et al., 2019).

From a biomechanics perspective for sports, 2D analysis has its limitations for motion capture and analysis by allowing for only 3 degrees of freedom to be captured. 2D motion capture is constrained to capturing motion only in a single plane at a time. Currently for the MTP joint only the PF and DF range of motion in the sagittal plane is captured. To truly study the range of motion in all planes of the MTP joint a 3D pose estimation system would be required to identify the adduction and abduction on the joint. Latest research has provided 3D pose estimation algorithms (Du, et al., 2016) for single view RGB images but could not be implemented because the dataset curated was limited to 2D coordinates.

A flaw in the ground truth coordinates of the dataset is that the camera intrinsic and extrinsic parameters are unknown. The ground truth values would have to be reconstructed to avoid error due to camera distortion (Dunn, Wheat, Miller, Haake, & Goodwill, 2012) for biomechanical applications. Lack of knowledge of the parameters of the cameras used to capture the dataset is due to nature of data acquisition. Using a controlled lab setup with known camera parameters (focal length , lens distortion (skew) , aspect ratio and principal point) to collect and reconstruct the 2D coordinates to overcome camera distortion would greatly strengthen the quality of the dataset and the study. Before the evaluation of the model is conducted, the predictions too would have to be subjected to coordinate reconstruction.

For 3D motion capture using OMS systems and retroreflective markers there is a reported error due to skin movement (Alexander & Andriacchi, 2001). Similarly, for the current project the ground truth data is prone to manual annotation error since the joints are occluded by footwear in the dataset. The ground truth error can not be quantified since the properties of the footwear were unknown. Formulating a dataset with an OMS based motion capture system will help avail a better quality of ground truth values for the data to be trained on. Being able to quantify and rectify the error due to occlusion of joints due to footwear would improve the results of the current 2D pose estimation model and any further markerless motion capture research for foot joints occluded by footwear.

Both the feature extraction architectures used for the 2D pose estimation implementation are lightweight feature extraction networks or small networks. These architectures were chosen to reduce computational complexity of the training process due to limited hardware resources. Using deeper convolution networks would greatly increase the number of parameters learnt during the training process. In the current implementation ResNet18 model learns 12.26 million parameters and MobileNetV2 learns 3.91 million parameters. Based on the results of average keypoint error, ResNet18 has learnt more features making it more accurate. For the current implementation, the input size of the image was 224 x 224 pixels. But, experimenting with different input sizes could lead to improved performance on the model (LARSSON, 2020).

For the project, a custom dataset had to be curated since existing datasets did not include ground truth values for the anatomical landmarks of interest namely left and right ankle, MTP and distal phalanges(toe). The dataset had 800 images extracted from videos on the internet. Studies report for supervised learning neural networks the models exposed to learning with larger dataset sizes produce better accuracy (Abdulraheem, Arshah, & Qin, 2015). State of the art pose estimation algorithms (Newell, Yang, & Deng, 2016) that report an accuracy of 99% percentage of correct keypoints were trained on larger datasets such as MPII (25,000 images with over 40,000 humans with keypoints) (Andriluka, Pishchulin, Gehler, & Schiele, 2014) and Common Objects in Context dataset(250,000 humans with keypoints) (Lin, et al., 2014). Therefore, expanding the dataset to include more datapoints (images) will improve the learning process of the model.

The custom dataset created is from images extracted from making it a sequential dataset. Sequential datasets have been criticized for not being true representations of the real world. (Torralba & Efros, 2011). Therefore, the exposure of the model to learning new features is limited and the testing dataset has high chances of being similar to the training dataset causing the model to overfit. Having a truly wild and random dataset for the training process will greatly improve the model's ability to generalize. The annotation of the keypoint is done manually for the dataset which means there could be human error involved thereby, reducing the quality of the dataset.

Contribution to Knowledge and Impact

The study has explored the feasibility of using 2D pose estimation for studying lower leg kinematics; In particular the range of motion (plantarflexion, dorsiflexion) of the metatarsophalangeal joint by identify three anatomical landmarks on the foot namely the ankle, metatarsophalangeal and distal phalanges. The two models trained on the custom-made dataset showed promising learning curve based on the training loss. Therefore, further research and creation of a comprehensive dataset is encouraged. The significant correlation value (0.726) obtained on performing inter class correlation implies that 2D pose estimation can be considered a valid method for obtaining range of motion of the MTP joint in the sagittal plane. Although, the current model does not meet the accuracy required for biomechanical applications, the study has provided insight into the ability of pose estimation algorithms to learn underlying characteristics of image to detect keypoints. With further research and lab-based validation 2D pose estimation algorithms can be considered for biomechanical application.

References

Abdulraheem, A., Arshah, R. A., & Qin, H. (2015). Evaluating the Effect of Dataset Size on Predictive Model Using Supervised Learning Technique. *International Journal of Computer Systems & Software Engineering*, 1:75-84.

Akio, S., & Jack, S. (October, 1991). Segmentation of people in motion. *In Proceedings of the IEEE Workshop on Visual Motion* , (pp. 325-332).

Akita, K. (1984). Image sequence analysis of real world human motion. *Pattern Recognition*, Vol 17, No. 1, pp. 78-83.

Alexander, E. J., & Andriacchi, T. P. (2001). Correcting for deformation in skin-based marker systems. *Journal of Biomechanics*, 34(3):355-361.

Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014). 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Bezodis, N. E., Salo, A. I., & Trewartha, G. (2015). kinematics and block phase performance in a cross section of sprinters. *EurJ Sport Sci.*, 15(2):118–24.

Bishop, M., Fiolkowski, P., Conrad, B., Brunt, D., & Horodyski, M. (2006). Athletic footwear, leg stiffness, and running kinematics. *Journal of athletic training*, 41(4), 387.

Bourgit, D., Millet, G. Y., & Fuchslocher, J. (2008). Influence of shoes increasing dorsiflexion and decreasing metatarsus flexion on lower limb muscular activity during fitness exercises, walking, and running. *The Journal of Strength & Conditioning Research.*, 22(3), 966-973.

Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2018). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *arXiv:1812.08008 [cs.CV]*.

Du, Y., Wong, Y., Liu, Y., Han, F., Gui, Y., Wang, Z., . . . Geng, W. (2016). Marker-Less 3D Human Motion Capture with Monocular Image Sequence and Height-Maps. *European Conference on Computer Vision*, (pp. 20-36).

Dunn, M., Wheat, J., Miller, S., Haake, S., & Goodwill, S. (2012). TECHNIQUES, RECONSTRUCTING 2D PLANAR COORDINATES USING LINEAR AND NONLINEAR TECHNIQUES. *30 International Conference on Biomechanics in Sports .*

Fong, D. T., & Chan, Y. Y. (2010). The use of wearable inertial motion sensors in human lower limb biomechanics studies: a systematic review. *Sensors*, 10(12), 11556-11565.

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).

Karaulova, I. A., Hall, P. M., & Marshall, D. (2000). A Hierarchical Model of Dynamics for Tracking People witha Single Video Camera. *Proceedings of the British Machine Vision Conference* , pp. 262–352.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

LARSSON, S. H. (2020). *MOBILEPOSE: REAL-TIME 3D HAND POSE ESTIMATION FROM A SINGLE RGB IMAGE.*

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., . . . Dollár, P. (2014). Context, Microsoft COCO: Common Objects in Context. *Computer Vision and Pattern Recognition (cs.CV)*.

Mehrizi, R., Peng, X., Tang, Z., Xu, X., Metaxas, D., & & Li, K. (2018). Toward Marker-Free 3D Pose Estimation in Lifting: A Deep Multi-View Solution. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition*, (pp. 485-491).

Mojica, M. N., & Early, J. S. (2019). Foot Biomechanics. In J. S. Miguel N. Mojica, *Atlas of Orthoses and Assistive Devices (Fifth Edition)* (pp. 216-228). Springer.

Newell, A., Yang, K., & Deng, J. (2016). Stacked Hourglass Networks for Human Pose Estimation. *arXiv:1603.06937* .

Newell, A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation. *European conference on computer vision.*, (pp. 483-499).

Ng, E. X., Monkhouse, C., Wong, P., Meyer, G., Aloni, Y., & Chong, D. Y. (2014). Assessment of the Impact of Positive Heels (Plantarflexion) and Negative Heels (Dorsiflexion) Shoes on Human Walking Gait. *15th International Conference on Biomedical Engineering*, 379-382.

Nobuyasu, N., Tetsuro, S., Kazuhiro, U., Leon, O., Arata, K., Yoichi, I., . . . Shinsuke, Y. (n.d.). Evaluation of 3D markerless motion capture accuracy. *bioRxiv, 842492.*

Pueo, B., & Jimenez-Olmedo, J. M. (2017). Application of motion capture technology for sport performance analysis.

Richards, J. G. (1999). The measurement of human motion: A comparison of commercially available systems. *Human Movement Science*, Volume 18, Issue 5, Pages 589-602.

Rohr, K. (1994). Towards model-based recognition of human movements in image sequences. *CVGIP: Image Understanding*, Vol 74, No. 1, pp. 94–115.

Roy, J. P., & Stefanyshyn, D. J. (2006). Shoe midsole longitudinal bending stiffness and running economy, joint energy, and EMG. . *Medicine & Science in Sports & Exercise*, 38(3), 562-569.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4510-4520.

Stefanyshyn, D., & Fusco, C. (2004). Athletics: Increased shoe bending stiffness increases sprint performance. *Sports Biomechanics*, 3(1), 55-66.

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. *CVPR*, 1521-1528.

van der Kruk, E., & Reijne, M. M. (2018). Accuracy of human motion capture systems for sport applications; state-of-the-art review. *European journal of sport science*, 18(6), 806-819.

Xiu, Y., Chen, Z., & Fang, Y. (2018). *MobilePose-pytorch*. Retrieved from GitHub: https://github.com/YuliangXiu/MobilePose-pytorch

Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., & Brox, T. (2019). FreiHAND: A Dataset for Markerless Capture of Hand Pose and Shape from Single RGB Images. *arXiv:1909.04349v3 [cs.CV]* .