PU5922 MSc Research Project
Academic Paper Cover Page

Student ID Number: 52212588

Title: Exploring the potential of cross-species transfer learning between humans and macaque monkeys to improve markerless motion capture for healthcare applications

Total Word Count: 2,988

Referencing Style: Nature Referencing Style

Acknowledgements

I am grateful to both of my supervisors, Bradley Scott and Dr Dimitra Blana, for their continued guidance and dedicated support during this research project. Their valuable feedback and expertise have been instrumental. I appreciate the opportunity to learn from them and their belief in my abilities as a researcher. Thank you for being excellent mentors.

Declaration of the generated material

No content generated by AI technologies has been used in this assessment.

Abstract

Background: Markerless motion capture through human pose estimation is a promising approach for quantifying human movement patterns without the need for specialised equipment. This study explores the potential of transfer learning from macaque monkey data to improve human pose estimation accuracy using deep learning.

Methods: We establish the ground work by setting up the computational environment and selecting appropriate datasets. Our approach aims to transfer knowledge gained from pre-training the model on macaque monkeys images to human image data. The performance of the modelling approaches is evaluated and compared on macaque and human datasets. Additionally, we examine the impact of different confidence level cutoffs to detect keypoints. We further investigate the limitations of data availability and model evaluation.

Results: Our findings indicate that the baseline model performs worse on human data compared to macaque monkey data, highlighting the potential benefits of transfer learning. Overall low performance at detecting present keypoints reveals the need for further adjustments to the confidence level cutoff. Despite computing resource unavailability preventing the optimisation of re-training parameters, the transfer learning model demonstrates notable improvement in prediction accuracy of existing keypoints in humans.

Conclusion: Transfer learning from macaques shows promise in enhancing human pose estimation. Improved markerless motion capture built on using transfer learning with deep learning models holds potential for more effective diagnosis and monitoring of individuals affected by physical impairment. Future research should focus on refining evaluation concepts and data, and explore hyperparameter tuning to unlock further potential of transfer learning in human pose estimation.

Contents

1 Introduction

Individuals living with physical impairments due to injury or disorder experience abnormal movement patterns [1, 2]. These are likely to have an impact on the well-being of the affected [3], who may be hindered in their daily functioning. To mitigate their struggles through close monitoring and early diagnosis of individuals at risk for movement pathologies, practitioners need a reliable way to objectively measure these movement patterns [4]. Patient visits for regular check-ups or monitoring by a physician can be costly and burdensome. Limitations worsen when using quantitative data from marker-based motion capture systems requiring specialised equipment [4, 5].

Markerless motion capture (MMC) poses a both practical and accessible alternative [5]. This technique utilises algorithms to analyse regular RGB video camera data aiming to detect abnormal movement from video. While this represents a cost-effective and non-invasive approach to quantifying human movement patterns, it constitutes a hard computational problem. However, leveraging artificial neural networks can simplify the process and enhance the efficiency of MMC systems. Deep learning (DL) models in MMC aim to predict the coordinates of key skeletal joints represented by keypoints. This process, known as human pose estimation (HPE) [6], is learned by the model through training on large datasets of images with manually labeled ground truth annotations. Thereby, the combination of DL and MMC enables the automated and objective analysis of movement patterns.

DL models can be trained effectively on large datasets, which poses a challenge when using human image data due to ethical constraints. Similarly, existing datasets are naturally

limited in the variation of poses humans are shown in. Relevant ethical considerations as to recording time and movement restrictions are less strict for animals. Macaque monkeys are considered an essential non-human primate model owing to their behavioural traits, sensory perception, and neural structure [7], being similar to the respective human counterpart. Further, macaques as primates closely resemble humans in their skeletal system, including joint configurations and the number and proportions of limbs [8]. The challenge of acquiring large human image datasets can be addressed through transfer learning (TL) [9]. TL generally aims to improve performance at a specific task by transferring information captured in model parameters from learning another task. By leveraging knowledge from macaque monkey images, this approach hopes to achieve the adaptation of learned features to human data for HPE. Notably, successful cross-species TL has been demonstrated in various fields [10]. The exploration of the potential of macaques to improve HPE aims to contribute valuable insights to advance MMC techniques and benefit individuals with movement pathologies.

This study centres on the application of DeepLabCut [11], a state-of-the-art DL framework designed for pose estimation. Within this scope, we set up the computational environment, curated suitable data to explore the TL approach, and developed tailored modelling workflows. The primary aim of this study is to lay the foundation for investigating the potential of TL from macaques for HPE. Thereby, we seek to enhance the current comprehension of movement pathology and advance MMC for healthcare applications.

2 Methods

2.1 Setup

2.1.1 Data

Deep learning (DL) algorithms require large datasets with manually-devised labels to learn the locations of body parts and have their performance assessed. Ethical approval was not required for this study as only publicly-available human and macaque monkey data has been incorporated.

Common Objects in Context (COCO)

The Common Objects in Context (COCO) keypoint dataset [12] is a diverse collection of images initially compiled from various sources, e.g. Flickr. The widely-used dataset includes extensive annotations such as precise x and y coordinates for keypoint locations. These keypoints represent human body parts, e.g. major joints. The ground truth annotations are essential for training pose estimation (PE) models and benchmarking performance. Within this study, using COCO for re-training and evaluating DL models permits to explore the potential for cross-species transfer learning (TL) between humans and macaque monkeys.

MacaquePose

In addition to COCO, this study leverages the MacaquePose dataset [13], consisting of images obtained from Google Open Images, zoo environments, and the Primate Research

Institute of Kyoto University. The images capture macaque monkeys in diverse settings and poses of naturalistic scenes. Likewise to the COCO images, they are annotated with 17 keypoints that correspond to specific monkey body parts. These labels are compatible with COCO, enabling seamless exploration of cross-species TL. The MacaquePose dataset serves for both pre-training and assessing the model on macaque monkey data before adding human data for TL purposes.

2.1.2 Software and Computational Requirements

DeepLabCut [11] is an open source toolbox providing functionalities to develop deep neural networks learning keypoint locations from regular RGB camera video data. It provides a framework for exploring DL techniques for PE while being accessible to researchers without extensive knowledge of advanced machine learning software libraries, e.g. crude TensorFlow [14]. Additionally, DeepLabCut has the capability to facilitate cross-species TL thanks to the availability of weights of pre-trained models in the DeepLabCut ModelZoo [15]. By offering suitable models for a multitude of animal species including macaques, the DeepLabCut framework eliminates the need for extensive manual data annotation and training of a macaque monkey model within this study.

Access to a GPU-enabled High-Performance-Computing (HPC) cluster offering GPU acceleration and high-speed storage has been crucial for the successful handling of large datasets, as well as training and storing DL models. An overview of the cluster commands involved in this project is given in Figure 2.1.

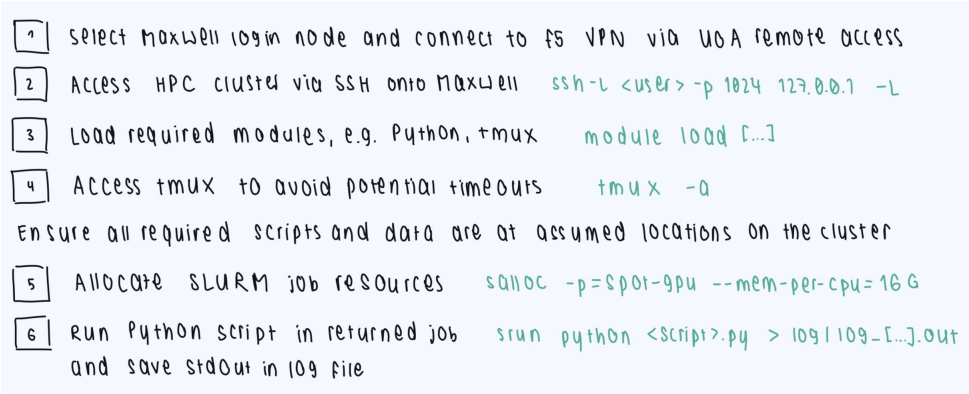


Figure 2.1: HPC cluster accessing and handling process, e.g. to perform deep learning model training.

2.2 Data Flow and Modelling

2.2.1 Data Preprocessing

Before feeding the data into the models, preprocessing steps were applied to ensure that the data was in a suitable format for training and evaluation. We used Python scripts to extract all relevant annotation information.

Since only keypoints were needed for this study, redundant annotations such as segmentation or dense-estimation information from the COCO dataset were discarded. Aligning with downstream clinical applications of individual patient care, the task of this study is predicting keypoint locations for a single person. As a consequence, we removed images depicting multiple humans. Since the original COCO dataset for training and validation contains more than 270k images, we applied a filter to disregard images with four or less visible keypoints, ensuring high training data quality. After preprocessing, the COCO validation set, test set, and training set contained 407, 406 and 19,517 images, respectively.

For the MacaquePose dataset, we likewise removed multi-monkey images and excluded redundant segmentation data. Apart from having served as pre-training data, this study sought to use the MacaquePose dataset for evaluation purposes only. As training would

therefore not be influenced, we did not apply any filter for the minimum number of keypoints visible. After preprocessing, the test set contained 779 images, in agreement with the original MacaquePose publication [13].

2.2.2 Modelling Workflow

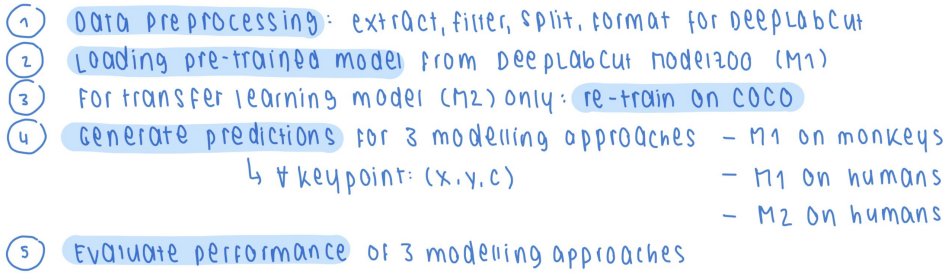


Figure 2.2: Overview of the workflow of this study.

The modelling workflow (see Figure 2.2) in this study can be summarised as follows. First, the model pre-trained on macaque monkey data was obtained from the DeepLabCut ModelZoo. Upon loading the model, we utilised DeepLabCut functionality to select annotations and corresponding images for the test sets, repeating this process for both MacaquePose and COCO data. The COCO test set is currently being held out for future comparisons to state-of-the-art methods and final performance assessments. For the TL part, the remaining COCO data was split into a validation set and a training set. The model used the latter to learn keypoint locations from human images. Predictions were generated further leveraging the DeepLabCut toolbox. The performance of the modelling approaches (see subsection 2.2.3) was estimated using a prediction-likelihood-based confusion matrix and selected performance metrics (see subsection 2.2.4). The code for this project was made available at `https://github.com/isjuao/pommes`.

2.2.3 Modelling Approaches

Pre-trained model M1

PE algorithms commonly build upon a backbone network for feature extraction trained on a large labeled dataset not necessarily related to the final task, thereby learning to recognise patterns in diverse and extensive data. On top of the backbone network, DeepLabCut-based models employ de-convolutional layers to upsample the extracted features for keypoint estimation. This approach enables DeepLabCut algorithms to perform effectively with smaller task-specific datasets, thus reducing the burden of data collection.

For this study, we used a 50-layer Residual Network (ResNet) [16] backbone, pre-trained on the ImageNet dataset [17] for object recognition. Subsequently, the final model was fine-tuned on the MacaquePose dataset (see 2.1.1) to acquire task-specific knowledge. The learned neural network weights were saved and made available in the DeepLabCut ModelZoo [15].

This modelling approach aimed to evaluate the baseline performance of the pre-trained network on the MacaquePose dataset. Since the primary objective of this research is to enhance human PE, predictions were additionally generated using the same model on the COCO validation subset. The model is expected to perform worse on COCO due to the inherent differences between macaque monkeys and humans, thereby highlighting the potential information gain to be achieved through the TL approach.

Transfer learning model M2

The TL model developed within the scope of this study builds upon the same backbone architecture as the pre-trained model to preserve comparability of the two networks. Here, TL is achieved through having the model re-learn the weights using human image data,

starting from weights already containing information extracted from a million pre-training iterations on macaque monkey data [13]. This process is anticipated to lead to a better performance of M2 than M1 on the COCO dataset, containing human data.

2.2.4 Model Evaluation

Data

For the performance evaluation of M1 and M2 on MacaquePose and COCO, we only retained those images with keypoints corresponding to both upper limbs (shoulder, elbow, and wrist) that are visible, due to particular research interest. For the COCO validation subset, 272 out of 406 images remained after the described filtering steps. For the COCO and MacaquePose test sets, 269 out of 407 and 558 of 779 passed, respectively.

Metrics

Confusion Matrix Each prediction of a keypoint location is accompanied by a confidence level $c \in [0, 1]$. To estimate how well the model performs at detecting keypoints, we define the confusion matrix based on a cutoff of 0.4 in accordance with Labuguen *et al.* [13]. Keypoints predicted with confidence level of at least 0.4 and existing in truth are defined as true-positive (TP). Based on the resulting confusion matrix of true-positives, false-positives (FPs), false-negatives (FNs), and true-negatives (TNs), common classification performance metrics such as accuracy, precision, recall, and F1-Score can be calculated. The respective definitions follow known conventions as provided by scikit-learn [18], a widely-used Python library for development and evaluation of machine learning models. To enable a comparison of performance on datasets of different sizes, we calculated the normalised matrix.

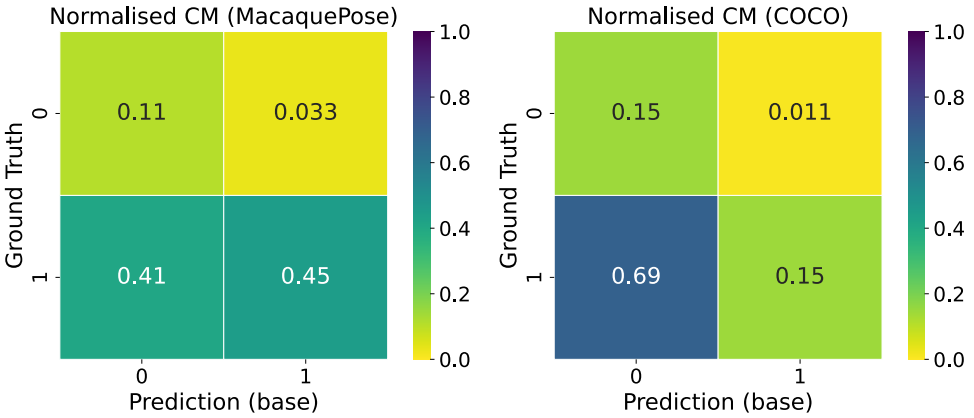
Root Mean Square Error Since the model not only classifies keypoints into being present or not, but predicts exact coordinates, it is appropriate to evaluate how much those predictions differ from the ground truth. Calculating the Root Mean Square Error (RMSE) is a crucial aspect of model evaluation, quantifying the average discrepancy between the model's predictions and the ground truth annotations, providing a comprehensive measure of the model's overall accuracy across the entire dataset. We define the RMSE for each bodypart in accordance with scikit-learn [18]. Equation 2.1 describes the calculation of the RMSE using two-dimensional ground truth (t) and prediction (p) coordinates on n images for body part b.

$$RMSE_b = \sqrt{\frac{\sum_{i=0}^{n-1}\left(\frac{\sqrt{(x_{p_i}-x_{t_i})^2+(y_{p_i}-y_{t_i})^2}}{s}\right)^2}{n}} \tag{2.1}$$

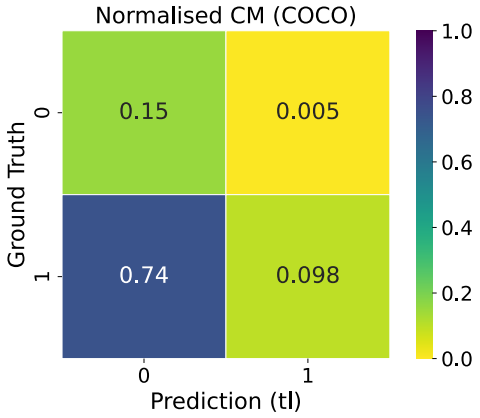
The difference between ground truth and prediction corresponds to the Euclidean distance in pixels. However, the images in the two source datasets as well as within MacaquePose itself can vary significantly in size. In addition to the subject's varying scale and distance, this results in high variation of the impact of keypoint instances on the final RMSE values. To mitigate this effect we included scaling factor s in equation 2.1. Since the primary focus of this research concerns the upper limb, we used the ground truth locations of the six corresponding keypoints to calculate the mean x and y coordinates. This point in two-dimensional space serves as a proxy for the true centroid to minimise the sum of the squared distances, whose calculation would require time-consuming minimisation algorithms. We define s as the mean of the distances of each of the upper limb keypoints to the centroid proxy. Thus, s allows to re-scale the impact of a keypoint on the final RMSE for a body part, depending on the characteristics of corresponding image.

3 Results

The performance of both modelling approaches was evaluated using predictions generated on the High-Performance-Computing cluster. Since this aims to serve as a scoping study, we limit ourselves to numerical comparisons and omitted the calculation of bootstrapping-based confidence intervals.



(a) Baseline model on MacaquePose. (b) Baseline model on COCO.



(c) Transfer learning model on COCO.

Figure 3.1: Normalised confusion matrices of performance of modelling approaches, confidence level $c = 0.4$.

As visualised in Figure 3.1, baseline model M1 correctly identified a higher proportion of keypoints (45%) as present for the monkey images than in images depicting humans

(15%). On COCO, the transfer learning model only detected 10% of all actually present keypoints. Both models performed equally well (15%) at identifying non-annotated keypoints correctly as missing. Notably, both modelling approaches had greater difficulties at predicting keypoints that exist in truth to be present than at correctly identifying when a keypoint is not visible. This is reflected in the low recall value for all modelling approaches (see Table 3.1), which corresponds to a comparatively small proportion of correctly detected keypoints out of all those truly existing. For a confidence level cutoff of $p = 0.1$, the respective proportion was notably higher without compromising precision, as reflected in the higher F1-score [18] (see Table 3.2) and the underlying confusion matrices (see Figure 3.2).

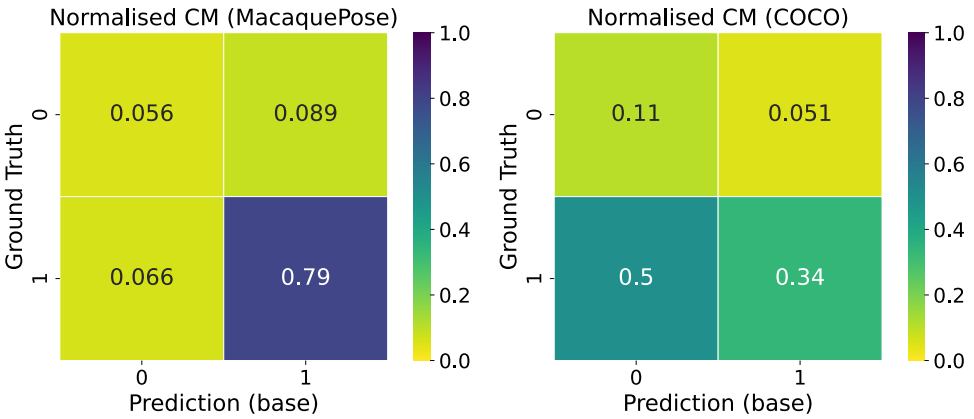
Metric	M1 on MacaquePose	M1 on COCO	M2 on COCO
Accuracy	0.56	0.30	0.25
Precision	0.93	0.93	0.95
Recall	0.52	0.18	0.12
F1-Score	0.67	0.30	0.21

Table 3.1: Performance of modelling approaches using classification metrics, rounded to two decimals, $c = 0.4$.

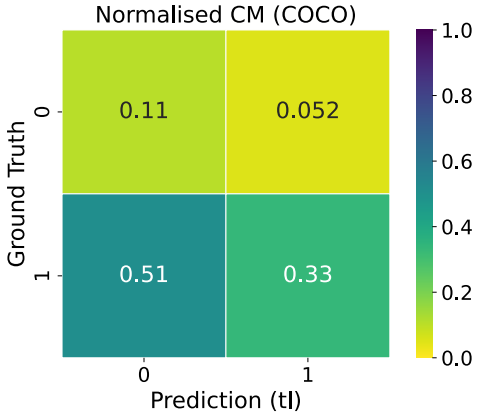
Metric	M1 on MacaquePose	M1 on COCO	M2 on COCO
Accuracy	0.85	0.45	0.44
Precision	0.90	0.87	0.87
Recall	0.92	0.40	0.40
F1-Score	0.91	0.55	0.54

Table 3.2: Performance of modelling approaches using classification metrics, rounded to two decimals, $c = 0.1$.

The prediction accuracy of the modelling approaches was further assessed. Overall, the mean Root Mean Squared Error (RMSE) value of all body parts was greater for the performance of the baseline model on human data than on macaques like it had been trained on (see Figure 3.3). When comparing the former to the transfer learning model, the latter had notably smaller RMSE values for all bodyparts, indicating a more accurate



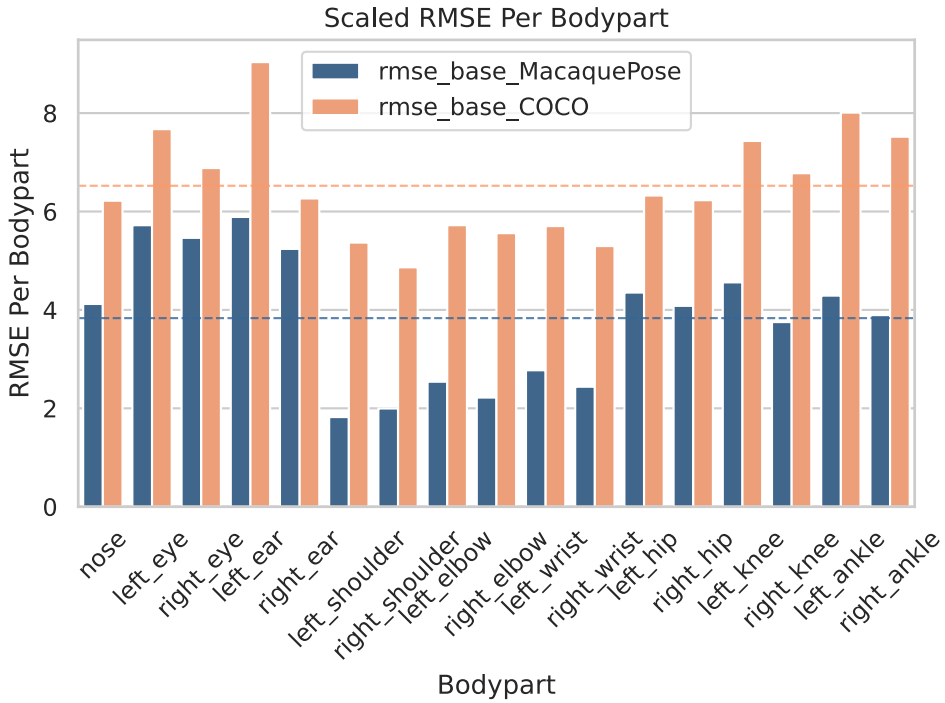
(a) Baseline model on MacaquePose. (b) Baseline model on COCO.



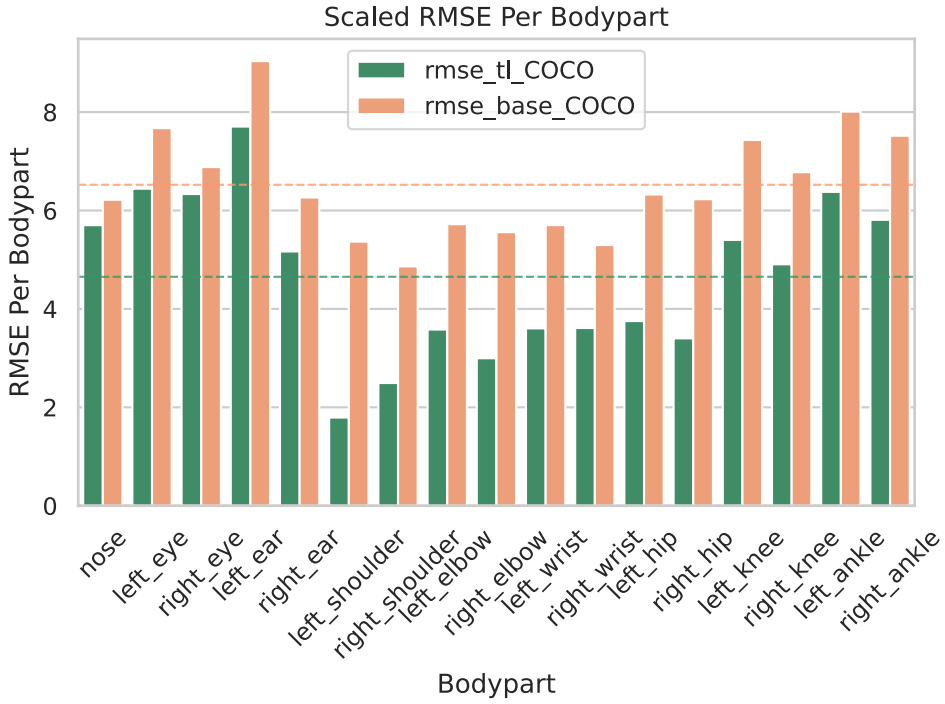
(c) Transfer learning model on COCO.

Figure 3.2: Normalised confusion matrices of performance of modelling approaches, confidence level $c = 0.1$.

performance across the whole dataset. For both comparisons, the more accurate modelling approach was found to perform best on upper limb keypoints.



(a) Baseline model on MacaquePose and COCO validation datasets.



(b) Baseline model and transfer learning model on COCO validation dataset.

Figure 3.3: Root Mean Square Error (RMSE) of predictions for locations of 17 keypoints. Horizontal lines correspond to mean RMSE values of the respective dataset.

4 Discussion

In this study, we confirmed our primary anticipation of the baseline model performing notably worse on humans compared to macaques. However, we observed that the overall performance in detecting present keypoints was low (Figure 3.1, Table 3.1). This is unlikely attributed to the imbalanced nature of the COCO dataset, which is biased towards present keypoints. To address this issue, primary investigation into model behaviour with a lower confidence level cutoff was conducted. Opportunities for other adjustments such as hyperparameter optimisation and extensive model training were limited, as due to external constraints, the High-Performance-Computing cluster was unavailable for the later stages of this study. Consequently, the current transfer learning (TL) model performed similarly to the baseline model in classifying the presence of keypoints (Figure 3.2, Table 3.2). Nevertheless, the TL model exhibited notable improvement in accuracy of coordinate prediction for existing keypoints (Figure 3.3), suggesting untapped potential that warrants exploration through hyperparameter tuning and further research.

4.1 Limitations

Data Limitations

A significant limitation is the lack of information about exact data splits in the MacaquePose dataset, potentially implying performance overestimation due to reuse of images involved in pre-training. Additionally, the lack of images originating as frames from video data in the datasets used for this study implies the absence of contextual information that is

typically present in real-life scenarios, captured in video frames. Such temporal contextual details can be highly beneficial for understanding human movement patterns [19]. Another popular HPE benchmarking dataset consisting of video frames, Human3.6M [20], was found inaccessible due to licensing constraints. However, COCO ensured keypoint compatability and varying image backgrounds, unlike those of Human3.6M out of a clean indoor setting. Thus, COCO images are more appropriate considering downstream usage of the model in diverse settings. Future work will incorporate curated video data focusing on upper limb movements to enable task-specific performance evaluation.

Limitations of Model Evaluation

The evaluation process of this study faced further limitations.

Firstly, although DeepLabCut [11] includes "RMSE" values as part of its evaluation functionality, the calculation of these values corresponds to the arithmetic mean of the Euclidean distances between ground truth and prediction, and does not match the known definition of the Root Mean Square Error (RMSE) [18]. Since Labuguen *et al.* [13] do not provide clarification on the calculation of their RMSE values, we assumed they followed DeepLabCut definitions to obtain their results. Consequently, we evaluated the pre-trained model's performance using our evaluation routine to reach values fit for comparison to the alternative modelling pathways.

Secondly, the variation in images sizes, especially within the MacaquePose dataset, might have a strong impact on the RMSE value and its comparability (see subsection 2.2.4). To mitigate this, a scaling factor was introduced into the RMSE calculation, however, this approach is limited. The proxy for the true centroid does not necessarily minimise the distance to all concerned keypoints and lacks a direct relationship with real-world distances such as upper limb lengths, making it more of an approximation for scaling.

4.2 Outlook

Future studies comparing a tuned TL model against a model pre-trained on humans should consider architectural differences such as high-level structure or the number of layers, impacting the learning capacity and performance of the model [21]. In downstream studies on the power of TL for human pose estimation (HPE), the performance impact of the model's architectural parameters in relation to their effect on the model's interpretability [22] is essential before application in healthcare [23]. Despite promising results, the resource-intensive nature of deep learning model storage and computation limits widespread adoption for markerless motion capture, particularly in healthcare settings.

This study provided the foundation for investigating the potential of TL from macaque monkey data to improve HPE using DeepLabCut. The fundamental TL model performs similarly to the baseline at detecting keypoints in human image data, but predicts coordinates of existing keypoints at higher accuracy. Future research will examine the statistical significance of the findings presented in this study and further explore the potential of TL from macaque monkey data for HPE.

Bibliography

1. Ozturk, A., Tartar, A., Ersoz Huseyinsinoglu, B. & Ertas, A. H. A clinically feasible kinematic assessment method of upper extremity motor function impairment after stroke. *Measurement* **80,** 207–216. doi:https://doi.org/10.1016/j.measurement.2015.11.026 (2016).

2. Subramanian, S. K., Yamanaka, J., Chilingaryan, G. & Levin, M. F. Validity of Movement Pattern Kinematics as Measures of Arm Motor Impairment Poststroke. *Stroke* **41,** 2303–2308. doi:10.1161/STROKEAHA.110.593368. eprint: https://www.ahajournals.org/doi/pdf/10.1161/STROKEAHA.110.593368 (2010).

3. Gane, E. M., McPhail, S. M., Hatton, A. L., Panizza, B. J. & O'Leary, S. P. The relationship between physical impairments, quality of life and disability of the neck and upper limb in patients following neck dissection. *Journal of Cancer Survivorship* **12,** 619–631. doi:10.1007/s11764-018-0697-5 (Oct. 2018).

4. Kidziński, Ł., Yang, B., Hicks, J. L., Rajagopal, A., Delp, S. L. & Schwartz, M. H. Deep neural networks enable quantitative movement analysis using single-camera videos. *Nature Communications* **11,** 4054. doi:10.1038/s41467-020-17807-z (Aug. 2020).

5. Mathis, A., Schneider, S., Lauer, J. & Mathis, M. W. A Primer on Motion Capture with Deep Learning: Principles, Pitfalls, and Perspectives. *Neuron* **108,** 44–65. doi:https://doi.org/10.1016/j.neuron.2020.09.017 (2020).

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

7. Gray, D. T. & Barnes, C. A. Experiments in macaque monkeys provide critical insights into age-associated changes in cognitive and sensory function. *Proceedings of the National Academy of Sciences* **116,** 26247–26254. doi:10.1073/pnas.1902279116 (Dec. 2019).

8. Liang, F., Yu, S., Pang, S., Wang, X., Jie, J., Gao, F., Song, Z., Li, B., Liao, W.-H. & Yin, M. Non-human primate models and systems for gait and neurophysiological analysis. *Frontiers in Neuroscience* **17.** doi:10.3389/fnins.2023.1141567 (2023).

9. Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. *Journal of Big Data* **3,** 9. doi:10.1186/s40537-016-0043-6 (May 2016).

10. Das, S. D. & Kumar, A. Bird Species Classification using Transfer Learning with Multistage Training. *CoRR* **abs/1810.04250.** arXiv: 1810.04250 (2018).

11. Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W. & Bethge, M. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience* **21,** 1281–1289. doi:10.1038/s41593-018-0209-y (Sept. 2018).

12. Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. Microsoft COCO: Common Objects in Context. *CoRR* **abs/1405.0312.** arXiv: 1405.0312 (2014).

13. Labuguen, R., Matsumoto, J., Negrete, S. B., Nishimaru, H., Nishijo, H., Takada, M., Go, Y., Inoue, K.-i. & Shibata, T. MacaquePose: A Novel "In the Wild" Macaque

Monkey Pose Dataset for Markerless Motion Capture. *Frontiers in Behavioral Neuroscience* **14.** doi:10.3389/fnbeh.2020.581154 (2021).

14. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Jia, Y., Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu & Xiaoqiang Zheng. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems* Software available from tensorflow.org. 2015.

15. Ye, S., Filippova, A., Lauer, J., Vidal, M., Schneider, S., Qiu, T., Mathis, A. & Mathis, M. W. *SuperAnimal models pretrained for plug-and-play analysis of animal behavior* 2023. arXiv: 2203.07436 [cs.CV].

16. He, K., Zhang, X., Ren, S. & Sun, J. *Deep Residual Learning for Image Recognition* 2015. arXiv: 1512.03385 [cs.CV].

17. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. *ImageNet: A large-scale hierarchical image database* in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), 248–255. doi:10.1109/CVPR.2009.5206848.

18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

19. Tian, W., Gao, Z. & Tan, D. Single-view multi-human pose estimation by attentive cross-dimension matching. *Frontiers in Neuroscience* **17.** doi:10.3389/fnins.2023.1201088 (2023).

20. Ionescu, C., Papava, D., Olaru, V. & Sminchisescu, C. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36,** 1325–1339 (July 2014).

21. Shrestha, A. & Mahmood, A. Review of Deep Learning Algorithms and Architectures. *IEEE Access* **7,** 53040–53065. doi:10.1109/ACCESS.2019.2912200 (2019).

22. Hohman, F., Head, A., Caruana, R., DeLine, R. & Drucker, S. M. *Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models* in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, Glasgow, Scotland Uk, 2019), 1–13. doi:10.1145/3290605.3300809.

23. De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., Hughes, C. O., Raine, R., Hughes, J., Sim, D. A., Egan, C., Tufail, A., Montgomery, H., Hassabis, D., Rees, G., Back, T., Khaw, P. T., Suleyman, M., Cornebise, J., Keane, P. A. & Ronneberger, O. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine* **24,** 1342–1350. doi:10.1038/s41591-018-0107-6 (Sept. 2018).