Primate Pose Estimation with OpenMonkeyChallenge Proposal

Chahyon Ku, Gustav Baumgart, Josh Spitzer-Resnick, Maximilian Scheder-Bieschin
University of Minnesota
{ku000045,baumg260,spitz123,sched088}@umn.edu

Abstract

In 2021, the Park Lab at the University of Minnesota released the benchmark challenge OpenMonkeyChallenge. This challenge is the leading effort facilitating the creation and collection of models to automatically track non-human primate poses through various environments, and has seen about 20 teams to the competition. Our team proposes to participate in this challenge with the aim of assisting researchers in fields such as biology and biomedicine, and improve their capabilities to gather insights into populations of non-human primates through pose and, by extension, gait analysis. After having reviewed existing literature on both human and non-human pose estimation, our team proposed to implement the recently released transformer architecture, ViTPose. Our team hopes that we can contribute a competitive submission to the OpenMonkeyChallenge by applying this architecture which achieved 81% AP on the MS COCO dataset, and claims to be easily transferable to animal pose estimation [18].

1. Introduction

OpenMonkeyChallenge is a benchmark challenge for 2D non-human primate pose estimation [19]. It consists of 111,529 photographs labeled with 17 body landmarks and is the largest non-human primate image dataset, both in number of images and number of species included. Non-human primate pose estimation is seen as more challenging than human pose estimation because non-human primates have more variation in their joint ranges and body geometry [2]. Despite this, some papers have been able to build robust models against these challenges, and achieve comparable performances to pose estimation on primates [12] [9].

There have been efforts to reconstruct the pose of macaques in 3D, as compared to the 2D in this challenge [2], however, the resources required to gather this data are significant and therefore limit the practicality of these techniques. Because of this limitation, being able to obtain accurate pose analysis from a single 2D image is critical for real-world applications.

Pose estimation has been applied to a wide breath of applications for humans, such as healthcare [4, 14], assisted driving [3], and video games [13]. Some of these applications require near-real time decisions to be made with small compute. Progress has been made in this area through lightweight architectures such as Fast Pose Distillation (FPD) [21]. Other applications such as AR manipulation of fine objects require more precise estimation [5]. Models created as part of the OpenMonkeyChallenge can be applied to study effects of drugs, infectious diseases, and mental illnesses on monkeys. Additionally they can be used for studies ”in the wild” such as automated monitoring of the health of wild primates [2], or understanding their social behaviors.

2. Related Work

2.1. 2D Human Pose Estimation

Human pose estimation was pioneered by Google in 2014 [16], and progress has been supported by two popular datasets: MPII [1] and COCO [10]. These datasets, the improvement of available compute, and expanding applications has lead to human pose estimation gaining increased interest and performance over the years.

More recent work, such as the convolutional pose machine, focuses on end-to-end training fully convolutional networks to classify pixels as landmark locations [17]. HRNet connected parallel high- and low-resolution convolution streams to combine the spatial and semantic information from respective streams and achieved then state of the art results in many downstream tasks including human pose estimation [15].

Transformers are an emerging architecture which has seen strong performance when applied to vision tasks. ViTPose, for example, achieved state-of-the-art performance by attaching a convolutional decoder head on top of the vision transformer to directly regress the heatmap of landmark locations [18]. These transformer architectures continue to push the limits of what information can be extracted from image data.

2.2. 2D Non-human Pose Estimation

The work done on 2D non-human pose estimation is much more limited. A substantial part of pose estimation research on primates has been done on macaque monkeys. For example, an attention-refined light-weight high-resolution network (HR-MPE) aimed at reducing the computing resources required [11].

The data set we will work with is not limited to this species. DeepLabCut has been used in order to perform pose estimation on different species of non-human primates, where they showed a slight improvement on previous work with CNNs [8].

3. Dataset

The OpenMonkeyChallenge has 111,529 total RGB images with 17 possible landmarks per sample. The dataset builds on the scope of OpenMonkeyPose [2] and MacaquePose [9] datasets. OpenMonkeyPose has 195,228 images with 13 annotated landmarks, but is specifically designed for understanding 3D movement, and only contains rhesus macaques. Additionally it was compile using a single, controlled environment potentially making it difficult to apply to broad applications. The MacaquePose dataset consists of only 13,083 images with 17 annotated landmarks, sourced from an number of different environments but likewise only for macaques [19].

The OpenMonkeyChallenge dataset is divided into 60/20/20 train/validation/test splits, namely, 66,917, 22,306, and 22,306 images respectively [19]. Each image is cropped to contain one or more monkeys, with each crop having a resolution of at least 500x500 pixels. These landmarks consist of the nose, left and right eyes, head, neck, left and right shoulder, left and right elbow, left and right wrist, hip, left and right knee, left and right ankle, and tail. Notably, this dataset is comprised of a wide variety of primate species (26 in total), which the authors categorize as New World (6) and Old World (14) monkeys, as well as apes (6). Their environments are also varied, with sources including Flickr, YouTube, three National Primate Research Centers, and the Minnesota Zoo among others. Because of this the OpenMonkeyChallenge provides the strongest publicly available dataset for the study of non-human primate images across domains.

Previous datasets have limitations with regards to species diversity, environments, complexity of image-capture, and dataset size, which hinders generalization of primate pose estimation. OpenMonkeyChallenge is designed to increase the ability to generalize pose estimation performance across primates, regardless of their environment [6, 20].

4. Evaluation Metrics

We will use probability of correct keypoint (PCK@ϵ with $\epsilon = 0.2$) and mean per joint position error (MPJPE$_i$) as performance metrics. PCK@ϵ is a way of measuring how likely it is for the model to predict any joint's position. MPJPE$_i$ calculates the mean distance of the predicted joint position from the actual joint position. Both of these metrics do normalize by the size of the bounding box W.

$$\text{PCK@}\epsilon = \frac{1}{17J}\sum_{j=1}^{J}\sum_{i=1}^{17}\delta\left(\frac{||\hat{x}_{ij} - x_{ij}||}{W} < \epsilon\right) \tag{1}$$

$$\text{MPJPE}_i = \frac{1}{J}\sum_{j=1}^{J}\frac{||\hat{x}_{ij} - x_{ij}||}{W} \tag{2}$$

These are the metrics proposed by OpenMonkeyChallenge to evaluate model performance.

5. Baseline Method

For the baseline we will implement HRNet. This model was chosen as it is one of the best baseline models from the original OpenMonkeyChallenge paper, with average precision of 0.78. Its implementation will allow us to confirm that our pipeline is working properly. Additionally, its implementation will provide us with a more gentle learning curve of implementing our own neural network architecture in Google Colab. The code and models used in the HRNet paper are available through their Github repository and have been broadly implemented, providing additional troubleshooting resources if needed.

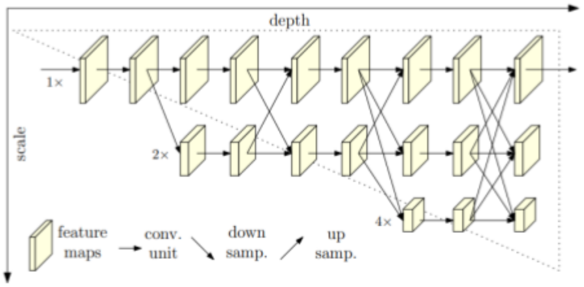
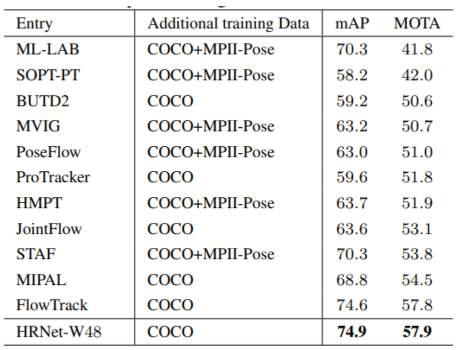


Figure 1. HRNet visual from Deep High-Resolution Representation Learning for Human Pose Estimation. [15]

HRNet was initially proposed for human pose estimation in 2019 [15]. Unlike previous methods that traversed low-to-high or high-to-low resolutions, [7], HRNet proposed a new architecture that maintained high-resolution representations through the entire model. This led to two benefits as compared to existing models (1) improved spatial precision due to not needing to recover resolution (2) improved capabilities for pose estimation due to repeated multiscale fusions [15]. Figure 2 shows that HRNet performs well against a number of state-of-the-art models.



Entry	Additional training Data	mAP	MOTA
ML-LAB	COCO+MPII-Pose	70.3	41.8
SOPT-PT	COCO+MPII-Pose	58.2	42.0
BUTD2	COCO	59.2	50.6
MVIG	COCO+MPII-Pose	63.2	50.7
PoseFlow	COCO+MPII-Pose	63.0	51.0
ProTracker	COCO	59.6	51.8
HMPT	COCO+MPII-Pose	63.7	51.9
JointFlow	COCO	63.6	53.1
STAF	COCO+MPII-Pose	70.3	53.8
MIPAL	COCO	68.8	54.5
FlowTrack	COCO	74.6	57.8
HRNet-W48	COCO	**74.9**	**57.9**

Figure 2. Performance of HRNet against other popular pose identification architectures. [15]

The HRNet model architecture and weights from training on ImageNet have been ported into PyTorch and TensorFlow allowing for easy transfer into our own model. To implement our model we will modify the original HRNet by adding a 17 channel regression head. These channels provide output a heatmap of for each of the 17 landmarks used by OpenMonkeyChallenge. After transfering in the weights of the original HRNet Model we will incrementally unfreeze the layers train against the OpenMonkeyChallenge dataset to update the weights so that they are refined to our specific use. To keep the initial complexity low We will start with the shallowest variant of HRNet before training deeper models on Google's Collab machines.

6. Proposed Method

We propose to implement the ViTPose model. This model was selected as it has not yet been applied to the task of non-human primate pose estimation and has recently state-of-the-art performance for human pose estimation [18]. Our goal with this model is to improve upon the current best performing model submitted to OpenMonkeyChallenge. The paper was submitted to Arxiv in April 2022, to document the performance of applying the excellent performance seen from transformers in visual recognition tasks to pose estimation. As their paper shows they achieved a significant performance improvement against other state-of-the-art methods on the COCO validation set.

The ViTPose architecture consists of non-hierarchical vision transformers as backbones which serve to extract feature maps for the given instances of people. The backbones they employed were trained on masked image modeling pretext such as MAE to provide a strong weight initialization. Once the feature maps are extracted they are passed into a lightweight decoder which processes them by upsampling the feature maps and regressing the associated heatmaps with respect to the landmarks. This decoder consists of just two deconvolution layers and one prediction layer. An overview of this architecture is shown in Figure 4.

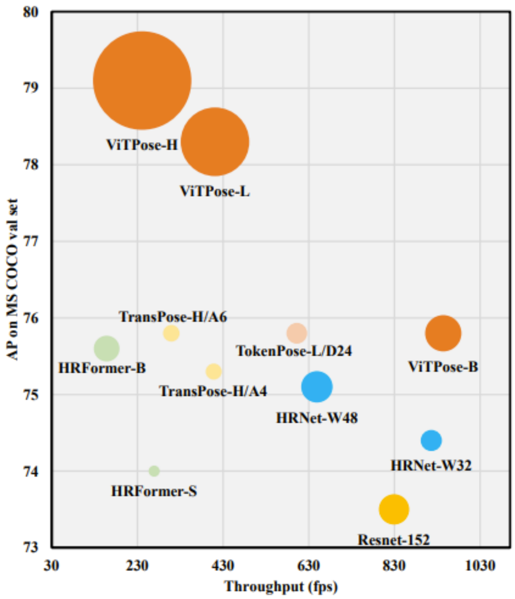


Figure 3. Comparison of ViTPose against other state-of-the-art methods using the MS COCO validation set. The size of each bubble represents the number of model parameters, the horizontal axis throughput, and the vertical axis precision. [18]

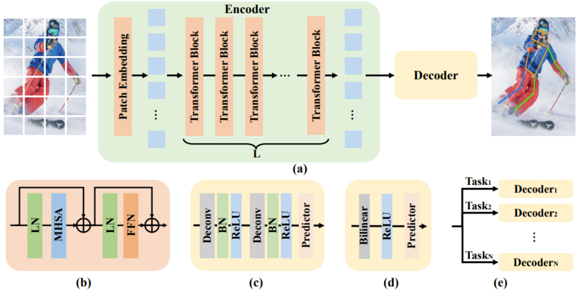


Figure 4. Visuals from ViTPose (a) The ViTPose model. (b) The transformer block. (c) The classic decoder. (d) The simple decoder. (e) The decoders for multiple datasets. [18]

Outside of its performance the ViTPose shows that it is adjustable in regards to its simplicity, scalability, flexibility, and transferability. These traits should help us be able to cater our implementation better to our application and available compute. In the discussion section of their paper Xu et. al. share their belief in the success of ViTPose when applies to animal pose estimation; one of the reasons our group decided to pursue this as our proposed method.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 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. 1

[2] Praneet C. Bala, Benjamin R. Eisenreich, Seng Bum Michael Yoo, Benjamin Y. Hayden, Hyun Soo Park, and Jan Zimmermann. Openmonkeystudio: Automated markerless pose estimation in freely moving macaques. *bioRxiv*, 2020. 1, 2

[3] Guido Borghi, Marco Venturelli, Roberto Vezzani, and Rita Cucchiara. Poseidon: Face-from-depth for driver pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4661–4670, 2017. 1

[4] Vandad Davoodnia, Saeed Ghorbani, and Ali Etemad. In-bed pressure-based pose estimation using image space representation learning. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Jun 2021. 1

[5] Niklas Gard, Anna Hilsmann, and Peter Eisert. Combining local and global pose estimation for precise tracking of similar objects. 2022. 1

[6] Benjamin Y Hayden, Hyun Soo Park, and Jan Zimmermann. Automated pose estimation in primates. *American journal of primatology*, page e23348, 2021. 2

[7] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model, 2016. 2

[8] Rollyn Labuguen, Dean Karlo Bardeloza, Salvador Blanco Negrete, Jumpei Matsumoto, Kenichi Inoue, and Tomohiro Shibata. Primate markerless pose estimation and movement analysis using deeplabcut. In *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, pages 297–300. IEEE, 2019. 2

[9] Rollyn Labuguen, Jumpei Matsumoto, Salvador Blanco Negrete, Hiroshi Nishimaru, Hisao Nishijo, Masahiko Takada, Yasuhiro Go, Ken-ichi Inoue, and Tomohiro Shibata. Macaquepose: A novel “in the wild” macaque monkey pose dataset for markerless motion capture. *Frontiers in behavioral neuroscience*, 14:581154, 2021. 1, 2

[10] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 1

[11] Sicong Liu, Qingcheng Fan, Shanghao Liu, Shuqin Li, and Chunjiang Zhao. An attention-refined light-weight high-resolution network for macaque monkey pose estimation. *Information*, 13(8):356, 2022. 2

[12] Salvador Blanco Negrete, Rollyn Labuguen, Jumpei Matsumoto, Yasuhiro Go, Ken-ichi Inoue, and Tomohiro Shibata. Multiple monkey pose estimation using openpose. *bioRxiv*, 2021. 1

[13] Manisha Patel and Nilesh Kalani. A survey on pose estimation using deep convolutional neural networks. *IOP Conference Series: Materials Science and Engineering*, 1042(1):012008, Jan 2021. 1

[14] Jan Stenum, Kendra M. Cherry-Allen, Connor O. Pyles, Rachel D. Reetzke, Michael F. Vignos, and Ryan T. Roemmich. Applications of pose estimation in human health and performance across the lifespan. *Sensors*, 21(21), 2021. 1

[15] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 1, 2, 3

[16] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. 2013. 1

[17] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. *CoRR*, abs/1602.00134, 2016. 1

[18] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation, 2022. 1, 3

[19] Yuan Yao, Abhiraj Mohan, Eliza Bliss-Moreau, Kristine Coleman, Sienna M. Freeman, Christopher J. Machado, Jessica Raper, Jan Zimmermann, Benjamin Y. Hayden, and Hyun Soo Park. Openmonkeychallenge: Dataset and benchmark challenges for pose tracking of non-human primates. *bioRxiv*, 2021. 1, 2

[20] Haozheng Yu. Segmentation and dense keypoints estimation of monkeys. 2021. 2

[21] Feng Zhang, Xiatian Zhu, and Mao Ye. Fast human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1