Pose Estimation in Non-Human Primates - OpenMonkeyChallenge

Hari Veeramallu, Harrison Russell-Pribnow, & Thomas Moreau
University of Minnesota
veera047@umn.edu, russe823@umn.edu, morea024@umn.edu

Abstract

Pose estimation is an important and rapidly growing field in computer vision. The applications of pose estimation range from detecting figures in security cameras to studying the behavior of animals. Over the past few years, many accurate methods and benchmarks for human pose estimation have been introduced. But these techniques were specific to humans and couldn't be generalized to other species. Human pose estimation benefits from the fact that humans have limited pose configurations and similar anatomy to one another. Non-human primate pose estimation is a step towards generalizing pose estimation to be more robust to non-standard situations. We propose to improve the accuracy of the Convolutional Pose Machine by training separate pose machines for three classes of monkeys apes, old world, and new world. We propose that When Pose Machine is specialized for a subset of monkeys, they will have increased efficacy over more generalized models.

1. Introduction

Over recent years, there have been significant advancements in systems that can estimate the pose of animals. With the support of deep learning and sophisticated pose estimation tools, research has further been promoted. However, animal pose estimation is still in the preliminary stage compared to the ever long human pose estimation, which has high accuracy and applicability [7]. In particular, the ability to track non-human primates has greatly lagged due to their homogeneous body texture and potentially large pose configurations [13].

The OpenMonkeyChallnege [13] presents a challenge to design an algorithm to estimate the pose and detect the pose key points of various species of primates. These pose key points correspond to a physical feature of the primates (like eye, shoulder, tail, etc.).

This work presents a non-human primate pose estimation technique based on Convolutional Pose Machine (CPM) [12], originally developed for human pose estimation.

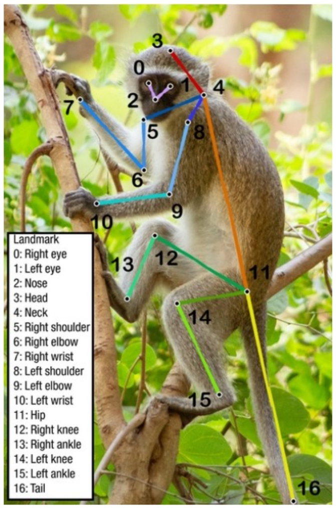


Figure 1. **Monkey Landmark Annotations.** [13]

1.1. Motivation

The motivation behind this project is that the OpenMonkeyChallenge is both challenging and requires skills applicable to other computer vision projects. This challenge is also relevant to many modern computer vision problems. If a pose estimator can be created to allow for the variance between different species of primates, then a more general pose estimator that allows for greater variance in human subjects can be made. Solving this problem will provide insight into the problems and solutions involved in generalizable pose estimation.

2. Related Work

2.1. OpenMonkeyChallenge

The OpenMonkeyChallenge [13] (http://openmonkeychallenge.com/) is a benchmark challenge for non-human primate pose estimation and tracking. It is an open and ongoing challenge that encourages the community to build generalizable on-human primate pose, estimation models. The performance of each submitted model is evaluated using standard evaluation metrics: Mean Per Joint Position Error (MPJPE) [6], Probability of Correct Keypoint (PCK) [3], Average Precision (AP) [13], and Object Keypoint Similarity (OKS) [9].

2.1.1 Dataset

The OpenMonkeyChallenge provides a diverse dataset of 111,529 annotated images of 26 species (6 New World monkeys, 14 Old World monkeys, and 6 apes as shown in Fig. 2) of non-human primates in natural contexts with 17 landmark annotations. This dataset is made up of cropped images containing at least one primate extracted from

1. different images and videos from the internet
2. photographs from National Primate Research Centers
3. videos of 27 Japanese macaques in the Minnesota Zoo

There are 17 landmarks int total which comprise a pose - Nose, Left eye, Right eye, Head, Neck, Left shoulder, Left elbow, Left wrist, Right shoulder, Right elbow, Right wrist, Hip, Left knee, Left ankle, Right knee, Right ankle, and Tail (as shown in Fig. 1).

This dataset is split into training, validation and test datasets (60%, 20%, and 20% respectively) [13] and are made publicly available for use. This dataset has been demonstrated as qualitatively effective by comparing the dataset with existing datasets based on seven state-of-the-art pose estimation models.

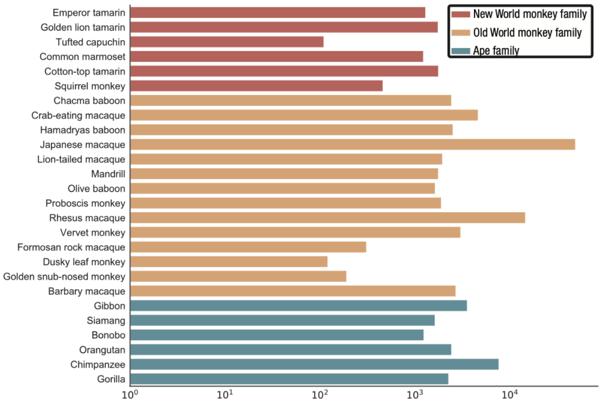


Figure 2. **Composition of OpenMonkeyChallenge Dataset.** [13]

2.2. Metrics

In accordance with the OpenMonkeyChallenge, we evaluate both methods using Mean Per Joint Position Error (MPJPE) [6]. It measures the normalized error between the ground truth and estimated landmark locations. That is,

$$MPJPE_i = \frac{1}{J}\sum_{j=1}^{J}\frac{\|\hat{x_{ij}} - x_{ij}\|}{W} \quad (1)$$

for the i$^{\text{th}}$ landmark, J number of images, and W width bounding box. The ground truth and estimated landmark locations are $\hat{x_{ij}}$ and x_{ij} respectively. Smaller MPJPE values indicate better performance.

We also use Probability of Correct Keypoint (PCK) [3]. It measures the probability that the normalized error between the ground truth and estimated landmark locations is within a given error tolerance ϵ. That is,

$$PCK@\epsilon = \frac{1}{17J}\sum_{j=1}^{J}\sum_{i=1}^{17}\delta\left(\frac{\|\hat{x_{ij}} - x_{ij}\|}{W} > \epsilon\right) \quad (2)$$

where $\delta(.)$ returns 1 when the condition is true and 0 when false.

2.3. Convolutional Pose Machine

Wei *et al*. [12] developed a method for human pose estimation that involves the use of a Convolutional Pose Machine (CPM). In their paper, they describe a convolutional architecture for Pose Machines that provides state-of-the-art pose estimation. Over multiple stages, they create increasingly refined landmark belief maps (shown in Fig. 3) using image features and spatial context features of preceding belief maps.

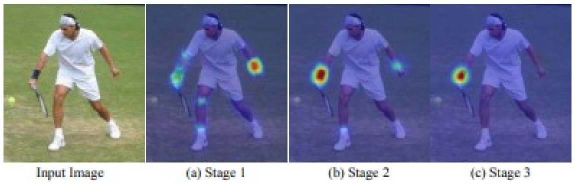


Figure 3. **Human Pose Estimation with a CPM.** [12]

As a result of this architecure, the CPM generates increasingly refined estimates for the landmarks. The model was evaluated on three human pose datasets: MPII Human Pose Dataset (more than 28000 training images), Leeds Sports Pose Dataset (11000 training images and 1000 testing images), and FLIC Dataset (3987 training images and 1016 testing images). The model demonstrated state of the art performance using PCK for all landmarks.

2.4. Non-Human Primate Pose Estimation

In the context of pose estimation in non-human primates (*e.g*. monkeys, macaques, baboons, *etc*.), several markerless approaches have been realized through the use of multiple deep sensing cameras, where a skeleton model is fitted into the 3D reconstructed monkey. However, this approach depends on several factors, such as the number of cameras, their calibration, controlled laboratory setup and a limited number of subjects [1, 5].

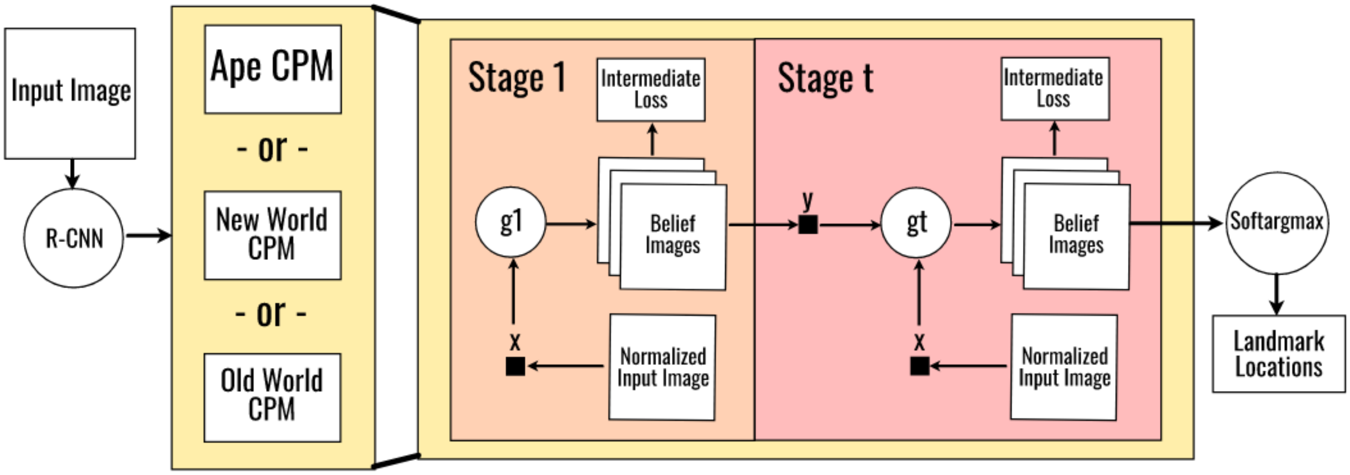


Figure 4. **Overview of our method.** We show the flow of our method through a R-CNN, one of three CPMs, and the softargmax function. First, the R-CNN classifies the input image and estimates a bounding box. Second, a CPM creates belief images for its corresponding species class. Third, the softargmax function extracts landmark locations.

Yao *et al*. [14] developed Multiview Optical Supervision Network (MONET), an end-to-end semi-supervised learning framework to detect monkey pose using multi-view image streams. This framework relies upon spatial and temporal consistency across image streams and does not use any deep learning models. Gaurav *et al*. [4] used deep convolutional neural networks (CNN) as means to locate key feature points of monkey towards understanding its behavior in its natural habitat. In the recent works, DeepLabCut (DLC) [8, 10] and OpenPose [2, 11] were used for monkey pose estimation.

3. Method

3.1. Baseline

Our method builds upon Wei *et al*. [12], who use a Convolutional Pose Machine to estimate landmark locations on humans. The CPM is divided into three stages that generate 18 belief images, 17 for landmarks and 1 for the background. The belief images convey the confidence that a landmark is at a location. So, $b_t^i(u)$ is the confidence that the i^{th} landmark in stage t is at location u. In the first stage, image features x are extracted from the input image through a sequence of convolution, ReLu, and pooling layers. The classifier g_1 then uses the image features to calculate landmark confidence values for each location:

$$g_1(x_u) \rightarrow \{b_1^i(u) \mid 1 \leq i \leq 18\} \qquad (3)$$

In later stages, context features y are extracted from the belief images of the previous stage through convolution. The classifier g_t uses both image features and context features to calculate refined landmark confidence values for each location:

$$g_t(x_u, y_u) \rightarrow \{b_t^i(u) \mid 1 \leq i \leq 18\} \qquad (4)$$

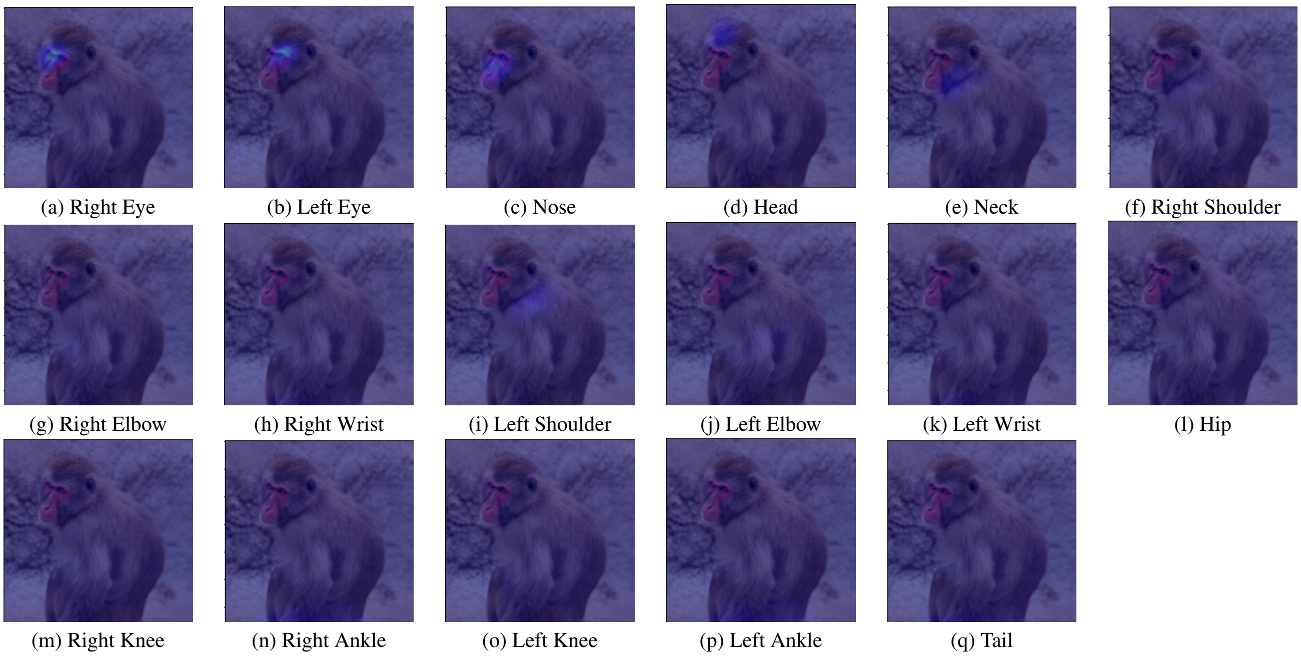
The use of context features allows the model to learn spatial context between landmarks.

3.2. Proposed

The diverse anatomy and behaviour of non-human primate species make generalizing the relationships between image and context features for pose estimation difficult. To handle the differences between species, a CPM will be trained on each class of species (Old World, New World, and Ape). Non-human primates will be classified into one of the classes, and the corresponding CPM will estimate its landmark locations. An overview of the method is illustrated in Figure 4.

3.2.1 Primate Classification

To classify each image as Old World, New World, or Ape, we used a faster-r-CNN object detector. This object detector will calculate bounding boxes and classify each image to sort them into their appropriate model. To train the faster-r-CNN model, we will use transfer learning. We will start with pre-trained weights and fine-tune the model to detect and classify monkeys into their species. There are 26 species classes that the classifier will consider, each of which belongs to one of the three monkey groups. Faster-r-CNN will detect multiple monkeys in some validation images. To account for this during evaluation, we will store the images by the detected monkey with the highest confidence. Sometimes, the detector will not detect monkeys in images with them. In these cases, we will assign images



(a) Right Eye (b) Left Eye (c) Nose (d) Head (e) Neck (f) Right Shoulder

(g) Right Elbow (h) Right Wrist (i) Left Shoulder (j) Left Elbow (k) Left Wrist (l) Hip

(m) Right Knee (n) Right Ankle (o) Left Knee (p) Left Ankle (q) Tail

Figure 5. **Belief Images.** Our method can confidently estimate the location of prominent landmarks (eyes, nose, head, etc.) but struggles wth others (wrists, ankles, tail, etc.).

to a random species. This overall is an effective method of sorting images as they are received.
The open monkey challenge training set does not contain a balanced distribution of the species classes; there is a disproportionately large amount of Japanese Macaques. Of the 60,917 training images, 29,424 of them are of Japanese Macaques. Almost half of the images are of a single of the 26 species. Such a disparity in training classes may cause issues with classification accuracy as the detector will mostly train a single class. We will trim 20k images from the Japanese Macaque group to rectify this. The hope is that the other groups will get proportionally more training time, but just in case this lowers the overall accuracy, we will train two detectors, one based on the reduced training set and one on the full, and use the higher accuracy model in the final version.

3.3. Training

Each input image is cropped to the bounding box of the non-human primate, resized to 368 x 368 pixels, and normalized. The CPM estimates belief images from the normalized image. The ground truth belief image for each landmark is created by placing a Gaussian peak at the ground truth landmark location. Then, the ground truth belief image for the background is created by inverting the sum of all landmark belief images. At the end of each stage, the intermediate L2 norm squared loss between estimated and ground truth belief images is calculated. Thus, the overall loss is the sum of intermediate losses. All stages of the CPM are trained jointly using stochastic gradient descent to minimize the overall loss.

3.4. Landmark Location Extraction

A belief image provides confidence values to indicate the locations where landmarks are likely to appear. To choose the most likely location z, we use the softargmax function on the third stage belief images b_3^i. So,

$$z = \sum_u \frac{\exp(\alpha b_3^i(u))}{\sum_v \exp(\alpha b_3^i(v))} u \tag{5}$$

where u and v are locations and α is a constant determining how heavily confidence should be weighted.

4. Results

We evaluated our proposed and baseline methods on their efficacy in detecting the poses of the images in the validation set. We measured each model based on its MPJPE and PCK.

4.1. Classification

The classification was tested on two faster-r-CNN models, one trained on a reduced dataset and the other on the full. Both models had difficulty detecting monkeys on some of the more difficult validation images. Some images were

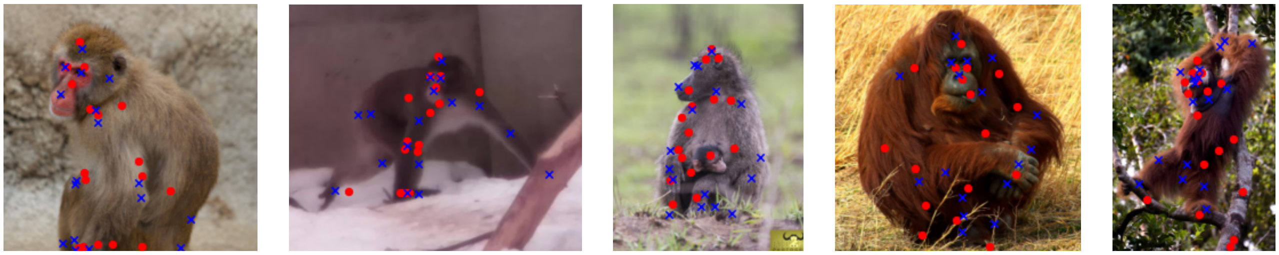


Figure 6. **Landmark Location Estimations.** The blue X's are the ground truth locations and the red circles are our estimated locations. Our method can reasonably estimate the landmark locations for a variety of non-human primate species and poses.

occluded or zoomed out, so the model had difficulty detecting these cases. The two models did miss images at different rates, the model trained on the reduced data set missed 3.4% of validation images, and the full data set missed 4%. The reduced model had an overall classification accuracy of 77.5%. And the full had an accuracy of 72.2%. The reduced dataset model missed fewer images and classified more images correctly, so we used the reduced trainset model in the final proposed pose detector. Figure 7 shows a confusion matrix of the reduced trainset model.

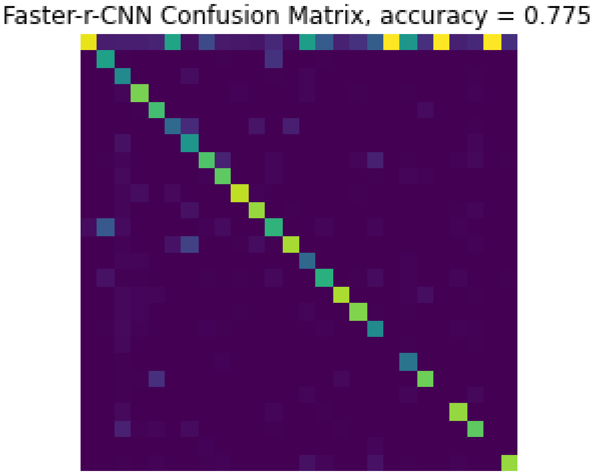


Figure 7. **Confusion Matrix of Faster-R-CNN object detector.** Classification accuracy for species is 0.775 and 0.973 for species class.

4.2. Pose Estimation

Table 1. **Overall Metrics**

Method	MPJPE	PCK@0.2	PCK@0.5
Baseline	**0.2121**	**0.6527**	0.8967
Ours	0.2178	0.6363	**0.8984**

Table 2. **MPJPE by Landmark**

Method	Baseline	Ours
Right Eye	**0.1275**	0.1371
Left Eye	**0.1267**	0.1349
Nose	**0.1345**	0.1446
Head	**0.1362**	0.1425
Neck	**0.1471**	0.1548
Right Shoulder	**0.1737**	0.1816
Right Elbow	**0.2183**	0.2187
Right Wrist	0.2955	**0.2925**
Left Shoulder	**0.1727**	0.1808
Left Elbow	**0.2100**	0.2200
Left Wrist	0.2953	**0.2903**
Hip	**0.2938**	0.3039
Right Knee	**0.2349**	**0.2349**
Right Ankle	**0.2366**	0.2370
Left Knee	**0.2323**	0.2403
Left Ankle	0.2368	**0.2336**
Tail	**0.3335**	0.3553

Table 3. **MPJPE by Species Class**

Method	Ape	Old World	New World
Baseline	0.2316	**0.2101**	0.1880
Ours	**0.2266**	0.2188	**0.1818**

According to Overall Metrics 1, our method achieves comparable overall performance to the baseline. According to MPJPE by Landmark 2, our method tends to perform slightly better on harder landmarks and slightly worse on easier ones.The proposed model has a better MPJPE for the right wrist, left wrist, right knee and left elbow. According to MPJPE by Species Class 3, our method performs slightly better on underrepresented species classes, Apes and New World Monkeys, and slightly worse on Old World Monkeys. Figure 6 illustrates these results. The red predicted locations are generally near the blue ground truth locations, but some ground truth locations are missed. Overall, our method is effective at estimating landmark locations across landmarks and species classes.

5. Conclusion

Summarizing our work, we trained three separate CPM models, each trained to learn the landmarks for each species class. To classify each monkey image into its respective species category, we trained an R-CNN object detector using transfer learning. Overall, we achieved comparable results as compared to the baseline method. Our model performed better in identifying crucial landmarks like the wrists and knees. Additionally, our model gave better results for Apes and New World monkeys but could not outperform the baseline for Old World monkeys. We could observe closer landmark detection using our approach, as shown in Figure 6.

Overall, our method is largely comparable to the baseline method. Our method does achieve better performance on the underrepresented species classes, Apes and New World Monkeys. However, the dataset is dominated by Old World Monkeys, so the increase in overall performance is likely negligible. The performance of our method could be further improved by performing additional data augmentation to limit the number of poor training samples (multiple or occluded non-human primates). Additionally, the number of stages in each CPM could be raised to increase the range of learned context features. Finally, more samples of New World Monkeys and Apes could be added so that Old World Monkeys do not dominate the dataset.

References

[1] Praneet C Bala, Benjamin R Eisenreich, Seng Bum Michael Yoo, Benjamin Y Hayden, Hyun Soo Park, and Jan Zimmermann. Automated markerless pose estimation in freely moving macaques with openmonkeystudio. *Nature communications*, 11(1):1–12, 2020. 2

[2] Salvador Blanco Negrete, Rollyn Labuguen, Jumpei Matsumoto, Yasuhiro Go, Ken-ichi Inoue, and Tomohiro Shibata. Multiple monkey pose estimation using openpose. 2021. 3

[3] Zhe Cao, G Hidalgo Martinez, Tomas Simon, and S Wei. Ya, sheikh. openpose: Realtime multi-person 2d pose, estimation using part affinity fields. *IEEE Transactions on Pattern, Analysis and Machine Intelligence*, 4, 2019. 2

[4] Vishal Gaurav, Salvador Negrete Blanco, Jumpei Matsumoto, Kenichi Inoue, Tomohiro Shibata, et al. Monkey features location identification using convolutional neural networks. *bioRxiv*, page 377895, 2018. 3

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

[6] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7718–7727, 2019. 2

[7] Le Jiang, Caleb Lee, Divyang Teotia, and Sarah Ostadabbas. Animal pose estimation: A closer look at the state-of-the-art, existing gaps and opportunities. *Computer Vision and Image Understanding*, 222:103483, 2022. 1

[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. 3

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[10] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018. 3

[11] Daniil Osokin. Real-time 2d multi-person pose estimation on cpu: Lightweight openpose. *arXiv preprint arXiv:1811.12004*, 2018. 3

[12] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 1, 2, 3

[13] Yuan Yao, Praneet Bala, Abhiraj Mohan, Eliza Bliss-Moreau, Kristine Coleman, Sienna M Freeman, Christopher J Machado, Jessica Raper, Jan Zimmermann, Benjamin Y Hayden, et al. Openmonkeychallenge: Dataset and benchmark challenges for pose estimation of non-human primates. *International Journal of Computer Vision*, pages 1–16, 2022. 1, 2

[14] Yuan Yao, Yasamin Jafarian, and Hyun Soo Park. Monet: Multiview semi-supervised keypoint detection via epipolar divergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 753–762, 2019. 3