A Case Study in Fairness Evaluation: Current Limitations and Challenges for Human Pose Estimation

Julienne LaChance*, William Thong*, Shruti Nagpal, Alice Xiang

* equal contribution

firstname.lastname@sony.com

Sony AI

Motivation

- Computer vision models are affected by sensitive attributes (e.g., gender, skin tone, age)
- Current studies focus mostly on face-related tasks
- Operationalizing fairness evaluation in practice is challenging

Contributions

- Fairness evaluation of 2D pose human estimation
- Highlight current limitations and challenges in fairness evaluations
- Provide future recommendations towards a better operationalization

Challenge 1: Lack of Demographic Annotations

- We query Google Scholar for pose estimation datasets
- Most datasets have not considered demographic labels during collection
- Fairness evaluation for pose estimation is limited due to the lack of annotations

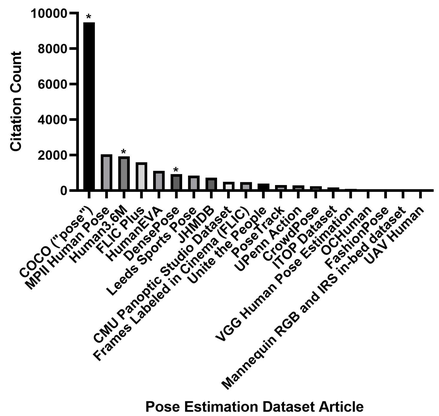


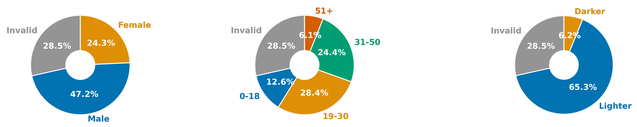
Figure 1: **Citation count** for articles introducing popular pose datasets, per Google Scholar as accessed Aug. 2, 2022. We reduce the COCO citation count by querying only retrieved results with the search term "pose", given that this dataset is utilized for multiple tasks. Asterisks (*) indicate datasets with limited demographic information. The utilization of COCO is shown to be widespread for pose estimation and additional visual tasks.

Takeaways and Recommendations

- The absence of demographic labels harms the operationalization of fairness evaluations:
  - Practitioners might not perform a fairness assessment
  - Practitioners might utilize inappropriate datasets (e.g., too small or not representative)
  - Only apparent labels could then be inferred from existing datasets
- We recommend to collect demographic annotations from the start

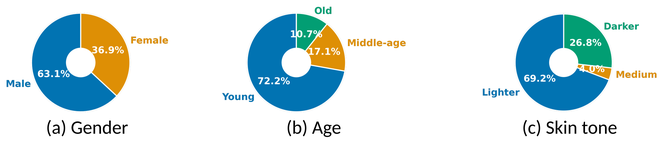
Challenge 2: Imbalanced Demographic Labels

- Following the MoveNet model card [1], we introduce **COCO-Keypoints-Demographics**
  - **657 images** with **demographic labels** derived from COCO2017-val [2] (we only keep the ones where demographic labels can be meaningfully inferred)
  - **Demographic labels**: *female* and *male* genders; *0-18*, *19-30*, *31-50* and *51+* age categories; and *lighter* and *darker* skin tones
  - Link to the annotations: https://github.com/SonyAI/coco_kd
- Annotations of demographic labels require a manual check to ensure their validity:
  - We started from the 919 images identified in the MoveNet model card and label semi-automatically every image for demographic labels and flag any inappropriate content
  - Automatic labeling has errors, which affects the fairness evaluation
- Annotating COCO2017-val reveals demographic imbalances:
  - Only 17 darker-skinned females out of 657 images (while lighter-skinned males represent 394)



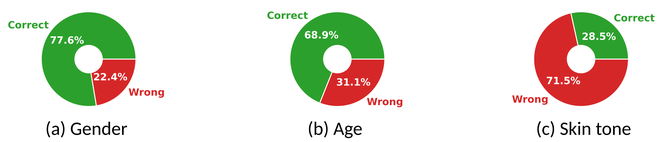
(a) Gender (b) Age (c) Skin tone

Figure 2: **COCO-KD demographic** distribution over 919 images. One third of the images are deemed to be invalid because demographic labels cannot be inferred. For the other two thirds, the distributions show an over-representation of males and light-skinned subjects, as well as an under-representation of older subjects.



(a) Gender (b) Age (c) Skin tone

Figure 3: **COCO MoveNet demographic** distribution over 919 images, taken from the original model card. While these distributions also show imbalanced demographics, the proportions differ from Figure 2, which highlights the importance of manual checks.



(a) Gender (b) Age (c) Skin tone

Figure 4: **COCO-KD annotation correctness** over 657 images after manually verifying the demographic annotations.

Takeaways and Recommendations

- The imbalanced demographic labels question the applicability of COCO for fairness evaluations and call for better evaluation datasets
- We recommend future datasets to avoid automatic annotations, and rely either on self-reported attributes or manual quality checks
- We recommend future datasets to incorporate from the start a balanced demographic representation with enough samples (especially for intersection sub-groups)

Challenge 3: Fairness Evaluation with Limited Samples

- We consider three pose estimation models: OpenPose [3], MoveNet Thunder [4] and PoseNet [5]
- We follow the pre-precessing done in MoveNet (i.e., cropping the subject according to the keypoint locations and make sure the center of the image is the middle point of the hip area)
- We report the mAP and mAR with object keypoint similarity

Overall results

Images	OpenPose		MoveNet		PoseNet	
	mAP	mAR	mAP	mAR	mAP	mAR
657	79.3	83.2	77.1	80.6	55.9	62.4

Table 1: **2D human pose estimation results** on COCO-KD. OpenPose achieves the highest mAP and mAR, slightly above MoveNet, while PoseNet is far below.

Disaggregated results

Gender	Images	OpenPose		MoveNet		PoseNet	
		mAP	mAR	mAP	mAR	mAP	mAR
Female	223	78.2	81.7	75.9	79.7	57.1	62.6
Male	434	79.9	83.9	77.9	81.1	55.3	62.3

(a) **Gender**. OpenPose and MoveNet achieve a lower performance for the female group than the male group; while it is the opposite for PoseNet.

Age	Images	OpenPose		MoveNet		PoseNet	
		mAP	mAR	mAP	mAR	mAP	mAR
[0, 18]	116	80.3	83.2	80.4	83.4	59.1	65.0
[19, 30]	261	80.4	84.0	77.4	80.6	51.5	58.0
[31, 50]	224	78.2	82.3	75.3	79.0	59.1	65.0
[51+]	56	78.2	82.9	79.6	81.4	61.4	66.8

(b) **Age**. While discrepencies exist among age groups, they differ with the selected models.

Skin tone	Images	OpenPose		MoveNet		PoseNet	
		mAP	mAR	mAP	mAR	mAP	mAR
Lighter-	600	79.3	83.2	77.1	80.6	56.8	63.0
Darker-	57	78.7	82.6	77.4	80.9	47.0	56.0

(c) **Skin tone**. Lighter skins achieve a high performance than darker skins, except for MoveNet where no difference is observed.

Table 2: **Breakdown by demographic labels** on COCO-KD. Models perform differently depending on the demographic sub-group. Such model bias could lead to potential discrimination.

- Gender and age affect the performance of all three models
- Performance seems to be on-par for skin tone
- When reducing the number of samples for lighter-skinned images to 57, we observe a high variance in the results

Takeaways and Recommendations

- The low sample size affects the validity of the results interpretation
- We recommend future evaluation to report multiple results on multiple splits
- We recommend future evaluation to consider datasets with enough samples for every sub-group to confirm the presence of a bias

References

[1] Francois Beletti, Yu-Hui Chen, Ard Oerlemans, and Ronny Votel. MoveNet Single Pose: Model Card. https://storage.googleapis.com/movenet/MoveNet.SinglePose%20Model%20Card.pdf, 2022. Accessed: 2022-08.

[2] 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 *ECCV*, 2014.

[3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *TPAMI*, 2019.

[4] Yu-Hui Chen, Ard Oerlemans, Francois Belletti, Andrew Bunner, and Vijay Sundaram. MoveNet Thunder. https://tfhub.dev/google/lite-model/movenet/singlepose/thunder/tflite/float16/4, 2022. Accessed: 2022-08.

[5] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *ECCV*, 2018.