A fairness evaluation of pose detectors for pedestrian attention

Maria Eduarda Mota
Centro de Informática
Universidade Federal de Pernambuco
Av. Jorn. Aníbal Fernandes, Recife, Brazil
mebm@cin.ufpe.br

Maria Luísa Lima
Centro de Informática
Universidade Federal de Pernambuco
Av. Jorn. Aníbal Fernandes, Recife, Brazil
mlll@cin.ufpe.br

I. Abstract

Pedestrian attention detection is an interesting and possibly lifesaving technique that can be used in autonomous cars to understand whether the pedestrian has noticed the car, and therefore, is less likely to put themselves in danger in front of it. It is important however, to understand the fairness limitations this technology may have, as it is known in the literature that many vision models carry biases from their training data. This work focuses on evaluating gender and age biases present on pose detectors used to extract poses for pedestrian attention detection.

II. Introduction

There has been interest in developing autonomous vehicles for many years, and with the rise of computer vision, many tasks being researched for autonomous driving rely on information extracted from images. One of the dangers of autonomous vehicles is the possibility of a crash that can result in an accident or a casualty. With this in mind, researchers train their models to be as accurate as possible. However, recent research has shown implicit biases in these models, making the detection of underprivileged groups worse and therefore endangering them more than privileged groups

One of the main computer vision tasks used in autonomous vehicles is pedestrian detection, which entails using images to identify any pedestrians inside the field of view of the vehicle. Extending pedestrian detection, an interesting task is pedestrian attention detection, which tries to understand if the pedestrian is aware of the car, which could indicate a lesser chance of an accident occurring. [1] proposes a pedestrian attention detection technique that uses a 2D pose estimation from the pedestrian to determine if they have or not paid attention to the vehicle.

In this work, we choose to test whether the detector of choice for [1], OpenPifPaf [2], presents bias against any gender or age group present in the JAAD [3] and PIE [4] datasets. We also compare with the pose detectors ViTPose [5] and HRNet.

III. Related Works

Object detection is an important computer vision task for Autonomous Driving Systems (ADS) when trying to avoid crashes and fatalities. One of the areas of object detection is pedestrian detection, which involves the detection pedestrians on the streets. It is already been documented in the literature that there is bias on pedestrian models, largely because of the data used in their training

Recent works like [6] focus on understanding the biases present in object and pedestrian detection models commonly used in ADS. They find a difference in performance between age groups, with children being consistently poorly detected. They also find that low brightness conditions influence the detectors performances, lowering their performance and enhancing their bias against females.

Some works, like [7], study fairness in pedestrian intention by trying to understand bias in their trajectory estimation and find that there is specially bias against yound and old age.

As the prediction of pedestrian intention can be a key safety cue for ADS, [1] brings a method that focuses on detecting pedestrian attention. While gaze estimation is an important task used mainly to gather the attention level a person is paying to something, but is not specifically focused on autonomous driving scenarios, where the person may be far away from the camera, the author [1] experiments on using eye contact of the pedestrian along with their body pose estimation for the attention estimation.

Pose estimation is expected to have bias according to the training data, much like other tasks in computer vision. As found by [8], there are differences in performance across age and gender groups, depending on the model being evaluated (and the data used in training it) and the data being used for evaluation.

IV. Proposed Approach

The Looking detector [1] works by taking a $2D$ representation of a skeleton $\mathbf{K} \in \mathcal{R}^{17\times 2}$, in which $\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_{17}$, each $\mathbf{k}_i$ represents body joint of the detected person. The skeletons are normalized and input into the network as described in Figure 2. This first pipeline is described as "Body Keypoints (joints)". Another detection method provided by [1] includes the cropped head of the detected person alongside the body keypoints as this is an information the authors thought could improve the detection of pedestrian in Autonomous Vehicles, in which case the detections could occur more far away from the camera recording the images.

Figure 1: Examples from the JAAD dataset [3]. For each video there is an annotation of perceived gender (Male and Female) and age (Child, Adult, Senior).

We propose (*i*) evaluating [1] across different demographics, age and gender, as these annotations are made available on the dataset [3], which brings videos of traffic scenarios, whose examples can be found in Figure 1. This allows for a study on the fairness of the pedestrian attention detection method proposed by [1]. We also propose (*ii*) an evaluation of the impact of different pose detectors on the fairness of the pedestrian attention detection, as the extracted poses are a crucial part of the detector, as shown in Figure 2. For *ii* we will be retraining the network and evaluating the model, calculating the overall performance as well as performing a statistical analysis using the Chi-Squared independence test [9] to try to identify any biases across demographic groups of gender and age.

A. Evaluation Metrics

For the detection we will use Average Precision (AP). For the fairness evaluation, we will perform statistical analysis using the Chi-Square test of independence to determine if there was a significant difference on the performance between groups.

V. Experiments

A. Pose Detection

The following detectors are going to be used to extract the pedestrian poses: OpenPifPaf [2], as the default detector also used by [1], ViTPose [5] and HRNet [10].

The extracted poses will be converted to a common format, the OpenPifPaf [2] format, for the insertion on the Looking [1] pipeline. The inferences of the looking pipeline will be collected and used for the subsequent fairness analysis, using the metrics described in Section IV-A. Detectors like MediaPipe [11], despite its wide usage in pose detection, were excluded since they are not suited for use in crowded or faraway scenes.

B. Implementation

The statistical tests were performed using the Scipy Python library [12].

The generation of the pose detections with OpenPifPaf was done with version $0.13.0$, and with HRNet and ViTPose was done using the MMCV [13] library version $2.1.0$ and MMPose [14], on PyTorch $1.10.0$ and the models available in the model zoo. The HRNet model used was the HRNet UDP COCO[1] ([10], [15], [16]) model and the ViTPose model used was the ViTPose Large COCO[2] model ([5], [17]).

VI. Results and Discussion

For the first phase of the experiments, we generated a baseline of the basic Looking results using just the body joints, aiming at trying to reproduce the results reported by the authors of [1]. We obtained an AP of 85.6 against the 85.9 reported by the authors.

A. Baseline Statistical Analysis

By taking the test dataset, we evaluated the model's baseline joints performance on the samples with categories Male and Female for gender and Child, Adult and Senior for age. The total number of evaluated frames from the test split of the JAAD dataset is 16998 and the number of evaluated frames split per demographic and its resulting correct or incorrect attention detections are detailed in Table I.

By taking a Chi-Square test of independence [9] we evaluate the dependence of the correctness of detection with relation to either gender or age. We provide the following hypotheses and its subsequent results, with a significance of 5%:

1) (H_1) **Gender - Male x Female**.

[1]HRNet-w48+UDP Model
[2]ViTPose-L Model

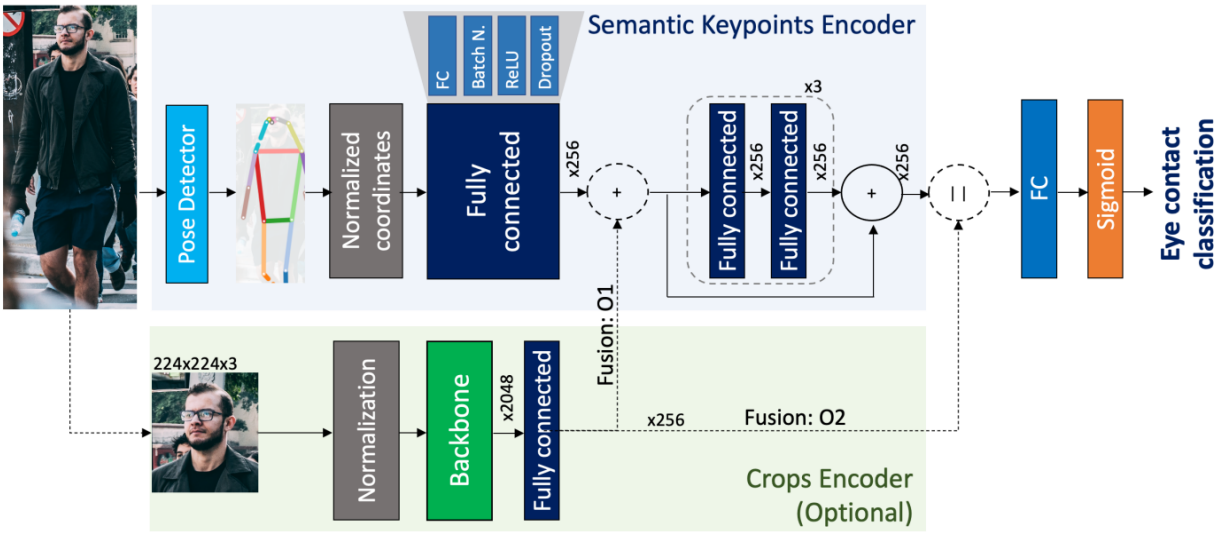

Figure 2: Full pipeline of the pedestrian attention detector [1].

Table I: JAAD dataset demographics for a total of 16998 frames with defined demographics.

	Demographic				
	Age			Gender	
Detections	Child	Adult	Senior	Female	Male
Correct	588	10505	2123	7612	5604
Incorrect	190	3082	510	2137	1645

We define the success detection rate for Male pedestrians as (P_M) and for Female pedestrians as (P_F). We then provide the following null and alternative hypotheses:

$$\mathbf{H_0} : P_M = P_F$$
$$\mathbf{H_a} : P_M \neq P_F$$

For this, the p-value of the test was 0.2383, not offering enough evidence for a null hypothesis rejection, indicating no significant difference in performance between perceived males and females.

2) (H_2) **Age - Child x Adult**.
We define the success detection rate for Child pedestrians as (P_C) and for Adult pedestrians as (P_A). We then provide the following null and alternative hypotheses:

$$\mathbf{H_0} : P_C = P_A$$
$$\mathbf{H_a} : P_C \neq P_A$$

For this, the p-value of the test was 0.2800, not offering enough evidence for a null hypothesis rejection, indicating no significant difference in performance between children and adults.

3) (H_3) **Age - Child x Senior**.
We define the success detection rate for Child pedestrians as (P_C) and for Senior pedestrians as (P_S). We then provide the following null and alternative hypotheses:

$$\mathbf{H_0} : P_C = P_S$$
$$\mathbf{H_a} : P_C \neq P_S$$

For this, the p-value of the test was 0.0025, offering enough evidence for the null hypothesis to be rejected. Hence, we detect statistically relevant disparity in model performance between child and senior pedestrians.

4) (H_4) **Age - Adult x Senior**.
We define the success detection rate for Adult pedestrians as (P_A) and for Senior pedestrians as (P_S). We then provide the following null and alternative hypotheses:

$$\mathbf{H_0} : P_A = P_S$$
$$\mathbf{H_a} : P_A \neq P_S$$

For this, the p-value of the test was 0.0002, offering enough evidence for the null hypothesis to be rejected. Hence, we detect statistically relevant disparity in model performance between adult and senior pedestrians.

From these tests we take the following conclusion: there is a different performance among the senior pedestrians. In addition to the above tests, which merely indicate a disparity in performance without indicating the direction in which the better performance occurs, we performed an odds ratio test to check in which direction the better performance leaned on. The test yielded the following results:

1) **Age - Child x Senior**. Odds of 0.7435, indicating that the odds of being correctly detected for pedestrians belonging to the child category is 0.7435 times the odds of pedestrians belonging to the senior category.
2) **Age - Adult x Senior**. Odds of 0.8188, indicating that the odds of being correctly detected for pedestrians belonging to the adult category is 0.8188 times the odds of pedestrians belonging to the senior category.

These results indicate that the model would perform better on senior citizens, which seems to contradict some common assumptions. As discussed in [7], the data capture for this type of dataset heavily favors adults over children and seniors, and is expected to yield a better performance on adults. However, our work goes against this expectation, as with [7], who also observed a better performance on the elderly. Table 1 in [7] exposes the demographic breakdown by age and Table 2 by

gender, for the whole JAAD dataset and the numbers indicate to limited representativity, specially with the data coming from Ukraine, Canada, Germany and United States. All of this supports the need for testing on other datasets as well, such as PIE [4] and LOOK [1], which is composed of KITTI [18], nuScenes [19] and JRDB [20].

B. Data Augmentation Experiments

Figure 3: Example of an image modification. The first image shows the original version, the second displays the darkened version, and the final image presents the darkened version with OpenPifPaf applied.

Following the experiments described in Section VI-A, we focus on trying to increase the variety of JAAD dataset. We conducted augmentation experiments on the original images. Variations such as images distortions, rotations, brightness changes and cropping were considered and tested. However, as we used the keypoints annotations for comparison with those generated by OpenPifPaf, we opted to maintain the original position of the image. By that, in this project, we focused on reducing the brightness and contrast of the images to simulate nighttime conditions. Figure 3 shows an example of this image modification.

Additionally, it is important to mention that these new augmented images have been added to the existing dataset, maintaining the original version.

An augmentation pipeline was created to automate this process. The main steps of this new image processing pipeline include:

- Relating images attributes annotations in a dictionary, in order to maintain a structured dataset;
- Randomly selecting the images to be augmented;
- Processing the image augmentation and updating the annotations by adding the new data paths;
- Passing the augmented images through OpenPifpaf to generate new pose detections.

One necessary adjustment was made to adapt the problem context in the selection of the best bounding box correspondences, to be used in the train, validation and test datasets. We lowered the Intersection over Union (IoU) threshold from 0.3 to 0.1, allowing more data to be incorporated into the augmentation pipeline. However, comparing to the Looking [1] pipeline, the core structure was largely preserved. This approach was essential for a clearer evaluation of the impacts of the changes made using the data augmentation.

Regarding data volume, the initial idea was to incorporate image augmentation of, at least, 20% of the original dataset, to ensure a sufficiently diverse augmented data in the training set. However, due to computational limitations, it was not possible to execute this. To ensure to have at least a minimal application, only a small portion of the data was processed and passed through the IoU threshold. Unfortunately, it was less than 1% of the entire dataset, which strongly compromised our analysis.

Table II shows the results obtained in a reduced version of the JAAD dataset (about 70,000 images) with and without the augmentation set. This reduction was done due to computational limitations of the authors. On it, it is possible to observe that without augmentation, H_1 shows that there is no relevant statistical difference in the correct detection between male and female pedestrians, however the augmentation seems to have induced a statistically relevant gender bias towards males. Another relevant discovery is that H_3 shows a statistically relevant bias in the no augmentation dataset, which is corrected in the model trained with the augmentation data. It is also noteworthy that H_4 showed a reduction in bias, even if it was not totally corrected.

These findings show the importance data augmentation can have on the outcome of a model, specially considering inducing and correcting biases. The augmentation process should be done with the utmost care, considering the demographic distribution of the original data, along with other possible bias inducers along the method's pipeline. Besides that, we find the model trained with the augmented images had an AP of 76.7, against 85.3 of the model trained without the augmented images. This shows that the augmentation experiments were ineffective in their attempt at improving the AP value. It is

possible that this decrease stems from the induced gender bias, but more experiments need to be made to further investigate this.

C. Pose detection Comparison

Table III shows that OpenPifPaf produces poses that yield better pedestrian attention detection than HRNet and ViTPose. This result is not surprising when comparing OpenPifPaf with HRNet, an older model, but the ViTPose results are suprising, for it being a newer Vision Transformer Model. Further investigation is needed to find if there are better conditions for its use, but one of the limitations could be necessity of use of a pedestrian detector in the ViTPose and HRNet detection pipelines. This is case as they are both top-down models, which indicates that they work by first having the person's location and then detecting the body's keypoints to form the skeleton. In this case, as OpenPifPaf is a bottom-up method [2], this additional step, which is also susceptible to failure and biases, is not present. Overall, we find that the pose detection model can influence heavily on the AP of the pedestrian attention detector, as evidenced by the difference between OpenPifPaf's result and ViTPose's.

D. Limitations

Due to computational limitations, we used a subset of the original data and this is a limitation we intend to correct in the future, repeating the experiments with the full datasets. Our time constraints also limited us in the amount of experiments and analysis we could perform, for example, with the number of pose detectors and the fairness analysis of the models. This constraint also limited our ability to test more thoroughly the augmentation process.

Another limitation in this work is that the techniques for body pose estimation that were used to compare with OpenPifPaf are top down, which implicates a need for a person detector to find the pedestrian location first, so the body pose estimation can be made. As previouly explained in this work on Section III, pedestrian detectors can induce biases themselves and could be influencing on the results.

VII. Conclusion

We find that the Looking pedestrian attention detector does show difference in performance across age groups, largely influenced by the training data. We also find that data plays a large role in how these biases present themselves, as shown by the augmentation experiments, where a gender bias was introduced into the model. We did not make any definite conclusions about the pose detectors themselves given the limitations listed previously and more experiments are needed. However, we can state that the pose detectors can have a large influence on the overall performance.

A. Future Works

To improve this work, we intend to do the following:

- Re-do the experiments on the whole datasets.
- Test more pose detectors, such as MotionBERT [21], AlphaPose [22], OpenPose [23], and Lite-HRNet [24].
- Evaluate a fairness performance comparison between top-down and bottom-up pose detectors.
- Evaluate the impact of pedestrian detectors on the pose detectors, by either including them as a variable or by also using bottom-up methods, which do not need a pedestrian detector to indicate the person's position first.
- Test the impact of the CNN used on the head and head and joints pipelines, comparing performance and fairness with others.
- Execute the fairness statistical analysis on all experiments and use other established fairness metrics to assess the bias more thoroughly.
- Improve the augmentation experimentation, targeting intended demographic groups to understand better the impact of this increased distribution on the training.

References

[1] Y. Belkada, L. Bertoni, R. Caristan, T. Mordan, and A. Alahi, "Do pedestrians pay attention? eye contact detection in the wild," *arXiv preprint arXiv:2112.04212*, 2021.

[2] S. Kreiss, L. Bertoni, and A. Alahi, "OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association," *arXiv preprint arXiv:2103.02440*, Mar. 2021.

[3] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Joint attention in autonomous driving (jaad)," *arXiv preprint arXiv:1609.04741*, 2016.

[4] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6262–6271.

[5] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38 571–38 584, 2022.

[6] X. Li, Z. Chen, J. M. Zhang, F. Sarro, Y. Zhang, and X. Liu, "Bias behind the wheel: Fairness testing of autonomous driving systems," *ACM Transactions on Software Engineering and Methodology*, 2024.

[7] A. Bae and S. Xu, "Discovering and understanding algorithmic biases in autonomous pedestrian trajectory predictions," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 2022, pp. 1155–1161.

[8] J. LaChance, W. Thong, S. Nagpal, and A. Xiang, "A case study in fairness evaluation: Current limitations and challenges for human pose estimation," in *Association for the Advancement of Artificial Intelligence 2023 Workshop on Representation Learning for Responsible Humancentric AI (R2HCAI), Washington, DC*, 2023.

[9] M. L. McHugh, "The chi-square test of independence," *Biochemia medica*, vol. 23, no. 2, pp. 143–149, 2013.

Table II: Statistical analysis of the correct and incorrect predictions on the subset of the JAAD dataset per demographic group according to the hypothesis defined in VI-A.

	Method			
	Augmentation		**No Augmentation**	
Group Comparison	**P-value**	**Odds Ratio**	**P-value**	**Odds Ratio**
(H_1) Gender - Male x Female	1.8003e-6	1.4714	0.0124	1.2562
(H_2) Age - Child x Adult	0.1101	0.7720	0.9872	1.017
(H_3) Age - Child x Senior	0.1960	1.3035	2.5702e-10	3.5405
(H_4) Age - Adult x Senior	5.1981e-5	1.6891	3.5e-24	3.4852

Table III: AP results of the results on pedestrian attention detection using different pose detection models on subset of the PIE dataset.

Pose Detector	**AP**
OpenPifPaf	88.2
HRNet	81.1
VitPose	80.0

[10] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[11] C. Lugaresi, J. Tang, H. Nash, *et al.*, *Mediapipe: A framework for building perception pipelines*, 2019. arXiv: 1906.08172 [cs.DC]. [Online]. Available: https://arxiv.org/abs/1906.08172.

[12] P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.

[13] M. Contributors, *MMCV: OpenMMLab computer vision foundation*, https://github.com/open-mmlab/mmcv, 2018.

[14] M. Contributors, *Openmmlab pose estimation toolbox and benchmark*, https://github.com/open-mmlab/mmpose, 2020.

[15] J. Huang, Z. Zhu, F. Guo, and G. Huang, "The devil is in the details: Delving into unbiased data processing for human pose estimation," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020.

[16] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *European conference on computer vision*, Springer, 2014, pp. 740–755.

[17] S. Jin, L. Xu, J. Xu, *et al.*, "Whole-body human pose estimation in the wild," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[18] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[19] H. Caesar, V. Bankiti, A. H. Lang, *et al.*, "Nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[20] R. Martin-Martin, M. Patel, H. Rezatofighi, *et al.*, "Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 6, pp. 6748–6765, 2021.

[21] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, "Motionbert: A unified perspective on learning human motion representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 085–15 099.

[22] H.-S. Fang, J. Li, H. Tang, *et al.*, "Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7157–7173, 2022.

[23] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *CoRR*, vol. abs/1812.08008, 2018. arXiv: 1812.08008. [Online]. Available: http://arxiv.org/abs/1812.08008.

[24] C. Yu, B. Xiao, C. Gao, *et al.*, "Lite-hrnet: A lightweight high-resolution network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 440–10 450.