Accuracy and Fairness in Pupil Detection Algorithm

Omkar N. Kulkarni[1], Vikram Patil[1], Vivek K. Singh[2] and Pradeep K. Atrey[1]
[1]Albany Lab for Privacy and Security, College of Engineering and Applied Sciences
University at Albany, State University of New York, Albany, NY, USA
[2] Rutgers University, USA
Email: {onkulkarni, vpatil, patrey}@albany.edu, v.singh@rutgers.edu

2021 IEEE Seventh International Conference on Multimedia Big Data (BigMM) | 978-1-6654-3414-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigMM52142.2021.00011

***Abstract*—Despite being highly accurate, widely used multimedia analysis algorithms can suffer from bias, affecting users' trust in them. For instance, algorithms for face recognition, pedestrian detection, and image search have recently been reported to be biased. In this paper, we move the discussion on algorithmic fairness to a new domain, namely, pupil detection. We audit a widely used OpenCV algorithm for pupil detection from the perspective of fairness and accuracy. The algorithm is audited using two different datasets: a single-person image dataset (CelebA), and a group image dataset (Images of Groups). In both datasets, we found the OpenCV pupil detection algorithm to provide reasonably high accuracy but also yield statistically significant bias with respect to gender. The results provide the first empirical evidence for the existence of gender bias in pupil detection algorithms in both single person as well as group image settings. These results could have a deleterious impact on downstream multimedia applications such as identity authentication, attention assessment, and e-learning. Finally, we discuss the process of appropriately choosing the pupil detection algorithm parameters that may help reduce bias while maintaining high accuracy in various multimedia applications.**

***Keywords*-Algorithms, Accuracy, Algorithmic Fairness, Face Detection, Gaze Tracking, Computer Vision, Pupil Detection, Machine Learning, Statistical Analysis.**

I. Introduction

In the last decade, we have seen a dramatic increase in multimedia-based applications and services. Applications like facial recognition systems, smart surveillance systems, self-driving cars, and numerous other services that we use every day have benefited from the recent advances in machine learning technologies. Several business processes are being automated [1] [2] and such automation has resulted in a reduction in cost and processing time and has improved quality.

The primary focus of improvements in these algorithms has been to improve their performance, run time complexity, accuracy, etc. However, with increased adoption, the issue of algorithmic bias has come to the forefront [3]. Algorithmic bias refers to the notion that a given algorithm behaves differently for different demographic groups. (Conversely, an algorithm is considered to be 'fair' when there is no difference in performance across different demographic groups.) These different groups can be based on different social descriptors such as gender, ethnicity, age, or their combination. For instance, Buolamwini et al., have reported that face detection algorithms work poorly for women and people of color compared to men and light-skinned people [4] and Wilson et al., have found pedestrian detection algorithms to perform differently for different skin tones [5]. Such biases, if left unreported or unmitigated, would have deleterious downstream effects on society.

The bias in an algorithm can originate from several factors, including but not limited to an unbalanced training dataset and the use of features implicitly connected with demographic descriptors (e.g., zip code) [6]. Because of the bias in algorithms, the outcome is often favorable for one group (privileged group) as compared to others (unprivileged group). This results in the denial of equal opportunity to the unprivileged group and can also lead to discrimination and prejudicial treatment to certain parts of society without their fault.

Algorithmic bias can reduce the users' trust in technology and reduce the algorithm's acceptability even if it is highly accurate [7], [8]. From the perspective of technological organizations, algorithmic bias is a crucial problem since it limits their user base and reduces the utility of their products and services. Thus from various such perspectives like discrimination, loss of credibility, trust, and justice, determining the algorithmic bias and finding ways to correct it is of utmost importance. Ideally, every algorithm should be checked for bias, and users should be educated about the bias, if there is any. Furthermore, steps should be taken to make the outcome of these algorithms fairer. Hence, auditing state of the art or widely used algorithms is of critical importance. In this paper, we have audited the OpenCV [9] implementation of the pupil detection algorithm for gender bias on two independent datasets. We also make an observation on parameter selection (rather than going with 'defaults' or selection solely based only on accuracy) to ensure that human-facing multimedia algorithms prioritize *both* fairness and accuracy.

The OpenCV [9] library is widely used by researchers and professionals in the industry. It contains implementations of several well-known algorithms in the computer vision and multimedia domain, e.g., face recognition, pupil detection, etc. In this paper, we focus on the pupil detection algorithm since its OpenCV implementation is widely used and is also an essential component of several advanced algorithms such as attention analysis, person identification, and gaze detection [10]. The gaze of a person is the direction in which the person is looking. In real world, gaze detection is an essential part of any nonverbal communication. It helps in understanding

978-1-6654-3414-0/21/$31.00 ©2021 IEEE
DOI 10.1109/BigMM52142.2021.00011

the mood, attention, attitude and helps in conveying emotions. These features can be further utilized in combination with other information to train the machine learning algorithms to develop various applications. One of the basic steps in gaze detection algorithms is pupil detection. Some of the popular downstream applications of such algorithms include unlocking smartphones, gaming, virtual and augmented reality-based applications.

Hence, gender bias (e.g., unequal performance for different groups based on gender) in pupil detection can significantly advantage (or disadvantage) certain sections of the society. In this paper, we evaluate the algorithmic bias or fairness based on gender. We consider the users to belong to two groups: privileged and unprivileged. Based on significant sociological as well as emerging fairness in AI literature, we consider males to be the privileged group [4], [11]. Then, based on statistical tests, we audit whether the algorithm performs better for the privileged group. We test this hypothesis on two different datasets, one with group images and the other with single images. A group photo entails more than one subject in the photo, whereas a single image represents a photo with only one subject. In this paper, we also make a note on parameter selection criteria for the algorithm to make it highly accurate and less biased.

The key contributions of this paper are:

1) Audit the OpenCV pupil detection algorithm for gender bias.
2) Provide empirical results for both single and group image settings.
3) Discuss the trade-off between accuracy and fairness based on parameter selection in the pupil detection algorithm.

We use the 'Images of Groups' and 'CelebA' datasets to perform our experiments [12], [13]. The goal was to audit gender bias in them and try to come up with a way of parameter tuning in order to make the OpenCV implementation of the pupil detection algorithm fair and accurate. The results provide the first empirical evidence for the existence of gender bias in pupil detection algorithms in both single person and group image settings and also a potential pathway for balancing fairness and accuracy via parameter tuning.

The rest of this paper's organization is as follows: In Section II we describe the related literature. Section III describes our implementation of the OpenCV algorithm for pupil detection along with definitions and assumptions. In Section IV, we explain the datasets used in experiments and discuss their results. Section V discusses the social implications of gender bias. Finally, in Section VI, we conclude the paper along with potential future work.

II. Related Work

A. Pupil Detection

Recently, gaze detection in an image has become a widely popular research field. In [14] Kar and Corcoran presented an extensive review of eye gaze detection techniques and compared the performances of various methodologies based on the use case. The paper summarized various methods of gaze tracking, namely, 2D regression, 3D modeling, shape-based and appearance-based. Every approach comprises of varying hardware setup and supports a variety of use cases. In [15], Yonetani et al. proposed a novel method of eye gaze probing to detect an object that is focused by the user. In [16], Okamoto et al. provide an application of gaze detection for the classification of pedestrian behavior using the data of pedestrians looking into the direction boards in a shopping mall. Similarly, there are multiple studies that focus on gaze detection and gaze following using the detected gaze and its applications [17], [18].

There has also been some research in gaze detection and gaze following in group images. In [19], Kodama et al. discussed a target localization technique based on gaze direction in group images. They used aggregation of each person's gaze direction and deduced the position of the target being gazed at by the group. However, the paper only achieves higher accuracy when the number of subjects in the group image is large. For an image with fever than five faces, the Mean Absolute Error (M.A.E) of target localization is higher. Similarly, in [20], Recasens et al. proposed an approach for gaze-following in a group image. They used a deep neural network based approach and worked on various head and gaze orientations. Some researchers have also used 3D modeling in order to detect the z position along with the traditional x, y positions of gaze on a 3D plane [21].

Pupil detection applications also remain in head-mounted pupil detection devices and static or remote image processing. In [22], Fuhl et al. have summarized eight pupil detection algorithms and tested them over three datasets with varying illumination, occlusion, head, and camera position. Haar cascading was used for detecting the face and eye region. They concluded from the extensive evaluations that the algorithms made for pupil detection on remote images are not better than the ones for head-mounted pupil detection. In [23], Fuhl et al. proposed a novel algorithm for robust pupil detection in real-world images incorporating edge filtering and Angular Integral Projection Function-based histogram calculations. They tested their method on thousands of images from the real world over multiple datasets and achieved very high detection rates compared to two state-of-the-art algorithms. Another approach was introduced in [24] by Yang et al. that used ultrasound images for pupil detection. In [25], George and Routray proposed a novel approach in gaze localization in low-resolution images. The multi-phased approach made use of the eye's geometrical characteristics and was tested on multiple datasets.

Another popular approach for pupil detection is by using the OpenCV library. This method has been widely used by researchers in computer vision applications, especially for the creation of more advanced pupil detection algorithms downstream. In [26], Luo et al. made use of a fast contour based approach using the binary pupil image to get the center of the pupil and calculate its diameter. Similarly, in [27], Bonteanu et al. used adaptive thresholding methodology on

TABLE I: Comparison of proposed method with the related works

Work	Aspect of Bias	Application
Alasadi et al. [35]	Gender	Cyberbullying Detectors
Bolukbasi et al. [36]	Gender	Word Embedding
Singh et al. [37]	Gender	Digital Media
Dwork et al. [38]	Membership in group	Classification
Caliskan et al. [39]	Gender, Race	Language Processing
Buolamwini and Gebru [4]	Gender, Race	Facial Analysis
Proposed Method	**Gender**	**Pupil Detection**

top of the ellipse-fitting on a binary image using contour to get the pupil center and achieved decent results for variable lighting conditions. In [28], Pimplaskar et al. incorporated the centroid method and connected component technique to track the eye position real time. This method was also used to check for blinks in real time.

There are numerous research efforts in literature that make use of OpenCV platform to perform pupil detection [29] [30] [31] [32] [33]. OpenCV remains one of the most popular libraries for computer vision and multimedia tasks. For instance, a book on using the OpenCV library has over 8,000 citations as of October 2021 [34]. Hence, understanding the working of algorithms in OpenCV, especially their fairness impact is important for the research community.

B. Bias in Algorithms

Bias or fairness in machine learning algorithms has been a popular research topic in the recent past. Many researchers have condemned the machine learning algorithms for biased results based on gender and ethnicity and proposed methods to overcome it. In [4], Buolamwini and Gebru discussed the bias in facial analysis algorithms based on gender and skin type. The paper audited bias in detecting darker males, darker females, lighter males, and lighter females with varying error rates. The bias in an algorithm can also surface due to the fact that the trained data was biased towards a particular group. In [39], Caliskan et al. explained the bias in ordinary language on the basis of gender and race, which, when used as training data for an application, can result in an unfair algorithm. They used GloVe word embedding model trained over a sample of text from the web. Similarly, in [36], Bolukbasi et al. discussed the male and female stereotypes in word embedding even when trained on well-known Google News Articles data. They evaluated this bias based on occupational words and by producing analogies that humans use to derive a gender-based stereotype. In [38], Dwork et al. discussed fairness in the classification algorithms for admission of students in a university, forbidding discrimination based on their memberships in groups.

In [37], Singh et al. discussed the presence of gender bias in image search results coming from different digital media platforms. Alasadi et al. studied the problem of bias in multimodal cyberbullying detectors and proposed a fairness-aware fusion framework for the same [35]. Almuzaini et al., [11] discussed gender-based bias in multiple sentiment detection algorithms and proposed an approach to reduce such bias. In another effort Alasadi et al., discuss fairness in face matching algorithms and propose an adversarial network based approach for countering it [40].

There have been some other studies about gender bias and gaze in various other fields as well. For instance, in [41], Bayliss et al. explained the gender bias in infants with autism. They evaluated the 'extreme male brain' theory of autism, which states that, in the general population, males should display more autism-like traits than females. Similarly, [42], and [43], reported that the difference in gender affects the user's perception of emotion and their preferences psychologically. Although these studies study the interplay between gender and gaze, our work is focused on gender bias in the gaze detection algorithms as applied to digital images.

As can be seen from the above discussion, both pupil detection and bias detection in algorithms have been studied extensively in the literature. However, there has been no previous work at their intersection. Table I summarizes the bias detection related work from literature. Although there are various studies in the literature on detecting the bias and fairness of algorithms, to the best of our knowledge, this is the first pursuit to analyze gender bias in the pupil detection algorithms.

III. Proposed Work

Our work uses a well-known pupil detection algorithm and a bias evaluation methodology for that algorithm for two groups: male and non-male. We have used the Dlib-ml [44] library for face detection and eye isolation. OpenCV [45] was used for binary masking and detection of pupil center.

A. Definitions

In this subsection, we provide definitions as follows:

Definition 1: (Accuracy (α)) Accuracy is defined as the ratio of the number of faces with correctly detected pupil center over the total number of faces.

$$\alpha = \frac{n}{N} \tag{1}$$

Where N is the total number of faces and n is the total number of faces with correct pupil center prediction such that, $n \leq N$. The detection of pupil center is considered correct when the detected and the ground truth center coordinates of the pupil are within a pre-defined threshold τ.

Definition 2: (Bias) Bias is defined as the difference in accuracies between the privileged and unprivileged group.

$$Bias = \alpha_{priv} - \alpha_{unpriv} \tag{2}$$

α_{prev} and α_{unprev} denote the accuracies for privileged (male) and unprivileged (non-male) group respectively.

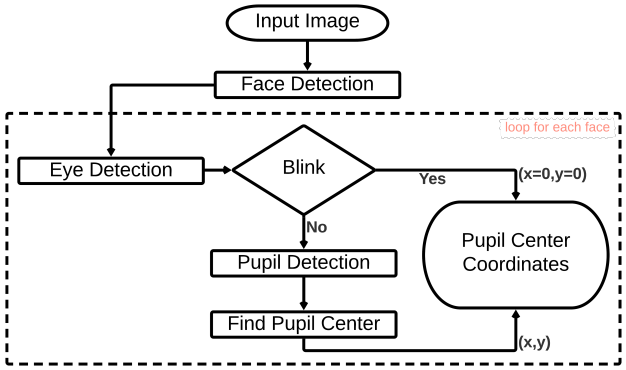
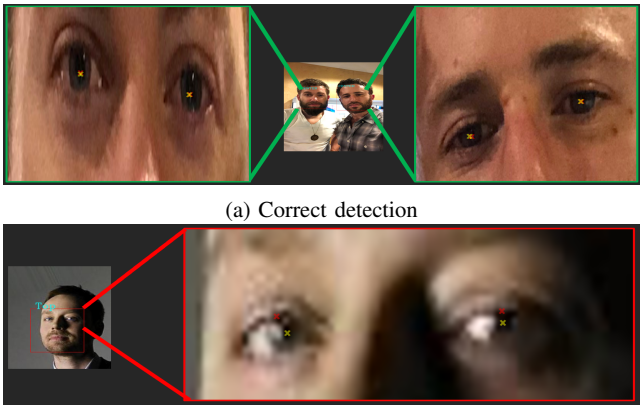


Fig. 1: Workflow of the proposed method



(a) Correct detection

(b) Incorrect detection

Fig. 2: Pupil center coordinates detection example

B. Assumptions

Similar to our past work in this area [10], we assume the following points in the paper:

- A1 The state-of-the-art face detection and eye isolation algorithms used work seamlessly.
- A2 Photo is taken in either group setting, i.e., more than one face exists in it, or is a single photo, i.e, only one face exists in it.
- A3 Each face is without any disrupting accessories like sunglasses or a hat.
- A4 Both eyes are visible for each subject in the photo.

C. Method

Figure 1 depicts the workflow of the OpenCV implementation of the pupil detection algorithm. Specifically, it uses the following steps to find the center of the pupil:

Step 1. Apply face detection algorithm on the input photo I, and get N detected faces, N $\geq$ 1.

Step 2. For each face $n \in N$, apply an eye isolation algorithm and isolate the eye area.

Step 3. Apply blink detection on the eye area to detect if there is a blink.

Step 4. If Yes, then finalize $x = 0$ and $y = 0$ as the coordinates for the pupil center of the face and move onto the next face (Step 2). Else, move to Step 5.

Step 5. Pupil detection mechanism is applied to the eye region to find the pupil area.

Step 6. Find the coordinates of the center of the pupil area.

Output: The center coordinates for pupil for all the n faces detected in Step 1.

The proposed method begins by applying a face detection algorithm to the given input image. For each face detected, an eye detection algorithm is applied, which provides an eye frame to be passed onto the blink detector. Blink detector makes use of Eye-Aspect-Ratio (EAR), as discussed by Soukupova and Cech in [46]. A blink is detected if and only if the EAR drops below a specific threshold value. If a blink is detected, the location of the center of the pupil is set to zero for that face, and the algorithm moves onto the next face in the input image. If not, then the eye frame is passed on to the pupil center detection module. Binary masking coupled with adaptive thresholding was used in order to detect the pupil for the eye. The center of the pupil was detected by passing the image through an erosion-dilation-blurring mechanism. Along these lines, the pupil center coordinates are recorded for each face in the input image. The coordinates for the pupil center consider a threshold (τ) with respect to the size of the face (γ). This threshold was calculated as follows:

$$\tau = 0.0001 \times area(\gamma) \tag{3}$$

where $area$ is calculated using following formula:

$$area = w \times h \tag{4}$$

where w is the width of face area and h is the height of face area.

This was done in order to take the face size into consideration when working with the group images in one of our datasets. The images in our datasets were from the real world and hence had a varying size of faces depending on the number of subjects and the distance of the subjects from the camera.

Figure 2 depicts the eye area regions of examples with correct and incorrect detection of pupil center. In the figure, the yellow marker represents the ground truth coordinates of the pupil center, and the red marker represents the coordinate points detected by the program. As can be seen in Figure 2a, the yellow and the red marker almost overlap, i.e., the detected pupil center coordinates and the ground truth pupil center coordinates have a difference value that lies within the

TABLE II: Summary of results for Images of Groups dataset

	Male	**Non-Male**	**Total**
Number of faces (N)	284	338	622
Correct detection (n)	254	280	534
Incorrect detection	30	58	88
Accuracy (α) **%**	89.43	82.84	85.85

threshold τ. This exhibits a correct detection of the pupil center. Similarly, for the second face in the figure, the red marker and yellow marker overlap making the difference between them less than the threshold τ. Whereas, in Figure 2b, the red marker and the yellow marker do not coincide with each other. The difference in their positions is more than the threshold τ. This exhibits incorrect detection.

IV. Experiments and Results

In this section, we first describe the datasets, then provide experimental results for bias audit in the pupil detection algorithm using statistical analysis.

A. Dataset

1) Images of Groups: For our experiments, we used a subset of the 'Images of Groups' dataset proposed by Gallagher and Chen [12]. The dataset consists of group images from Flickr with multiple faces or subjects in the image. The face count ranges from 2 to 10 over a total of 200 group images. The images have varying combinations of the number of male and non-male faces. The dataset had a total of 606 faces, out of which 280 were male, and 326 were non-male faces. Table II summarize the statistics about the dataset. All the faces in each group image were labeled with gender and pupil center coordinate by manual observation as the ground truth as described previously in [10].

2) CelebA: We have used a subset of the popular CelebA dataset proposed by Liu et al. [13]. Our dataset consists of a total of 67,234 face images of popular celebrities, with each having more than 40 attributes. Out of which, 28,020 were male, whereas 39,214 were non-male. Table III summarize the statistics about the dataset. The gender attribute for the people (celebrities) was provided as part of the dataset, and the pupil center coordinate was annotated manually for ground truth.

We note the limitations with gender inference based on visual markers and that they may vary from the individuals' disclosed gender [37]. A face in the dataset had two possible gender values: male and non-male.

B. Bias Auditing

We approached to test the OpenCV pupil detection algorithm's accuracy by comparing the pupil center coordinates of a person given by the algorithm with the "ground truth" manual observation, and accuracy was calculated using Eq. 1. The accuracy was calculated for two groups: male and non-male. In the Images of Groups dataset, we observed that the accuracy for correctly detecting the pupil center for males was 89.43%, whereas that for non-males was 82.84%. Overall, an accuracy of 85.85% was obtained. Similarly, for the CelebA dataset, the accuracy for males was 92.59%, whereas that for non-males was 90.45%. The overall accuracy of 91.34% was obtained. Hence, there was a difference in the performance of the pupil detection for the male and non-male groups in both the datasets.

TABLE III: Summary of results for CelebA dataset

	Male	**Non-Male**	**Total**
Number of faces (N)	28020	39214	67234
Correct detection (n)	25944	35469	61413
Incorrect detection	2076	3745	5821
Accuracy (α) **%**	92.59	90.45	91.34

In order to determine the statistical significance of the abovementioned performance difference in the pupil detection algorithm, we used bootstrapping method and performed the Welsh's t-test. Bootstrapping is a metric used in resampling methods by incorporating random samplings. It assigns various accuracy measures like confidence intervals, variance, etc., to the sample estimates. This technique allows estimation of the sampling distribution of almost any statistic using random sampling methods [47]. We used bootstrapping to randomly select samples to calculate the mean and variance of accuracies for both groups (privileged and unprivileged) on both datasets. For each group (male and non-male), we took 200 samples from the Images of Groups dataset randomly 100 times, and similarly, we took 20000 samples from the CelebA dataset randomly for 1000 times. The mean ($\overline{X1}$ and $\overline{X2}$ for male and non-male group respectively) and standard deviation (σ_1 and σ_2 for male and non-male group respectively) were calculated. With the Images of Groups dataset, the mean accuracy for male faces over a sample size of N=100 was 89.56%, and that for the non-male group was 81.27% with the same sample size. The variance for the male group was calculated to be 0.08%, whereas for non-male was 0.07%. With the CelebA dataset, the mean accuracy for male faces over a sample size of N=1000 was 92.60%, and that for the non-male group was 90.54% with the same sample size. The variance for the male group was calculated to be 0.05%, whereas for non-male was 0.09%. The delta (Δ) between these two groups was used for auditing of bias.

The Δmean = 8.29% for Images of Groups dataset and the Δmean = 2.06% for CelebA dataset here shows that the pupil detection algorithm works better for the privileged male group as opposed to the unprivileged non-male group. Next, we used an independent, two sample Welch's t-testing approach for equal sample sizes and unequal variances. The results are summarized in Tables IV and V.

The null and alternate hypothesis is defined as follows:
Null Hypothesis [H_a]: There is no significant difference between the accuracies of the pupil detection algorithm for male and non-male groups.
Alternate Hypothesis [H_a]: There is a significant difference between the accuracies of the pupil detection algorithm for male and non-male groups.

As observed from Table IV, when tested on the Images of Groups dataset, the two-tailed p-value for $t = 740.45$ and *degree of freedom* = 197 was $p < 0.0001$. Therefore, the null hypothesis (H_o) was rejected and alternate hypothesis (H_a), was accepted. Similarly, from Table V, when tested on the CelebA dataset, the two-tailed p-value for $t = 632.7240$ and *degree of freedom* = 1562 was $p < 0.0001$. Therefore, the

TABLE IV: Performance comparison for privileged (male) and unprivileged group (non-male) for Groups of Images dataset

Test Parameters	Male	Non-Male
Mean ($\overline{Xi}$) **(%)**	89.56	81.27
Variance (σ_i) **(%)**	0.08	0.07
Sample size	100	
p**-value**	<0.0001	
Result	H_0 Rejected and H_1 Accepted	

null hypothesis (H_o) was rejected, and the alternate hypothesis (H_a) was accepted as well. This means that the algorithm appears to favor the privileged group (males) over the unprivileged group (non-males) in a statistically significant manner, and the results hold true for single versus group images as tested on two different independent datasets.

C. Parameter Selection and Tradeoff

An integral part of the OpenCV implementation of pupil detection algorithm is the *adaptive_threshold()* function. According to the documentation of OpenCV, this function takes six parameters as input: source image, the maximum non-zero value assigned to the pixels for which the condition is satisfied, the adaptive method to be used for thresholding, threshold type, block size b, i.e., the size of neighborhood pixels in order to calculate the threshold, and a constant c which is subtracted from the mean or weighted mean calculated [9]. Out of these parameters, the only tunable ones that typically affect the algorithm's accuracy are b and c.

In one of the cases, while testing with the Images of Groups dataset, the accuracy of around 85% was obtained with $b = 115$ and $c = 1.8$, chosen at random in the implementation of the OpenCV for adaptive thresholding function. As the results of both datasets were eventually aimed to be compared with each other, we decided to keep these as the base values for evaluating the accuracy for both datasets.

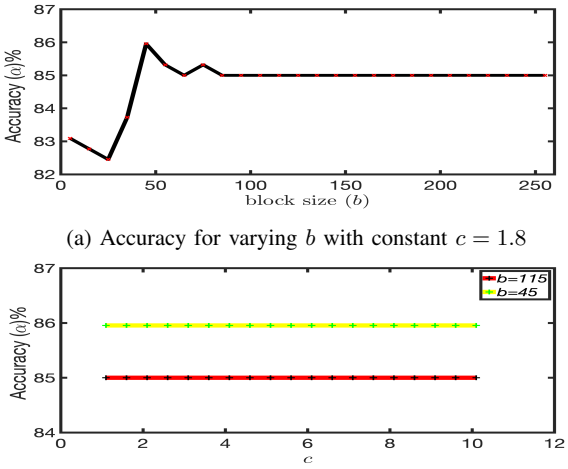
We decided to evaluate various combinations of these two parameters with an aim to find a combination of values that may yield an accurate and fair algorithm on group images from the Images of Groups dataset. This was achieved empirically. Figures 3 and 4 represent the results obtained in this process.

As can be observed in Figure 3a, the pupil detection algorithm's accuracy on the group images from the Images of Groups dataset varies with different values of b, for a constant value of $c = 1.8$. The algorithm achieves the highest accuracy of 86% for $b = 45$ and steadies at 85% for the b values from 95 through 255. On the contrary, when we keep the value of b constant at 115, as in Figure 3b, for varying c values, the accuracy is not affected at all. The same result holds for various values of b kept as constant with varying c. We experimented with changing the value of b until the highest accuracy was obtained. Highest accuracy of 86% was obtained with $b = 45$. Next, the value of b was kept constant, and the value for c was varied as can be seen in Figure 3b.

The accuracy of the algorithm does not change with the change in values of c. This trend of values remaining constant

TABLE V: Performance comparison for privileged (male) and unprivileged group (non-male) for CelebA dataset

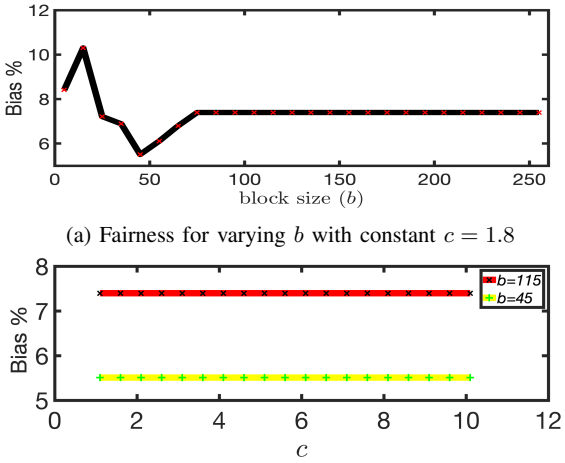
Test Parameters	Male	Non-Male
Mean ($\overline{Xi}$) **(%)**	92.60	90.54
Variance (σ_i) **(%)**	0.05	0.09
Sample size	1000	
p**-value**	<0.0001	
Result	H_0 Rejected and H_1 Accepted	



(a) Accuracy for varying b with constant $c = 1.8$

(b) Accuracy for varying c with constant b

Fig. 3: Parameter selection for Accuracy



(a) Fairness for varying b with constant $c = 1.8$

(b) Fairness for varying c with constant b

Fig. 4: Parameter selection for Fairness

for varying c value with a constant b value can be seen for fairness results as well. The bias or fairness of an algorithm was calculated by using Equation 2. The bias between the accuracy of privileged and unprivileged groups remains constant for any value of constant b and varying c, as can be seen in Figure 4b. When the c was kept constant, the bias dropped from 10.29% at b = 15 to 5.5% at b = 45. The bias then increased to 7.1% and remains steady after that for values ranging from $b = 75$ through $b = 255$ as can be seen in Figure 4a. Hence, different applications could easily see a range of bias from 5.50% to 10.29% depending on the value of b chosen.

As these are preliminary results, auditing bias is the paper's main focus and the onus of correcting it is with the application developers who use the aforementioned pupil detection algorithm as a base for their applications. The combination of Figures 3a and 4a can be used by system designers to identify the suitable thresholds that work in their particular setting. For instance, they can choose to prioritize fairness or accuracy, or as is a possibility here, pick a combination of parameters that provides a good balance between fairness and accuracy. In the presented analysis, a setting of $b = 45, c = 1.8$ gives both high accuracy (86%) and low bias (5.5%) in the Images of Groups dataset, and the same was kept constant for the CelebA dataset for fair comparative analysis. An interesting future study could include comparing the results when the values are changed on both datasets.

V. Discussion

Our results based on statistical analysis over multiple datasets indicate that the pupil detection algorithm in OpenCV may yield biased results across gender groups. We take inspiration from the work by Boulamwini and Gebru [4] who were the first to undertake fairness audit in face analysis algorithms in a systematic way. Their work has helped motivate multiple algorithmic audit and algorithmic adaptation efforts [35], [48]. As the first systematic empirical effort in the area of fairness of pupil detection, this work aims to motivate further research in the area of fairness of pupil detection. Some of the limitations of the current work (e.g., use of binary gender, two datasets) can be countered with follow up work in this important area of fairness in pupil detection.

Creation of fair and accurate pupil detectors can impact the life of users in multiple ways, including:

1) Gaze based identity authentication [49]: These days many smartphones and other devices use face recognition in order to authenticate the user. Along with that, many applications also use persons' gaze towards the camera or screen for unlocking. In such applications, the aspect of gender-based bias provides an unfair advantage to the privileged group based on gender.
2) Attention assessment or evaluations [50]: With the increasing use of online platforms for conducting job interviews and research studies, attention assessment, and confidence assessment, using the candidate's eye contact has been tough. In such applications where a person's gaze can influence decisions, gender bias could play an important role.
3) E-Learning [51]: With the advent of online learning methodologies, fair assessment of the student's gaze towards the screen (e.g, in online examinations) has become critical. A gender bias in this application could hinder the opportunities for the unprivileged group (e.g., based on gender).

As mentioned earlier, an important limitation of this paper is the usage of binary visual gender labels, which for the "Images of Groups" dataset was based on human observations to establish ground truth. Connecting with the subjects or obtaining the disclosed gender of the subjects in the dataset was not practically possible, and hence as a starting point we have decided to use these gender labels to test for bias in gaze detection algorithms. The labels in CelebA (celebrity images) were provided by the creators of the dataset, which was in turn based on resources such as IMDB and Wikipedia. While not perfect, this process complements the manual gender labeling process adopted for "Images of Groups". The results for gender bias in gaze detection were found to be consistent in both datasets. Future work with self-described gender labels in image datasets should be welcomed.

VI. Conclusion

Our audit of OpenCV's pupil detection algorithm, tested on two separate datasets suggests that it has a statistically significant bias favoring male faces compared to non-male faces. This bias can impact opportunities in downstream multimedia applications in fields such as authentication, education, and employment assessment. Initial experimental results on the 'Images of Groups' dataset suggest that this bias can be reduced by carefully choosing the algorithmic parameters (block size, in this case), with minimal loss of overall accuracy of the algorithm. Future work includes auditing other state-of-the-art algorithms for fairness and creating a generic framework to ensure high accuracy and fairness in similar settings.

References

[1] A. Malinowski, J. Chen, S. Mishra, S. Samavedam, and D. Sohn, "What is killing moore's law? challenges in advanced finfet technology integration," in *26th International Conference Mixed Design of Integrated Circuits and Systems*, Rzeszów, Poland, 2019, pp. 46–51.

[2] D. Citron and F. Pasquale, "The scored society: Due process for automated predictions," *Wash L. Rev*, vol. 89, p. 1, 2014.

[3] V. Singh, E. André, S. Boll, M. Hildebrandt, and D. Shamma, "Legal and ethical challenges in multimedia research," *IEEE MultiMedia*, vol. 27, no. 2, pp. 46–54, 2020.

[4] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*, New York, USA, 2018, pp. 77–91.

[5] B. Wilson, J. Hoffman, and J. Morgenstern, "Predictive inequity in object detection," *arXiv preprint arXiv:1902.11097*, 2019.

[6] S. Hajian, F. Bonchi, and C. Castillo, "Algorithmic bias: From discrimination discovery to fairness-aware data mining," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, California, USA, 2016, pp. 2125–2126.

[7] O. Asan, A. Bayrak, and A. Choudhury, "Artificial intelligence and human trust in healthcare: focus on clinicians," *Journal of medical Internet research*, vol. 22, no. 6, p. e15154, 2020.

[8] M. Arnold, R. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. Ramamurthy, A. Olteanu, D. Piorkowski *et al.*, "Factsheets: Increasing trust in ai services through supplier's declarations of conformity," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 6–1, 2019.

[9] "Opencv library," https://opencv.org/.

[10] O. Kulkarni, V. Patil, S. Parikh, S. Arora, and P. Atrey, "Can you all look here? towards determining gaze uniformity in group images," in *2020 IEEE International Symposium on Multimedia (ISM)*. Naples, Italy: IEEE, 2020, pp. 100–103.

[11] A. Almuzaini and V. Singh, "Balancing fairness and accuracy in sentiment detection using multiple black box models," in *Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia*, Washington, USA, 2020, pp. 13–19.

[12] A. Gallagher and T. Chen, "Understanding images of groups of people," in *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami Beach, FL, USA, 2009, pp. 256–263.

[13] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, Santiago, Chile, December 2015.

[14] A. Kar and P. Corcoran, "A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms," *IEEE Access*, vol. 5, pp. 16 495–16 519, 2017.

[15] R. Yonetani, H. Kawashima, T. Hirayama, and T. Matsuyama, "Gaze probing: Event-based estimation of objects being focused on," in *2010 20th International Conference on Pattern Recognition*. Istanbul, Turkey: IEEE, 2010, pp. 101–104.

[16] K. Okamoto, A. Utsumi, H. Yamazoe, T. Miyashita, S. Abe, K. Takahashi, and N. Hagita, "Classification of pedestrian behavior in a shopping mall based on lrf and camera observations," in *MVA*, Nara, Japan, 2011, pp. 1–5.

[17] T. Hirayama, Y. Sumi, T. Kawahara, and T. Matsuyama, "Info-concierge: Proactive multi-modal interaction through mind probing," in *The Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2011)*, Xián, China, 2011.

[18] L. Fridman, P. Langhans, J. Lee, and B. Reimer, "Driver gaze region estimation without use of eye movement," *IEEE Intelligent Systems*, vol. 31, no. 3, pp. 49–56, 2016.

[19] Y. Kodama, Y. Kawanishi, T. Hirayama, D. Deguchi, I. Ide, H. Murase, H. Nagano, and K. Kashino, "Localizing the gaze target of a crowd of people," in *Asian Conference on Computer Vision (ACCV)*. Perth, Australia: Springer, 2018, pp. 15–30.

[20] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba, "Where are they looking?" in *Twenty-ninth Conference on Neural Information Processing Systems (NeurIPS)*, Montreal, Canada, 2015, pp. 199–207.

[21] J. Lee, C. Cho, K. Shin, E. Lee, and K. Park, "3d gaze tracking method using purkinje images on eye optical model and pupil," *Optics and Lasers in Engineering*, vol. 50, no. 5, pp. 736–751, 2012.

[22] W. Fuhl, D. Geisler, T. Santini, W. Rosenstiel, and E. Kasneci, "Evaluation of state-of-the-art pupil detection algorithms on remote eye images," in *UbiComp '16: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, Heidelberg Germany, 2016, pp. 1716–1725.

[23] W. Fuhl, T. K¨ubler, K. Sippel, W. Rosenstiel, and E. Kasneci, "Excuse: Robust pupil detection in real-world scenarios," in *International Conference on Computer Analysis of Images and Patterns*. Valletta, Malta: Springer, 2015, pp. 39–51.

[24] Y. Yang, C. Feng, and R. Wang, "Ultrasound pupil image segmentation based on edge detection and detection operators," in *Proceedings of the 2020 12th International Conference on Machine Learning and Computing*, Shenzhen, China, 2020, pp. 271–275.

[25] A. George and A. Routray, "Fast and accurate algorithm for eye localisation for gaze tracking in low-resolution images," *IET Computer Vision*, vol. 10, no. 7, pp. 660–669, 2016.

[26] Y. Luo, Y. Zou, and X. Xia, "Real-time pupil parameter detection for infrared image using opencv [j]," *International Journal of Computer Science and Telecommunications*, vol. 6, pp. 71–75, 2013.

[27] P. Bonteanu, A. Cracan, R. Bozomitu, and G. Bonteanu, "A robust pupil detection algorithm based on a new adaptive thresholding procedure," in *E-Health and Bioengineering Conference (EHB)*. Iasi, Romania: IEEE, 2019, pp. 1–4.

[28] D. Pimplaskar, M. Nagmode, and A. Borkar, "Real time eye blinking detection and tracking using opencv," *technology*, vol. 13, no. 14, p. 15, 2015.

[29] L. Schwarz, H. Gamba, F. Pacheco, R. Ramos, and M. Sovierzoski, "Pupil and iris detection in dynamic pupillometry using the opencv library," in *5th International Congress on Image and Signal Processing*. Chongqing, China: IEEE, 2012, pp. 211–215.

[30] R. Lupu, F. Ungureanu, and V. Siriteanu, "Eye tracking mouse for human computer interaction," in *2013 E-Health and Bioengineering Conference (EHB)*. Iaşi,Romania: IEEE, 2013, pp. 1–4.

[31] H. Lee, S. Heo, S. Lee, and Y. Yu, "Real-time pupil motion recognition and efficient character selection system using fpga and opencv," in *Proceedings of the Korean Institute of Information and Commucation Sciences Conference*. The Korea Institute of Information and Commucation Engineering, 2018, pp. 393–394.

[32] C. Li, C. Kim, and J. Park, "The indirect keyboard control system by using the gaze tracing based on haar classifier in opencv," in *2009 International Forum on Information Technology and Applications*, vol. 2. Chengdu,China: IEEE, 2009, pp. 362–366.

[33] L. Hong-xia, "The human eye pupil localization algorithm based on opencv," *Electronics Quality*, no. 11, p. 6, 2012.

[34] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.", 2008.

[35] J. Alasadi, R. Arunachalam, P. Atrey, and V. Singh, "A fairness-aware fusion framework for multimodal cyberbullying detection," in *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*. New Delhi, India: IEEE, 2020, pp. 166–173.

[36] T. Bolukbasi, K. Chang, J. Zou, V. Saligrama, and A. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in neural information processing systems*, Barcelona, Spain, 2016, pp. 4349–4357.

[37] V. Singh, M. Chayko, R. Inamdar, and D. Floegel, "Female librarians and male computer programmers? gender bias in occupational images on digital media platforms," *Journal of the Association for Information Science and Technology*, 2020.

[38] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, Massachusetts, USA, 2012, pp. 214–226.

[39] A. Caliskan, J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[40] J. Alasadi, A. Al Hilli, and V. Singh, "Toward fairness in face matching algorithms," in *Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia*, 2019, pp. 19–25.

[41] A. Bayliss, G. Di Pellegrino, and S. Tipper, "Sex differences in eye gaze and symbolic cueing of attention," *The Quarterly Journal of Experimental Psychology Section A*, vol. 58, no. 4, pp. 631–650, 2005.

[42] M. Slepian, M. Weisbuch, R. Adams Jr, and N. Ambady, "Gender moderates the relationship between emotion and perceived gaze." *Emotion*, vol. 11, no. 6, p. 1439, 2011.

[43] S. Shimojo, C. Simion, E. Shimojo, and C. Scheier, "Gaze bias both reflects and influences preference," *Nature neuroscience*, vol. 6, no. 12, pp. 1317–1322, 2003.

[44] D. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[45] G. Bradski and A. Kaehler, "Opencv," *Dr. Dobb's journal of software tools*, vol. 3, 2000.

[46] J. Cech and T. Soukupova, "Real-time eye blink detection using facial landmarks," *Cent. Mach. Perception, Dep. Cybern. Fac. Electr. Eng. Czech Tech. Univ. Prague*, pp. 1–8, 2016.

[47] B. Efron and R. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.

[48] D. Pessach and E. Shmueli, "Algorithmic fairness," *arXiv preprint arXiv:2001.09784*, 2020.

[49] J. Weaver, K. Mock, and B. Hoanca, "Gaze-based password authentication through automatic clustering of gaze points," in *2011 IEEE International Conference on Systems, Man, and Cybernetics*. Alaska, USA: IEEE, 2011, pp. 2749–2754.

[50] N. Chen and P. Clarke, "Gaze-based assessments of vigilance and avoidance in social anxiety: a review," *Current psychiatry reports*, vol. 19, no. 9, pp. 1–9, 2017.

[51] S. Asteriadis, P. Tzouveli, K. Karpouzis, and S. Kollias, "Estimation of behavioral user state based on eye gaze and head pose—application in an e-learning environment," *Multimedia Tools and Applications*, vol. 41, no. 3, pp. 469–493, 2009.