Unravelling the Effect of Image Distortions for Biased Prediction of Pre-trained Face Recognition Models

Puspita Majumdar\textsuperscript{1,2}, Surbhi Mittal\textsuperscript{2}, Richa Singh\textsuperscript{2}, Mayank Vatsa\textsuperscript{2}
\textsuperscript{1}IIIT-Delhi, India \textsuperscript{2}IIT Jodhpur, India
pushpitam@iiitd.ac.in, \{mittal.5, richa, mvatsa\}@iitj.ac.in

Abstract

Identifying and mitigating bias in deep learning algorithms has gained significant popularity in the past few years due to its impact on the society. Researchers argue that models trained on balanced datasets with good representation provide equal and unbiased performance across subgroups. However, can seemingly unbiased pre-trained model become biased when input data undergoes certain distortions? For the first time, we attempt to answer this question in the context of face recognition. We provide a systematic analysis to evaluate the performance of four state-of-the-art deep face recognition models in the presence of image distortions across different gender and race subgroups. We have observed that image distortions have a relationship with the performance gap of the model across different subgroups.

1. Introduction

Over the past few years, there has been a growing focus on understanding bias in deep learning models. Researchers have attempted to realize the sources of bias and analyze the performance of pre-trained deep models across different demographic subgroups in face analysis problems (e.g., male and female are subgroups of gender) \cite{5, 28}. It has been shown that human bias incorporated during the collection and curation of data \cite{13}, and imbalance in training data distribution with respect to a particular subgroup \cite{3} are some of the potential sources of bias that lead to unfair predictions. A model performing equally well across different subgroups is considered to be an unbiased model \cite{10}, while they are considered biased when the model favors one subgroup over the other. In this research, we demonstrate that an initially unbiased model may become biased under certain scenarios such as distortions which raises the doubts on models’ robustness.

Several researchers have analyzed the robustness of deep models under image distortions and designed algorithms to enhance robustness \cite{8, 14, 15, 21}. However, none of the studies analyze the effect of distortions on the performance of deep models across different demographic subgroups. With recent incidents of biased prediction of deep models towards particular demographic subgroups \cite{4, 9}, this analysis is crucial to provide insights towards understanding bias in model prediction.

As shown in Figure 1, experiments conducted to evaluate the performance of face recognition across different subgroups in presence of Gaussian blur show that the confidence of recognizing images is lower for some subgroups compared to others when the same intensity of ‘blur’ is applied. With the objective to unravel this effect with different real world distortions, we investigate two key aspects:

1. Do unbiased model predictions become biased in presence of image distortions?
2. How does the performance vary across different demographic subgroups when images undergo distortions?

We perform a detailed analysis to understand the reason for the difference in the performance of the models across different subgroups in the presence of distortions. Further, we analyze the changes in the regions-of-importance through feature visualization of the salient regions provided by the models. Finally, we discuss some of the possible solutions to overcome the problem of biased prediction of deep models in the presence of image distortions.

2. Related Work

This section is segregated into studies related to (i) the effect of image distortions on deep models and (ii) understanding bias.

2.1. Effect of Image Distortions on Deep Models

Karahan et al. [21] provided a systematic analysis to assess the effect of image quality on face recognition performance using three popular deep models. It is observed that blur, noise, and occlusion significantly deteriorate the performance of deep models. Dodge and Karam [8] have shown that image distortions result in overall degradation in the classification performance. Grm et al. [15] analyzed the effect of image distortions on face verification performance of four deep models using the LFW dataset [20]. The results indicate that noise, blur, missing pixels, and brightness have a significant effect on the overall model performance. Later, Richard Webster et al. [32] studied the behavior of deep models in the presence of image distortions through visual psychophysics and have shown that visual psychophysics makes face recognition more explainable. Two covariate related problems on unconstrained face verification are studied by Lu et al. [23]. First, the effect of covariates is analyzed and then it is utilized to improve verification performance. It is observed that pose variations and occlusions severely affect performance. Also, covariates such as age, gender, and skin tone have shown impacts on performance. Recently, Yang et al. [40] provided a detailed study of object and face detection in poor visibility conditions.

2.2. Understanding Bias

In the recent literature, several studies have focused on detection and mitigation of bias present in deep learning models [12, 24, 37]. However, very few focus on the analysis of this prevalent bias. In [4], the authors unveil the disparity in the performance of commercial-grade gender classification systems based on phenotypic subgroups - lighter-skinned males, darker-skinned males, lighter-skinned females, and darker-skinned females. In [6], analysis of real and apparent age differences and its correlation with other attributes is performed. In another work by Nagpal et al. [28], the authors attempt to answer how bias is encoded in facial recognition models. They perform analysis based on state-of-the-art deep learning models and suggest how bias encoded by models is comparable to human biases. Wang et al. [38] highlight how balanced datasets are not enough to ensure unbiased performance in deep learning models. They further show how existing biases are amplified by deep learning models in visual recognition tasks. In [5], the authors analyze latent representation of facial images to identify sources of bias. They show that protected attributes like race play a role in biasing latent representations. Recently, in [35], the authors provide an in-depth analysis of the correlation between the quality of face images and biased facial recognition systems. They show that the current definition of face quality transfers bias from face quality assessment systems to facial recognition systems for certain subgroups. In [34], the authors focus on understanding the feature space generated by deep models. Krishnapriya et al. [22] highlight the issues related to face recognition performance with varying skin tone and race. A detailed survey of demographic bias in biometric systems is presented in [9].

3. Proposed Evaluation Framework

In this research, we start with the hypotheses that (i) image distortions affect the performance of a model and cause biased predictions, and (ii) the performance varies differently across different demographic subgroups when images undergo real-world distortions. In order to validate and substantiate these hypotheses, the performance of pre-trained deep models is evaluated with and without the presence of distortions across different demographic subgroups. We analyze the regions used by the models for recognition and observe whether the regions of interest remain consistent under the effect of distortions. For this purpose, experiments are performed for face recognition across gender and race subgroups, using data corresponding to gender G1 (Male) and G2 (Female), and race R1 (light skin color) and R2 (dark skin color). We have also analyzed the performance of the models across the intersectional subgroups1 of gender and race - \{G1,R1\}, \{G1,R2\}, \{G2,R1\}, and \{G2,R2\}. For the experiments, four pre-trained deep face recognition models: (i) LightCNN-29 [39] (ii) SENet50 [19] (iii) ResNet50 [18], and (iv) ArcFace [7] are used. Experimental details are shown in Table 1. The details of the pre-trained models are given in the supplementary file. This section discusses the details of the datasets with the corresponding protocols, image distortions considered in this research, and the evaluation metrics.

1The performance of the models across the intersectional subgroups is reported in the supplementary file at: https://github.com/Puspitamajumdariiit/Unravelling-the-Effect-of-Image-Distortions.
3.1. Datasets and Protocols

Two publicly available constrained face recognition datasets are used to analyze the effect of image distortions on the performance of pre-trained models across different demographic subgroups. Constrained datasets are considered to solely understand the effect of one type of image distortion on the model performance across different subgroups. MORPH dataset (Album-2) [31] contains more than 54K images of 13K subjects. The dataset is pre-labeled with two gender subgroups, Male (G1) and Female (G2) and six race subgroups, White (R1), Black (R2), Hispanic, Indian, Asian, and Other. The dataset is imbalanced with respect to different subgroups. Therefore, equal subgroup-wise distribution is ensured during the experiments.

MUCT dataset [25] consists of 3,755 images of 131 male (G1) and 146 female (G2) subjects. For the experiments, equal subgroup-wise distribution is ensured.

Protocol Details: For evaluation, face verification is performed and the results are reported at the 0.01 False Accept Rate (FAR)\(^2\). Similar to the LFW dataset [20], we created 10 disjoint splits of image pairs, each having 300 positive and 300 negative pairs. Here, the positive and negative pairs in each split are created corresponding to each subgroup of a demographic group (e.g. male and female are the subgroups of the demographic group gender). The final evaluation is performed on 12000 pairs with 6000 positive and 6000 negative pairs for each demographic group. Also, for analyzing face recognition performance across gender subgroups, race-wise equal distribution is ensured, and vice versa.

3.2. Details of Image Distortions

To emulate the real-world scenario of matching unconstrained probe images with constrained gallery images, distortions of different levels/intensities are applied to one of the images in each pair (considering it as probe image) for the verification experiments. The following six image distortions are considered for analysis. Sample images of the MUCT dataset after applying image distortions of different intensities are shown in Figure 1 of the supplementary file.

Occlusion: Occluded images are generated by occluding seven facial regions: eyes, nose, mouth, forehead, left cheek, right cheek, and area typically covered by protective face masks. We first detect 68 facial keypoints [33] which are then utilized in selecting the region to be occluded.

Gaussian Blur: Images are blurred using Gaussian filters with varying standard deviations \(\sigma\). The size of the filter is decided as \(2 \times \lceil(2\sigma)\rceil + 1\). We vary \(\sigma\) from 2.0 to 4.0 with a constant step size of 0.2.

Brightness: To adjust brightness of images, we apply operations as in [15]. Each image is multiplied by a brightness factor \(\beta\) from 1.0 to 3.0 with a constant step size of 0.5 and the values are subsequently clipped to lie between the image pixel intensity range (0,255).

Gaussian Noise: To generate images with Gaussian noise, an additive Gaussian noise vector with dimensions equal to the size of the image is used. This vector is generated with values of \(\sigma\) varying from 10 to 40, with a step-size of 10.

Salt and Pepper Noise: To generate images with salt and pepper noise, an image pixel is set to zero with a probability of \(p/2\), or set to 255 with a probability of \(p/2\) across all image channels. The value of \(p\) is varied from 0.03 to 0.15 with a step size of 0.03.

Resolution: We reduce the resolution of the images using cv2 library [27] in Python with INTER_AREA interpolation. The resolutions are varied as 96 × 96, 64 × 64, 48 × 48, 32 × 32, and 28 × 28.

3.3. Evaluation Metrics

To evaluate the effect of distortions on face recognition performance for different subgroups, deep features ex-
Table 3. Verification accuracy and DoB (%) across different race subgroups under occlusion corresponding to the MORPH dataset. Occlusion of nose region significantly degrades the performance of models for subgroup R2.

<table>
<thead>
<tr>
<th>Role of facial regions in recognition across subgroups</th>
</tr>
</thead>
</table>
| We occlude different facial regions to investigate their importance in recognition of a particular subgroup. The verification performance across different gender subgroups is shown in Table 2. The results indicate a significant degradation in performance on occluding the eyes, nose, and facial region covered by a protective face mask. On the other hand, occlusion of mouth, forehead, and cheeks does not have a significant effect. Hence, we can conclude that eyes, nose, and facial mask regions are the most discriminative regions for recognition across gender subgroups. We also observe that all the models perform poorly for subgroup G2 resulting in a large performance gap between G1 and G2 upon occluding the discriminative regions. Similarly, for race subgroups (Table 3), eyes, nose, and facial mask regions are found to be the most discriminative regions. It is observed that the difference in performance between R1 and R2 is maximum when the nose region is occluded, for most models. The models perform poorly for subgroup R2 on occluding the nose region. On the other hand, occlusion of other facial regions almost equally affects the performance of the models. This indicates that nose is the most discriminative region for subgroup R2. To further investigate the underlying reasons for our observation, we have analyzed the regions used by the models for discrimination through feature visualization. The salient regions are obtained by interpolating the final convolution layer filter responses and superimposing on the input image. Figure 2 shows the visualization of salient regions used for feature extraction by the ArcFace model. It is observed that the model focuses predominantly on the eyes and nose regions for feature extraction. For subgroup R2, nose is observed to be the most salient region. Does model performance degrade equally across subgroups in presence of Gaussian blur?
Table 4 shows the variation in performance with varying intensities of blur across different gender subgroups. It is interesting to observe an initially unbiased model becomes biased in the presence of blur. The performance gap between $G1$ and $G2$ increases as we increase the intensity of blur. For instance, the accuracy for $G1$ and $G2$ is 100% and 99.83%, respectively, on original images, corresponding to the LCNN-29 model. However, it reduces to 84.70% and 72.27% when degraded with blur with $\sigma = 4.0$. As a result, the DoB increases from 0.12% to 8.79%, which indicates bias is introduced in model prediction. For the MUCT dataset, a similar trend is observed as shown in Table 2 of supplementary file. A huge disparity in performance across the intersectional subgroups of gender (Table 5 of supplementary file). We have also analyzed the performance degradation observed for subgroup $R1$ and $R2$ with higher performance degradation observed for subgroup $R1$ (Table 4 of supplementary file). We have also analyzed the performance across race subgroups, we observe that the performance gap increases between $R1$ and $R2$ with higher performance degradation observed for subgroup $R1$ (Table 4 of supplementary file). A huge disparity in performance across different subgroups is observed.

To analyze how blur impacts the model’s ability to recognize faces across different subgroups, we use feature visualization of salient regions. Figure 3(a-b) shows the feature visualization obtained by LCNN-29 on original and blurred images of varying intensities. It is interesting to observe that the regions of interest change when blur is applied. The model shifts the focus from nose to eyes region for subgroup $G1$ while it shifts from eyes to mouth region for subgroup $G2$. Among race subgroups, the focus shifts from the upper nose region to lower nose region for subgroup $R2$, and to mouth region for subgroup $R1$. As previously seen, mouth region is less discriminative than eyes and nose regions. This shows that model shifts the regions of interest from higher discriminative regions to lower discriminative regions for subgroups $G2$ and $R1$ in presence of blur. Thus, higher performance degradation is observed for these subgroups.

### Table 4

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>LCNN-29</th>
<th>SENet50</th>
<th>ResNet50</th>
<th>ArcFace</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>100.00</td>
<td>97.90</td>
<td>98.27</td>
<td>99.90</td>
</tr>
<tr>
<td>2.0</td>
<td>99.83</td>
<td>91.97</td>
<td>93.30</td>
<td>99.67</td>
</tr>
<tr>
<td>2.4</td>
<td>99.83</td>
<td>91.23</td>
<td>94.30</td>
<td>99.70</td>
</tr>
<tr>
<td>3.0</td>
<td>98.43</td>
<td>73.60</td>
<td>84.33</td>
<td>96.93</td>
</tr>
<tr>
<td>3.4</td>
<td>98.43</td>
<td>73.60</td>
<td>84.33</td>
<td>96.93</td>
</tr>
<tr>
<td>4.0</td>
<td>98.43</td>
<td>73.60</td>
<td>84.33</td>
<td>96.93</td>
</tr>
</tbody>
</table>

Table 5 shows the variation in performance with varying intensities of brightness across different race subgroups. Accuracy of the models for subgroup $R1$ deteriorates significantly in presence of brightness.

### Table 5

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>LCNN-29</th>
<th>SENet50</th>
<th>ResNet50</th>
<th>ArcFace</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>100.00</td>
<td>97.93</td>
<td>98.80</td>
<td>99.97</td>
</tr>
<tr>
<td>1.5</td>
<td>99.83</td>
<td>91.97</td>
<td>99.40</td>
<td>99.70</td>
</tr>
<tr>
<td>2.0</td>
<td>98.43</td>
<td>73.60</td>
<td>99.70</td>
<td>99.70</td>
</tr>
<tr>
<td>2.5</td>
<td>98.43</td>
<td>73.60</td>
<td>99.70</td>
<td>99.70</td>
</tr>
</tbody>
</table>

Figure 3. Visualizing the salient regions of the pre-trained LCNN-29 model for recognition with varying intensities of (a-b) Gaussian blur and (c-d) brightness. Left block: variation across gender subgroups and right block: variation across race subgroups. For gender subgroups, race is kept constant and vice versa.

Table 6 shows the variation in performance with varying intensities of brightness across different gender subgroups. Accuracy of the models for subgroup $R1$ deteriorates significantly in presence of brightness.

### Table 6

<table>
<thead>
<tr>
<th>Gender</th>
<th>LCNN-29</th>
<th>SENet50</th>
<th>ResNet50</th>
<th>ArcFace</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G1$</td>
<td>100.00</td>
<td>97.93</td>
<td>98.80</td>
<td>99.97</td>
</tr>
<tr>
<td>$G2$</td>
<td>99.83</td>
<td>91.97</td>
<td>99.40</td>
<td>99.70</td>
</tr>
<tr>
<td>$R1$</td>
<td>99.83</td>
<td>91.97</td>
<td>99.40</td>
<td>99.70</td>
</tr>
<tr>
<td>$R2$</td>
<td>98.43</td>
<td>73.60</td>
<td>99.70</td>
<td>99.70</td>
</tr>
<tr>
<td>$G1$, $R1$</td>
<td>98.43</td>
<td>73.60</td>
<td>99.70</td>
<td>99.70</td>
</tr>
<tr>
<td>$G2$, $R2$</td>
<td>98.43</td>
<td>73.60</td>
<td>99.70</td>
<td>99.70</td>
</tr>
</tbody>
</table>

Figure 3. Visualizing the salient regions of the pre-trained LCNN-29 model for recognition with varying intensities of (a-b) Gaussian blur and (c-d) brightness. Left block: variation across gender subgroups and right block: variation across race subgroups. For gender subgroups, race is kept constant and vice versa.

To analyze how blur impacts the model’s ability to recognize faces across different subgroups, we use feature visualization of salient regions. Figure 3(a-b) shows the feature visualization obtained by LCNN-29 on original and blurred images of varying intensities. It is interesting to observe that the regions of interest change when blur is applied. The model shifts the focus from nose to eyes region for subgroup $G1$ while it shifts from eyes to mouth region for subgroup $G2$. Among race subgroups, the focus shifts from the upper nose region to lower nose region for subgroup $R2$, and to mouth region for subgroup $R1$. As previously seen, mouth region is less discriminative than eyes and nose regions. This shows that model shifts the regions of interest from higher discriminative regions to lower discriminative regions for subgroups $G2$ and $R1$ in presence of blur. Thus, higher performance degradation is observed for these subgroups.

**Does performance gap between subgroups increase for brighter images?**

In this experiment, we analyze the effect of brightness. The performance is shown across gender (Table 6 of supplementary file) and race (Table 5 of main paper) subgroups of the MORPH dataset. Increasing the intensity of brightness...
significantly affects the performance of deep models for subgroups G2 and R1. The score distribution corresponding to the SENet50 model shown in Figure 2 of the supplementary file further validates this fact. It is observed that the overlap increases with increasing intensity of brightness. However, subgroups G1 and R2 still show more distinct margins compared to G2 and R1, respectively. This indicates that subgroups G1 and R2 are still recognizable when exposed to high brightness as compared to G2 and R1, respectively. On analyzing the bias in model prediction, we observe that DoB increases from 4.19% to 10.25%, and 0.52% to 8.29% for gender and race subgroups, respectively at $\beta = 2.0$. Similar observations are noted for the MUCT dataset as well (Table 7 of supplementary file).

We observe that subgroup R1 is highly susceptible to the effect of brightness. On increasing the brightness factor, the facial features of subgroup R1 are heavily affected, which in turn affects the overall performance. The accuracies of all the models for subgroup R1 drop below 20% beyond the brightness factor $\beta = 2.5$. We further strengthen the observation using the feature visualization shown in Figure 3(d), where we observe that the model is unable to extract features for subgroup R1 while it focuses on the nose region for subgroup R2 with increasing intensities of brightness. Similarly, for gender subgroups, the model’s focus changes from eyes to nose region for subgroup G1, while it shifts from nose to right cheek as shown in Figure 3(c). Cheeks are observed to be less discriminative in our occlusion experiment and thus, performance degradation is higher for subgroup G2.

**Do models perform differently across subgroups in presence of noise?**

The effect of *Gaussian Noise* and *Salt and Pepper Noise* are analyzed in the next set of experiments. Table 6 shows the effect of Gaussian noise on the performance of pre-trained models across different gender subgroups. It is observed that the overall performance decreases as the intensity of noise increases, but the performance gap between different subgroups is not significant. For example, the DoB of ResNet50 model increases from 3.51% to 4.48% when $\sigma$ is increased to 40. This indicates that model performance is equally affected for gender subgroups by Gaussian noise. Similar conclusions are drawn on observing the performance across race subgroups (Table 9 of supplementary file). For the MUCT dataset, a similar set of observations are drawn.

The performance of models under the effect of salt and pepper noise across gender subgroups is shown in Table 7. Here, we observe that unlike the models’ performance in presence of Gaussian noise, the performance gap between subgroups increases under the effect of salt and pepper noise. For the ResNet50 model, the DoB increases from 3.51% to 14.05%. Similar observations are drawn for race subgroups (Table 11 of supplementary file). Earlier studies [21, 15] have shown that deep models behave similarly for both types of noise. But, in this study, we have observed that salt and pepper noise affect the performance of most of the deep models for subgroups G2 and R1 more severely than subgroups G1 and R2, respectively. In order to investigate the difference in behavior of the models for both types of noise across different subgroups, we use the t-SNE visualization for gender subgroups, as shown in Figure 4. For Gaussian noise, it is observed that on increasing the intensity of noise, the overlap in the feature distribution of individual subgroups increases, making the clusters of each subgroup dense. The dense clusters indicate a high misclassification rate and overall performance degradation for each

---

4We have reported the values up to $p = 0.09$ due to the page limitation. However, a similar trend in the results is observed for higher values of $p$. ```

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>LCNN-29</th>
<th>SENet50</th>
<th>ResNet50</th>
<th>ArcFace</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>100.00</td>
<td>97.90</td>
<td>98.26</td>
<td>99.90</td>
</tr>
<tr>
<td>G2</td>
<td>99.83</td>
<td>91.96</td>
<td>93.30</td>
<td>99.66</td>
</tr>
<tr>
<td>DoB</td>
<td>0.12</td>
<td>4.19*</td>
<td>3.51*</td>
<td>0.16</td>
</tr>
<tr>
<td>G1</td>
<td>99.93</td>
<td>96.36</td>
<td>97.13</td>
<td>99.73</td>
</tr>
<tr>
<td>G2</td>
<td>99.70</td>
<td>90.86</td>
<td>89.03</td>
<td>98.03</td>
</tr>
<tr>
<td>DoB</td>
<td>0.16</td>
<td>3.89</td>
<td>5.30</td>
<td>1.20</td>
</tr>
<tr>
<td>G1</td>
<td>99.83</td>
<td>94.56</td>
<td>93.60</td>
<td>99.23</td>
</tr>
<tr>
<td>G2</td>
<td>99.36</td>
<td>86.46</td>
<td>82.80</td>
<td>96.66</td>
</tr>
<tr>
<td>DoB</td>
<td>0.33</td>
<td>5.73</td>
<td>7.64</td>
<td>1.82</td>
</tr>
<tr>
<td>G1</td>
<td>99.63</td>
<td>89.93</td>
<td>79.40</td>
<td>94.10</td>
</tr>
<tr>
<td>G2</td>
<td>98.53</td>
<td>84.46</td>
<td>73.06</td>
<td>90.66</td>
</tr>
<tr>
<td>DoB</td>
<td>0.78</td>
<td>3.87</td>
<td>4.48</td>
<td>2.43</td>
</tr>
</tbody>
</table>

Table 6. Verification accuracy and DoB (%) across different gender subgroups with varying intensities of Gaussian noise corresponding to the MORPH dataset. * represents a relatively high disparity in model performance across different subgroups on undistorted images.

<table>
<thead>
<tr>
<th>$p$</th>
<th>LCNN-29</th>
<th>SENet50</th>
<th>ResNet50</th>
<th>ArcFace</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>G1</td>
<td>100.00</td>
<td>97.90</td>
<td>98.26</td>
</tr>
<tr>
<td>G2</td>
<td>99.83</td>
<td>91.96</td>
<td>93.30</td>
<td>99.66</td>
</tr>
<tr>
<td>DoB</td>
<td>0.12</td>
<td>4.19*</td>
<td>3.51*</td>
<td>0.16</td>
</tr>
<tr>
<td>0.03</td>
<td>G1</td>
<td>99.83</td>
<td>92.84</td>
<td>94.16</td>
</tr>
<tr>
<td>G2</td>
<td>98.80</td>
<td>82.83</td>
<td>83.16</td>
<td>55.86</td>
</tr>
<tr>
<td>DoB</td>
<td>0.73</td>
<td>7.07</td>
<td>7.78</td>
<td>11.20</td>
</tr>
<tr>
<td>0.06</td>
<td>G1</td>
<td>99.46</td>
<td>83.90</td>
<td>81.06</td>
</tr>
<tr>
<td>G2</td>
<td>96.23</td>
<td>72.13</td>
<td>63.20</td>
<td>10.13</td>
</tr>
<tr>
<td>DoB</td>
<td>2.28</td>
<td>8.32</td>
<td>12.63</td>
<td>4.67</td>
</tr>
<tr>
<td>0.09</td>
<td>G1</td>
<td>98.26</td>
<td>71.03</td>
<td>55.60</td>
</tr>
<tr>
<td>G2</td>
<td>94.00</td>
<td>58.30</td>
<td>35.73</td>
<td>3.66</td>
</tr>
<tr>
<td>DoB</td>
<td>3.01</td>
<td>9.00</td>
<td>14.05</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 7. Verification accuracy and DoB (%) across different gender subgroups with varying intensities of salt and pepper noise corresponding to the MORPH dataset. The performance of subgroup G2 gets severely affected with increasing intensities of noise. * represents a relatively high disparity in model performance across different subgroups on undistorted images.
subgroup. On the other hand, with increasing intensities of salt and pepper noise, the cluster of subgroup G2 becomes denser compared to G1. Similar observations are obtained for race subgroups, where the cluster of subgroup R1 becomes denser compared to R2. This indicates high misclassification in subgroups G2 and R1, which in turn affects the performance of these subgroups.

We also analyze the salient regions used by the models for recognition under the effect of noise, as shown in Figure 5. It is observed that when Gaussian noise is applied, the region of interest shifts from nose to eyes region. Both these regions are observed to be discriminative in our occlusion experiments, and therefore, minimal performance degradation is observed across different subgroups for the LCNN-29 model. On the other hand, when salt and pepper noise is applied, the model’s focus shifts from nose to eyes for subgroup G1 and to the left side of forehead for subgroup G2. Similarly, among race subgroups, the focus changes from eyes to hair for subgroup R1, and nose to eyes for subgroup R2. As a result, higher performance degradation is observed for G2 and R1 in the presence of salt and pepper noise.

**Does model performance differ across subgroups with varying image resolution?**

Table 8 shows the effect of varying the resolution of the images on the performance of deep models across different gender subgroups. A sharp drop in accuracy is observed beyond $48 \times 48$ resolution for most models. It is also observed that a significant amount of bias is incorporated in predictions of SENet50 and ResNet50 models. As the resolution of the images is reduced to $28 \times 28$, the DoB reaches up to 14.59% and 13.10%, respectively. Similarly, for race subgroups, the DoB increases at lower resolutions for SENet50 and ResNet50 models (Table 13 of supplementary file). A huge performance gap of these models is also observed across the intersectional subgroups (Table 14 of supplementary file). On the other hand, a lesser amount of bias is introduced in the LCNN-29 and ArcFace models.

From Table 8, it is observed that ArcFace significantly degrades the performance at low resolution. On varying the resolution of the images from $48 \times 48$ to $28 \times 28$, the accuracy of ArcFace drops by 89.74% and 81.93% for subgroup G1 and G2, respectively. For the race subgroups, it results in 87.73% and 86.70% degradation for R1 and R2, respectively. This shows the vulnerability of ArcFace for low resolution image recognition. Figure 6 shows the shift in the re-
Figure 6. Visualization of salient regions used by the pre-trained ArcFace model for recognition with varying resolution across (a) gender and (b) race subgroups. For gender subgroups, race is kept constant and vice versa.

region of interest under the effect of resolution. It is observed that at low resolution the model is not able to focus on the facial region. Instead, the focus shifts to non-facial regions such as hair, which in turn leads to poor performance.

5. Discussion on Unbiased Model Predictions in the Presence of Image Distortions

The detailed experimental evaluation performed in this research highlights the idea of gender and racial bias as a consequence of degraded image quality. The results also provide important insights and we believe that the observations drawn from the experimental evaluation can open new research threads. To facilitate research in this direction, we discuss some of the possible solutions in the following subsections to ensure reliable and unbiased model predictions in the presence of real-world image distortions.

5.1. Image Quality Enhancement

Enhancing the quality of the images before providing to the face recognition models may reduce the disparity in model predictions. Generative approaches [17, 29] and denoising techniques [1] can be used to enhance image quality. We assert that matching high-quality images will decrease the performance gap of the model across different subgroups and enhance the overall model performance.

5.2. Improving Generalizability of Deep Models

Training deep models for generalized solutions is an important approach for bias mitigation [2, 30]. In this context, we believe that training deep face recognition models to extract discriminative features from different facial regions instead of focusing on specific facial regions (e.g., eyes, nose, mouth) for recognition can reduce the bias in model prediction. In other words, different facial regions must be equally discriminative for the models during recognition. In our occlusion experiments, we have observed that the nose is the most discriminative region for recognizing subgroup R2. The disparity in the discriminative regions used by models for recognition across different subgroups should be reduced for bias mitigation.

5.3. Utilizing Image Quality during Recognition

The quality of the images should be considered during recognition. In the past, various quality assessment metrics [26] and methods [11, 41] have been proposed to determine the image quality. Recently, the NTIRE challenge organized in CVPR 2021 focused on perceptual image quality assessment [16]. In the literature, researchers have shown the improvement in recognition performance by utilizing the quality score with model predictions [42, 43]. It is our assertion that fusing the quality score with the model prediction will impact the confidence of the model prediction, which may reduce the disparity of the model for recognition across different subgroups.

6. Conclusion

This paper analyzes the interplay and effect of bias and real-world image distortions on the performance of face recognition algorithms. The paper contributes in understanding how seemingly unbiased models produce biased predictions in the presence of real-world image distortions. We observe that eyes, nose, and mask are the most discriminative regions for recognition across race and gender subgroups. However, in the presence of distortions, the regions of interest used by the models shift towards less discriminative regions, thus resulting in unequal performance degradation. For instance, we observe that the models are biased against gender subgroup G2 and race subgroup R1. Moreover, different models introduce different amounts of bias in the predictions, and they largely favor (or disfavor) the same demographic subgroups. We assert that these understandings are important in building deep learning models that are unbiased under different scenarios.

Acknowledgements

P. Majumdar is partly supported by DST Inspire Ph.D. Fellowship. S. Mittal is partially supported by UGC-Net JRF Fellowship. M. Vatsa is partially supported through Swarnajayanti Fellowship. This research is also partially supported by Facebook Ethics in AI award.
References


