On The Nature of Misidentification With Privacy Preserving Algorithms

Sophie Noiret
snoiret@cvl.tuwien.ac.at
Technische Universität Wien
Vienna, Austria

Siddharth Ravi
Universidad de Alicante
Alicante, Spain

Martin Kampel
Technische Universität Wien
Vienna, Austria

Francisco Florez-Revuelta
Universidad de Alicante
Alicante, Spain

ABSTRACT

The ubiquitous use of computer vision and camera surveillance makes it increasingly easy to automatically recognize persons in visuals. In this context, obfuscation methods like blurring and pixelation can impart privacy by preventing facial recognition. But even in cases where these techniques successfully obscure the subject's identity, the question of who is recognized in their stead and what influences this misidentification is still open. As facial recognition is an area which is particularly prone to demographic bias, we analyse misidentifications along the lines of race and gender. We show that persons are most often mistaken for someone of their own gender. However, in terms of racial bias, white people tend to be under-represented among the misidentifications.

KEYWORDS

Fairness, privacy preservation, machine learning

ACM Reference Format:
Sophie Noiret, Siddharth Ravi, Martin Kampel, and Francisco Florez-Revuelta. 2022. On The Nature of Misidentification With Privacy Preserving Algorithms. In *The15th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '22), June 29-July 1, 2022, Corfu, Greece.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3529190.3534760

1 INTRODUCTION

Issues of fairness in facial recognition have been pointed out by previous studies such as Gender Shades [1] and the Face Recognition Vendor Test (FRVT) [7], with systems under-performing on demographics such as women and people of colour. In these studies, the emphasis is placed on error rates (false positive rate and false discovery rates in Gender Shades, false positive rate and false negative rates in FRVT) being higher on those demographics, or on classification accuracy and positive predicted value (Gender Shades) being lower. This assumes that less favourable outcomes are the ones in which people are not being recognized.

PETRA '22, June 29-July 1, 2022, Corfu, Greece

ACM ISBN 978-1-4503-9631-8/22/06…$15.00
https://doi.org/10.1145/3529190.3534760

However, in the context of large-scale video surveillance and the threat to privacy that it poses [5], it can be argued that not being identified - neither correctly nor incorrectly – is the favourable outcome. To do so, visual Privacy Preservation Algorithms (PPAs) can be deployed to protect bodily privacy. PPAs work by perceptually obfuscating sensitive regions of the visual feed to various degrees depending on the context. Some of the simplest and most commonly used visual PPAs include blurring and pixelation.

The specific question of the unequal performance of face recognition or face obfuscation across groups is left to other works such as [2] or [4]. This study focuses on the case where someone's identity has been successfully protected by a face obfuscation technique, in which case we examine the following questions :

- When people from a group are misidentified by facial recognition algorithms, are they misidentified as someone from the same group?
- Does the face obfuscation technique influence the answer to this question?

This work analyses a scenario in which bad actors gain access to unobfuscated images of people, possibly through a data leak, from a specific area being monitored, and where the visual feed is normally one that is privacy protected. They use it to subsequently train facial recognition models that then can be used on the obfuscated visuals coming from the feed.

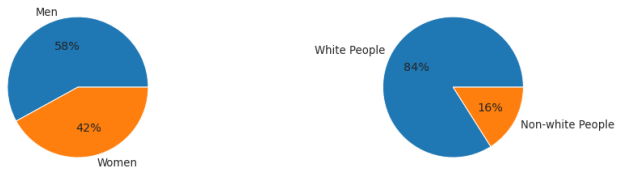
By training a facial recognition system on unobfuscated images and setting it to predict on obfuscated ones, we simulate the scenario under consideration, and subsequently analyse the nature of misidentification. We find that misidentifications happen within one's group with regards to gender. Regarding race, people are predominantly misidentified as white people, but this observation must be tempered by the imbalance in the original dataset. The rest of this paper is structured as follows. Section 2 relates the details of the experiment, the results of which are presented in Section 3. Section 4 concludes this paper, and clarifies some avenues for future work.

2 EXPERIMENT

Face detection is performed using the Histogram of Oriented Gradients (HOG) as implemented in the dlib [3] and face_recognition[1] libraries.

[1] Available at https://github.com/ageitgey/face_recognition

2.1 Dataset



(a) Distribution of subjects by gender **(b) Distribution of subjects by race**

Figure 1: Composition of Dataset

The experiments for this work are conducted on a modified version of the PubFig dataset [6]. The dataset is manually inspected by researchers and label errors are corrected. The original race labels in the PubFig dataset are *White*, *Black*, *Indian* and *Asian*. However, due to extreme label imbalances present in the dataset, the labels are mapped to white (168 people) and non-white (32 people). As for gender, the dataset consists of pictures of 116 men and 84 women. The composition of the dataset can be seen in Fig. 1.

2.2 Pipeline

(a) Original Image **(b) Blurred** **(c) Pixelated**

Figure 2: Original and modified image

For each person in the dataset, 20% of the pictures are randomly selected. On each of these pictures, a face is detected and obfuscated using a PPA. A machine learning classifier is then trained on the remaining 80% of (unobfuscated) images. To do so, face detection is performed and a 128-dimensional vector of face encodings is created. These encodings are then used as training data for a multi-class classifier, with each possible person being a class. This mirrors a scenario in which a data leak leads to a bad actor getting access to some unobfuscated images and uses them to identify people on obfuscated images. The experiment is run twice, once with the face blurred and once with the face pixelated. The effect of the PPAs are illustrated in Fig. 2.

To ensure variety in the machine learning algorithms used to classify facial encodings, classification is performed using Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Naive Bayes (NB), and Multi-Layer Perceptron (MLP) models.

3 RESULTS

The results of the experiment are presented in Table 1 and Table 2. As it is apparent that the nature of the classifier does not influence the general trends, the average results from the use of SVM, KNN, NB, and MLP are also reported.

3.1 Nature of misidentifications across gender

As can be seen in Table 1 and Table 2, persons are predominantly misidentified as persons of the same gender, meaning that men are miscategorised as other men and women as other women. Gender flipped misidentifications happen more for women than men. In blurred and pixelated images, men are mistaken for women in 10.9% and 2.4% of cases respectively, while women are mistaken for men in 12.9% and 16% of cases respectively.

The influence of the face obfuscation technique used is unclear. Using pixelation leads to misidentified men being categorized as other men at a higher rate than blurring (97.6% and 89.1%), but also to misidentified women being categorized as other women at a lower rate than blurring (87.1% and 84%).

3.2 Nature of misidentifications on across race

When misidentified, both white and non-white people are mainly misidentified as white people. This is especially true when using pixelation, which causes an increase of 29.6% of misidentified white people being mistaken for other white people and an increase of 26.8% of misidentified non-white people being mistaken for white people.

However, the imbalance of the dataset must be considered. Were the misidentifications random, the proportions of errors would be the same as the proportions of groups in the original dataset (i.e. 84% of white people and 16% of non-white people). We observe that for white people when using blurring, 72.7% of misidentifications are as other white people. While this is still the majority, it is lower than the 84% expected, which means that they are disproportionately being misidentified as non-white people. On the other hand, when using pixelation as the method of obfuscation, 92.3% are misidentified as other white people, meaning that in that case they are disproportionately being misidentified as other white people. For non-white people, when using either blurring or pixelation, white people are under-represented in the resulting misidentification (54% and 70.8% instead of 86%). Additionally, we observe that the proportion of non-white people among misidentifications is higher when the person being misidentified is not white, indicating that neither of those techniques are fully successful in obfuscating race.

4 CONCLUSION AND FUTURE WORK

This work analyses the nature of misidentification when various machine learning models are set to identify individuals in images obfuscated through two commonly used algorithms - blurring, and pixelation.

Regarding gender, both men and women are predominantly misidentified as people from their own group. This trend is accentuated when pixelation is used for face obfuscation instead of blurring. As for race, the results are less clear-cut. White people are a majority among the misidentifications, but are under-represented when considering the composition of the original dataset. While

Table 1: Nature of misidentification with blurred images

	SVM	KNN	NB	MLP	Average
White people are misidentified as :					
White People	75.3%	77.7%	72.4%	65.5%	72.7%
Non-white People	24.7%	22.3%	27.6%	34.5%	27.3%
Non-white people are misidentified as :					
White people	56.1%	58%	54.5%	47.2%	54%
Non-white people	43.9%	42%	45.5%	52.8%	46%
Men are misidentified as :					
Men	90%	89.9%	89.6%	86.8%	89.1%
Women	10%	10.1%	10.4%	13.2%	10.9%
Women are misidentified as :					
Men	12.7%	12.8%	12.7%	13.4%	12.9%
Women	87.3%	87.2%	87.3%	86.6%	87.1%

Table 2: Nature of misidentification with pixelated images

	SVM	KNN	NB	MLP	Average
White people are misidentified as :					
White people	92.9%	91.3%	93.1%	91.8%	92.3%
Non-white people	7.1%	8.7%	6.9%	8.2%	7.7%
Non-white people are misidentified as :					
White people	71.6%	68.9%	73.1%	69.5%	70.8%
Non-white people	28.4%	31.1%	26.9%	30.5%	29.2%
Men are misidentified as :					
Men	98.3%	96.7%	98.3%	97.1%	97.6%
Women	1.7%	3.3%	1.7%	2.9%	2.4%
Women are misidentified as :					
Men	16.7%	16.2%	15.3%	15.9%	16%
Women	83.3%	83.7%	84.7%	84.1%	84%

non-white people are disproportionately misidentified as people from their own group when using either pixelation or blurring, the proportions of white people among incorrect predictions of other white people is superior to the dataset proportions when using pixelation and inferior when using blurring.

Although this work analyses misidentification using the lens of blurring and pixelation algorithms, it remains to be seen as to whether the same patterns of misidentification repeat for other privacy preserving algorithms as well. Further analysis is also required to understand the reasons behind why misidentifications occur in the way in which it was observed.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union's H2020 research and innovation programme under grant agreement No. 861091 and FFG Grant 878730. The publication reflects the views only of the authors, and the European Union cannot be held responsible for any use which may be made of the information contained therein.

REFERENCES

[1] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[2] Raul Vicente Garcia, Lukasz Wandzik, Louisa Grabner, and Joerg Krueger. 2019. The Harms of Demographic Bias in Deep Face Recognition Research. In *2019 International Conference on Biometrics (ICB)*. 1–6. https://doi.org/10.1109/ICB45273.2019.8987334

[3] Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.

[4] Anoop Krishnan, Ali Almadan, and Ajita Rattani. 2020. Understanding Fairness of Gender Classification Algorithms Across Gender-Race Groups. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 1028–1035. https://doi.org/10.1109/ICMLA51294.2020.00167

[5] Jean Kumagai and Steven Cherry. 2004. Society: Sensors &; Sensibility. *IEEE Spectr.* 41, 7 (jul 2004), 22–28. https://doi.org/10.1109/MSPEC.2004.1309801

[6] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. 2009. Attribute and Simile Classifiers For Face Verification. In *2009 IEEE 12th International Conference on Computer Vision*. 365–372. https://doi.org/10.1109/ICCV.2009.5459250

[7] P. Jonathon Phillips, Patrick Grother, Ross Micheals, Duane M. Blackburn, Elham Tabassi, and Mike Bone. 2003. Face Recognition Vendor Test 2002. In *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG '03)*. IEEE Computer Society, USA, 44.