Paper 2646

# Facial Expressions as a Vulnerability in Face Recognition

**Alejandro Peña, Aythami Morales, Ignacio Serna, Julian Fierrez**

BiDA Lab, Universidad Autónoma de Madrid, Spain

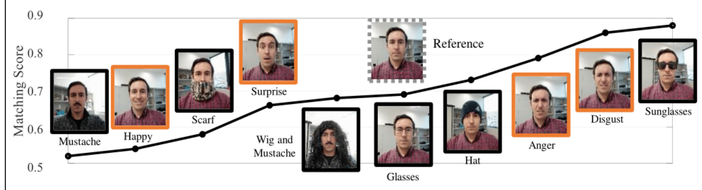
**Agata Lapedriza**

Computer Science Dpt./ eHealth Center
Universitat Oberta de Catalunya, Spain

## Abstract

This work explores **facial expression bias** as a security **vulnerability** of face recognition systems. Despite the great performance achieved by state-of-the-art face recognition systems, the algorithms are still **sensitive** to a large range of covariates. This work presents a comprehensive analysis of how facial expression bias impacts the performance of face recognition technologies. Our study analyzes: i) facial expression biases in the most popular face recognition databases; and ii) the impact of facial expression in face recognition performances. Our experimental framework includes **two face detectors**, **three face recognition models**, and **three different databases**. Our results demonstrate a huge facial expression bias in the most widely used databases, as well as a related **impact** of face expression in the performance of state-of-the-art algorithms. This work opens the door to new research lines focused on mitigating the observed vulnerability

## Contributions




- We study facial expressions as a vulnerability of face recognition systems.
- Our work focus on facial expressions related to 6 basic human emotions (Happy, Sad, Anger, Surprised, Disgusted and Fearful) plus Neutral expressions.
- We conducted experiments in both authentication and identification setups, and analyze the facial expressions bias in commonly used face datasets.

## Data and Methods

- We used 3 different state-of-the-art face recognition models in our experiments:
  (a) VGG16 [1] model, pretrained with VGGFace2, and the MTCNN detector.
  (b) ResNet-50 [2] model, pretrained on VGGFace2, and the MTCNN detector.
  (c) LResNet100E-IR [3] model, pretrained with MS-Celeb using ArcFace loss, and the RetinaFace detector.
- The three models were evaluated using 3 public databases: CFEE [4], CK+ and CelebA. While CelebA is a large-scale database collected using search engines, both CFEE and CK+ were collected in a controlled environment of illumination, distance and pose, which make them ideal to analyze the isolated effect of facial expressions.

## Key references

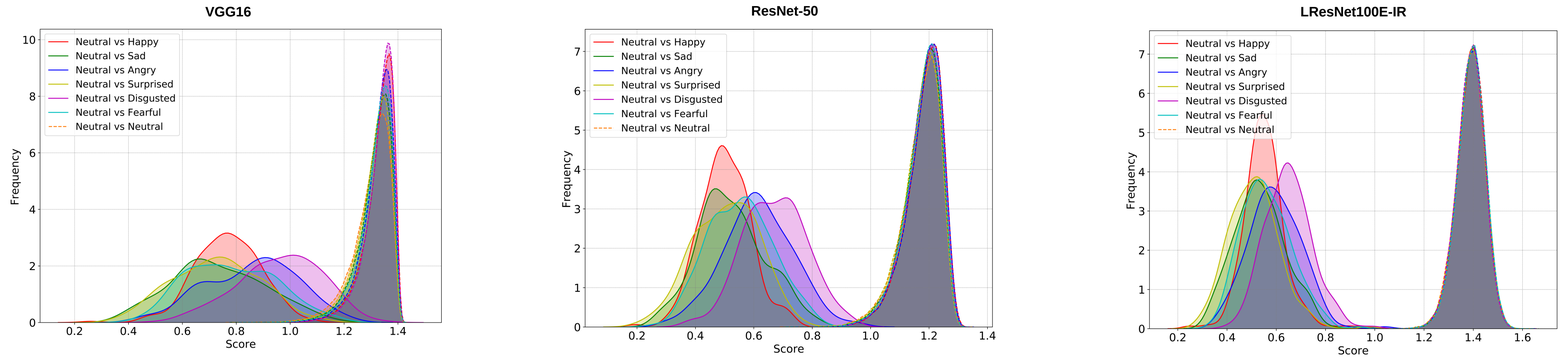
[1] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, in *International Conference on Learning Representations (ICLR),* 2015.

[2] K. He, X. Zhang, S. Ren and J. Sun, “Deep Residual Learning for Image Recognition”, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2015.

[3] J. Deng, J. Guo, N. Xue and S. Zafeiriou, “ArcFace: Additive Angular Margin Loss for Deep Face Recognition”, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2019.

[4] S. Du, Y. Tao and A. Martinez, “Compound Facial Expressions of Emotion”, in *Proceedings of the National Academy of Sciences,* 2014.

## Impact on the identification accuracy




- We used CFEE [4] and the three face matchers to extract both **genuine** and **impostor** score distributions using diferent **face expressions** as reference. We can observe that the genuine distributions are clearly influenced by facial expressions, while impostor distributions barely change across expresions in all three cases.
- This impact in the genuine distributions may suppose a **vulnerability**, as it may influence the probabilities to be identified just by changing the facial expression.
- We designed an **identification** experiment with the subjects of CFEE and CK+, adding CelebA images to the background set, where we selected subject images belonging to a specific facial expression class as query samples, and extract **Rank-1 statistics** individually for all the other facial expressions.
- The Rank-1 results show the same trend seen in the genuine score distributions, with some facial expressions having better affinity than others. While these differences in performance are more pronounced in both ResNet [1] and VGG [2] models, the state of the art LResNet100E-IR [3] model also suffers from this effect, despite the **high quality standards** of the CFEE and CK+ images.

| Average Genuine Score $_{\text{Rank-1 in \%}}$ (Reference = Neutral) | | | | | | |
|---|---|---|---|---|---|---|
| Method | Happy | Sad | Anger | Surprised | Disgusted | Fearful |
| VGG16 | $.77_{96.0}$ | $.73_{90.3}$ | $.86_{77.3}$ | $.73_{89.4}$ | $.96_{50.3}$ | $.78_{84.2}$ |
| ResNet-50 | $.50_{100}$ | $.52_{100}$ | $.61_{97.1}$ | $.50_{100}$ | $.67_{94.4}$ | $.55_{98.8}$ |
| LResNet100E-IR | $.55_{100}$ | $.53_{100}$ | $.59_{100}$ | $.51_{99.7}$ | $.65_{99.3}$ | $.56_{100}$ |
| Average Genuine Score $_{\text{Rank-1 in \%}}$ (Reference = Happy) | | | | | | |
| Method | Neutral | Sad | Anger | Surprised | Disgusted | Fearful |
| VGG16 | $.77_{96.9}$ | $.87_{84.3}$ | $.95_{60.0}$ | $.88_{76.4}$ | $.97_{55.2}$ | $.86_{85.8}$ |
| ResNet-50 | $.50_{100}$ | $.62_{97.6}$ | $.67_{97.3}$ | $.62_{97.6}$ | $.68_{93.0}$ | $.60_{98.4}$ |
| LResNet100E-IR | $.55_{100}$ | $.65_{99.6}$ | $.68_{99.6}$ | $.64_{99.6}$ | $.69_{99.3}$ | $.63_{100}$ |
| Average Genuine Score $_{\text{Rank-1 in \%}}$ (Reference = Sad) | | | | | | |
| Method | Happy | Neutral | Anger | Surprised | Disgusted | Fearful |
| VGG16 | $.87_{72.4}$ | $.73_{94.2}$ | $.77_{86.8}$ | $.84_{79.1}$ | $.92_{56.4}$ | $.74_{92.1}$ |
| ResNet-50 | $.62_{97.2}$ | $.52_{99.5}$ | $.56_{98.0}$ | $.61_{97.6}$ | $.66_{89.2}$ | $.54_{99.3}$ |
| LResNet100E-IR | $.65_{99.2}$ | $.54_{100}$ | $.55_{99.6}$ | $.60_{99.6}$ | $.65_{99.2}$ | $.56_{100}$ |
| Average Genuine Score $_{\text{Rank-1 in \%}}$ (Reference = Anger) | | | | | | |
| Method | Happy | Sad | Neutral | Surprised | Disgusted | Fearful |
| VGG16 | $.95_{54.1}$ | $.77_{88.0}$ | $.86_{80.4}$ | $.96_{44.4}$ | $.87_{69.1}$ | $.90_{78.4}$ |
| ResNet-50 | $.67_{95.3}$ | $.56_{97.1}$ | $.61_{97.4}$ | $.71_{86.6}$ | $.62_{83.5}$ | $.65_{95.6}$ |
| LResNet100E-IR | $.68_{100}$ | $.55_{100}$ | $.59_{100}$ | $.66_{99.7}$ | $.60_{99.2}$ | $.63_{100}$ |
| Average Genuine Score $_{\text{Rank-1 in \%}}$ (Reference = Surprised) | | | | | | |
| Method | Happy | Sad | Anger | Neutral | Disgusted | Fearful |
| VGG16 | $.88_{62.1}$ | $.84_{74.7}$ | $.96_{43.7}$ | $.72_{81.0}$ | $.99_{97.3}$ | $.73_{66.0}$ |
| ResNet-50 | $.62_{95.1}$ | $.61_{95.6}$ | $.71_{87.0}$ | $.50_{96.5}$ | $.71_{83.5}$ | $.52_{90.5}$ |
| LResNet100E-IR | $.64_{99.7}$ | $.60_{100}$ | $.66_{99.6}$ | $.51_{99.9}$ | $.70_{99.3}$ | $.53_{100}$ |
| Average Genuine Score $_{\text{Rank-1 in \%}}$ (Reference = Disgusted) | | | | | | |
| Method | Happy | Sad | Anger | Surprised | Neutral | Fearful |
| VGG16 | $.97_{41.9}$ | $.92_{53.9}$ | $.87_{65.5}$ | $.99_{35.8}$ | $.96_{40.4}$ | $.95_{43.0}$ |
| ResNet-50 | $.68_{86.0}$ | $.66_{85.0}$ | $.62_{93.5}$ | $.71_{77.8}$ | $.67_{90.3}$ | $.67_{83.4}$ |
| LResNet100E-IR | $.69_{99.3}$ | $.65_{99.6}$ | $.60_{99.2}$ | $.70_{99.3}$ | $.65_{99.7}$ | $.67_{100}$ |
| Average Genuine Score $_{\text{Rank-1 in \%}}$ (Reference = Fearful) | | | | | | |
| Method | Happy | Sad | Anger | Surprised | Disgusted | Neutral |
| VGG16 | $.86_{75.2}$ | $.74_{91.1}$ | $.90_{61.9}$ | $.73_{88.0}$ | $.95_{52.9}$ | $.78_{83.0}$ |
| ResNet-50 | $.60_{98.8}$ | $.54_{99.2}$ | $.65_{92.5}$ | $.52_{98.4}$ | $.67_{87.6}$ | $.55_{98.6}$ |
| LResNet100E-IR | $.63_{99.2}$ | $.56_{99.6}$ | $.63_{99.2}$ | $.53_{99.6}$ | $.67_{99.2}$ | $.56_{99.4}$ |

## Face expression bias in face datasets

- We use the COTS Affectiva to analyze the most popular face databases used to train current deep face recognition systems.
- The most used face datasets are biased towards Neutral and Happy expressions, which may result in models with heterogenous performance across expressions.

| Database | #Images | Neutral | Happy | Sad | Anger | Surprised | Disgusted | Fearful |
|---|---|---|---|---|---|---|---|---|
| MS-Celeb-1M | 8.5 M | 83.7% | 5.7% | 0.2% | 3.4% | 2.2% | 4.6% | ~0.0% |
| MegaFace | 4.7 M | 82.0% | 4.5% | 0.1% | 7.0% | 1.3% | 5.0% | ~0.0% |
| YTF | 621 K | 81-5% | 8.3% | 0.4% | 0.7% | 1.6% | 5.9% | ~0.0% |
| CASIA | 500 K | 64.5% | 30.4% | 0.1% | 0.4% | 1.9% | 1.2% | ~0.0% |
| CelebA | 203 K | 62.2% | 33.3% | 0.1% | 0.5% | 1.6% | 0.9% | ~0.0% |
| IJB-C | 21 K | 66.2% | 26.9% | 0.1% | 0.6% | 2.6% | 2.0% | 0.1% |
| Age-DB | 16 K | 76.7% | 16.1% | 0.4% | 1.5% | 1.9% | 1.6% | ~0.0% |
| LFW | 13 K | 61.2% | 28% | 0.3% | 1.8% | 3.1% | 4.4% | ~0.0% |
| VGGFace2 | 3.3 M | 64.5% | 28.2% | 0.2% | 0.4% | 3.3% | 2.0% | 0.1% |

## Conclusions

The main findings of this work can be summarized as follows:

1) The most popular face recognition databases **systematically** present huge **facial expression biases**.
2) The facial expression bias affects the performance of **genuine comparisons** with variations in the scores of up to 40 %.
3) Facial expression bias has no impact in the **impostor comparisons**.

As a result of this work, we strongly advocate for reducing the facial expression bias in future face recognition databases, and further development of bias-reduction methods applicable to existing databases and existing models already trained on biased datasets.