IDEMIA PUBLIC SECURITY

Technical Report

OFIQ Evaluation

Demographic factors

Stéphane Gentric

Chief AI Scientist
Research and Technology Unit
IDEMIA

CONTEXT

OFIQ is a public set of algorithms assessing quality of face images. It has been released in Q2 2024 and is proposed as reference implementation for an ISO standard for image quality assessment.

In the context of regulations recently passed in major geographies (namely and for example, the European Union Artificial Intelligence Act adopted on May 21st , 2024 and the United States President's Executive Order 14110 issued on October 30th, 2023) underlining the importance of the consideration and mitigation of potential bias in the deployment of artificial intelligence systems, it seemed important to the author to provide a quantified, fact-based assessment of OFIQ with respect to demographic effects.

The evaluated version is OFIQ 1.0.0 downloaded March 23th 2024

In this report we have evaluated OFIQ on different internal datasets:

- 8 medium sized datasets representing various scenarios.
- 1 large dataset including 4 million of Identities (8 million images).

This analysis aims to assess the fairness of OFIQ algorithms regarding two demographic factors: gender and skin tone.
On the 8 datasets, gender and skin tone are estimated automatically.
On the large dataset, gender and country of origin are extracted from passport information (ground truth).

EXECUTIVE SUMMARY

The unified quality shows bias regarding gender. As this unified quality leads the global quality assessment it will trigger at least twice as many discards for female subjects than for male subjects.
Other qualities also show bias including all illumination related qualities, leading to more than 20% of discard on darker skin subjects than on lighter skin subjects, at a threshold set for 1% on lighter skin.

As a result of these findings, we recommend that all deployers and users of OFIQ, in its quality of reference implementation of an ISO standard, be made aware of these results, so that they can implement in their application integrating OFIQ the appropriate measures to mitigate demographic effects.

Table of contents

Methodology

This evaluation is divided into two parts.
The first part is performed on datasets covering various scenario and image types. The goal is to check if there are consistent performance differences between demographic groups.
As no ground truth is available on these datasets, group assignments are done automatically with algorithms (gender estimation and skin tone estimation)
The second part is performed on a large dataset. The goal is to reduce the measurements uncertainty and check the performance at lower discard rates.
Here we have ground truth for genders and nationalities

Robusteness on various datasets

Topic description

When differences in discard rates are observed for a quality component, on different groups in a dataset, this can have two possible root causes: It can be linked to a bias the algorithm or to characteristics or behaviors of the group in this specific dataset.
To address this ambiguity, we compute discard rates on various datasets coming from different scenario. When a difference between discard rates is consistently observed on all datasets, we can consider that the algorithm is the root cause and call this bias.

Regarding fairness in face recognition, the most sensitive topics are discrimination toward females and discrimination toward darker skin. For that, we focus the analysis to assess these biases, setting the threshold to have a given discard rate on the other group (resp male and lighter skin).
The analysis is done independently for each factor.

Datasets description

We use 8 different image types coming from 5 different origins.
A – airport e-gate A1=Passport A2=Live
B – airport e-gate B1=Passport B2=Live
C – seaport e-gate C1=Passport C2=Live
D – mugshot
E – wild
All these datasets include around 10k face images.

All these datasets are of good quality and generated successful biometric recognition tasks. Nevertheless, as biometric algorithms are robust to a lot of criteria, it may happen that some images are not fully compliant with quality requirements defined in ISO/IEC, while still allowing successful biometric recognition tasks. However, for bias evaluation we can assume that false discard rates are correlated to discard rates.

Results

Table 1 shows the discard rate of each quality component, for female, when we set a threshold to have 1% of discard for male.

	GENDER									
	A1	A2	B1	B2	C1	C2	D	E	Mean	Min
UnifiedQualityScore	3,1%	2,5%	2,2%	3,1%	2,0%	1,8%	2,7%	1,7%	**2,3%**	1,67%
ExpressionNeutrality	1,6%	3,3%	2,5%	1,3%	3,0%	2,6%	3,4%	0,8%	**2,1%**	0,77%

Table 1 : Fairness on gender factor.

Table 2 shows the discard rate of each quality component, for darker skin, when we set a threshold to have 1% of discard for lighter skin.

	SKINTONE									
	A1	A2	B1	B2	C1	C2	D	E	Mean	Min
LuminanceMean	15,2%	10,6%	12,2%	7,2%	10,7%	23,3%	13,5%	17,5%	**13,0%**	7,22%
DynamicRange	4,8%	9,3%	4,7%	8,2%	5,4%	16,9%	8,6%	11,4%	**7,9%**	4,66%
LuminanceVariance	3,5%	5,3%	4,4%	5,2%	4,8%	13,3%	5,6%	6,1%	**5,6%**	3,48%
CompressionArtifacts	1,7%	1,1%	1,8%	2,2%	4,4%	0,7%	0,8%	1,2%	**1,5%**	0,67%

Table 2: Fairness on skin tone factor

These tables show that OFIQ algorithms,"unified_quality" and "expression neutrality" have bias to the detriment of female.
"Luminance", "dynamic range" severely underperform on darker skin.

Fairness on a large operational dataset

Qualities have been computed on a large dataset with 8 million of images, with passport country of origin and gender labels.
These images come from an operational system with two images per subject. One comes from a live acquisition and the other one comes from the chip of an ICAO compliant passport.

The paragraph studies the OFIQ unified quality score.

Results on gender factor

Figure 3 shows discard rate per country (6 countries with mainly Asian people and 7 countries with mainly Caucasian people). We can see a factor of up to 4 in discard rates when comparing Caucasian male with Asian female rates.

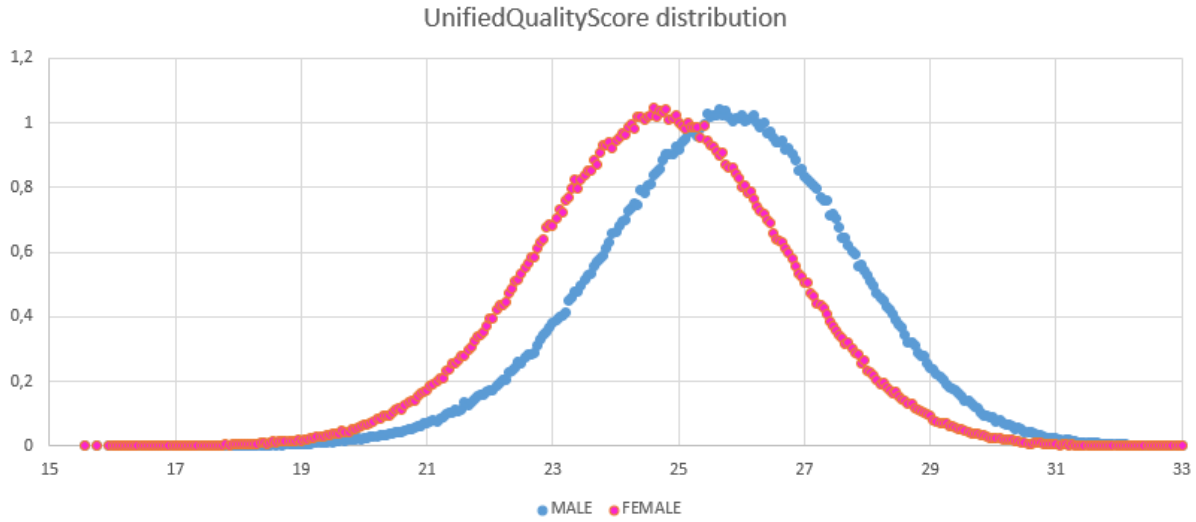

Figure 1: quality distribution

Figure 1 shows a global shift between the score distribution of Male and Female.

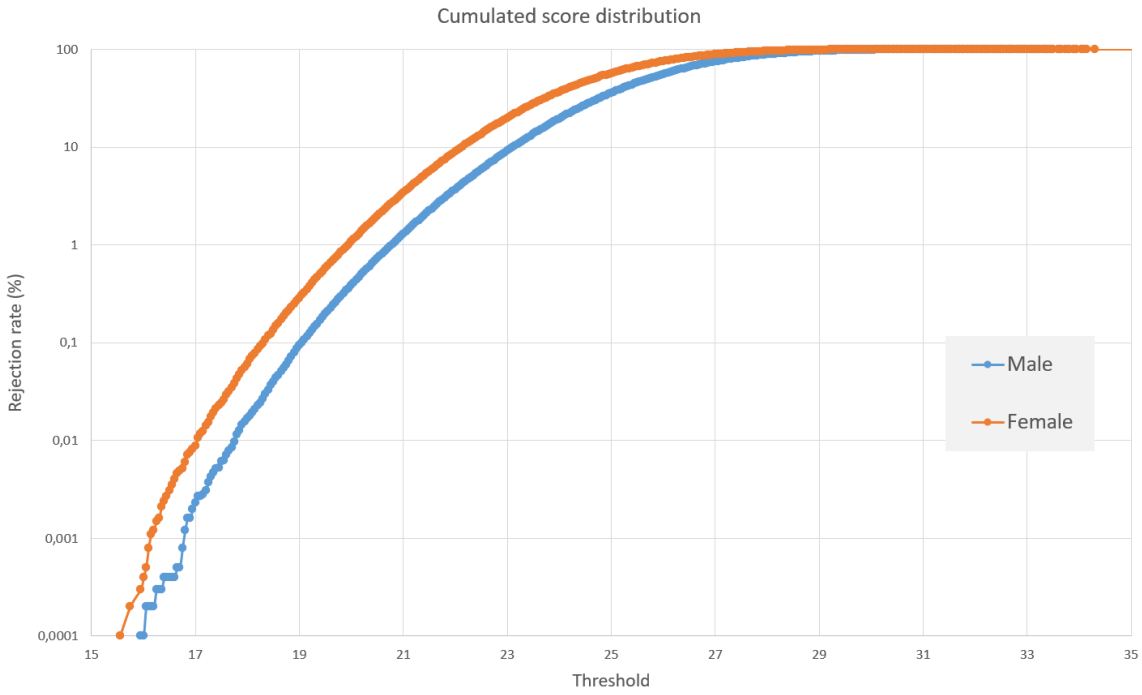

Figure 2: cumulated distribution

On Figure 2, we have

- 2.7% of discard on Female for 1% on Male
- 0.31% of discard on Female for 0.1% on Male.

OFIQ triggered around 3 times more discards on female than on male.

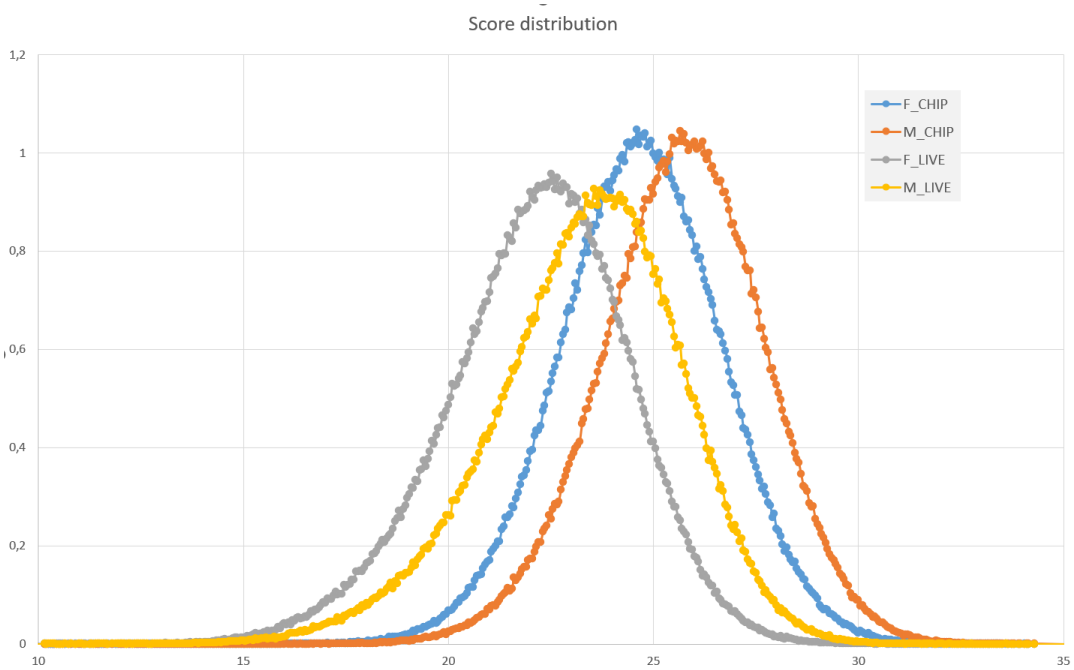

Figure 3: Score distribution by gender and image type

Figure 3 shows the unified quality score distribution for gender on live acquisitions as well as for passport images. As expected, the quality is lower for live acquisition than for passport images. The gender bias remains the same for both image type.

Results per passport country of origin

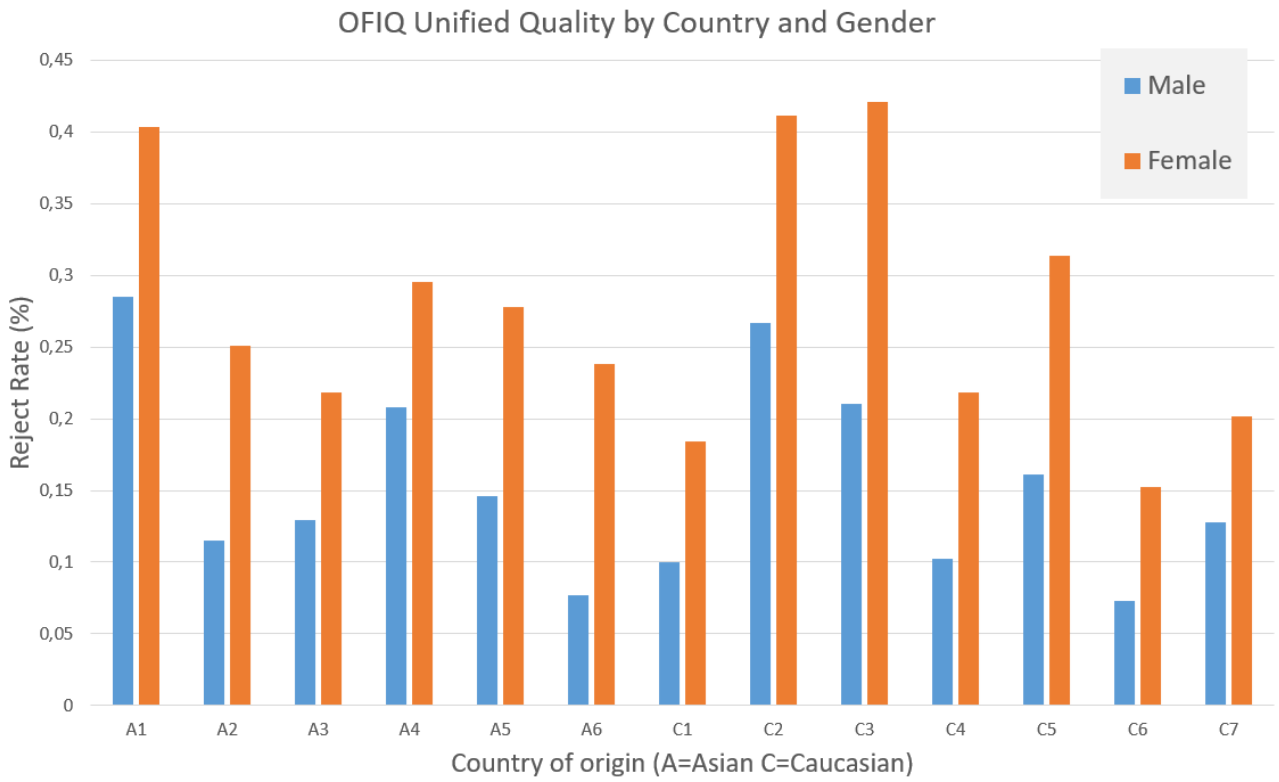

Figure 4: Discard rate (in%) per gender + passport country of origin.

Figure 4 shows discard rate for different country of origin. Countries are named by the majority ethnicity in that country. The threshold is set for 0.1% discard for male in C1, the most populated country.
There are not enough people with passport from African countries in this dataset to assess what happens with darker skin. There are no systematic differences in discard rates between Asian and Caucasian countries.
However, the gender bias is confirmed for all country.