



da/sec

BIOMETRICS AND INTERNET-SECURITY
RESEARCH GROUP



ATHENE

National Research Center
for Applied Cybersecurity

A Comparative Analysis of Demographic Variability Metrics: Face Image Quality Assessment in Operational Settings

A. Dörsch¹, C. Rathgeb^{1,3}, S. Utcke², S. Engelhardt², S. Kramp², T. Schlett¹, J. Merkle³, B. Tams³, O. Henniger⁴, C. Busch¹

¹da/sec – Biometrics and Security Research Group, Hochschule Darmstadt

²DERMALOG Identification Systems GmbH, Germany

³secunet Security Networks AG, Germany

⁴Fraunhofer Institute for Computer Graphics Research IGD

IEEE/IAPR IJCB 2026

When Biometrics Break (WBB): Security Implications of Performance Degradation, Misclassification, and Operational Drift

Motivation

- ▶ Recent studies have reported performance differences across face image quality assessment (FIQA) algorithms
 - ▶ Error rates may vary depending on the demographic characteristics of the captured biometric sample
 - ▶ Such disparities result in unequal treatment, formally referred to as *demographic differentials* (DDs)
- ▶ A quantitative comparison of DD metrics proposed for inclusion in ISO/IEC 29794-1 was conducted
 - ▶ The suitability of these DD metrics for operational border-control scenarios was evaluated
 - ▶ Evaluation was performed on a large-scale operational dataset comprising more than 100,000 facial images acquired for machine-readable travel documents (MRTD)
- ▶ Experimental results show that DD metrics solely relying on a single quality threshold may yield inconsistent results due to local threshold sensitivity

Background

- ▶ Biometric sample quality can be measured using two distinct concepts
 - ▶ Unified Quality (Unified Quality Score (UQS))
 - ▶ Component Quality Measure (CQM)
- ▶ OFIQ framework is an open-source FIQA software library
 - ▶ Reference implementation of ISO/IEC 29794-5
 - ▶ Consists of 27 CQMs, as well as a UQS

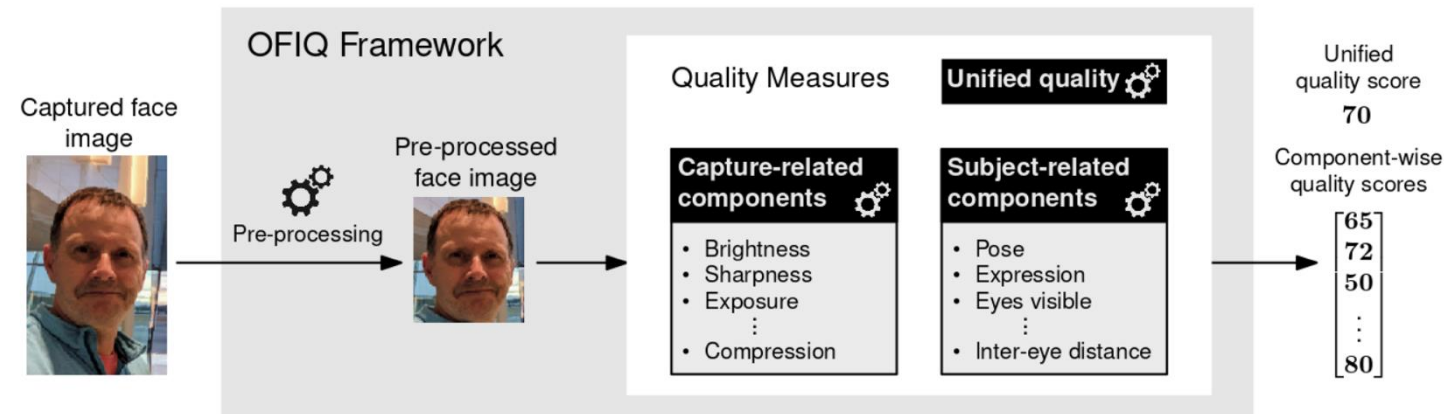
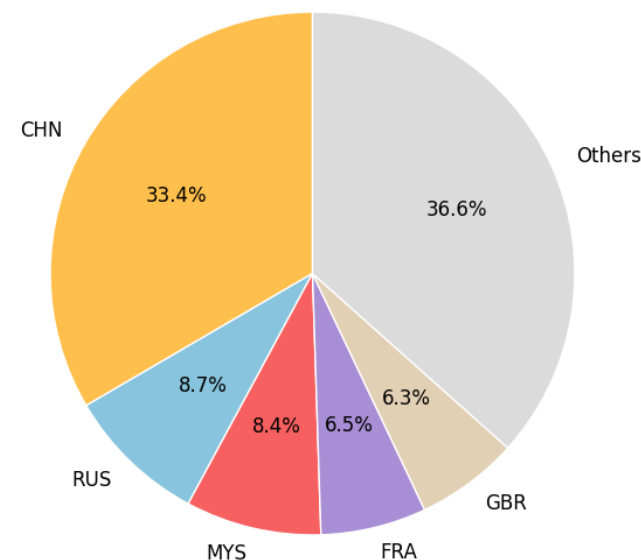


Image-Source: OFIQ-Public-Report: https://github.com/BSI-OFIQ/OFIQ-Project/blob/main/doc/reports/Public_Report_V1.2_2024_11_06.pdf

Operational Dataset

- ▶ The dataset used in this work was introduced by Utcke et al.
- ▶ Comprises biometric face image pairs acquired in operational border-control scenarios at multiple international checkpoints
- ▶ Each pair consists of a reference image extracted from the secure chip of an electronic passport and a mated probe image
- ▶ Only reference images are considered in this evaluation

Gender		Age Group				
Female	Male	0-20	20-30	30-40	40-50	50+
57,133	52,391	5,169	24,576	27,878	18,710	33,191

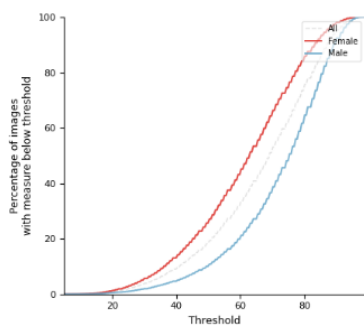


Demographic Variability Measures

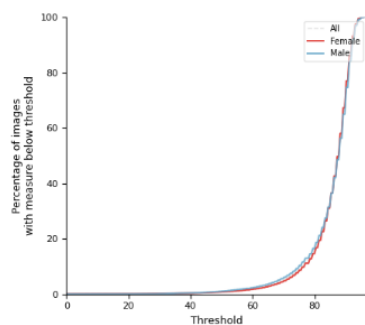
- ▶ **Failure-to-Acquire (FTA-DD)**
 - ▶ Quantifies demographic variability at a fixed operational threshold
 - ▶ The failure-to-acquire rate (FTAR) is computed as the proportion of samples below the selected threshold
- ▶ **Mean-Discard-Gap (MDG-DD)**
 - ▶ Quantifies demographic variability across a range of (percentile-based) quality thresholds
 - ▶ Averages the maximum group-wise discard-rate gap across thresholds
- ▶ **Low-Weighted-Mean (LWM-DD)**
 - ▶ Quantifies demographic variability by emphasising lower-quality samples
 - ▶ Assigns greater weight to lower quality scores
- ▶ **Normalized-Mean-Differences (NMD-DD)**
 - ▶ Quantifies demographic variability based on differences between group-wise mean quality scores
 - ▶ Normalizes the differences between the lowest and highest group mean by their variability



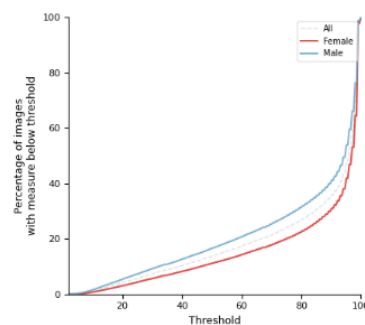
Cumulative distribution plots for gender



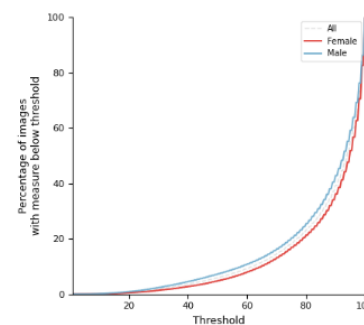
(a) UQS



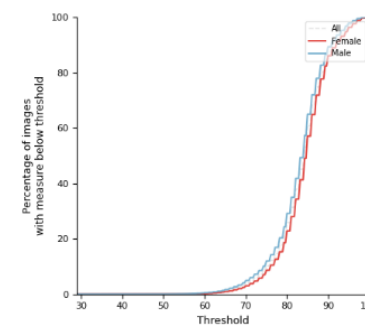
(b) Ill. Uni.



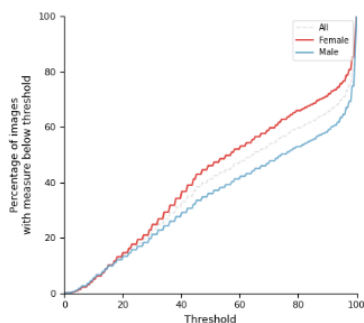
(c) Lum. Mean



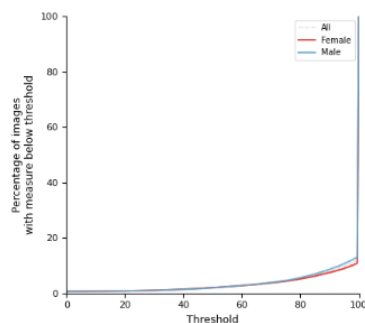
(d) Lum. Var.



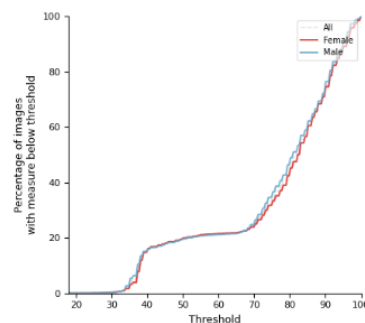
(e) Dyn. Range



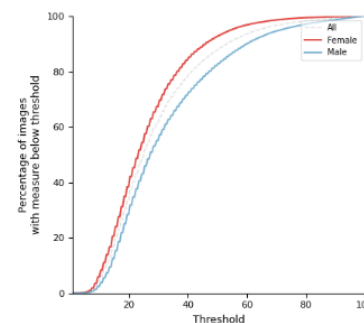
(f) Sharpness



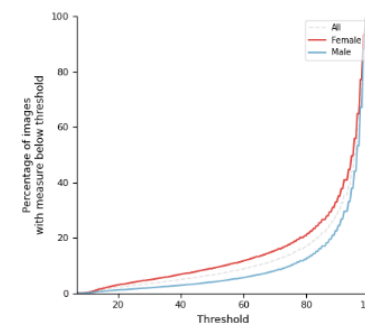
(g) Natural Colour



(h) Inter Eye Dist.



(i) Head Size



(j) Expr. Neutral.

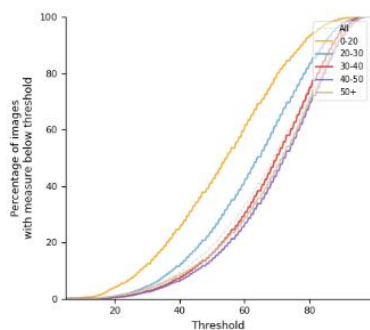
Demographic differential values for gender

Quality Measure	FTA-DD	NMD-DD	LWM-DD	MDG-DD	MDG-DD (0.1th–15th pctile)
UnifiedQualityScore	0.363 / 13	107.336	0.083	0.115	0.044
IlluminationUniformity	0.089 / 25	4.607	0.006	0.005	0.005
LuminanceMean	0.699 / 5	36.525	0.046	0.056	0.033
LuminanceVariance	0.553 / 10	23.905	0.029	0.026	0.017
DynamicRange	0.577 / 57	32.065	0.008	0.018	0.015
Sharpness	0.233 / 2	38.328	0.005	0.076	0.005
NaturalColour	– / –	3.649	0.032	0.003	0.003
InterEyeDistance	0.200 / 27	5.603	0.000	0.010	0.008
HeadSize	0.514 / 4	62.177	0.076	0.064	0.032
ExpressionNeutrality	0.283 / 10	45.825	0.083	0.054	0.043

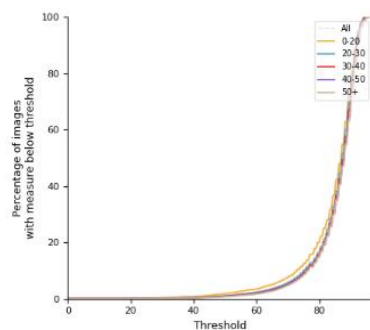
Gender Evaluation Results

- ▶ Several OFIQ CQM show demographic variability trends across gender
 - ▶ Cumulative curve separations (consistent discard trends) are observable over a broad range of quality thresholds
- ▶ FTA-DD shows sensitivity to local threshold effects
 - ▶ Single threshold evaluation may lead to inconsistent interpretations
 - ▶ Dynamic Range: FTA-DD = 0.577 vs. LWM-DD = 0.008 / MDG-DD = 0.018 despite nearly overlapping cumulative distributions
- ▶ Global DD metrics provide more consistent results with the cumulative-distribution trends
 - ▶ LWM-DD and MDG-DD better reflect the cumulative-distribution curves
 - ▶ NWM-DD, LWM-DD, and MDG-DD consistently quantify UQS as showing the strongest gender differential

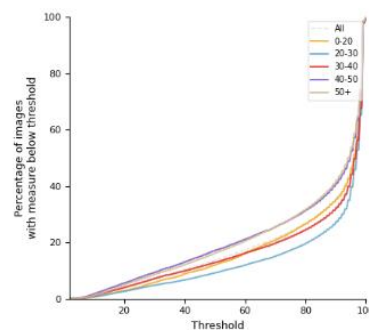
Cumulative distribution plots for age



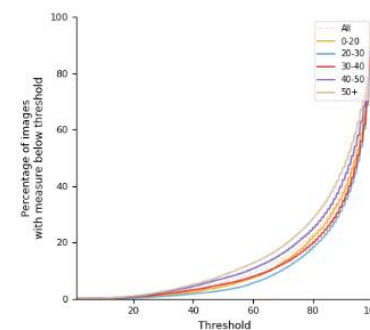
(a) UQS



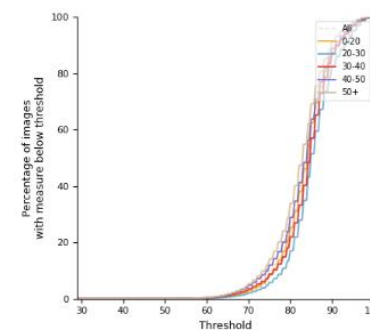
(b) Ill. Uni.



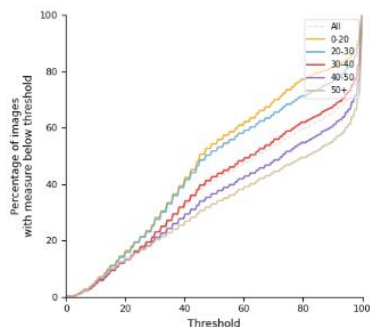
(c) Lum. Mean



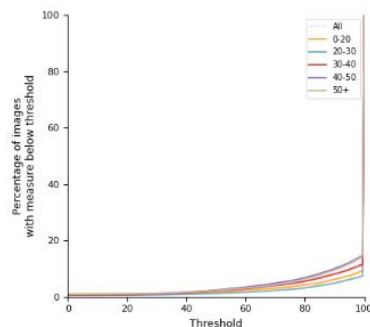
(d) Lum. Var.



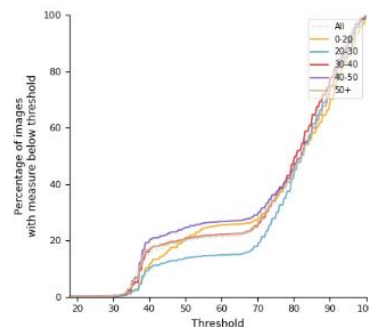
(e) Dyn. Range



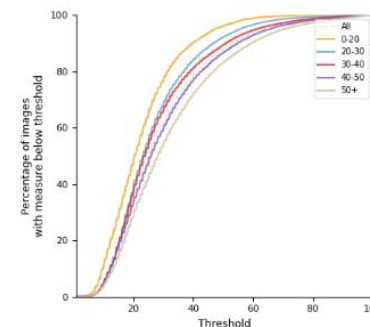
(f) Sharpness



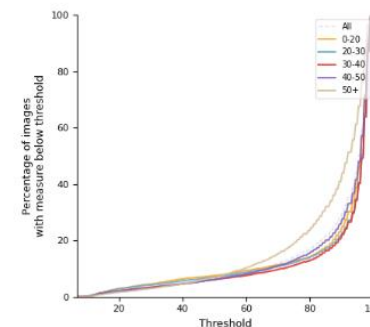
(g) Natural Colour



(h) Inter Eye Dist.



(i) Head Size



(j) Expr. Neutral.

Demographic differential values for age

Quality Measure	FTA-DD	NMD-DD	LWM-DD	MDG-DD	MDG-DD (0.1th–15th pctl)
UnifiedQualityScore	0.727 / 13	47.754	0.059	0.170	0.105
IlluminationUniformity	0.206 / 25	10.372	0.010	0.019	0.016
LuminanceMean	0.264 / 5	15.868	0.039	0.075	0.045
LuminanceVariance	0.258 / 10	17.819	0.033	0.056	0.038
DynamicRange	0.262 / 57	24.930	0.009	0.042	0.032
Sharpness	0.097 / 2	27.342	0.010	0.158	0.013
NaturalColour	– / –	8.355	0.041	0.020	0.020
InterEyeDistance	0.208 / 27	11.839	0.028	0.069	0.030
HeadSize	0.436 / 4	35.823	0.061	0.099	0.059
ExpressionNeutrality	0.309 / 10	13.744	0.033	0.060	0.024

Age Evaluation Results

- ▶ Several OFIQ CQM show demographic variability trends across age
 - ▶ Cumulative curve separations between younger and older subjects are observable across a broad range of quality thresholds
- ▶ FTA-DD shows sensitivity to local threshold effects
 - ▶ Sharpness: FTA-DD=0.097 despite showing one of the strongest separations in the cumulative distributions
- ▶ Global DD metrics provide more consistent results with the cumulative-distribution trends
 - ▶ UQS and Sharpness show the strongest overall age-related DDs
 - ▶ MDG-DD quantifies UQS and Sharpness as the two CQMs with the highest DDs
 - ▶ LWM-DD reports a less prominent gap for Sharpness due to its greater weighting of lower quality scores
 - ▶ NMD-DD generally aligns with MDG-DD and LWM-DD

Conclusion

- ▶ Demographic differentials are observable in operational FIQA across both gender and age groups
- ▶ DD evaluation should consider a broad range of operational quality thresholds rather than isolated thresholds
- ▶ Single-threshold metrics such as FTA-DD are sensitive to local threshold effects and may overemphasize or report no demographic variability
- ▶ Global DD metrics provide a more robust representation of the overall demographic variability trend



da/sec

BIOMETRICS AND INTERNET-SECURITY
RESEARCH GROUP

A Comparative Analysis of Demographic Variability Metrics



ATHENE

National Research Center
for Applied Cybersecurity

Thank you for your attention!

Questions?