EDC Considerations

Corresponding paper (under review):
"Considerations on the Evaluation of Biometric Quality Assessment Algorithms"

Torsten Schlett, Christian Rathgeb, Juan Tapia, Christoph Busch

da/sec - Biometrics and Security Research Group
Hochschule Darmstadt

2023-11-08

Overview

Introduction

Error versus Discard Characteristic (EDC):

- Standardised in the next edition of ISO/IEC 29794-1.
- Previously more commonly known as the "Error versus Reject Characteristic" (ERC).
- Used to evaluate quality assessment (QA) algorithms.
  (Not just for face image QA, but following examples use face image data.)
- Usually involves multiple QA algorithms and one recognition system.

EDC computation:

- A comparison score per sample pair is computed by the recognition system.
- A quality score (QS) per sample is computed by each QA algorithm.
  (In this presentation higher QS values are meant to imply higher biometric utility.)
- An error value is computed as images/comparisons are discarded in order of the QSs.
  (In this presentation the False Non-Match Rate, FNMR, is used.)

Introduction

Face image experiments use one face detector, one recognition system, and five QA models:

- Face detector model: **RetinaFace-R50**
  - Images are excluded if the face detection step fails.
  - Facial landmarks are used for preprocessing.

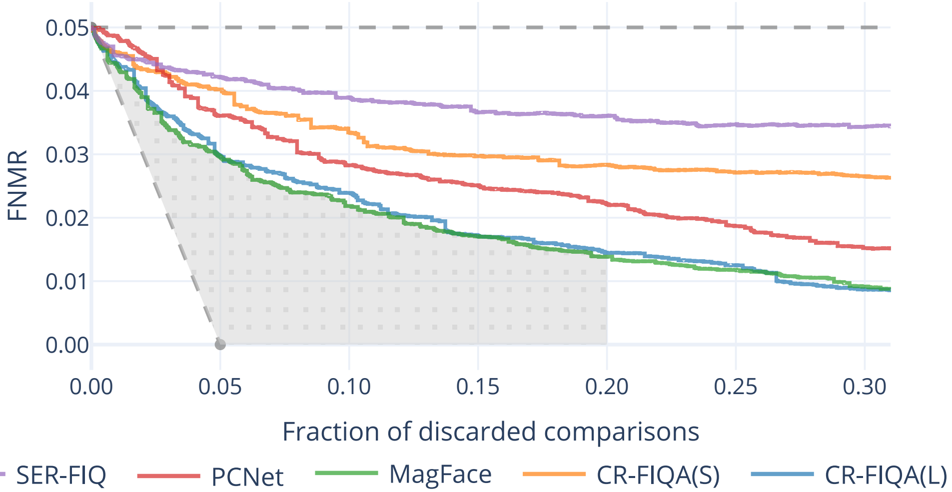
Original   Preprocessed

- Face recognition feature extraction model: **ArcFace-R100-MS1MV2**
- QA models: **CR-FIQA(L)**, **CR-FIQA(S)**, **MagFace**, **PCNet**, **SER-FIQ** (ArcFace)

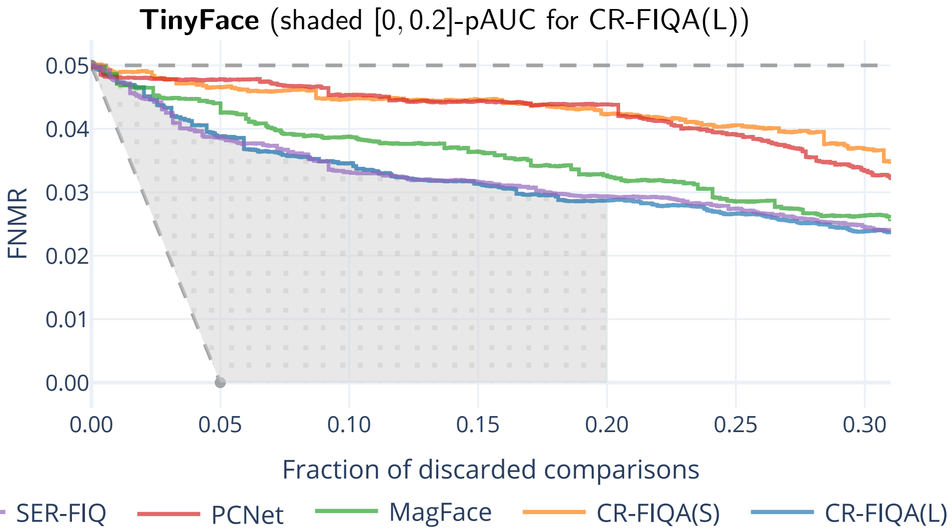
Used face image datasets:

- **LFW** (Labeled Faces in the Wild)
- **TinyFace** (subsets Testing_Set/Gallery_Match and Testing_Set/Probe)

LFW (shaded $[0, 0.2]$-pAUC for MagFace)



Introduction



Introduction

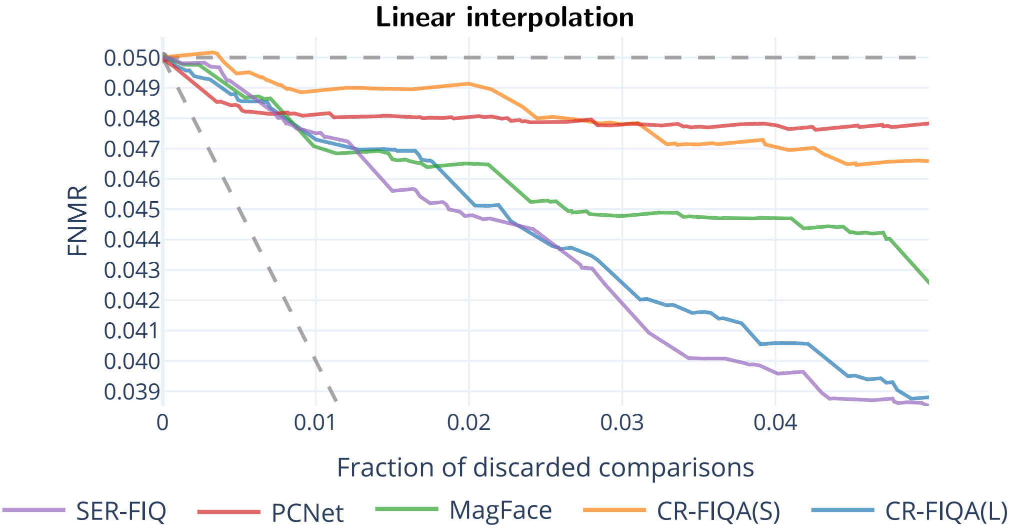
LFW

QA algorithm	$[0, 0.2]$-*pAUC value*	*Discrete ranking*	*Relative ranking*
MagFace	0.00362	1	0.00
CR-FIQA(L)	0.00383	2	0.07
PCNet	0.00506	3	0.46
CR-FIQA(S)	0.00572	4	0.68
SER-FIQ	0.00672	5	1.00

TinyFace

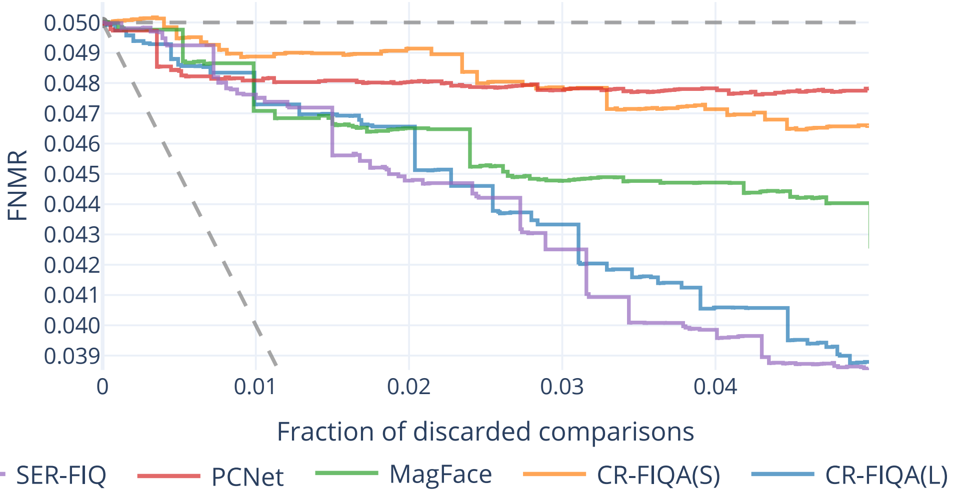
QA algorithm	$[0, 0.2]$-*pAUC value*	*Discrete ranking*	*Relative ranking*
CR-FIQA(L)	0.00588	1	0.00
SER-FIQ	0.00589	2	0.00
MagFace	0.00666	3	0.38
CR-FIQA(S)	0.00787	4	0.97
PCNet	0.00793	5	1.00

Curve interpolation

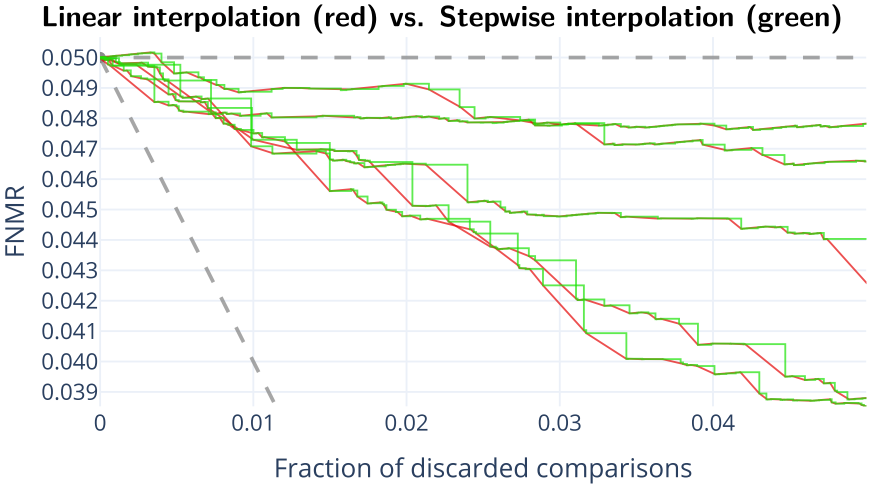


Curve interpolation

Stepwise interpolation



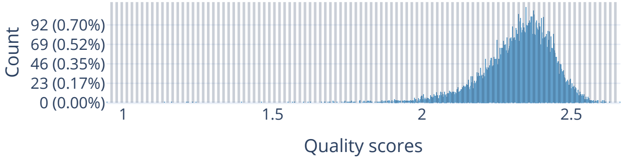
Curve interpolation



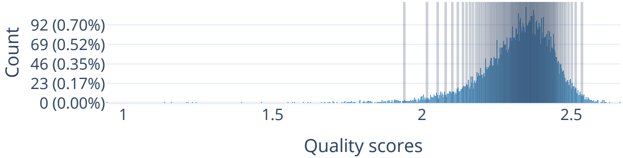
Quality score normalization

"Raw" (e.g. floating-point) QSs can be mapped to "normalized" QSs. ISO/IEC 29794-1 in particular requires a **[0,100] integer range (i.e. 101 bins)** for the data interchange format. Different **calibration functions** and **calibration data sources** can be used for this.

MinMax calibration on LFW quality scores from CR-FIQA(L)



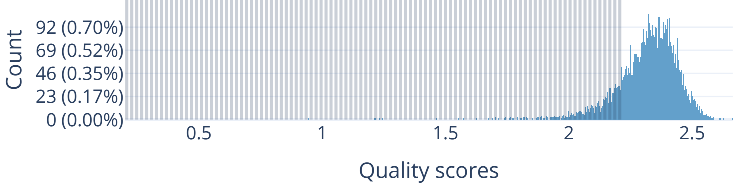
Proportional calibration on LFW quality scores from CR-FIQA(L)



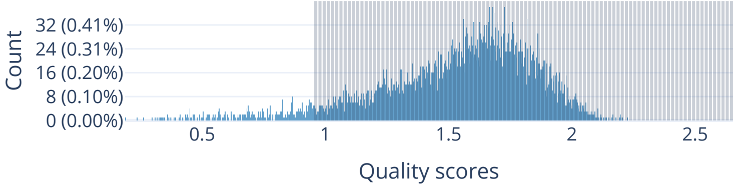
Quality score normalization

An example for bad calibration due to the used calibration data:

MinMax calibration on **TinyFace** quality scores over **LFW** quality scores

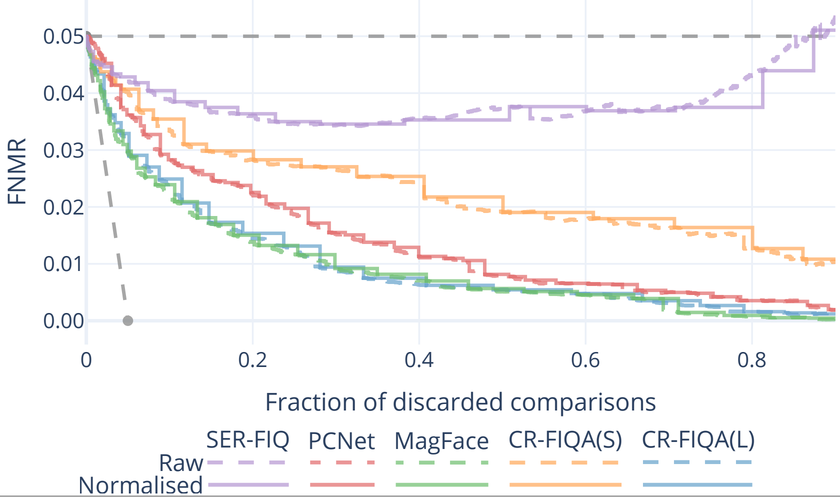


MinMax calibration on **LFW** quality scores over **TinyFace** quality scores



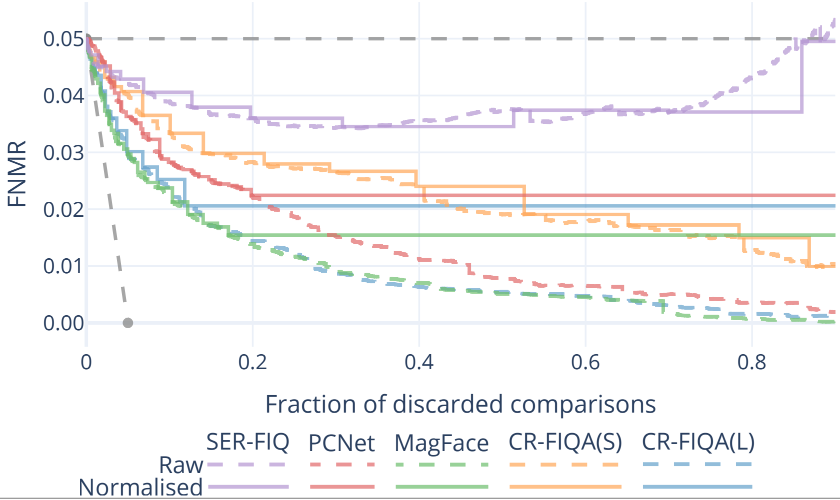
Quality score normalization

EDC plot on LFW, using the **same dataset (LFW)** as MinMax calibration source



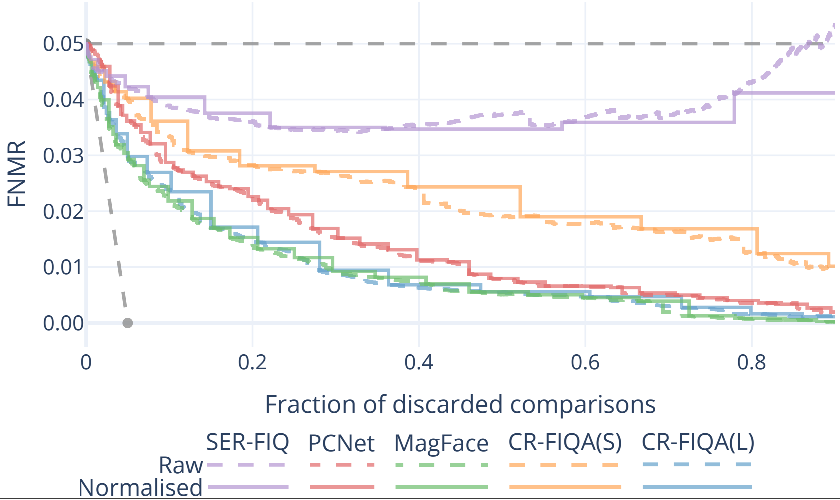
Quality score normalization

EDC plot on LFW, using the **other dataset (TinyFace)** as MinMax calibration source



Quality score normalization

EDC plot on LFW, using the **combined dataset** as MinMax calibration source



Ranking stability

The ranking "stability" is examined across different starting errors & pAUC discard limits:

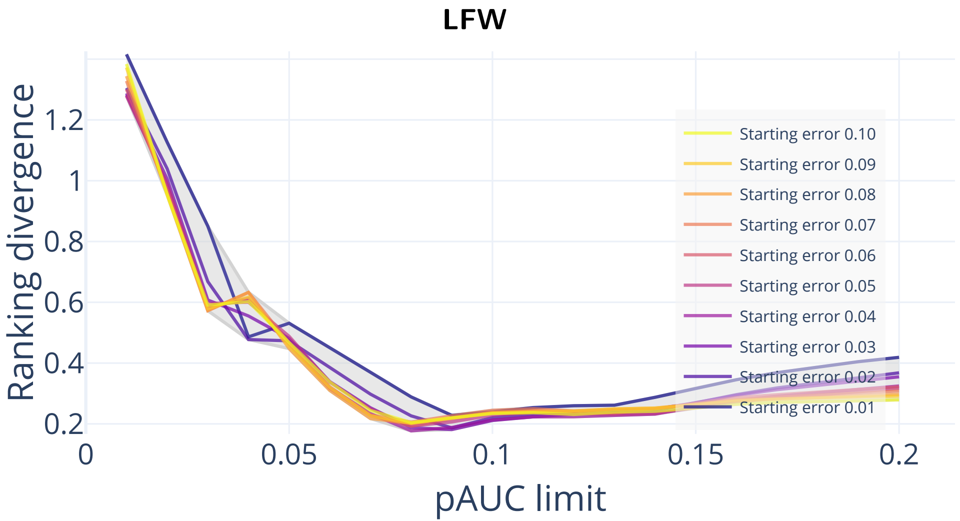
- Starting error: Range $[0.01, 0.10]$ with a 0.01 step (10 steps).
- pAUC discard limit: Range $[0.01, 0.20]$ with a 0.01 step (20 steps).

For each of these 200 configurations the ranking divergence is computed:

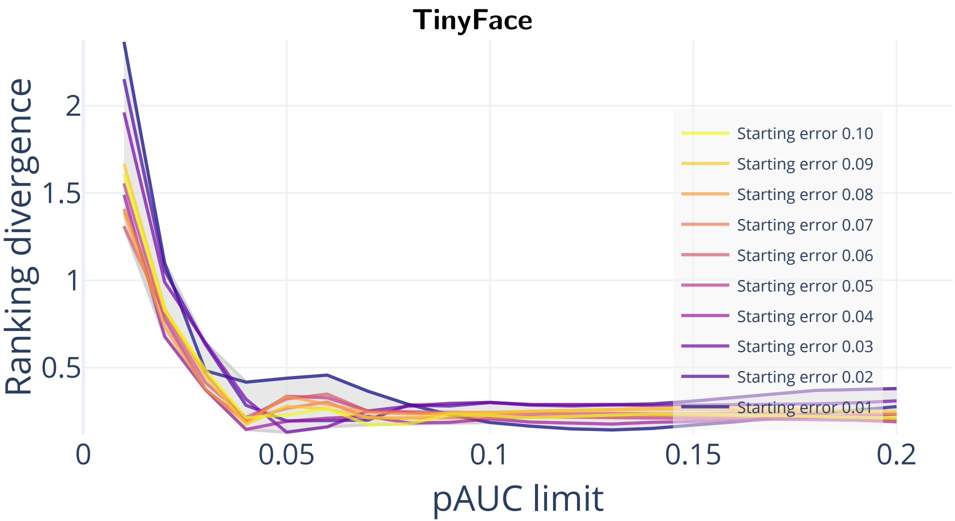
$$RankingDivergence = \Sigma_i^n |p_i - \bar{p}_i|$$

- n is the number of QA algorithms, i.e. 5.
- p_i is the relative ranking "placement" of one QA algorithm, i.e. a value in $[0, 1]$.
- $\bar{p}_i$ is the mean placement of one QA algorithm across all 200 configurations.
- The ranking divergence then is the sum of the distances between p_i and $\bar{p}_i$.
- A lower value implies greater "stability" (with respect to the other configurations).

Ranking stability



Ranking stability



Summary

Main points:

- Relative rankings (i.e. min-max normalized pAUC values) can be used to show how close each QA algorithm is to being the best or worst performing one.
- Stepwise curve interpolation should be preferred,
  to reflect the actual behaviour of the error with respect to the discard steps.
- QS normalisation depends on the calibration and will affect EDC curves/rankings.
  Even a simple min-max range calibration can be effective with the right values.
- For pAUC-based rankings, very low discard limits may not be reliable.

More can be found in the corresponding paper (currently under review):
"Considerations on the Evaluation of Biometric Quality Assessment Algorithms"
Preprint: `https://arxiv.org/abs/2303.13294`

Thank you!

Questions?