ATHENE
National Research Center
for Applied Cybersecurity

Automated data annotation for training face image quality assessment algorithms

Olaf Henniger

ATHENE is a research center
of the Fraunhofer-Gesellschaft
with the participation of

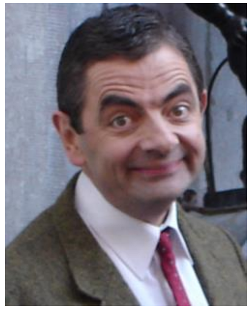
Fraunhofer SIT

Fraunhofer IGD

TECHNISCHE UNIVERSITÄT DARMSTADT

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Face image quality assessment



Feature extraction

- Handcrafted or
- Learnt

Prediction of utility, i.e., »degree to which a biometric sample supports biometric recognition performance« [ISO/IEC FDIS 29794-1]

Quality score in [0; 100] (the higher, the better)

- Narrow down the problem to the prediction of the degree to which any face image supports automated recognition against ICAO-compliant reference face images for ePassports
- For supervised machine learning, training images must be labelled (annotated) with target values

Training data set

- Consists of subsets with at least one ICAO-compliant reference face image per subject (known to be of high utility for automated recognition) from
  - Color FERET Version 2
  - Face Recognition Grand Challenge (FRGC) 2.0
  - NIST Special Database 32 – Multiple Encounter Dataset II (MEDS-II)
  - Multi-PIE (Pose, Illumination and Expression)
  - VGGFace2
- 18,674 probe images and 121 ICAO-compliant reference images

Face comparison algorithms used

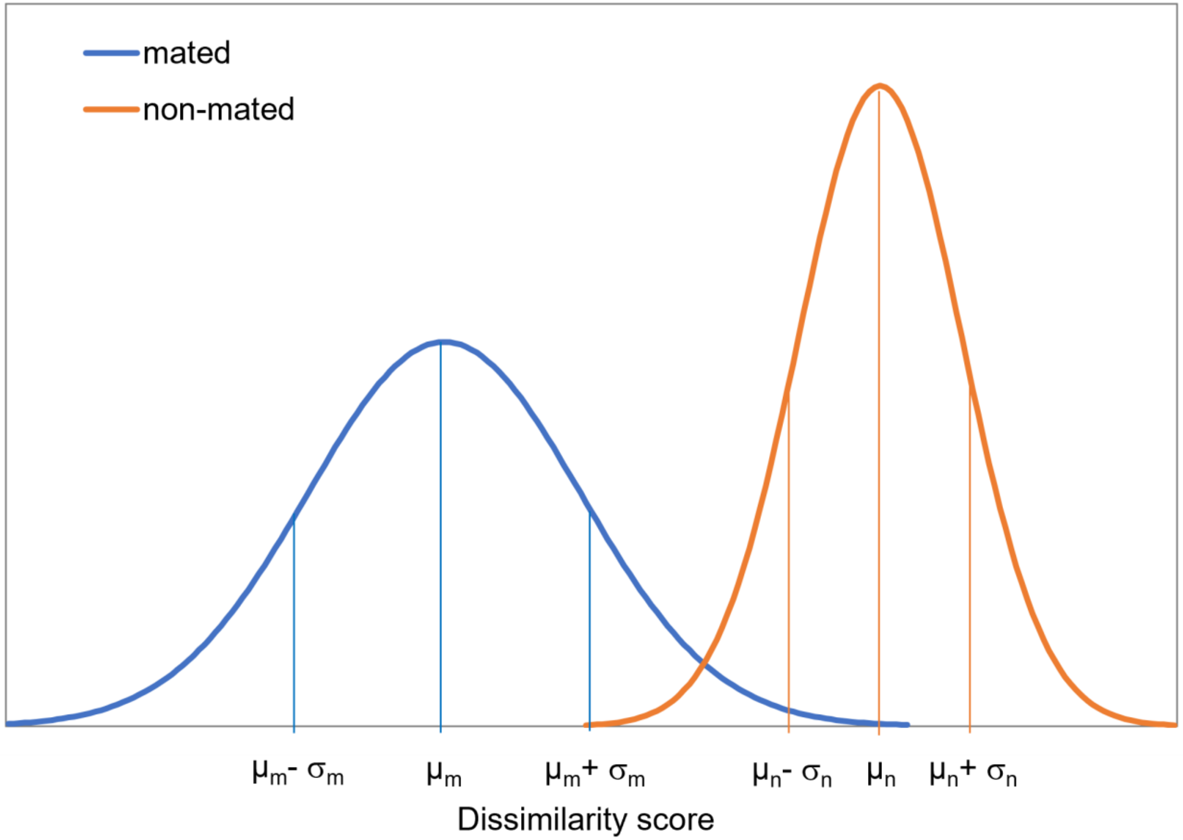
- To calculate comparison scores between each probe and each mated or non-mated reference image as a basis for assessing the utility of each probe image in the training data set
  - ArcFace
    - J. Deng, J. Guo, N. Xue, S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition CVPR, 2019
  - FaceNet
    - F. Schroff , D. Kalenichenko, J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition CVPR, 2015

Utility of a biometric sample for automated recognition

- Corresponds to distance between sample-specific mated and non-mated comparison score distributions within a biometric data set representative of normal use
  - Normalized difference between the means of mated and non-mated comparison scores for a biometric sample i:

$$u_i = \frac{\mu_{n_i} - \mu_{m_i}}{\sqrt{\sigma_n^2 + \sigma_m^2}}$$

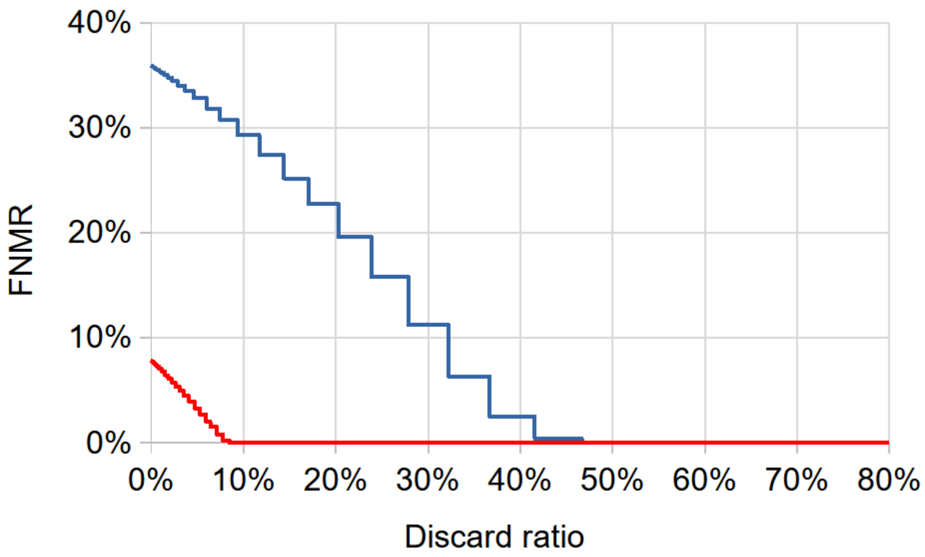
  - Mapped to [0; 100] to obtain utility scores u_i^*
    - 0 to 25: deficient quality
    - 26 to 50: marginal quality
    - 51 to 75: adequate quality
    - 76 to 100: excellent quality

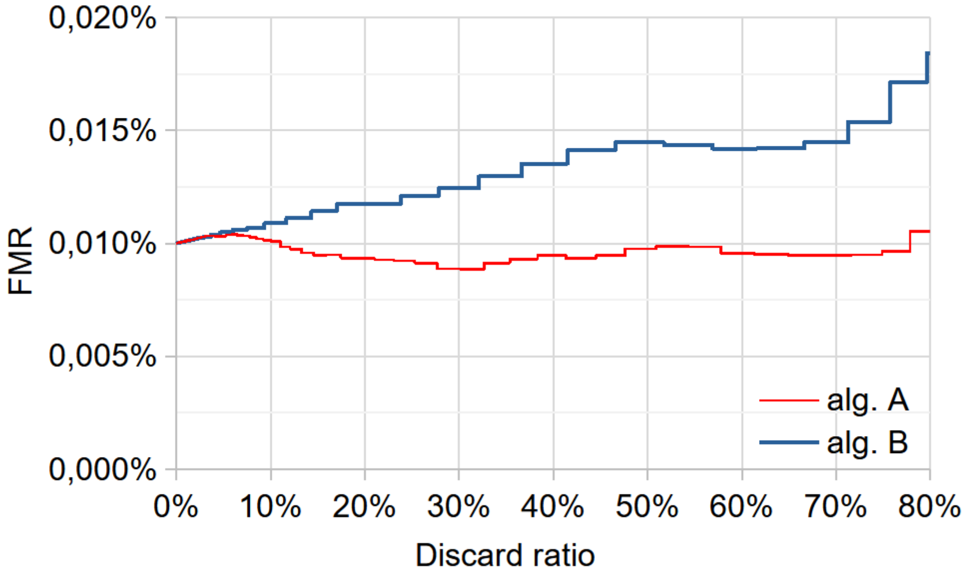


Examples of face images of deficient quality

- On average, more similar to non-mated reference images than to their mated reference image
- Potential quality issues
  - Capture-related such as poor lighting, blur
  - Subject-related such as pose, facial expression, face occlusion
- Examples are from VGGFace2 data set

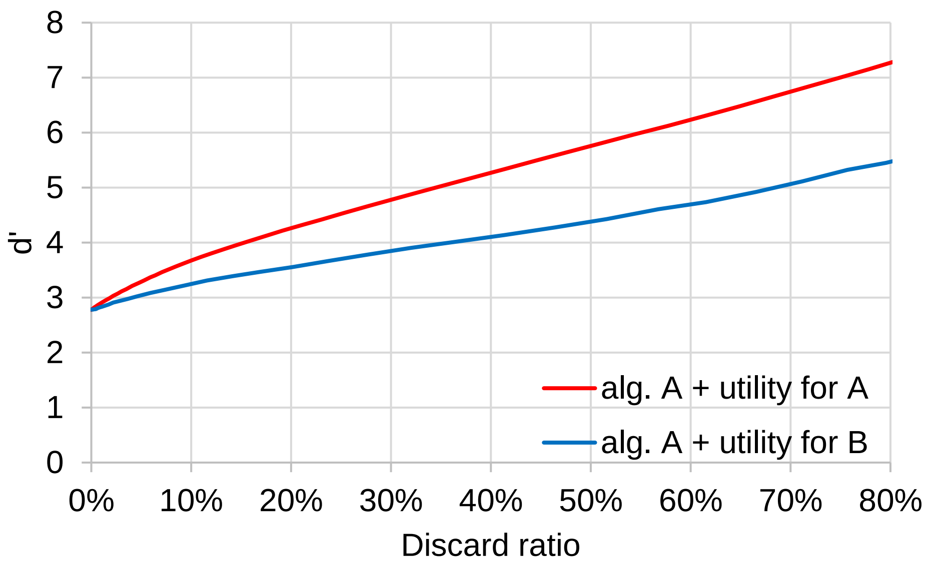
Error vs. discard characteristics with respect to comparator-specific utility scores

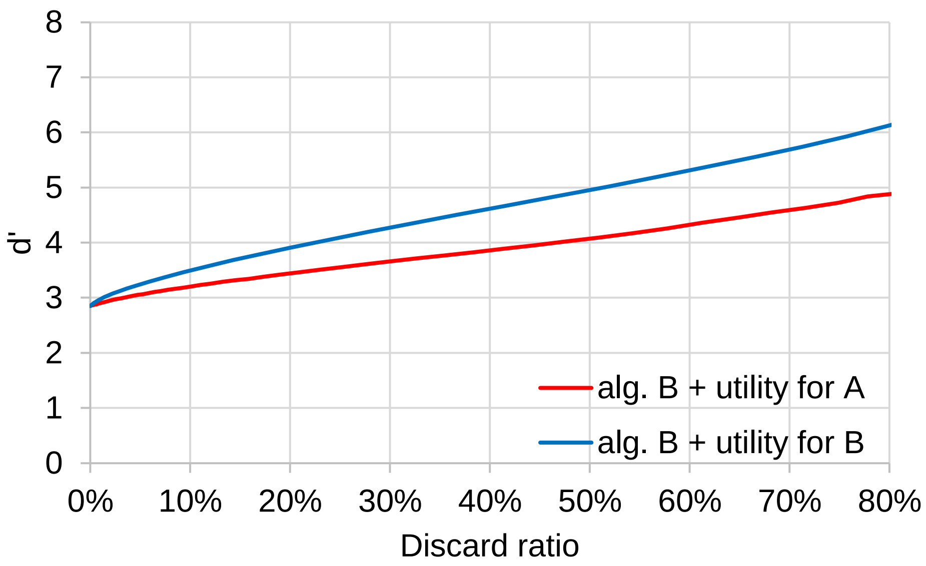




- False non-match rate (FNMR) and false match rate (FMR) over percentage of comparisons discarded due to low quality score of one of the compared samples; decision threshold fixed
- The steeper FNMR decreases with increasing discard ratio and without significantly increasing FMR, the better.
- Utility scores appear to be a good basis for training face image quality assessment algorithms.

d’ vs. discard characteristics with respect to comparator-specific utility scores



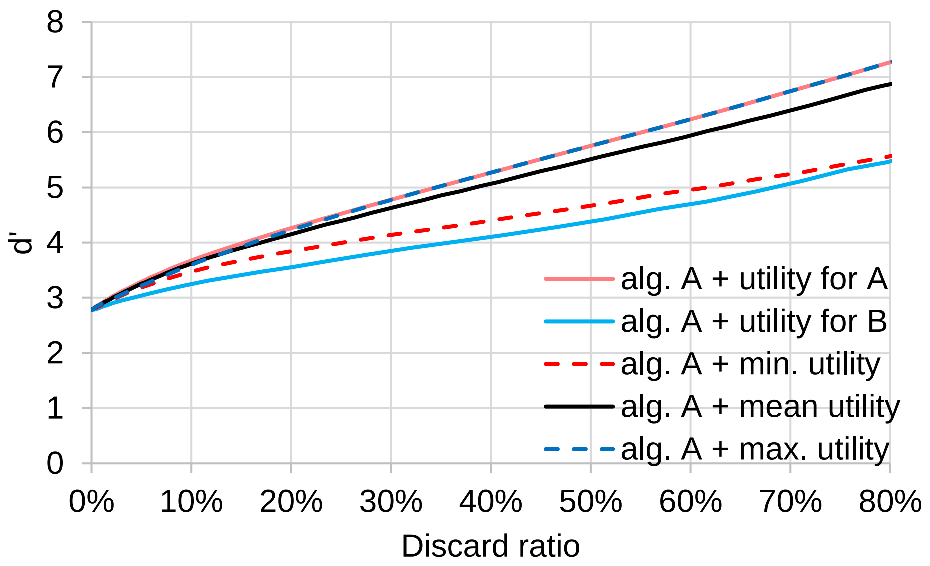


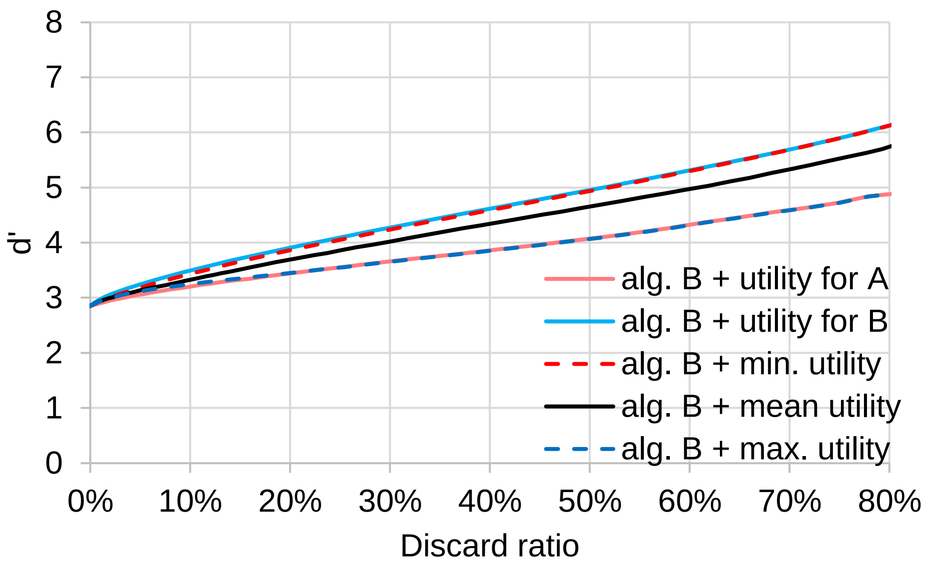
- $d' = \frac{\mu_n - \mu_m}{\sqrt{\sigma_n^2 + \sigma_m^2}}$ over percentage of comparisons discarded due to low quality score of the compared samples
- Summarize the utility-prediction performance of a face image quality assessment algorithm in a single plot
- The steeper d’ increases with increasing discard ratio, the better.

Fusion strategies for comparator-specific utility scores

- Utility for automated biometric recognition depends on the comparison algorithm used.
- At the time of capturing a face image and checking its quality, it is not known which comparison algorithm will be used for biometric recognition
- The question is which utility scores should we use for labelling the training data?
  - Comparator-specific utility scores for algorithm A
  - Comparator-specific utility scores for algorithm B
  - Minimum of the comparator-specific utility scores
  - Maximum of the comparator-specific utility scores
  - Arithmetic mean of the comparator-specific utility scores

d’ vs. discard characteristics with respect to fused utility scores





- If the arithmetic mean of the comparator-specific utility scores is used for discarding low-quality images
  - Performance improves significantly for each of the comparison algorithms used for face recognition

Summary and outlook

- Utility of a biometric sample depends on the comparison algorithm used
- To avoid dependence on a particular comparison algorithm,
  use the arithmetic mean of comparator-specific utility scores to label the training images
- In a next step, we intend to use the annotated training data to build
  - Support vector regression model
  - Random forest binary classification model
  - Deep-learning-based model

Fraunhofer
IGD

Dr. Olaf Henniger

Competence Center for Smart Living & Biometric Technologies
Fraunhofer Institute for Computer Graphics Research IGD

Fraunhoferstrasse 5 · 64283 Darmstadt · Germany
Phone +49 6151 155-526 · Fax -480
olaf.henniger@igd.fraunhofer.de