Federal Office
for Information Security

Open Source Face Image Quality (OFIQ)

Implementation and Evaluation of Algorithms

Version 1.1

2024-09-30

Authors

Johannes Merkle[1], *Christian Rathgeb*[1], *Benjamin Herdeanu*[1], *Benjamin Tams*[1],
Day-Parn Lou[1], *André Dörsch*[1], *Maxim Schaubert*[1], *Jonas Dehen*[1],
Liming Chen[2], *Xiangnan Yin*[2], *Di Huang*[3], *Anna Stratmann*[4], *Marcel Ginzler*[4],
Marcel Grimmer[5,6], *Christoph Busch*[5,6]

1) secunet Security Networks AG,
Kurfürstenstr. 58, 45138 Essen, Germany

2) Liris Laboratory UMR CNRS 5205, Ecole Centrale de Lyon,
36 Avenue Guy de Collongue, 69134 Ecully Cedex, France

3) School of Computer Science and Engineering, Beihang University,
Xueyuan Road, Haidian District, 100191Beijing, China

4) Federal Office for Information Security,
P.O. Box 20 03 63, 53133 Bonn, Germany

5) Hochschule Darmstadt, Fachbereich Informatik,
Schöfferstrasse 3, 64295 Darmstadt, Germany

6) Norwegian University of Science and Technology,
Teknologiveien 22, 2815 Gjøvik, Norway

Federal Office for Information Security
P.O. Box 20 03 63
53133 Bonn
E-Mail: OFIQ@bsi.bund.de
Internet: https://www.bsi.bund.de

Table of Contents

1 Executive Summary

This report summarizes the development of Open Source Facial Image Quality (OFIQ), a library of open-source algorithms for face image quality assessment. OFIQ is the reference implementation for the international standard ISO/IEC 29794-5 and comprises various quality assessment algorithms:

- An algorithm giving a unified quality score which aims to predict the utility of the facial image for recognition purposes.
- Various quality components algorithms, each of which assessing the conformance of the facial image to a certain requirement, e.g. from ISO/IEC 39794-5:2019 [1]. Some of these quality components depend on the environment used for imaging acquisition (capture-related), while others primarily depend on aspects of the presentation of the face (subject-related).

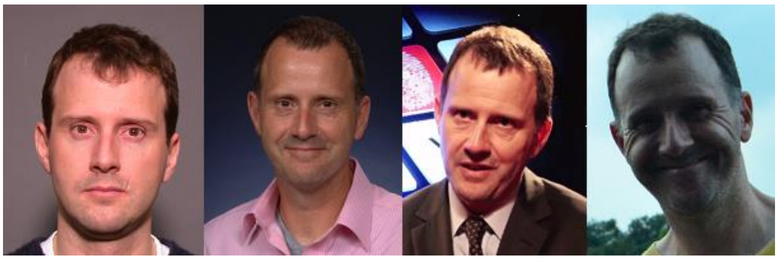
Consequently, OFIQ outputs a vector of quality measures, each of which being an integer in the range [0,100], where higher values signify higher quality, i.e., higher utility for face recognition or higher conformance to the requirements.

The report briefly summarizes the OFIQ.

2 Introduction

2.1 Motivation

With recent technological advances, in particular the introduction of algorithms based on deep learning, face recognition error rates dropped massively. However, they remain significant, depending on several factors, such as the imaging process or the level of cooperation of the biometric capture subject. Accordingly, face image quality assessment algorithms are designed to calculate a (unified) quality score for a captured facial image, which indicates its utility for recognition purposes, see Figure 1.



QS = 95	QS = 85	QS = 62	QS = 42
a	b	c	d

Figure 1: Examples of face images of a single subject with example quality scores (QS).

The estimated quality scores can be used to ensure that only facial images of sufficient quality are fed into a face recognition system. Specifically, the quality scores are compared with a quality threshold value, and if a quality score exceeds the quality threshold, the corresponding facial image is accepted. Otherwise, a re-capture might be initiated, where actionable feedback is needed for users and operators. This is achieved by computing, in addition to the unified quality score, various quality components assessing the conformance of the facial image to specific requirements. By assessing specific quality components, decisions of quality assessment algorithms become more transparent for users and operator, which is essential in biometric systems like the EES.

2.2 Objective and Approach

The objective of the project “Open Source Facial Image Quality” (OFIQ) is to develop, implement, and publish a library of open-source algorithms for the quality assessment of facial images used in face recognition. This library, named OFIQ, aims to be applicable for various biometric applications, particularly in border control scenarios. OFIQ is written in the C/C++ programming language. Its source code is available from https://github.com/BSI-OFIQ/OFIQ-Project.

OFIQ serves as a reference implementation for the standard ISO/IEC 29794-5. The recent revision of this standard was performed in parallel to OFIQ, so that findings obtained during the development of OFIQ were incorporated into this standard, and vice versa. In addition, OFIQ aims to allow checking facial images for conformance with the quality requirements defined ISO/IEC 19794-5:2011 and ISO/IEC 39794-5:2019.

In order to become a universally usable, OFIQ needs to meet various requirements. Based on a thorough research of the current state-of-the-art, suitable candidate algorithms have been shortlisted, implemented, tested and improved, and, for each task, the best algorithm was selected. Additionally, it was ensured that the licences of used algorithms allow commercial use, which is a key requirement for OFIQ.

Furthermore, OFIQ shall provide a good trade-off between state-of-the-art accuracy, runtime and resource consumption. To facilitate its use, OFIQ must be compatible with relevant platforms, including Windows, Linux, MacOS, Android and iOS.

3 Open Face Image Quality (OFIQ)

3.1 OFIQ Framework

An overview of the OFIQ framework[1] is depicted in Figure 2. OFIQ takes as input a single 2D face image. First different pre-processing steps (common computations) are performed. Subsequently, the unified quality and different capture- as well as subject-related quality components are assessed, resulting in an output vector consisting of the unified quality score and various quality component values.

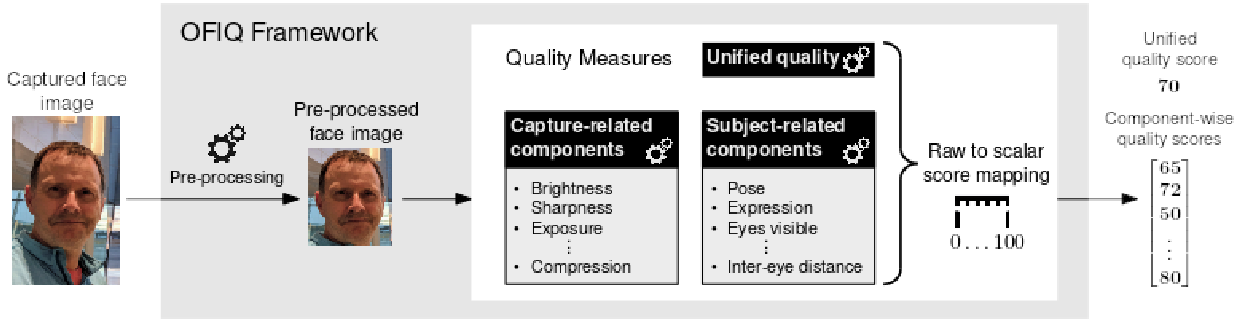


Figure 2: Overview of the OFIQ framework.

The assessed quality components are listed in Table 1, each of which is computed by a distinct algorithm.

Capture-related Quality Components	Subject-related Quality Components
• Background uniformity • Illumination uniformity • Moments of the luminance distribution • Over-exposure prevention • Under-exposure prevention • Dynamic range • Sharpness • No compression artifacts • Natural colour	• Single Face Present • Eyes open • Mouth closed • Eyes visible • Mouth occlusion prevention • Face occlusion prevention • Inter-eye distance • Head size • Crop of the face • Head pose • Expression neutrality • No head coverings

Table 1: List of capture- and subject-related quality components in OFIQ.

The algorithm for the unified quality score and the algorithms for the individual quality components all output a native quality measure, which is a floating-point value without uniform value range or semantic. Each of these native quality measures is then mapped to the unified quality score or the quality component value, respectively, in the target range [0,100] having a higher-is-better semantics. This means that a unified quality score or a quality component of 100 refers to optimal quality.

The algorithms for computing the quality measures (unified quality score and quality components) are described in Section 4.

[1] https://github.com/BSI-OFIQ/OFIQ-Project

4 Evaluation of Quality Measures

The evaluation of the algorithms for the quality measures was conducted through several means:

- Using custom test sets with ground truth labels for the image aspects assessed by the quality component. In most cases, binary (2-class) labels have been used (e.g. image is under-exposed or not), with notable exceptions being assessment of occlusions (of eyes, mouth, face) and head pose, where numerical labels were available.
- By evaluating the reduction of the error rate of face recognition algorithms when using the quality measure values for quality filtering. This was done using error-versus-discard characteristic (EDC) curves. This evaluation employs different open-source and commercial face recognition algorithms, along with a large test set derived from the VGGFace2 face database[2].
- Whenever possible, the most promising algorithms were submitted to the Specific Image Defect Detection (SIDD) track of the Face Analysis Technology Evaluation (FATE) Quality[3]. NIST evaluated these algorithms using their own sequestered test data.

4.1 Unified quality score

A unified quality score might can be derived from a set of component-based quality scores, as it is for instance done in the current version of the NIST Fingerprint Image Quality algorithm (NFIQ 2). In contrast, so-called monolithic algorithms are designed to directly extract a unified quality score from a biometric sample. For such an end-to-end solution, deep learning-based methods have been shown to achieve competitive performance. In particular, it was proposed to train loss functions with additional parameters and constraints with an explicit goal to encode the face quality within the face feature representation magnitude. A prominent example of such an approach is the MagFace algorithm[4]. The extracted unified quality score is supposed to assess the utility of the image for face recognition and other applications of face biometrics. This unified quality score is supposed to depend on all quality components, which are relevant for face recognition performance.

In the conducted evaluations, algorithms based on the MagFace approach achieved the best results. Specifically, the MagFace100 model[5] decreased the false non-match error rates from 10% to 6%, in case 10% of the face images with the lowest estimated quality are discarded.

4.2 Capture-related quality components

For capture-related quality components (see Table 1), different hand-crafted and deep learning-based algorithms have been considered. While certain quality components are estimated based on exact metrics (for instance Moments of the luminance distribution) the accuracy of other algorithms has been confirmed on suitable test databases. Depending on how pronounced the quality deficiencies were in the respective test databases, the algorithms show good detection performance, generally, above 90% accuracy (below 10% equal error rate).

Among the evaluated capture-related quality components, Sharpness and No compression artifacts were found to have the highest impact on face recognition accuracy. On the aforementioned test database, discarding 10% of face images with lowest quality scores according to Sharpness and No compression artifacts (two components that are obviously related) reduced the false non-match error by 2% (from 10% down to 8%). For the latter quality component, a neural network was trained to detect artifacts resulting from JPEG and JPEG 2000 image compression.

[2] https://www.robots.ox.ac.uk/~vgg/data/vgg_face2/
[3] https://pages.nist.gov/frvt/html/frvt_quality.html
[4] https://github.com/IrvingMeng/MagFace
[5] iResNet100 model trained on MS1MV2 from https://github.com/IrvingMeng/MagFace

Several quality components were found to only have marginal impact on face recognition accuracy, including Illumination uniformity, Moments of the luminance distribution, Under-exposure prevention, Over-exposure prevention, Dynamic range, and Natural colour. On the used test database, discarding 10% of face images with lowest quality according to the respective quality component reduced the false non-match error marginally from 10% down to 9.5%. It can be concluded that state-of-the-art face recognition systems are largely robust to (non-severe) defects caused by these quality components.

Finally, the quality component Background uniformity was found to have negligible effects on face recognition performance.

4.3 Subject-related quality components

Many subject-related quality components (see Table 1), such as Inter-eye distance or Mouth closed, are directly derived from the landmarks estimated during the pre-processing stage. Further subject-related quality components can be estimated based on the performed face segmentations, for instance Face occlusion prevention and No head coverings. For the estimation of other subject-related quality components like Pose or Expression neutrality, deep learning-based algorithms have been adapted. Generally speaking, most of the applied classification algorithms achieve high accuracy, above 95% accuracy (below 5% equal error rate) with some exceptions. In particular, significantly worse results are obtained for the detection of motion blur, which is a non-trivial problem. For motion blur detection, results reported in the scientific literature could not be re-produced. In some cases, the accuracy of the algorithms varied significantly between the different test sets, which can be attributed to a heterogeneous difficulty of the test sets. For instance, for the quality component Natural colour, the detection performance dropped on a self-created test set where slight variations of image colour were hardly detected, as opposed to more drastic changes present in another publicly available test set.

Regarding occlusions, discarding 10% of low-quality face images according to the quality components Face occlusion prevention, Eyes visible and Mouth occlusion prevention reduces the false non-match error up to 2% (from 10% down to 8%) on the used test database. This confirms the well-known fact that facial occlusions negatively affect face recognition performance, in particular in the ocular region. Related to the ocular region of the face, similar results were observed for the quality component Eyes open (up to 2% reduction of false non-match error for a 10% discard rate). Also, the quality component Inter-eye distance, which relates to the overall size of the face in the captured images, was found to reduce the false non-match error from 10% to around 8% if 10% of face images estimated to exhibit lowest quality are discarded.

Variations in head pose are commonly described by three angles: the pitch, yaw, and roll angles, which are the rotations about the vertical (y), the horizontal side-to-side (x), and the horizontal back to front (z) axes, respectively. Pitch, yaw, and roll angles of the head pose are defined in ISO/IEC 39794-5:2019. A frontal face image has 0° for all three angles. In case these angles strongly deviate from 0°, the visible facial parts change substantially. State-of-the-art head pose estimation algorithms are based on deep learning techniques. Accordingly, different deep learning-based methods have been evaluated among which the 3DDFAv2[6] algorithm has been identified as the most suitable. The effect of each pose angle was separately evaluated on used test database. It was found that the roll and pitch angles have only marginal impact face recognition accuracy. For discarding 10% of face images with lowest quality with respect to pitch and roll angles, false non-match rate errors were reduced by 1% to 2%, depending on the used face recognition system. This means that, some face recognition systems were observed to be more robust against variations in the roll and pitch angle. The same holds for the yaw angle, for which a higher impact on face recognition performance was observable. More specifically, for some face recognition systems, the false non-match rates could be reduced by more than 4% (from 10% down to below 6%) for discarding 10% of face images with the lowest quality scores according to the estimated yaw angle. For other face recognition system, this reduction in false non-match errors was less pronounced.

[6] https://github.com/cleardusk/3DDFA_V2

For discarding 10% of face images with lowest quality the quality component Motion blur prevention, a 2% reduction in false non-match rate was achieved (from 10% down to 8%). However, it is assumed that the used algorithm not only detects motion blur but also normal blur present in unsharp face images.

Further subject-related quality components were found to have only marginal impact on face recognition accuracy, namely Mouth closed, No head coverings, Expression neutrality. Discarding 10% of low-quality face images according to these quality components reduced the false non-match errors by up to 0.5%. Since the forehead is less relevant for face recognition systems, head covering rarely impact their accuracy unless they occlude more relevant facial regions. Similarly, the Mouth closed quality component was expected to have marginal impact on face recognition performance, since state-of-the-art systems tend to put more focus on the ocular region of a face. More interestingly, it was found that face recognition systems are largely robust against variation in facial expressions. To assess whether the face in a captured image has a neutral expression, the deep learning-based HSEmotion[7] method was adapted. This approach estimates probabilities for predefined expression classes, such as happiness, anger or sadness. To map the output of this network to a quality score, it was combined with a machine learning-based classifier (AdaBoost).

4.4 FATE quality SIDD

Algorithms for the quality components listed in Table 2 were submitted to the NIST FATE quality SIDD:

Capture-related Quality Components	Subject-related Quality Components
• Background uniformity • Over-exposure prevention • Under-exposure prevention • Sharpness • No compression artifacts	• Motion blur prevention • Face occlusion prevention • Eyes open • Mouth closed • Inter-eye distance • Pose

Table 2: List of capture- and subject-related quality measures submitted to NIST FATE quality SIDD.

Overall, the NIST FATE quality SIDD evaluation confirmed that the algorithms contained in OFIQ generally achieve state-of-the-art performance. It should be noted that, the overall number of submissions varied according to the quality component. For instance, only a few submissions have been made for the quality component Face occlusion prevention, most likely due to its complexity. In contrast, several submissions have been made for quality components that can be assessed easily, for instance inter-eye distance, which only requires the localization of facial landmarks.

In the context of capture-related quality components, moderate performance is achieved for detecting Background uniformity. Specifically, the OFIQ algorithm was outperformed by different commercial algorithms. However, as mentioned earlier, background uniformity was not found to significantly impact face recognition performance. For the quality component No compression artifacts, only JPEG image compression was considered in the NIST FATE quality SIDD. This means that, submitted algorithms are only required to detect artifacts stemming from JPEG compression, while other compression algorithms like JPEG 2000 would be of relevance, too. The submitted OFIQ algorithm is designed to detect both, JPEG and JPEG 2000 compression artifacts, but achieves inferior detection performance compared to other algorithms submitted to NIST FATE quality SIDD. Regarding capture-related quality components, several defects are created synthetically in the NIST FATE quality SIDD, including Over- and Under-exposure and Sharpness. While the performance of the respective OFIQ algorithms was comparable to other submission, results obtained on synthetic test data should be treated with care. Also, from example test images, it was observed that the test databases included images with irrelevant defects. For instance, in the NIST FATE quality SIDD slight degradations of contrast and brightness (to simulate under-exposure) that are not expected to impact face recognition accuracy are considered as defects.

[7] https://github.com/av-savchenko/face-emotion-recognition

Similar to some capture-related quality components, the motion blur was synthetically generated in order to evaluate the Motion blur prevention component in the NIST FATE quality SIDD. However, in the evaluation phase of the EESFM project, it was found that it is difficult to simulate motion blur. Consequentially, the accuracy of detection algorithm implemented in OFIQ were found to obtain different detection performance rate on face images containing real and synthetically generated motion blur. While the submitted algorithm for the Motion blur prevention component did not achieve good detection performance, this result should also be treated with care. For several other subject-related components, the OFIQ algorithms achieved the best performance or were on par with the best performing algorithms, in particular Face occlusion prevention, Eyes open, Mouth closed, Inter-eye distance. With respect to the Pose quality component, the accuracy of algorithms for estimating yaw, pitch and roll angles are separately evaluated in NIST FATE quality SIDD. For this purpose, two test sets are used for yaw as well as pitch angles and a single test set is used for roll angles. The OFIQ algorithm showed more errors for calculating strong variations in yaw angles, which are expected to be less relevant in cooperative capture scenarios like the EES. However, for smaller variations in pose angle the OFIQ algorithm showed good results. Competitive performance was also achieved for estimating pitch angles where on one test set the OFIQ algorithm obtained the lowest error rates. For estimating the roll angle, the OFIQ obtained good results comparable to other algorithms submitted to the NIST FATE quality SIDD.

5 Summary and Outlook

The OFIQ framework provides an open-source reference implementation of the standard ISO/IEC 29794-5, consisting of state-of-the-art algorithms that allow its commercial use. The ISO/IEC 29794-5 standard was developed by international experts in the field of face recognition. In parallel, OFIQ was developed and findings obtained in two prototyping and evaluation phases were fed into the corresponding standard, which was co-authored by OFIQ developers. The OFIQ framework contains algorithms to assess the unified quality of a face image. In addition, capture- and subject-related quality components that have been identified as relevant for face image quality assessment by international experts. This set of quality components makes it possible to provide explainable and actionable feedback to users and operators of face recognition systems.

The OFIQ framework fulfils requirements relevant for practical applications, such as platform compatibility. Throughout the development of OFIQ, resource consumption has been minimized. Hence, OFIQ is expected to become a de-facto standard that allows a transparent transnational assessment of face image quality. This is of utmost importance for large-scale face recognition systems like in the upcoming EES.