

# Face Recognition Vendor Test Ongoing

## Face Recognition Quality Assessment Concept and Goals

VERSION 0.2

Patrick Grother  
Mei Ngan  
Kayee Hanaoka  
*Information Access Division  
Information Technology Laboratory*

Contact via [frvt@nist.gov](mailto:frvt@nist.gov)

DRAFT  
FOR COMMENT

March 20, 2019

## 1. Scope

While standards exist for interchange of face images [ISO/IEC-2005 superseded by ISO/IEC-2019 which includes , ICAO-Portrait, and ANSI-NIST Type 10] and those standards additionally regulate the capture of images, there are no standards for how face image quality must be assessed<sup>1</sup> nor are there performance evaluations for automated quality assessment algorithms.

This document is intended to support accurate face recognition by:

- Establishing specifications for face image quality assessment algorithms that return scalar quality values, particularly by requiring image quality assessment algorithms to judge quality in reference to ISO/IEC 19794-5 full frontal and the ICAO Portrait Quality standards;
- Describing NIST's performance evaluation of such algorithms.

## 2. Applications of quality scalars

The primary use cases for scalar image quality assessments are:

- **Photo acceptance:** Foremost, scalar image quality values can be used to make an acceptance or rejection decisions. If an image's quality is too low, a system will reject the image and initiate collection of a new image. Such a process could be implemented in a camera, in a client computer, or on a remote server. Such a capability is most useful during initial enrollment, when a prior reference image of the subject is not available. It is also useful when forwarding the image to a remote recognition service would be time consuming or expensive.
- **Quality summarization:** Scalar image quality values are useful as a management indicator. That is, in some enterprise where face images are being collected from many subjects, say by different staff, at different sites, under different conditions, the quality values can be used to summarize the effectiveness of the collection. This might be done using some statistic such as average quality, or proportion with low quality. Such summarization can be used to reveal site-specific problems, population effects, as a response variable in A-B tests, and to reveal trends, diurnal or seasonal variation.
- **Photo selection:** Given  $K > 1$  images of a person, select the best image. This operation is useful when a receiving system expects exactly one image, and the capture subsystem must determine which of the several collected images should be transmitted. This application of quality is useful when a capture process includes some variation e.g. due to unavoidable motion of the subject or camera.

NOTE Ordinarily this function should not be used in place of recognition. A recognition application should generally enroll all  $K$  images of a person rather than select one. This recommendation is made because quality assessment infrastructure is an imperfect predictor of recognition outcome and it may arise that an enrolled image with lower quality might be successfully matched to a probe image due to certain characteristics of the image e.g. view angle or facial expression. That said, if some images may have been collected decades ago, then ageing may well reduce the utility of the image to a recognition against a recent image even if quality is excellent.

---

<sup>1</sup> The document ISO/IEC 29794-5:2010 is a technical report that, as such, does not establish any requirements that a formal standard would do. Its title is "ISO/IEC 29794 Biometric sample quality — Part 5: Face image data". It gives terminology, base concepts, and examples of how specific quality degradations might be measured.

### 3. Quality Assessment

#### 3.1. Prior standardization

Table 1 in technical report ISO/IEC 29794-5:2010 characterizes two aspects of face quality. The first distinguishes between subject-specific factors, and environmental and capture system factors. The second decomposes persistent “static” effects from those that occur temporarily. Table 1 is an excerpt of the table in the ISO document expressing that quality problems due to mis-presentation by the subject and those related to imaging are in many cases separable – for example photographs can be systematically mis-focused even when the subjects present perfectly.

*Table 1 – Characterization of Face Image Quality*

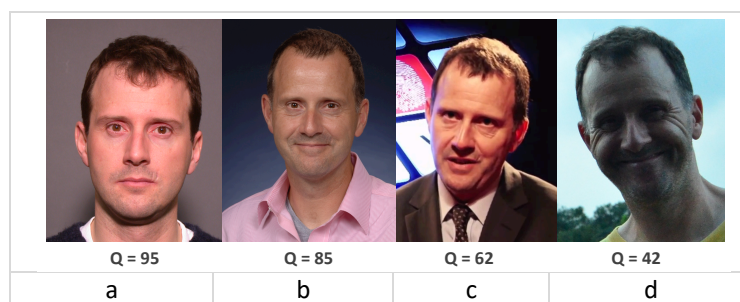
	Subject characteristics	Acquisition process
Static properties	Biological characteristics: <ul style="list-style-type: none"> <li>– injuries and scars</li> <li>– ...</li> </ul> Other static characteristics <ul style="list-style-type: none"> <li>– Thick or dark glasses</li> <li>– Permanent jewellery</li> </ul>	Acquisition process and capture device properties: <ul style="list-style-type: none"> <li>– image resolution</li> <li>– optical distortions</li> <li>– ...</li> </ul> Static properties of the background <ul style="list-style-type: none"> <li>– [textured] wallpaper</li> </ul>
Dynamic properties	Subject characteristics and behavior: <ul style="list-style-type: none"> <li>– exaggerated expression</li> <li>– hair across the eye</li> <li>– ...</li> </ul>	Scenery <ul style="list-style-type: none"> <li>– background moving objects</li> <li>– variation in lightning</li> </ul> Capture device variation <ul style="list-style-type: none"> <li>– mis-focus</li> <li>– poor exposure (due to bright sources)</li> </ul>

Note that in traditional live-scan fingerprint capture, quality problems related to imaging are essentially absent by virtue of the optical design and mode of operation of the sensor. For this reason, it was possible to build fingerprint quality assessment algorithms [NFIQ] that did not need to quantify quantities such as illumination non-uniformity and mis-focus. For face recognition, however, the distinctions inherent in the table influence what quality measurements should be made, as discussed next.

#### 3.2. Fundamental operations

##### 3.2.1. Scalar quality value

Given an image  $X$ , an image quality assessment algorithm,  $F$ , shall produce a scalar quality score,  $Q = F(X)$ . Four examples are shown in Figure 1. The progression, from left to right, implies that better images have higher quality values, where the term better here is the subject of this standard.



*Figure 1 –Four faces with example image quality values.*

### 3.2.2. Quality 2-tuples

**NOTE** Reporting of quality tuples is not part of the FRVT Quality Evaluation in 2019.

Given image  $X$ , a quality assessment algorithm,  $F$ , shall report  $(Q_{SUB}, Q_{SYS}) = F(X)$  where the scalar  $Q_{SUB}$  reflects subject-specific behavior, and  $Q_{SYS}$  summarizes properties inherent in the environment and imaging system.

- $Q_{SYS}$  should summarize quantities like resolution, compression, illumination amount, non-uniformity and sensor noise i.e. items which would be expected to affect all images collected from that system.
- $Q_{SUB}$  should summarize quantities like expression neutrality, pose, eye openness and eyeglasses.

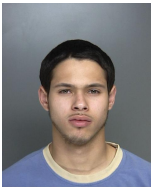



			
$Q_{SUB} = 98$ $Q_{SYS} = 90$	$Q_{SUB} = 94$ $Q_{SYS} = 40$	$Q_{SUB} = 20$ $Q_{SYS} = 95$	$Q_{SUB} = 28$ $Q_{SYS} = 23$
a	b	c	d

Figure 2 - Four faces with example quality 2-tuples

Figure 2b shows an image in which the subject presents almost perfectly to the camera, but photo quality is impaired by poor exposure. In contrast, Figure 2c shows an image in which the imaging is good, but the subject mis-presents to the camera. Figure 2d shows an image with both kinds of problem, and Figure 2a has neither.

### 3.3. Quantitative goal for quality scalars

ISO/IEC 29794-1 delineates three aspects of the umbrella term quality:

- *Character*: This is some statement of the normality of the anatomical biometric characteristic – thus a scarred fingerprint or a heavily bearded face may have poor character.
- *Fidelity*: This is any measurement that indicates how well a captured digital image faithfully represents the analog source – thus a blurred image of a face omits detail and has low fidelity.
- *Utility*: Finally, and most relevant in this standard, the term *utility* is used to indicate the value of an image to a receiving recognition algorithm.

This standard conceives of quality scalars as being measures of utility rather than, say, fidelity, because utility of a sample to a recognition engine is what drives outcome operationally and is of most interest to end-users<sup>2</sup>.

The standard, later, requires quality values to serve as predictors of true match outcome. Of course, recognition outcomes depend on the properties of at least two images, not just the sample being submitted to a quality algorithm. This apparent disconnect is handled by requiring sample quality to reflect expected comparison outcome of the target image with a canonical high-quality portrait image of the form given in Figure 3.

<sup>2</sup> The adoption of utility provides a quantitative goal for development of quality scalars, in the supervised machine learning sense. This approach was taken with the NIST Fingerprint Image Quality Algorithm. The ISO/IEC 29794-4 standard defines the NFIQ algorithm which was trained using a machine learning scheme to be a predictor of fingerprint true match accuracy. That algorithm, and its commercial analogues, have been run tens of billions of times in large scale identity operations in many global programs, including Aadhaar (India) and immigration (USA).

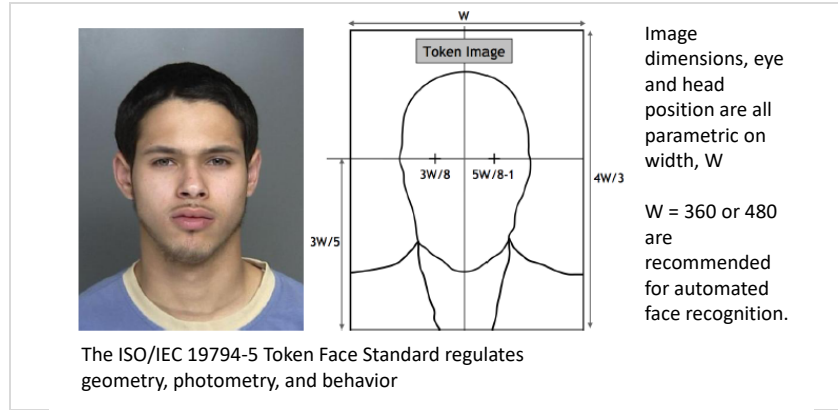


Figure 3 – Canonical Portrait Photograph, as standardized in ISO/IEC 19794-5

Formally, if a face verification algorithm,  $V$ , compares two samples  $X_1$  and  $X_2$ , to produce a comparison score

$$S = V(X_1, X_2) \quad [1]$$

this standard requires quality algorithms to predict  $S$  from  $X_1$  alone but under the assumption that  $X_2$  would be a canonical portrait image i.e. a pristine image of the same subject that is fully conformant to ISO and ICAO specifications<sup>3</sup>. Thus, a quality algorithm  $F$  operating on an image  $X_1$  produces value

$$Q = F(X_1) \quad [2]$$

that in the sense defined later predicts  $S$  because it implicitly assumes the comparison

$$V(X_1, X_{\text{PORTRAIT}}) \quad [3]$$

This goal respects the ISO/ICAO specification as the reference standard for automated face recognition. The light grey text indicates that quality assessment must be done blind<sup>4</sup>, targeting a hidden virtual portrait image.

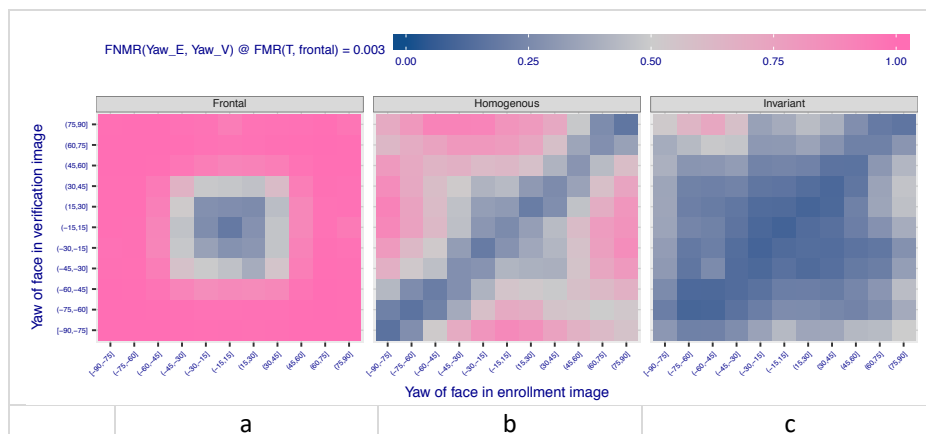
Without this formulation of the quality problem the position, noted in the academic literature, that quality assessment cannot be done on a single image - that quality should “come in pairs” - would be correct. Such assertions note that recognition outcomes (that are the result of comparing two images) depend on the properties of both images. For example, consider Figure 4. It presents the false non-match rates (FNMR) from three face verification algorithms executed on a database of images where facial pose (yaw) differs between the two images used in a comparison. Figure 4a corresponds to an algorithm that gives high FNMR except when the two images are frontal.

<sup>3</sup> A reasonable question here would be why the target must be a portrait. The answer is that it doesn’t have to be, that quality assessment might be done also referencing some other standard view of a face. This might in fact be desirable once we recall that forensic face examiners have preferred views where the ear is visible. Indeed, the immigration agencies in the United States used to require a quarter-left view on identity document for just this reason. For now, however, the target must be the ISO/ICAO portrait because the face recognition industry is currently capitalized on the basis of frontal face recognition. This standard could be extended to adopt quality assessment against some other standardized view.

<sup>4</sup> The term “blind” is borrowed from the image fidelity literature in which a “blind PSNR” i.e. peak signal to noise ratio is computed from, for example, a JPEG image or a video clip as a statement of quality. Such techniques may have applicability here.

Figure 4 – The classes of algorithm response to comparison of pairs of images that differ in the yaw angle of the face.

This is the common case. Figure 4b shows an algorithm that is capable of matching images of a face with the same yaw angle, even if non-frontal. Finally, Figure 4c represents the (rare) case of an algorithm that offers considerable pose invariance<sup>5</sup>.



The point of this example is that recognition outcome may actually depend on the pair of images, but quality assessment, run on a single image potentially long before any recognition occurs, must assume a reference standard, here the ISO/ICAO portrait.

### 3.4. Quality value as predictor of true matching performance

Quality values are most useful as predictors of false negative outcomes, arising from low genuine scores. The alternative, as predictors of false positives, is considered less feasible because these arise from high impostor scores which should result only from facial (e.g. anatomical) similarity of the input image pair. However, some recognition algorithms do yield spurious high impostor scores from certain images. Examples are from similar eye-glasses, or hair styles. Such effects are unwelcome but are not relevant to a quality standard.

### 3.5. Recognition algorithm dependence

This standard requires quality algorithms to predict false negative recognition outcomes. Of course, recognition algorithms extract various proprietary features from face images and have different accuracies and tolerance of quality problems. However, given extreme degradations they all fail: Sufficiently over- or under-exposed images will cause false negatives; blurred faces, likewise; faces presented at high pitch or yaw angles will generally cause failure<sup>6</sup>. The approach in building a quality algorithm, and in testing it, is to predict failure from a set of recognition algorithms.

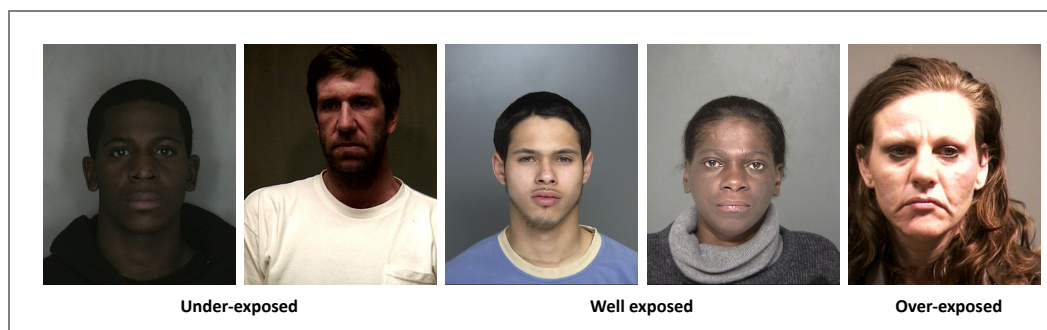


Figure 5 – High-resolution non-frontal views for forensics

<sup>5</sup> The figure is extracted from P. Grother, M. Ngan, K Hanaoka, *Ongoing Face Recognition Vendor Test (FRVT) Part 1: Verification*, NIST Interagency Report, 2019.

<sup>6</sup> The algorithm in Figure 3C shows wide pose invariance. However, this is a result for a recent (2018) prototype from a single developer, and frontal pose gives higher genuine scores even for this recognition algorithm.

## 4. Evaluation of image quality assessment algorithms

### 4.1. Overview

This section describes evaluation of algorithm submitted to NIST FRVT Image Quality Assessment Evaluation.

The evaluation is based on the execution of each quality assessment algorithm on large numbers of images for which reference target quality values are available.

### 4.2. Image and reference quality datasets

NIST will use several sets of images, initially reference portrait images. See NIST Interagency Report 8238 for recognition results using mugshot images.

For each image, NIST will establish reference quality values based on genuine recognition similarity scores obtained using that image. This assigns the lowest target scores to those images that are involved in false non-match errors. The annotation procedure might be based on an image quality oracle [Phillips13]. The target scores form the ideal performance of quality measures for a given data set.

**NOTE** Ageing causes face appearance to change and this causes genuine similarity scores to decline. This will occur even if all the images are perfectly captured with high quality. For this reason, the image quality assessment datasets will exclude image pairs for which there is large elapsed time between captures.

### 4.3. Performance metrics

The quality values should be predictors of the target scores. That is, the ordering of the quality values should be identical to that of the target scores, as required by [Grother07]. In general, this prediction will be imperfect, as shown in Figure 6.

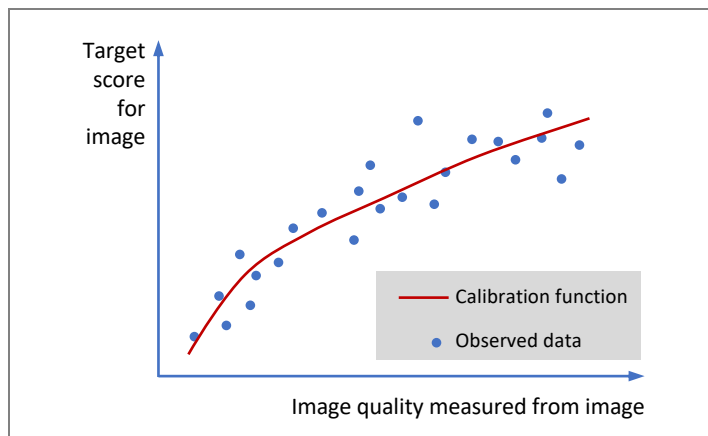


Figure 6 – Example association of quality scores with targets

Given  $N$  genuine image pairs,  $x_i$ , and  $N$  reference recognition scores,  $t_i$ , NIST will execute each image quality assessment algorithm to produce  $2N$  quality values,  $q_{1i}$  and  $q_{2i}$  from which NIST will compute  $N$  values  $q_i = \min(q_{1i}, q_{2i})$ . The use of  $\min()$  embeds the assumption that a low comparison score will be caused by the image with the lower image quality.

NIST will relate quality to reference recognition scores several methods such as:

- Scalar measures of association, such as Kendall's correlation coefficient, particularly at low ranks.
- Error vs. reject plots [Grother07] computed by taking proportions of the lowest computed quality values and graphing<sup>7</sup> how closely they correspond to the lowest target scores.

<sup>7</sup> Specifically, when a proportion  $0 < r \leq 1$  of the lowest quality values i.e. the set  $Q = \{i : 1 \leq i \leq L, q_i \leq q_{rN}\}$  are rejected this should lead to rejection of the lowest associated target values i.e. those that cause false rejections. Formally, compute

$$E(r) = 1 - L^{-1} \sum_Q H(t_i - T)$$

where  $T$  is the  $rN$ -th lowest target score;  $H$  is the unit step function;  $t_i$ , is the  $i$ -th target value; and index  $i$  runs over the  $rN$  indices in the set  $Q$ .

#### 4.3.1. Handling failure to process

Given an IQAA, NIST will execute the image quality assessment algorithm on all  $2N$  images in the reference dataset. This will generally produce  $M \leq 2N$  quality values,  $q_i$ . We will assign  $q_i = 0$  to the  $M$  failure cases.

The test report will disclose the number of failures,  $2N - M$ .

#### 4.3.2. Calibration

While quality values must exist on the range  $[0,100]$ , their distribution within that range will vary between algorithms. For example, one IQAA might give most values on  $[60,100]$  while another might assign values on  $[10,90]$ . This implies a need to do calibration.

NIST will explore calibration by computing, for example, the function, shown in red in Figure 6, that results from isotonic regression [Han12] of target score against quality score. That function,  $F$ , minimizes  $\sum (t_i - F(q_i))^2$  while requiring  $F$  to be monotonic. This can be achieved via the Pool Adjacent Violators algorithm. Once this function is available it can be used to map raw quality measurements,  $Q$ , to a calibrated quality  $F(Q)$  by simple lookup.  $F$  will generally not be linear.

NIST will report calibration functions.



## Bibliography

1.	[Grother07]	Patrick Grother, Elham Tabassi, <i>Performance of Biometric Quality Measures</i> , IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 29, Issue 4, April 2007. <a href="https://ieeexplore.ieee.org/document/4107559/">https://ieeexplore.ieee.org/document/4107559/</a>
2.	[Han12]	Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu, <i>Monotonic Regression: A New Way for Correlating Subjective and Objective Ratings in Image Quality Research</i> , IEEE Transactions on Image Processing, Volume: 21, Issue: 4, April 2012, Page(s): 2309 – 2313, 06 October 2011, DOI: <a href="https://doi.org/10.1109/TIP.2011.2170697">10.1109/TIP.2011.2170697</a>
3.	[Phillips13]	P. J. Phillips, J. R. Beveridge, D. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, S. L. Cheng, M. N. Teli, and H. Zhang, <i>On the Existence of Face Quality Measures</i> , In proc. IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS 2013), September 30, 2013. <a href="http://ws680.nist.gov/publication/get_pdf.cfm?pub_id=914258">http://ws680.nist.gov/publication/get_pdf.cfm?pub_id=914258</a>
4.	[ANSI-NIST Type 10]	NIST Special Publication 500-290 Edition 3, <i>Data Format for the Interchange of Fingerprint, Facial &amp; Other Biometric Information</i> , ANSI/NIST-ITL 1-2011, August 22, 2016 <a href="https://dx.doi.org/10.6028/NIST.SP.500-290e3">https://dx.doi.org/10.6028/NIST.SP.500-290e3</a>
5.	[ISO/IEC-2005]	ISO/IEC 19794-5:2005 Biometric Data Interchange Formats – Face Image data
6.	[ISO/IEC-2019]	ISO/IEC 39794-5:2019 Extensible biometric data interchange formats – Face Image data
7.	[ISO-Geometry]	ISO/IEC 1974-5:2005/AMD 1:2007 Conditions for taking photographs for face image data
8.	[ISO-Quality]	ISO/IEC TR 29794-5 Biometric sample quality -- Part 5: Face image data Technical report for aspects of quality specific to facial images. It <ul style="list-style-type: none"> <li>– specifies terms and definitions that are useful in the specification, use and testing of face image quality metrics;</li> <li>– defines the purpose, intent, and interpretation of face image quality scores.</li> </ul> Performance assessment of quality algorithms and standardization of quality algorithms are outside the scope of ISO/IEC TR 29794-5:2010.
9.	[ISO-Conform]	ISO/IEC 29109-5:2014 Conformance testing methodology for biometric data interchange formats defined in ISO/IEC 19794 -- Part 5: Face image data. This specifies conformance tests for the syntax of records defined in the ISO/IEC 19794-5:2005 biometric data interchange format standard.
10.	[ICAO-Portrait]	PORTRAIT QUALITY - REFERENCE FACIAL IMAGES FOR MRTD Version: 1.0 Date – 2018-04, International Civil Aviation Organization
11.	[ISO-3D]	ISO/IEC 19794-5:2005/AMD 2:2009 Three-dimensional face image data interchange format