Draft NISTIR XXXX
(For Public Comment)

Ongoing Face Recognition Vendor Test (FRVT)

Part 5: Face Image Quality Asssessment

Patrick Grother
Austin Hom
Mei Ngan
Kayee Hanaoka
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing

2020/03/06

National Institute of Standards and Technology
U.S. Department of Commerce

ACKNOWLEDGMENTS

The authors would like to thank the U.S. Department of Homeland Security Office of Biometric Identity Management (DHS OBIM) for their collaboration and contributions to this activity. The authors are also grateful to staff in the NIST Biometrics Resesarch Laboratory for infrastructure supporting rapid evaluation of algorithms.

DISCLAIMER

Specific hardware and software products identified in this report were used in order to perform the evaluations described in this document. In no case does identification of any commercial product, trade name, or vendor, imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

INSTITUTIONAL REVIEW BOARD

The National Institute of Standards and Technology's Research Protections Office reviewed the protocol for this project and determined it is not human subjects research as defined in Department of Commerce Regulations, 15 CFR 27, also known as the Common Rule for the Protection of Human Subjects (45 CFR 46, Subpart A).

Any comments submitted will be released in response to Freedom of Information Act (FOIA) requests.

Contents

ACKNOWLEDGMENTS 1

DISCLAIMER 1

INSTIUTIONAL REVIEW BOARD 1

1 THE FRVT QUALITY ASSESSMENT ACTIVITY 3
1.1 STATUS . . . 3
1.2 INTRODUCTION . . . 3
1.3 QUALITY VALUE AS PREDICTOR OF TRUE MATCHING PERFORMANCE . . . 4
1.4 SHOULD THE QUALITY ALGORITHM PREDICT FALSE POSITIVES? . . . 4
1.5 RECOGNITION ALGORITHM DEPENDENCE . . . 5

2 QUANTITATIVE GOAL FOR QUALITY SCALARS 5

3 ALGORITHMS 6

4 IMAGE DATASETS 6
4.1 APPLICATION IMAGES . . . 6
4.2 WEBCAM IMAGES . . . 6
4.3 WILD IMAGES . . . 7

5 METRICS 7
5.1 QUALITY AS PREDICTOR OF RECOGNITION OUTCOMES . . . 7
5.2 HANDLING FAILURE TO PROCESS . . . 8
5.3 CALIBRATION . . . 8

6 RESULTS 8
6.1 APPLICATION VERSUS WEBCAM IMAGES . . . 10
6.2 WILD IMAGES . . . 13

List of Tables

1 FRVT QUALITY ASSESSMENT PARTICIPANTS . . . 6

List of Figures

1 QUALITY EXAMPLES . . . 3
2 VISUAL QUALITY EXAMPLE . . . 4
3 CANONICAL PORTRAIT PHOTOGRAPH . . . 5
4 IMAGE SAMPLES . . . 7
5 EXAMPLE ASSOCIATION OF QUALITY SCORES WITH TARGETS . . . 7
6 PERFORMANCE SUMMARY: QUALITY SCORE DISTRIBUTION . . . 9
7 APPLICATION VS. WEBCAM IMAGES: FNMR VS. REJECT . . . 10
8 APPLICATION VS. WEBCAM IMAGES: MATCH SCORE VS. QUALITY SCORE . . . 11
9 APPLICATION VS. WEBCAM IMAGES: MATCH SCORE VS. QUALITY SCORE . . . 12
10 WILD IMAGES: FNMR VS. REJECT . . . 13
11 WILD IMAGES: MATCH SCORE VS. QUALITY SCORE . . . 14
12 WILD IMAGES: MATCH SCORE VS. QUALITY SCORE . . . 15

1 The FRVT Quality Assessment Activity

1.1 Status

This report is a draft NIST Interagency Report, and is open for comment. The FRVT Quality Assessment evaluation is an ongoing test that remains open to new participation. This report will be updated as new algorithms are evaluated, as new datasets are added, and as new analyses are included. Comments and suggestions should be directed to frvt@nist.gov.

1.2 Introduction

Face recognition accuracy has improved markedly due to development of new recognition algorithms and approaches. Current NIST FRVT results show increased error rates in applications where photography of faces is difficult or when stringent thresholds must be applied to recognition outcomes to reduce false positives. FRVT results also show that controlled capture, good portrait quality images provide the lowest error rates in face recognition applications. While standards exist for interchange of facial images (ISO-IEC-2005 [1] superseded by ISO-IEC-2019 [3] which includes ICAO-Portrait [9] and ANSI-NIST Type 10 [6]) and those standards additionally formalize the capture of images, there are no standards for how face image quality must be assessed [1] nor are there performance evaluations for automated quality assessment algorithms.

The NIST FRVT Quality Assessment activity establishes specifications for face image quality 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. Fundamentally, 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 right to left, implies that better images have higher quality values, where the term better here is the subject of this activity.

Figure 1: Four faces with example image quality values. The subject in the photos is a NIST employee.

A number of academic methodologies and commercial tools exist that report quality scalars. One such published quality assessment implementation [10] visualizes the outcomes of their quality algorithm on a set of wild images where face capture is non-cooperative, very unconstrained, with wide yaw, pitch, and roll pose variation, as presented in Figure 2. While only three levels of quality are reported, we can observe that the range of quality in the wild dataset used is very large, much larger than is evident in images collected in cooperative environments such as visa or port of entry settings, which suggests that binning wild imagery by quality might be an easier problem than the latter.

The efficacy of quality assessment algorithms is important, because they can make two kinds of error: false rejection - saying an image is poor when it is not, which can affect costs; false acceptance - saying an image is good when it is not,

[1] 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 "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.

Figure 2: Visualization of visual quality from [10] on wild imagery.

which can affect future recognition errors. Implicit in this statement is that image quality should predict recognition failure, and this gives the basis of the evaluation documented in this report.

1.3 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 similarity of the input image pair. However, some recognition algorithms do yield spurious high impostor scores from certain images that have similar eye-glasses, or hair styles. Such effects are unwelcome but are not part of a scalar quality standard.

1.4 Should the quality algorithm predict false positives?

This question arises because it has been reported, anecdotally, that some recognition algorithms produce false positives when either or both the images are of poor quality. This report will be updated to show examples of such behavior if and when they're observed. However, currently, we do not require algorithms to predict false positive outcomes, because we hypothesize that it will usually be the case that quality problems that cause false positives will also cause false negatives and that, therefore, it is sufficient for a quality algorithm to be assessed only on the basis of false negative prediction. Thus if overexposure, for example, caused a particular recognition algorithm to produce high impostor scores, we hypothesize that it will also cause low genuine scores. There will certainly be counter-examples to this, but for the purposes of putting quality assessment on a quantitative footing, we target false negatives which are the dominant (most likely) source of error in cooperative 1:1 and 1:N applications. This is formalized in section 2.

1.5 Recognition algorithm dependence

The evaluation 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. The approach in building a quality algorithm, and in testing it, is to predict failure from a set of recognition algorithms.

2 Quantitative goal for quality scalars

ISO/IEC 29794-1 [2] 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 partially occluded 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, 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.

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.

Figure 3: Canonical Portrait Photograph, as standardized in ISO/IEC 19794-5. The subject in the photo is from NIST Special Database 32 [5].

Formally, if a face verification algorithm, V, compares two samples X1 and X2, to produce a comparison score

$$S = V(X1, X2) \tag{1}$$

the standard requires quality algorithms to predict S from X1 alone but under the assumption that X2 would be a canonical portrait image of the same subject that is conformant to ISO and ICAO specifications. Thus, a quality algorithm F operating on an image X1 produces value

$$Q = F(X1) \tag{2}$$

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

$$V(X1, X_{PORTRAIT}) \tag{3}$$

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, targeting a hidden virtual portrait image.

3 Algorithms

The FRVT Quality Assessment activity is open to participation worldwide. The participation window opened in May 2019, and the test will evaluate submissions on an ongoing basis. There is no charge to participate. The process and format of algorithm submissions to NIST are described in the FRVT Quality Assessment Application Programming Interface (API) document. Participants provide their submissions in the form of libraries compiled on a specific Linux kernel, which are linked against NIST's test harness to produce executables. NIST provides a validation package to participants to ensure that NIST's execution of submitted libraries produces the expected output on NIST's test machines.

This report documents the results of all algorithms submitted for testing to date. Table 1 lists the participants who submitted algorithms to FRVT Quality Assessment.

Participant Name	Short Name	Submission Sequence	Submission Date
China Electronics Import-Export Corp	ceiec	001	2019.06.12
Guangzhou Pixel Solutions Co Ltd	pixelall	000	2020.01.15
Lomonosov Moscow State University	intsysmsu	000	2019.08.19
Paravision (EverAI)	paravision	001	2019.12.23
Rank One Computing	rankone	000	2019.06.03
Rank One Computing	rankone	001	2019.11.12
Universidad Autonoma de Madrid	uam-jrc-faceqnet	000	2019.08.19

Table 1: FRVT Quality Assessment Participants

4 Image Datasets

4.1 Application Images

The images are collected in an attended interview setting using dedicated capture equipment and lighting. The images, at size 300x300 pixels, are somewhat smaller than normally indicated by ISO. The images are all high-quality frontal portraits collected in immigration offices and with a white background. As such, potential quality related drivers of high false match rates (such as blur) can be expected to be absent. The images are encoded as ISO/IEC 10918 i.e. JPEG. Over a random sample of 1000 images, the images have compressed file sizes (mean: 42KB, median: 58KB, 25-th percentile: 15KB, and 75-th percentile: 66KB). The implied bit-rates are mostly benign and superior to many e-Passports. When these images are provided as input into the algorithm, they are labeled with the type "ISO".

4.2 Webcam Images

These images are taken with a camera oriented by an attendant toward a cooperating subject. This is done under time constraints, so there are roll, pitch, and yaw angle variation. Also, background illumination is sometimes bright, so the face is under exposed. Sometimes, there is perspective distortion due to close range images. The images are in poor

conformance with the ISO/IEC 19794-5 Full Frontal image type. The images have mean interocular distance of 38 pixels. The images are all live capture. When these images are provided as input into the algorithm, they are labeled with the type "WILD". Examples of such images are included in Figure 4 and Figure 4 in NIST Interagency Report 8271.

4.3 Wild Images

These images include many photojournalism-style photos. Images are given to the algorithm using a variable but generally tight crop of the head. Resolution varies very widely. The images are very unconstrained, with wide yaw, pitch, and roll pose variation. Faces can be occluded, including hair and hands. When these images are provided as input into the algorithm, they are labeled with the type "WILD".

(a) Application *(b) Webcam* *(c) Wild*

Figure 4: Samples of images used in this report. The subject in the photos is a NIST employee.

5 Metrics

5.1 Quality as predictor of recognition outcomes

The quality values should be predictors of the target scores. That is, the ordering of the quality values should be closely aligned with the target scores, as required by [7]. In general, this prediction will be imperfect, as shown in Figure 5.

Figure 5: 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, $q1_i$ and $q2_i$ from which NIST will compute N values $q_i = min(q1_i, q2_i)$. 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 using several methods such as:

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

5.2 Handling failure to process

During testing, 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, $2N - M$.

5.3 Calibration

While quality values must exist on the range $[0, 100]$, their distribution within that range will vary between algorithms. For example, one image quality assessment algorithm 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 that results from isotonic regression [8] 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.

6 Results

The results in this section assess the quality assessment algorithms on application, webcam, and wild imagery. Figure 6 plots algorithm quality score distribution for each image type. Correlation of quality scores with match scores is conducted using the quality assessment algorithms submitted to this test and 1:1 face verification algorithms from the NIST Ongoing FRVT 1:1 Evaluation [4]. For each developer that submitted to this test, where available, we also selected a 1:1 matcher from the same developer to use in our analysis.

Application vs. Webcam Images: The results of Section 6.1 correlate quality scores with match scores generated by comparing application images with webcam images. The quality scores used for analysis are from the webcam images used for verification.

Wild Images: The results of Section 6.2 correlate quality scores with match scores generated by comparing wild images with wild images. The quality score selection method here, is the minimum (or lower) quality score between the enrollment and verification images.

Figure 6: This density plot shows the quality score distribution for each quality assessment algorithm, on all the types of imagery evaluated.

6.1 Application versus Webcam Images

Figure 7: This plot shows FNMR vs. Reject, showing how FNMR reduces when the worst quality data is thrown away. Each row corresponds to a recognition algorithm, and the lines within each panel correspond to quality assessment algorithms. A perfect quality algorithm would predict which images are implicated in false non-matches. The quality score is from the verification (webcam) image. The line labeled "PERFECT" is generated using $\max((FNMR - x), 0)$*. The closer the quality algorithm line is to the "PERFECT" line, the better the quality prediction performance is relative to recognition outcomes. The solid grey horizontal line represents the starting FNMR, and the dotted grey horizontal line represents when FNMR has been reduced by half. The point where the quality algorithm line intersects with the FNMR(0)/2 line indicates the fraction of lowest quality scores that need to be rejected in order to reduce FNMR by half.*

Figure 8: This plot shows boxplots of match scores of various 1:1 face verification algorithms plotted against quality scores. The quality score is from the verification (webcam) image.

Figure 9: This plot shows scatterplots of match scores of various 1:1 face verification algorithms plotted against quality scores. Each dot represents the quality score of the verification (webcam) image, plotted against the median match score.

6.2 Wild Images

Figure 10: This plot shows FNMR vs. Reject, showing how FNMR reduces when the worst quality data is thrown away. Each row corresponds to a recognition algorithm, and the lines within each panel correspond to quality assessment algorithms. A perfect quality algorithm would predict which images are implicated in false non-matches. The quality score was generated using the minimum score between the enrollment (wild) and verification (wild) image. The line labeled "PERFECT" is generated using $\max((FNMR - x), 0)$. *The closer the quality algorithm line is to the "PERFECT" line, the better the quality prediction performance is relative to recognition outcomes. The solid grey horizontal line represents the starting FNMR, and the dotted grey horizontal line represents when FNMR has been reduced by half. The point where the quality algorithm line intersects with the FNMR(0)/2 line indicates the fraction of lowest quality scores that need to be rejected in order to reduce FNMR by half.*

Figure 11: This plot shows boxplots of match scores of various 1:1 face verification algorithms plotted against quality scores. The quality score was generated using the minimum score between the enrollment (wild) and verification (wild) image.

Figure 12: This plot shows scatterplots of match scores of various 1:1 face verification algorithms plotted against quality scores. Each dot represents the minimum quality score between the enrollment (wild) and verification (wild) image, plotted against the median match score.

References

[1] ISO/IEC 19794-5:2005 Biometric Data Interchange Formats Face Image data.

[2] ISO/IEC 29794-1:2016 Information Technology - Biometric sample quality - Part 1: Framework.

[3] ISO/IEC 39794-5:2019 Extensible biometric data interchange formats Face Image data.

[4] NIST Ongoing Face Recognition Vendor Test (FRVT) 1:1 Verification. https://pages.nist.gov/frvt/html/frvt11.html.

[5] NIST Special Database 32 - Multiple Encounter Dataset (MEDS-II). https://www.nist.gov/itl/iad/image-group/special-database-32-multiple-encounter-dataset-meds.

[6] NIST Special Publication 500-290 Edition 3, Data Format for the Interchange of Fingerprint, Facial and Other Biometric Information, ANSI/NIST-ITL 1-2011. https://dx.doi.org/10.6028/NIST.SP.500-290e3, August 2016.

[7] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):531–543, April 2007.

[8] Y. Han, Y. Cai, Y. Cao, and X. Xu. Monotonic regression: A new way for correlating subjective and objective ratings in image quality research. *IEEE Transactions on Image Processing*, 21(4):2309–2313, April 2012.

[9] International Civil Aviation Organization. Portrait quality - reference facial images for mrtd version: 1.0, April 2018.

[10] Weidi Xie and Andrew Zisserman. Multicolumn networks for face recognition. In *British Machine Vision Conference*, 2018.