Draft NISTIR XXXX
(For Public Comment)

Ongoing Face Recognition Vendor Test (FRVT)
Part 5: Face Image Quality Assessment

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

2022/07/13

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).

FOIA

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

STATUS

2022-07-12 The fifth draft of the report adds results for a new and different algorithm from Rankone computing. It also introduces a new visualization: simple plot of estimated FNMR for samples of a given quality - see the example in the Executive Summary and Figure 7.

We have announced a new quality-related track of FRVT dedicated to detection of specific image defects. The draft test plan is open for comment until 2022-08-18.

2021-12-22 This fourth draft of the report adds results for the first algorithm from a new developer: Tevian. It also fixes two bugs: it uses the correct formula for the ideal rejection line in Figure 10; it corrects equation 6 in respect to the denominator and the mate score acceptance vs. rejection.

2021-09-24 This third draft of the report simplifies prior versions by focusing assessment of quality algorithms on their ability to predict recognition errors of a pooled collection of 33 leading face verification algorithms (instead of individual algorithms).

- ▷ We added a leaderboard to the quality home page.
- ▷ We added a PDF chart to the home page depicting achievable FNMR reductions. The graph allows a better determination of the most effective quality algorithms, and the absolute potential for their use.

2020-10-19 This second draft of the report includes a major update to much of the content, as detailed below.

- ▷ We rewrote the introduction for clarity and added text on use-cases for quality scalars in section 1.1.
- ▷ We include a new error-tradeoff metric. By using thresholded verification scores to define accept-reject ground-truth, we compute two new quantities - the Incorrect Sample Acceptance and Rejection Rates - expressing occurence of, respectively, of how many samples are deemed to have high quality but ultimately do not match, and of how many samples are assigned low quality but then are matched. These rates are plotted as a function of quality value.
- ▷ We started plotting of error-vs-reject metrics with a logarithmic rejection axis to emphasize effect of rejection of small proportions of low quality data.
- ▷ We changed the way recognition algorithm failure-to-template occurences were handled. Previously we ignore all verification comparisons for which one or both of the input templates were missing. Now, instead, we regard such occurences as producing a low score. This affords quality algorithms an opportunity to correctly predict these outcomes.
- ▷ We changed the denominator in the error-versus-reject computation of FNMR to be the number of genuine samples below score threshold after quality rejection divided by the number of genuine samples left after quality rejection. The denominator had previously been the total number of genuine samples without rejection. This update renders error-versus-reject curves less favorable.
- ▷ We added a summary in the next section.

SUMMARY

OVERVIEW

This report summarizes the ongoing Quality Assessment track of the FRVT. Face image quality assessment is a less mature field than face recognition, and so NIST regards this work as a development activity rather than an evaluation. In particular, as performance metrics remain under-development - new ones were introduced in this edition of the report - we encourage submission of both new algorithms and comments toward improved formulation and analysis of the problem. Questions, comments and suggestions should be directed to frvt@nist.gov.

QUALITY MEASUREMENT FOR PREDICTING FAILURE

The most important use-case for quality asssessment is to exclude poor photos at the time of collection. This can be done by detection of specific image defects or by computing an overall quality score and comparing it with an acceptance threshold. The definition of "poor" used in this work is based on the likelihood that an image will not match a prior good quality photo, i.e. a false non-match. We seek algorithms that assign low quality to samples that will not match (and high values to samples that will). As Figure 1 shows the assignment of a low quality value by this particular quality algorithm (rankone-004) means a high likelihood that the sample will fail. The trend downward shows the expected behavior - that higher quality indicates fewer expected failures. Note too that the algorithm assigns 99.3% of samples the highest quality value, including to some samples that do not match.

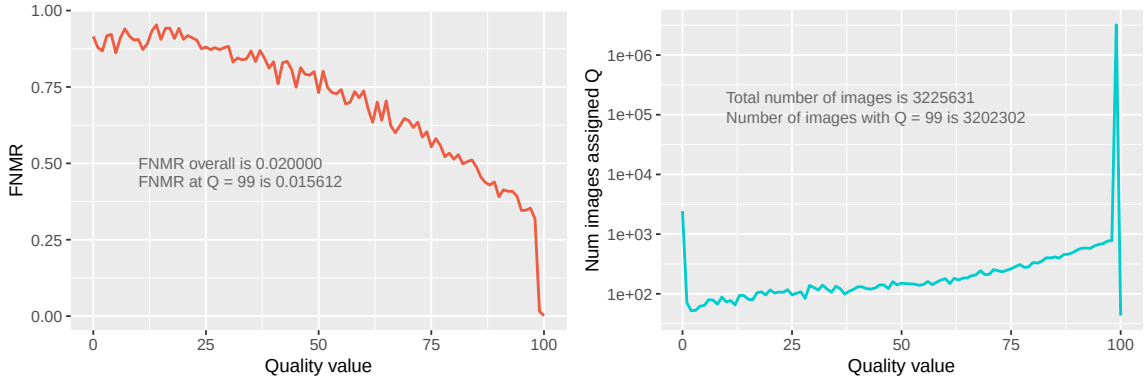
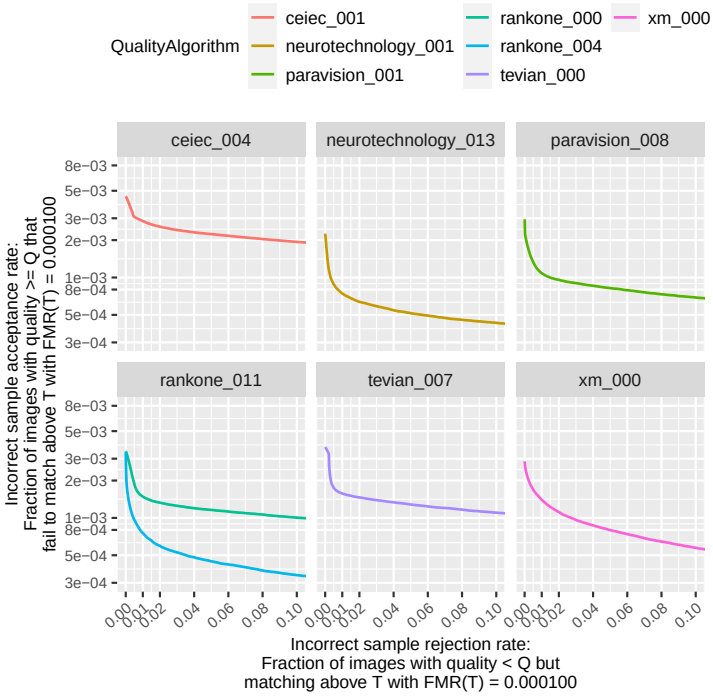


Figure 1: For one quality algorithm, the upper panel shows the proportion of samples of each quality value that fail when proceessed by a set of 22 recent recognition algorithms. The lower panel simply shows how many border crossing photos (out of 3.2 million) are assigned a particular quality algorithm.

The effectiveness of the quality algorithm is discussed extensively in section 5 using metrics of section 4.

QUALITY MEASUREMENT FOR SAMPLE ACCEPTANCE

The most exacting role for quality values is for making photo acceptance decisions on individual images. As such the quality algorithm will make Type I and II errors i.e. incorrectly assigning high quality to samples that ultimately do not match, and assigning low quality values to samples that do. We use FRVT 1:1 verification examples for matching.



We find quality assessment algorithms, from Neurotechnology, Paravision and RankOne Computing, that can predict false negative decisions produced by their respective face recognition algorithms. The inset figure shows the tradeoff between the two kinds of error. Each panel pairs one or more quality algorithms from a developer with a recognition algorithm from the same developer. Each trace corresponds to sweeping a quality threshold over its possible values. The horizontal axis quantifies incorrect sample relection - in, for example, a passport application process high values would annoy cooperating applicants. The vertical axis quantifies the reason for using a quality algorithm, namely reductions in dowwnstream recognition error rate. For example, with the Neurotechnology algorithm, if we can tolerate incorrect rejection of 1% of good samples, the expected false non-match rate falls from 0.004 to below 0.001.
Specifically, as the inset figure shows, they are capable of simultaneously correctly assigning low quality values to 1% of 3 225 633 border crossing photos that subsequently produce false negative outcomes when compared with high quality visa-like images. The result is that false non-match rates are reduced by an order of magnitude. See the metrics in sec. 4.2 for details.

However, the algorithms are not effective at prediction across-developer. This means the current quality algorithms are unsuited to the case where a quality control algorithm is used during capture for samples to be sent to a receiving system that employs a face recognition algorithm from a different manufacturer. See Figure **??**.

QUALITY FOR SURVEYS

As discussed in the use-cases text of section 1.1 quality values can be used to survey over large collections of images collected at certain sites or times, for example. As shown in the inset figure, the results show that quality values, on aggregate, have a higher-is-better relationship with recognition scores. This will make quality assessment meaningful in a survey role. The variance, however, is quite high so the distributions overlap so that any given image may be assessed incorrectly - as discussed in the prior section. The Neurotechnology and, particularly, the Xiamen quality algorithms, exhibit a consistent positive trend relationship with recognition scores from four algorithms. The trends on these plots is likely consistent with the worthwhile use of these algorithms as a survey tool, at least on this kind of data. For Paravision and Rankone the relationship is also positive increasing except for the lowest quality values which give higher higher recognition scores. This "whole distribution" view is not relevant to the issue of using quality for rejection of specific difficult to recognize samples, discussed below.

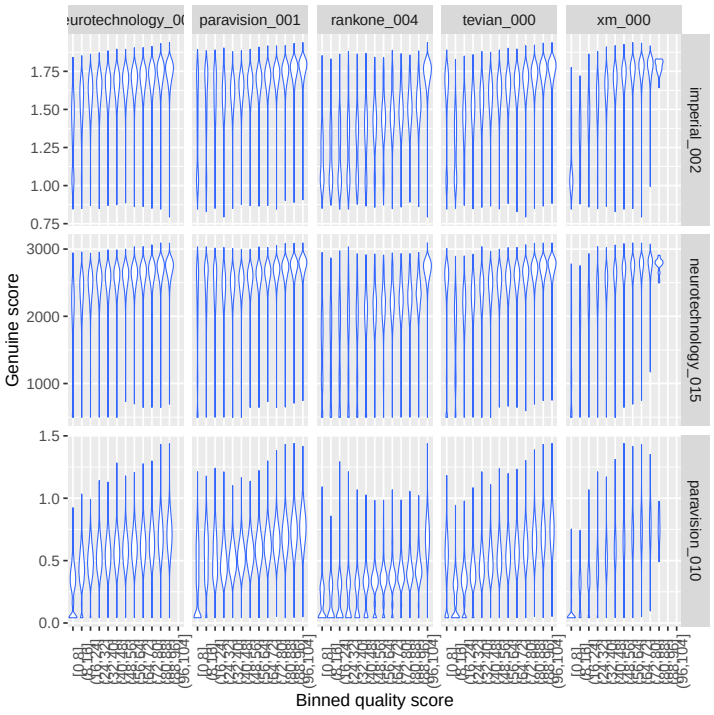


Figure 2: The figure shows four panels each containing thirteen boxes, produced by binning quality values. The columns show quality algorithms, and the rows show recognition scores from the algorithms they're intended to predict. The assignment of low quality scores to images that produce low match scores is present especially for the case where the algorithms hail from the same developer.

However, as shown in the main body of the report - see Figures 14 and 24 - some algorithms are inferior to the example shown here, particularly for the case when the developer of the recognition and quality assessment algorithms is different.

Contents

List of Tables

List of Figures

1 Introduction

As documented in FRVT verification and identification trials face recognition accuracy has improved markedly due to the development of new recognition algorithms and approaches. Simultaneously accuracy has been supported by improved compliance to the appearance-related requirements written into standards for interchange of facial images i.e. ISO-IEC-19794-5:2005 [1], as superseded by ISO-IEC-39794-5:2019 [3] which includes ICAO-Portrait [9] specifications, and ANSI-NIST Type 10 [6].

Nevertheless Recent NIST FRVT results show higher 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. Error rates increase when conformance to the frontal view standard is not achieved.

The quality assessment track of FRVT seeks to improve automated detection of poor images by evaluating algorithms that report scalar quality values. Given an image X, an image quality assessment algorithm, F, produces a scalar quality score, $Q = F(x)$. Examples of this are shown in Figure 3. 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 3: Four faces with example image quality values. The subject in the photos is a NIST employee.

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.

FRVT conceives of quality scalars as being measures of utility rather than fidelity because utility of a sample to a recognition engine is what drives outcome operationally and is of most interest to end-users.

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 4. 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.

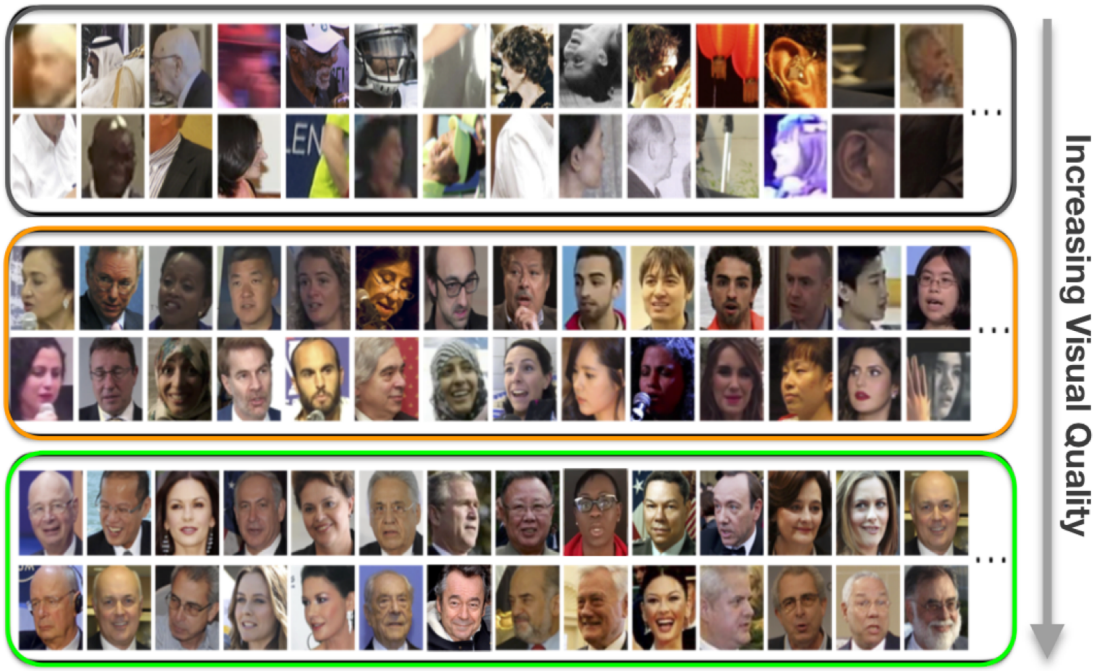


Figure 4: Visualization of visual quality from [10] on wild imagery.

1.1 Use-cases

In fingerprints, quality algorithms are applied during initial enrollment in applications where the goal is to retain images as authoritative reference samples against which future recognitions are done. For face recognition, this is the first of three uses-cases:

- **1. Photo acceptance**: Foremost, a scalar image quality value can be used to make an acceptance or rejection decision. 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.

  NOTE Ordinarily the photo acceptance function should not replace, or be used in place of, recognition because recognition outcomes are usually of primary interest. Thus rather than assess quality of a verification sample, it would be more reliable to just match it against the claimed reference sample: a match result is the ultimate quality indicator.

- **2. 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.

  NOTE In cases where samples are collected from the same persons regularly for example in a frequent traveler system, aggregated results from the matching of genuine image pairs will be an excellent indicator of expected recognition performance and will reveal image quality variations across time, collection sites etc.

- **3. Photo selection**: Given $K > 1$ images of a person, compute their quality values and select the best. 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 An alternative to selection would be to retain multiple samples for later recognition. Thus, in an identification application, a system might generally enroll all K images of a person rather than select just one. This recommendation is appropriate if the quality algorithm an imperfect predictor of recognition outcome and it may arise that an enrolled image with lower quality might be successfully matched to particular probe images due to certain idiosyncratic 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.

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

1.2 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 discussed in the next section.

The ISO/IEC 29794-1 and -5 standards conceive of quality values serving 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 an enrolled canonical high-quality portrait image of the form given in Figure 5.

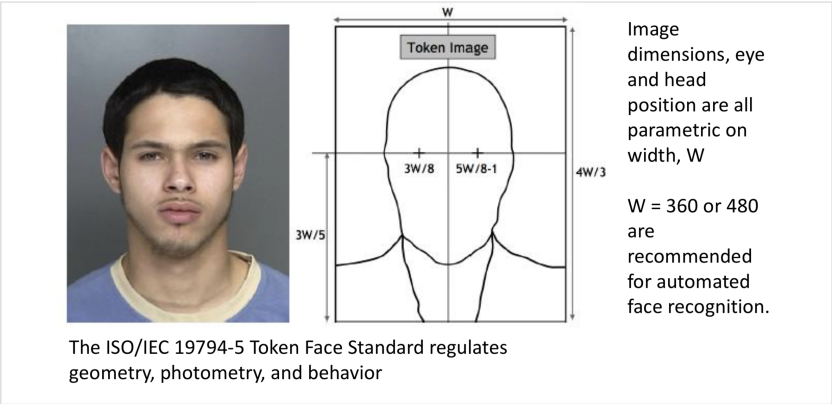


Figure 5: Canonical Portrait Photograph, as standardized in ISO/IEC 19794-5 now superseded by ISO/IEC 39794-5. The subject is taken from NIST Special Database 32 [5].

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

$$S = V(X_1, X_2) \tag{1}$$

the standard requires quality algorithms to predict S from X_1 alone but under the assumption that X_2 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 X_1 produces value

$$Q = F(X_1) \tag{2}$$

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

$$V(X_1, X_{PORTRAIT}) \tag{3}$$

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

1.3 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 was formalized in section 1.2.

1.4 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 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 [PDF]. 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
Neurotechnology	neurotechnology	001	2021.03.22
Paravision (EverAI)	paravision	001	2019.12.23
Rank One Computing	rankone	000	2019.06.03
Rank One Computing	rankone	001	2019.11.12
Rank One Computing	rankone	002	2020.06.26
Tevian	tevian	000	2021.09.23
Universidad Autonoma de Madrid - EC Joint Research Centre [8]	uam-jrc-faceqnet	000	2019.08.19
Xiamen University	xm	000	2020.11.04

Table 1: FRVT Quality Assessment Participants

3 Image Datasets

3.1 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".

3.2 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 6 and Figure 4 in NIST Interagency Report 8271.

3.3 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 6: Samples of images used in this report. The subject in the photos is a NIST employee.

4 Evaluation and metrics

We conduct two recognition tests, one with Application-Webcam comparisons, and another with Wild-Wild comparsions. Here we formalize measures of how well quality scores predict the comparison scores.

Consider application of a face verification algorithm to N genuine image pairs, x_{i1}, x_{i2}. to produce N genuine scores, s_i. We adopt these as a target for assessment of image quality assessment algorithms - because we have posed the quality problem as a predictor of genuine similarity measures. A quality algorithm, F, converts images to quality scalars:

$$q_{i1} = F(x_{i1})$$
$$q_{i2} = F(x_{i2})$$

For the Application-Webcam photos we form a vector of quality values

$$q_i = q_{i2} \tag{4}$$

by taking simply the quality of the probe image alone. We do this because that dataset compares almost pristine frontal reference images (see section 3.1), with markedly lower and variable quality probes (see section 3.2).

For the wild image dataset, both images are of widely varying quality (see section 3.3). We therefore evaluate a quality algorithm on the relationship between the score and the minimum of the two quality scores

$$q_i = \min(q_{i1}, q_{i2}) \tag{5}$$

on the assumption that a low comparison score will be caused by the image with the lower image quality[1].

From the above we have two vector q_i and s_i, and we now address how well the former predict the latter. We could report correlation measures (Spearman ρ or Kendall τ) for example but these don't acknowledge that we're usually only interested in prediction of low genuine scores, not all scores. Instead we produce two metrics showing the effect of rejecting images with low quality values.

4.1 Error vs. reject curve

Given N genuine scores and N quality values from equation 4 or 5, we construct the error vs. reject curve as follows. We set a recognition threshold, T, for example by referencing a table of false match rates, FMR(T), for some value say FMR = 0.00001. This partitions the scores into true accepts, $s_i \geq T$, and false rejects $s_i < T$. We then ask how does FNMR change by excluding a fraction, r, of low quality images from the computation. Using the step function, $H(x)$, the quantity

$$\text{FNMR}(r) = \frac{\sum_{i=1}^{N} H(q_i - Q)(1 - H(s_i - T))}{\sum_{i=1}^{N} H(q_i - Q)} \tag{6}$$

has a numerator that counts recognition false negatives (below threshold T) from images that have good quality (at or above Q), and a denominator that is the count of the images with good quality i.e. that are not discarded. The quality threshold is obtained from the inverse of the empirical cumulative distribution function of the N quality values $Q = F^{-1}(r)$.

We additionally normalize that equation by dividing by r to produce an efficiency, η, that should ideally be 1.

$$\eta(r) = \frac{1}{r}\left(\frac{\text{FNMR}(0) - \text{FNMR}(r)}{\text{FNMR}(0)}\right) \tag{7}$$

In the calculation of FNMR(r) and η, we perturb the quality values by adding random uniformly distributed noise on

[1] This is likely the case with a variable like blur or contrast, but may not be the case with variables like expression or pose where a similarity score may be high if both images have the same pose or expression. This is discussed further in the FRVT Quality Concept Document.

the interval $[-0.2, 0.2]$. This breaks ties without reordering results.

4.2 Sample acceptance error tradeoff

In the formulation above we have quality values as predictors of genuine scores. We seek to use quality algorithms to make decisions about whether or not to accept a photo for further processing. As such they are subject to Type I/II error tradeoff analysis from decision theory. We need to be careful with language here because we already have recognition error rates (FNMR and FMR) and we need to define two error rates: First, an error rate expressing incorrect rejection of a photo i.e. assignment of low quality when the image would be matched by a face recognition engine correctly; and second an error rate expressing incorrect acceptance of a photo when it ultimately gives a false negative in recognition. Thus given ground-truth match / non-match decisions from a recognition engine against some score threshold, T, we define Incorrect Sample Rejection Rate

$$\text{ISRR}(Q) = \frac{1}{N}\sum_{i}^{N}(1 - H(q_i - Q))H(s_i - T) \tag{8}$$

i.e. the proportion of samples with quality below quality threshold, Q, and genuine score at or above recognition threshold. We also define Incorrect Sample Acceptance Rate

$$\text{ISAR}(Q) = \frac{1}{N}\sum_{i}^{N}H(q_i - Q)(1 - H(s_i - T)) \tag{9}$$

i.e. the proportion of samples with quality above threshold, Q, but genuine score below recognition threshold T. This metric directly supports use-case (1), sample acceptance. Note that ISAR is similar but not the same as the error vs. reject quantity of equation (6).

In the calculation of ISAR and ISRR, we perturb the quality values by adding random uniformly distributed noise on the interval $[-0.2, 0.2]$. This breaks ties without reordering results.

The two error rates can be plotted against each other as an error tradeoff characteristic, or against threshold, to allow threshold setting.

4.2.1 Handling failure to process

When a recognition algorithm fails to execute a comparison - for example because one the algorithm failed to produce a template from an image - we assign a synthetic score value equal to the lowest observed genuine score.

When a quality algorithm fails to produce a quality value from an image, we assign a default value of zero.

5 Results

The results in this section assess the quality assessment algorithms on application, webcam, and wild imagery. Figure 9 plots algorithm quality score distribution for each image type. Note that some algorithms (paravision-001, intsysMSU-000) concentrate quality values in narrow ranges. This doesn't impede evaluation, but is contrary to the idea of a standardized range $[0, 100]$. Secondly note the quantization of quality values reported by the pixelall-000 algorithm for application and webcam images.

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 we also selected one or more verification algorithms from the same developer to use in our analysis. To assess cross-developer interoperability we added other verification algorithms from developers not participating in

the FRVT Quality track.

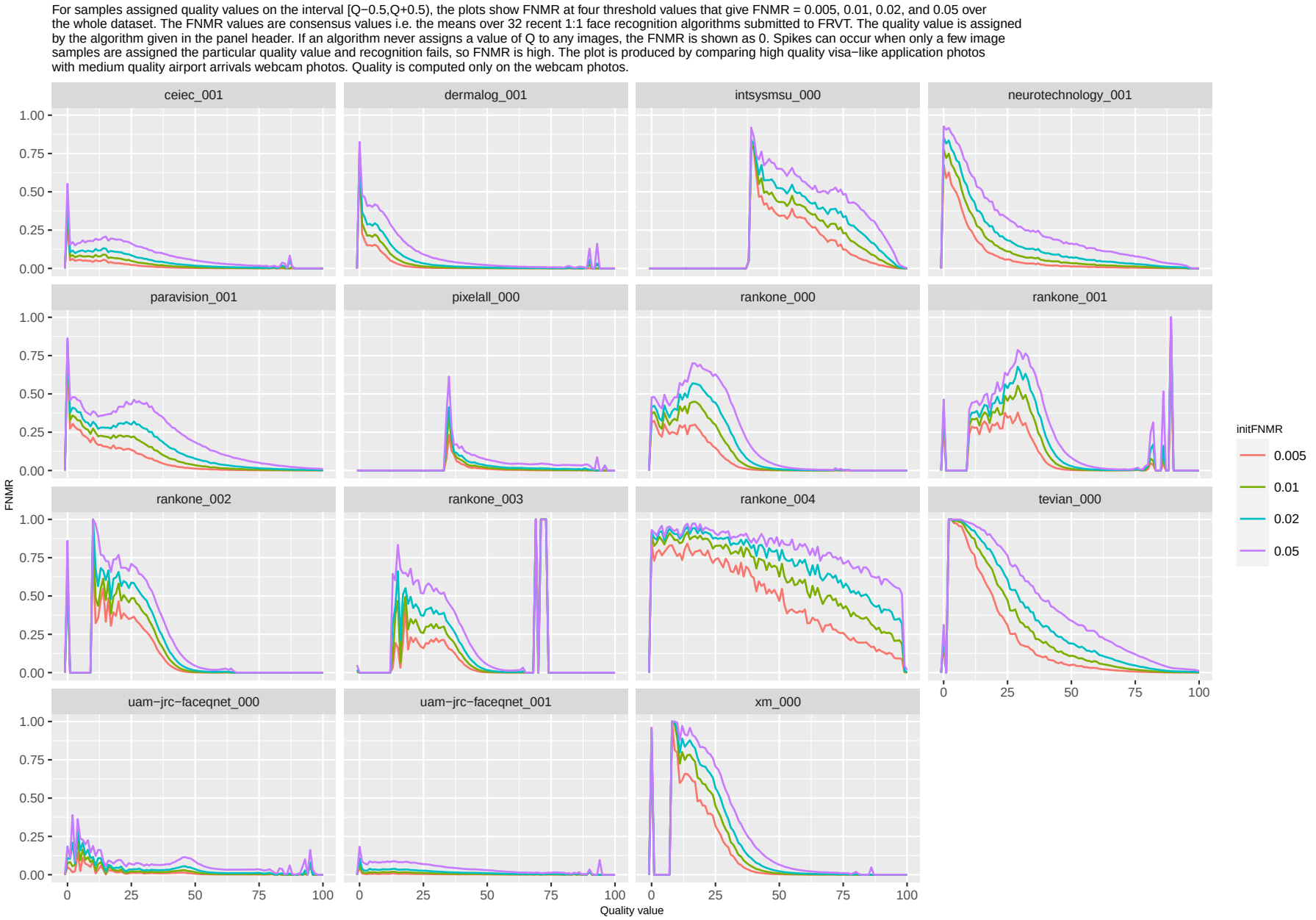


Figure 7: The plot shows the measured false non-match rate (for 22 contemporary FRVT verification algorithms) for each level of quality $-1 \ldots 100$. *A quality value of -1 indicates the algorithm did not return a value. The four traces correspond to for four different recognition thresholds. Each panel corresponds to one quality algorithm. assessment algorithm. The images are border-crossing images. The counts are shown on the next page.*

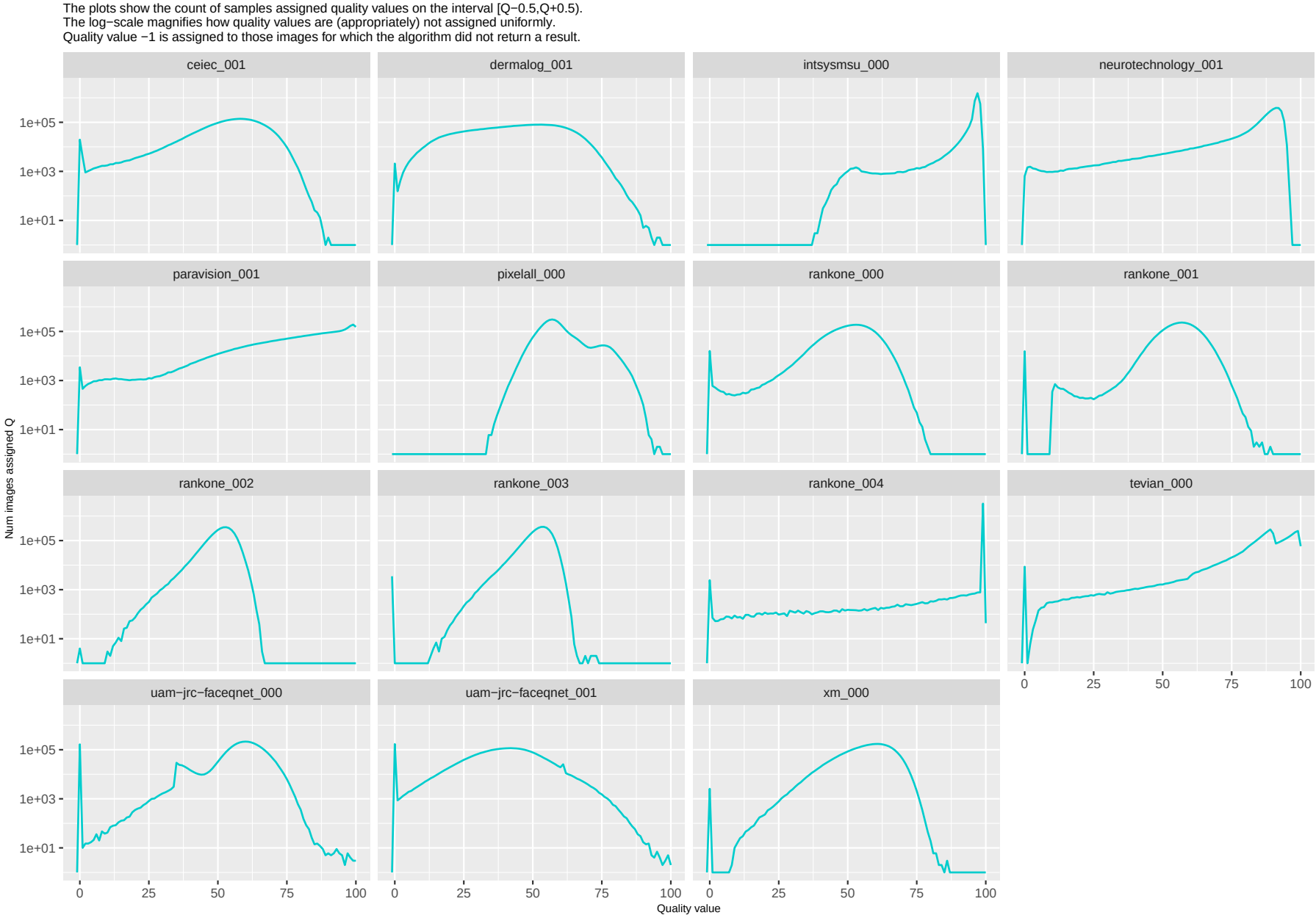


Figure 8: The plot shows the numbers of samples assigned each level of quality $-1 \ldots 100$. *A quality value of -1 indicates the algorithm did not return a value. Each panel corresponds to one quality algorithm. assessment algorithm. The images are border-crossing images.*

Application vs. Webcam Images: The results of Section 5.1 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 5.2 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.

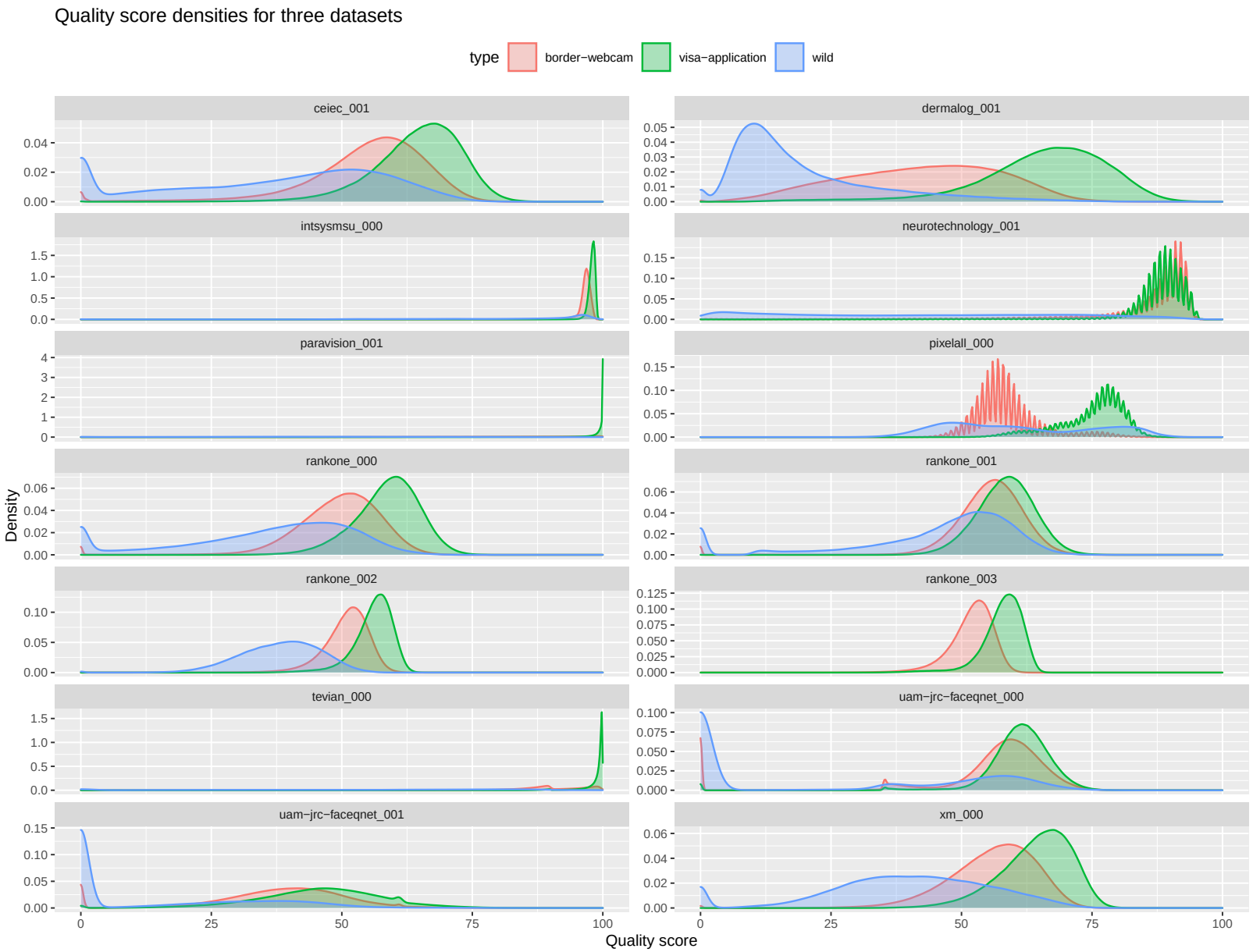


Figure 9: This density plot shows the quality score distribution for each quality assessment algorithm, on all the types of imagery evaluated. Visually the visa images are of the best quality, followed by the border crossing images and then the wild images.

5.1 Dataset 1: Application versus Webcam Images

Figure 10 shows the error vs. reject performance described in section 4.1. The notable results are:

- ▷ The quality algorithms that reject the lowest-scoring samples most efficiently are from RankOne and Paravision, when predicting low scores from their respective verification algorithms.
- ▷ The performance is considerably worse when algorithms are used to predict low-scores from other developers' algorithms. This implies quality algorithm interoperability is difficult.
- ▷ Figure 11 shows the "normalized" version of error vs. rejection i.e. efficiency in equation 7 for four developer who submitted both quality assessment and recognition algorithms. The impressive initial efficiency of the RankOne algorithms is due to the correct rejection of images that were not enrolled by the feature extraction function of the recognition algorithm.

Figures 12 and 13 goes further in plotting the incorrect sample rejection and acceptance rates. ISRR is a measure of inconvenience caused by rejecting matchable samples, and ISAR quantifies the benefit to reducing matching error rates (FNMR) by excluding low quality samples. The best result is for the Paravision algorithm predicting Paravision scores: at a Q threshold of 38, ISRR is 0.01, the ISAR value is 0.0009 vs. the 0.0076 at Q = 0.

Figure 14 simply shows genuine score distributions for values of quality quantized into bins of width 8. In the ideal case the variance within a bin would be low, consistent with quality predicting matching score. It is conventionally assumed that in cases where notches do not overlap in adjacent boxes the distributions are significantly separate. Further quantization can better induce this separation of the score distributions but gives coarser-grained quality bins.

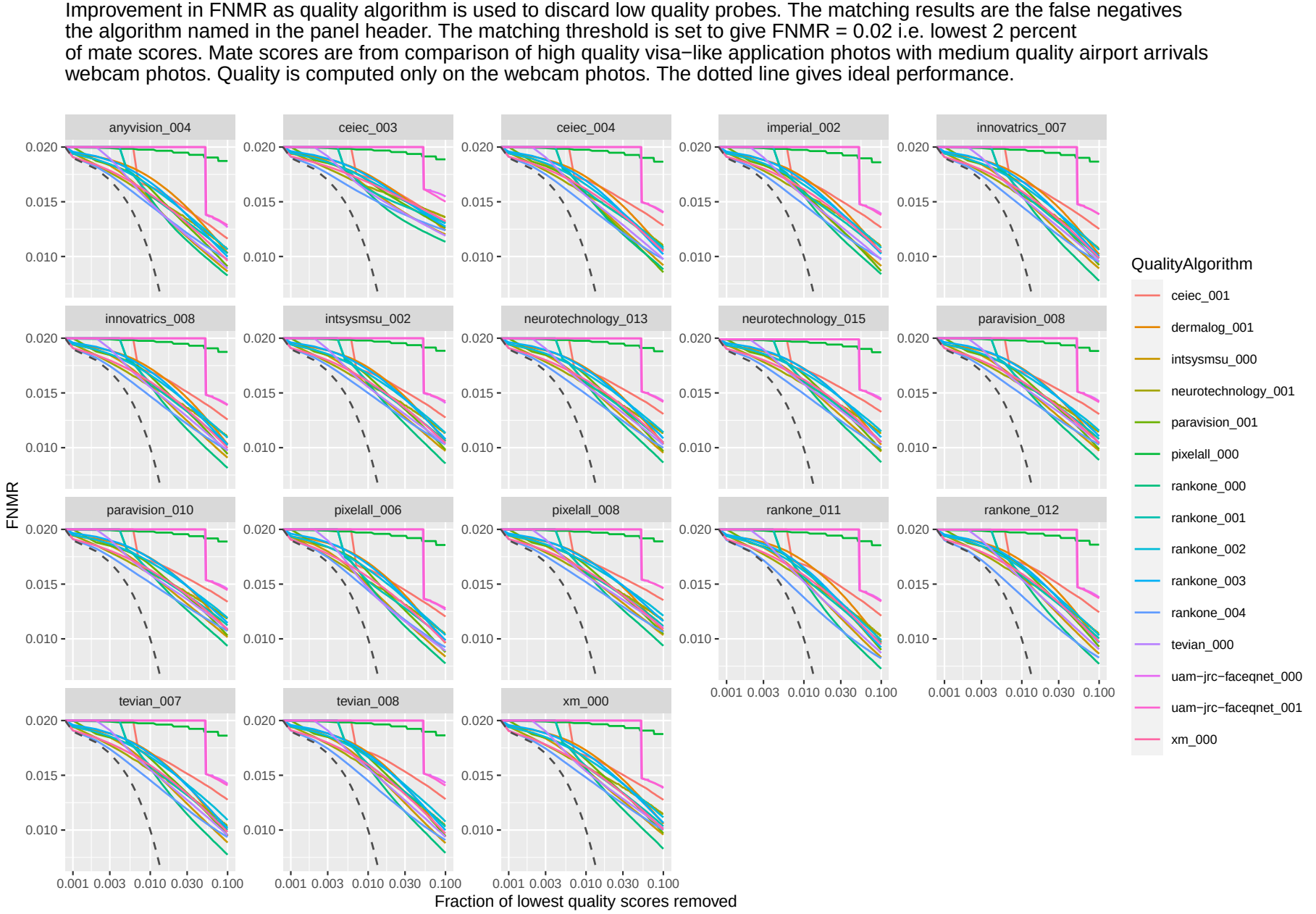


Figure 10: This plot shows FNMR vs. Reject, showing how FNMR reduces when the worst quality data is discarded. Each panel corresponds to a recognition algorithm, and the lines within each panel correspond to quality assessment algorithms. A perfect quality algorithm would predict which images are implicated in false non-matches. The quality score is from the verification (webcam) image. The red line labeled "PERFECT" is generated using $\max((FNMR - x), 0)$*. It is curved because of the log x-axis. The closer the quality algorithm line is to the "PERFECT" line, the better the quality prediction performance is relative to recognition outcomes.*

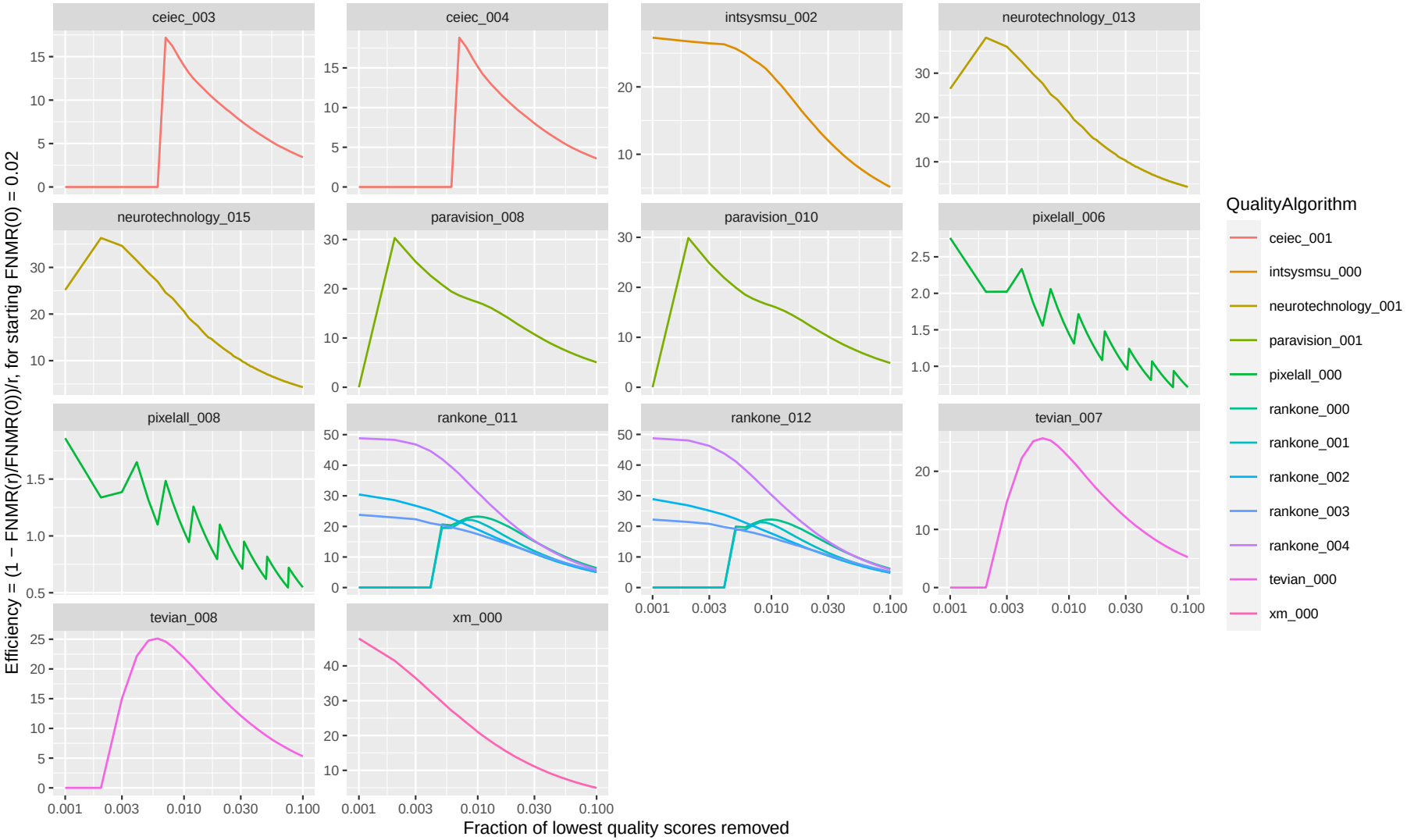


Figure 11: This plot is a simple modification of the prior figure. It shows efficiency vs. rejection, where the y-axis shows FNMR divided by the perfect FNMR i.e. how efficient the quality reject mechanism is. A value of 1 indicates perfect rejection of the low quality images. Each panel corresponds to a recognition algorithm, and the lines within each panel correspond to quality assessment algorithms. The quality score is from the verification (webcam) image.

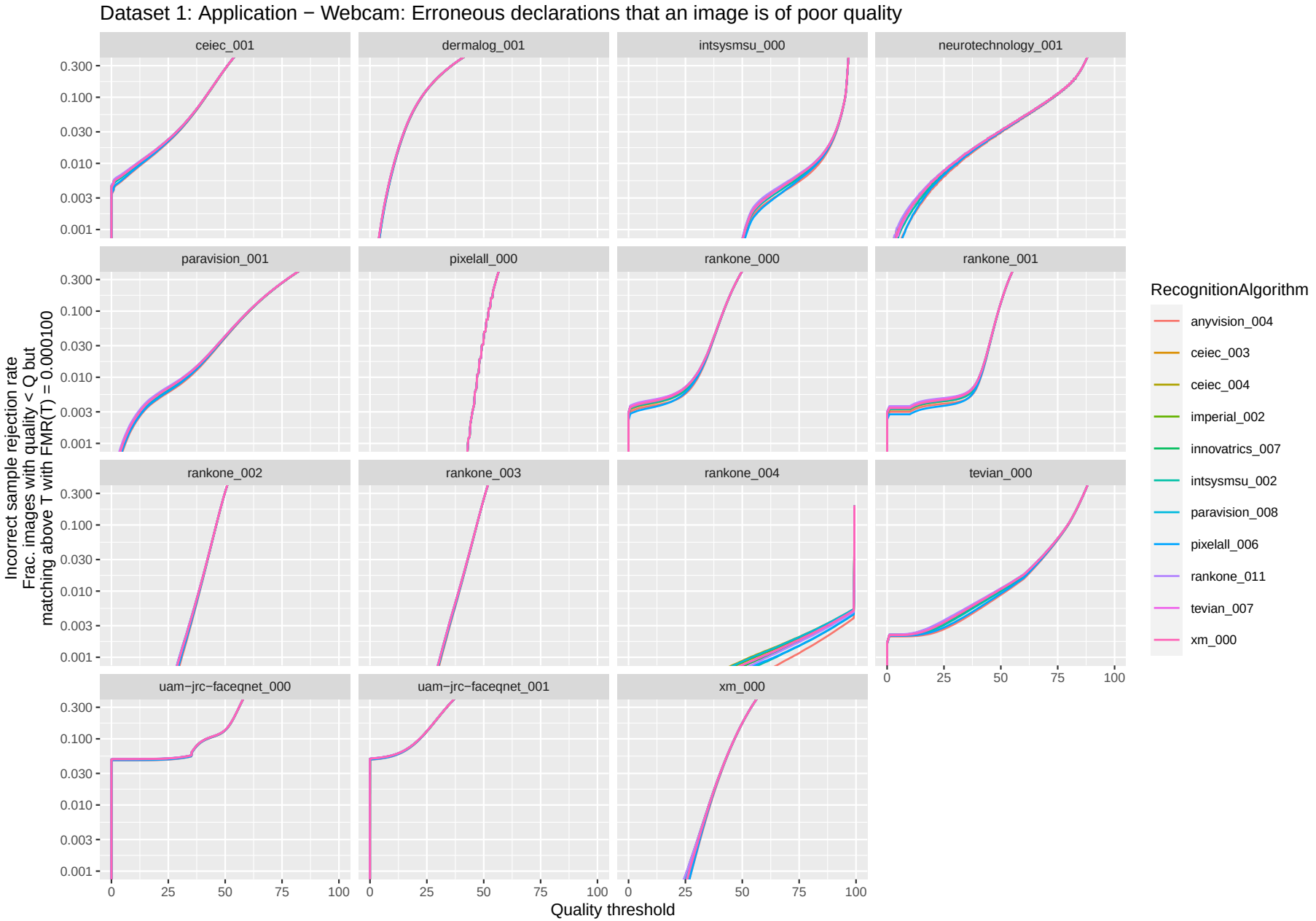


Figure 12: This plot shows the incorrect sample rejection rate, ISRR, as a function of quality threshold, Q.

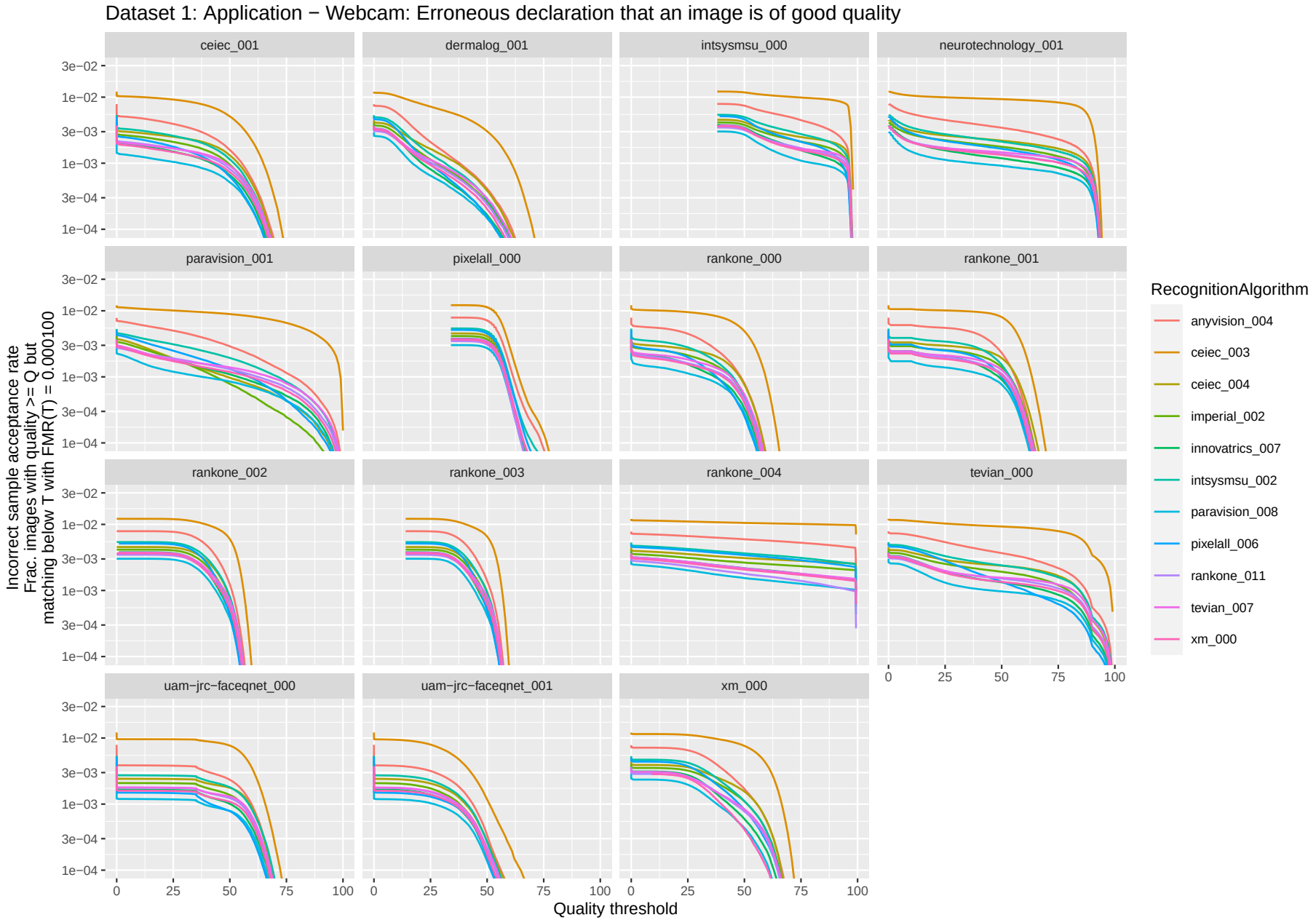


Figure 13: This plot shows the incorrect sample acceptance rate, ISAR, as a function of quality threshold, Q. The ISAR is higher for a less accurate recognition engine.

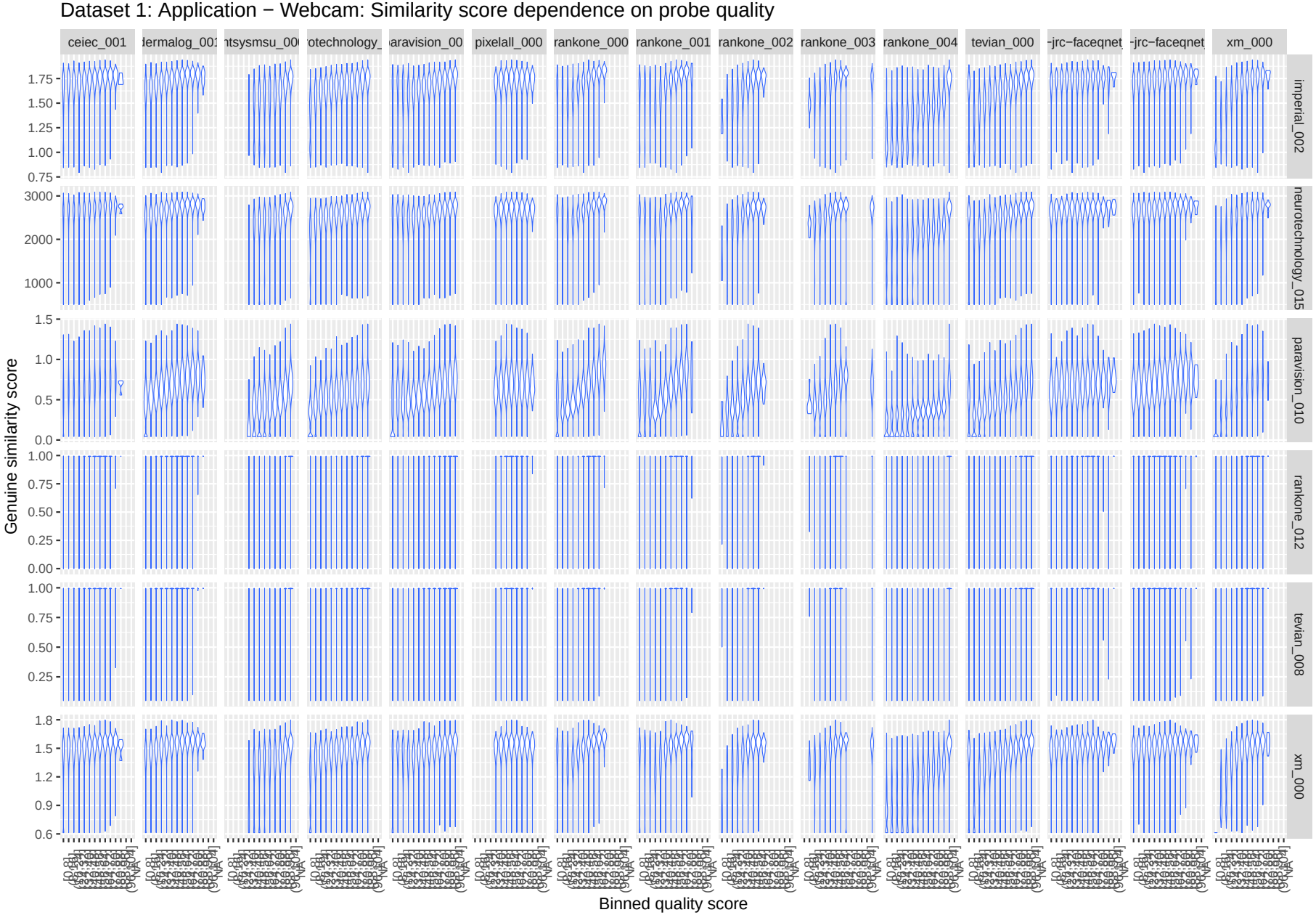


Figure 14: This plot shows boxplots of match scores of various 1:1 face verification algorithms plotted against quality scores. The quality score is from the verification (webcam) image.

5.2 Dataset 2: Wild Images

Wild images are, by definition, collected with the constraints implied by a photographic standard - sxamples are shown in Figure 4. The consequence of this is high recognition error rates (compared to the cooperative Dataset 1 images), and greater variation in quality.

We include three figures

- Figure 15 shows error vs. reject performance.
- Figure 16 shows the normalized version of that, $\eta(r)$.
- Figure 24 shows score distributions vs. binned quality.

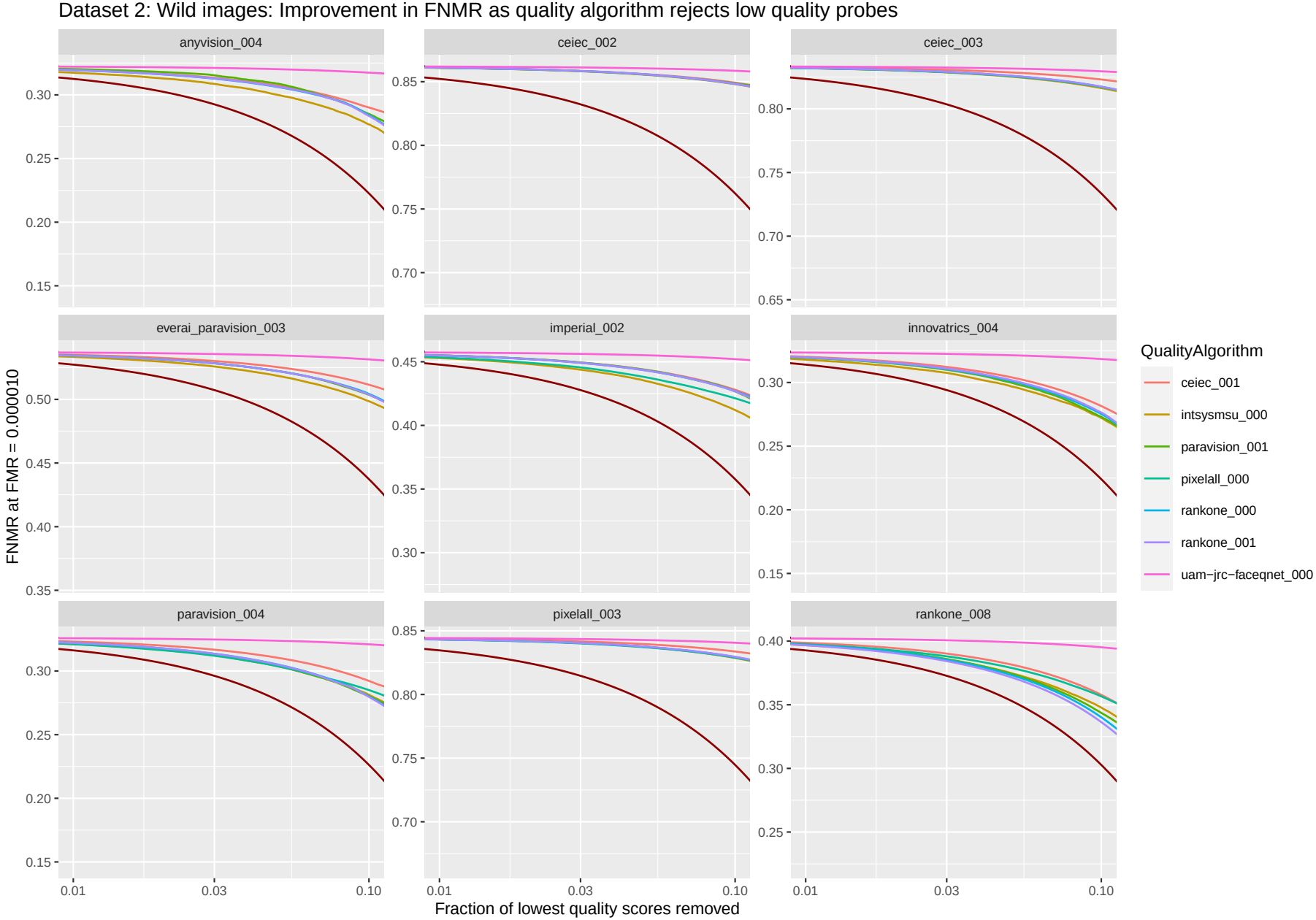


Figure 15: FNMR vs. Reject, showing how FNMR reduces when the worst quality data is thrown away. Each panel corresponds to a recognition algorithm, and the lines within each to a quality assessment algorithm. A perfect quality algorithm would predict which images are implicated in false non-matches. The quality score was generated using the minimum score between the enrollment (wild) and verification (wild) image. The lower dark-red line corresponds to perfect rejection i.e. max$((FNMR - x), 0)$ such that a line close to that gives better predictions of recognition outcome. The log-scale means that line is curved; it also prevents showing the zero-rejection FNMR, which varies by recognition algorithm.

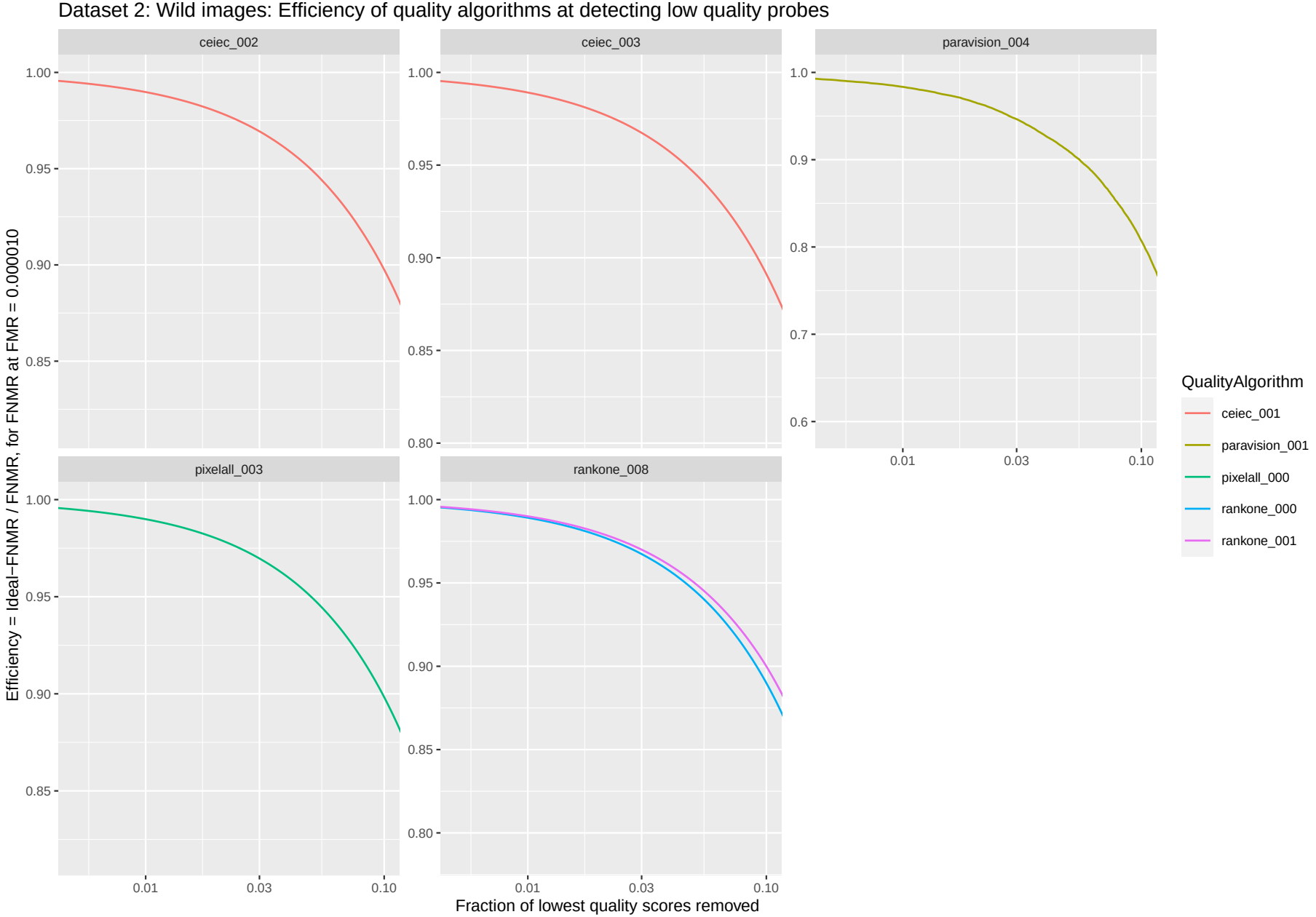


Figure 16: This plot is a simple modification of the prior figure. It shows efficiency vs. rejection, where the y-axis shows FNMR divided by the perfect FNMR i.e. how efficient the quality reject mechanism is. A value of 1 indicates perfect rejection of the low quality images Each row corresponds to a recognition algorithm, and the lines within each panel correspond to quality assessment algorithms. The quality score is from the verification (webcam) image.

Figure 24 spans several pages, each applying to a single quality algorithm. Each page includes several panels, one per recognition algorithm, and the boxplots show the distribution of genuine similarity scores for binned quality values. The overall form is the intended increasing dependence of score on quality. The ideal behavior would be for the boxes to not overlap and for variance to be low. This depends on the number of boxes - here we set twenty - and well separated box notches would suggest significance in the result.

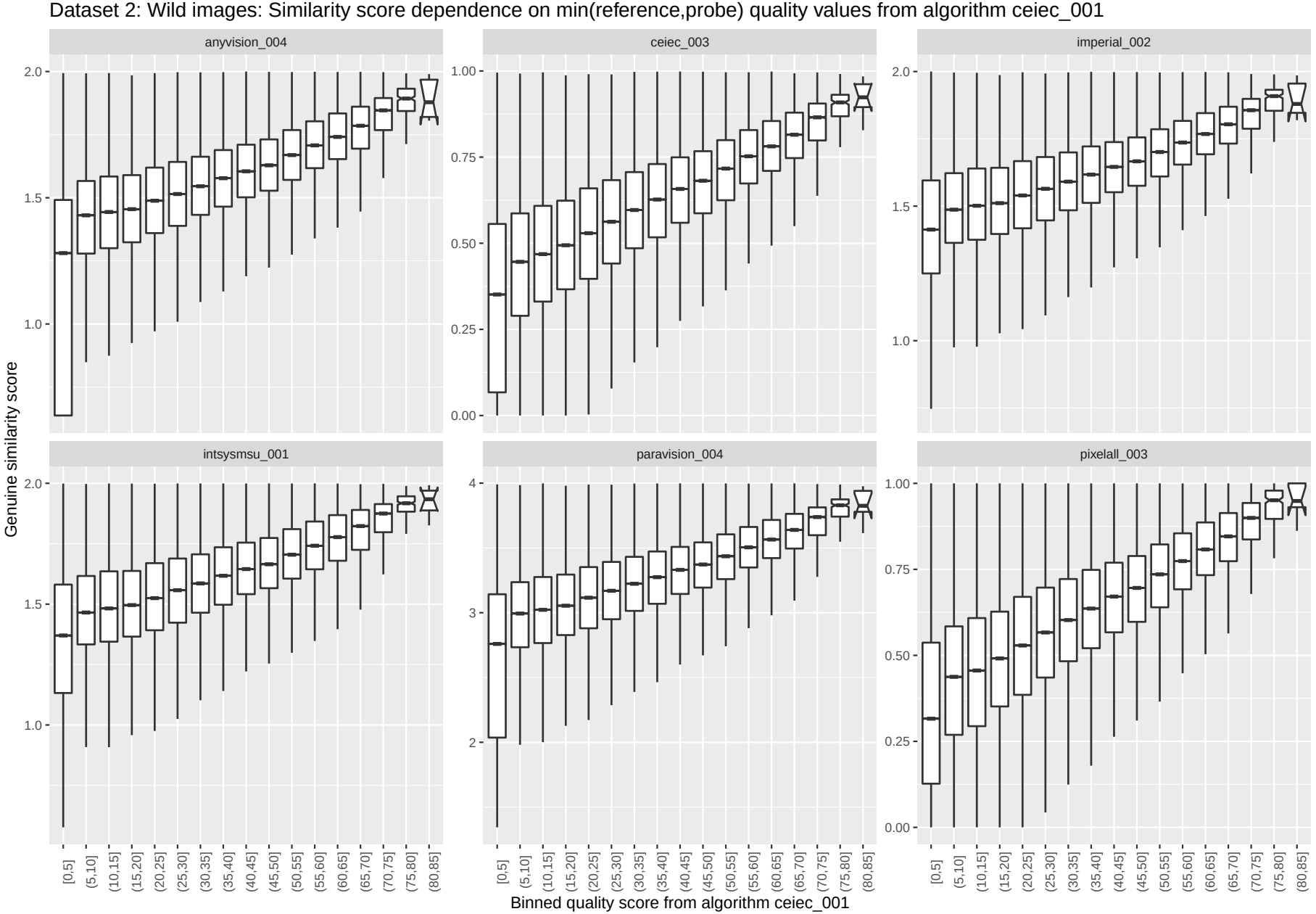


Figure 17: This plot shows boxplots of match scores of various 1:1 face verification algorithms plotted against quality scores. The quality score was generated using the minimum score between the enrollment (wild) and verification (wild) image.

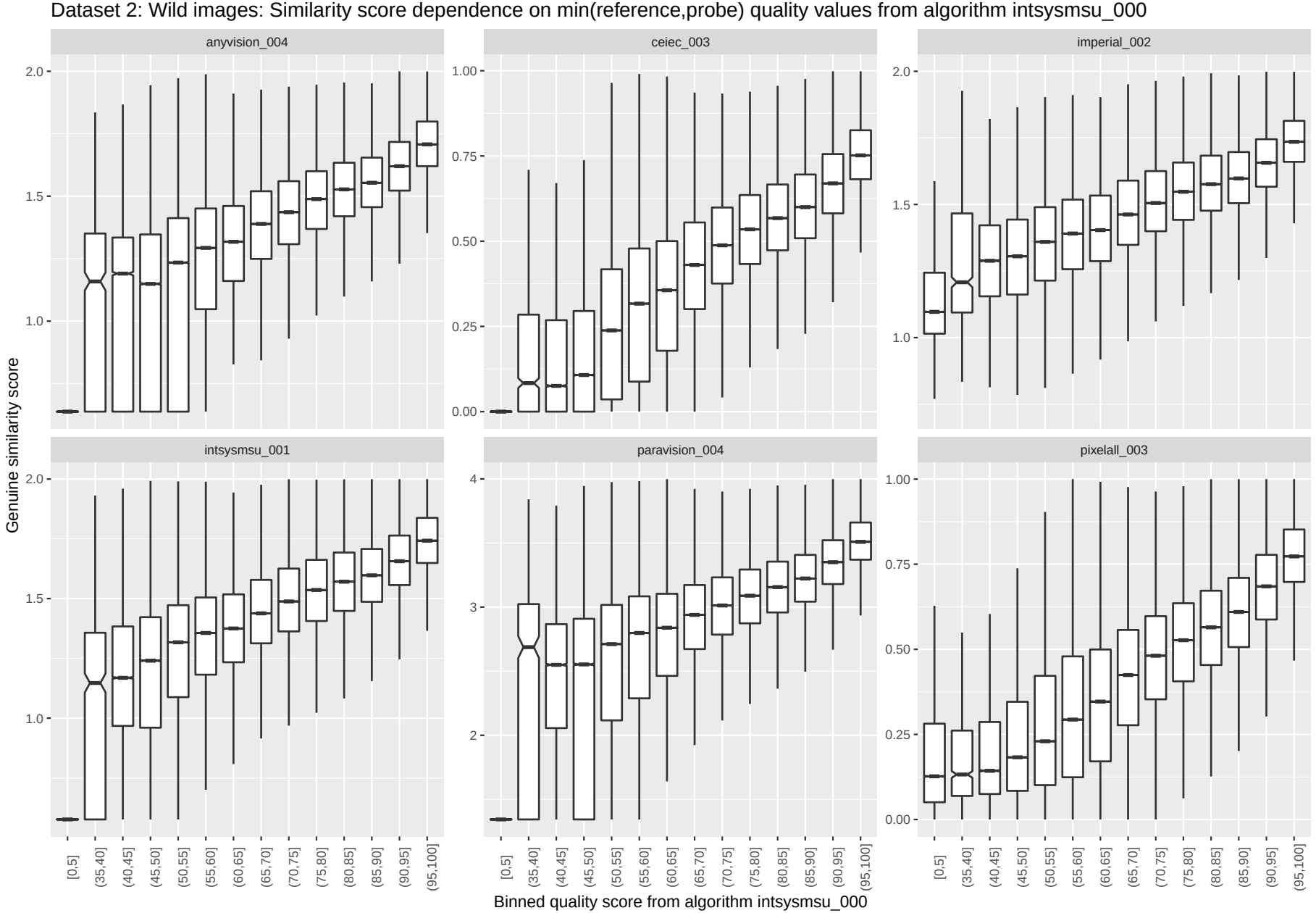


Figure 18: This plot shows boxplots of match scores of various 1:1 face verification algorithms plotted against quality scores. The quality score was generated using the minimum score between the enrollment (wild) and verification (wild) image.

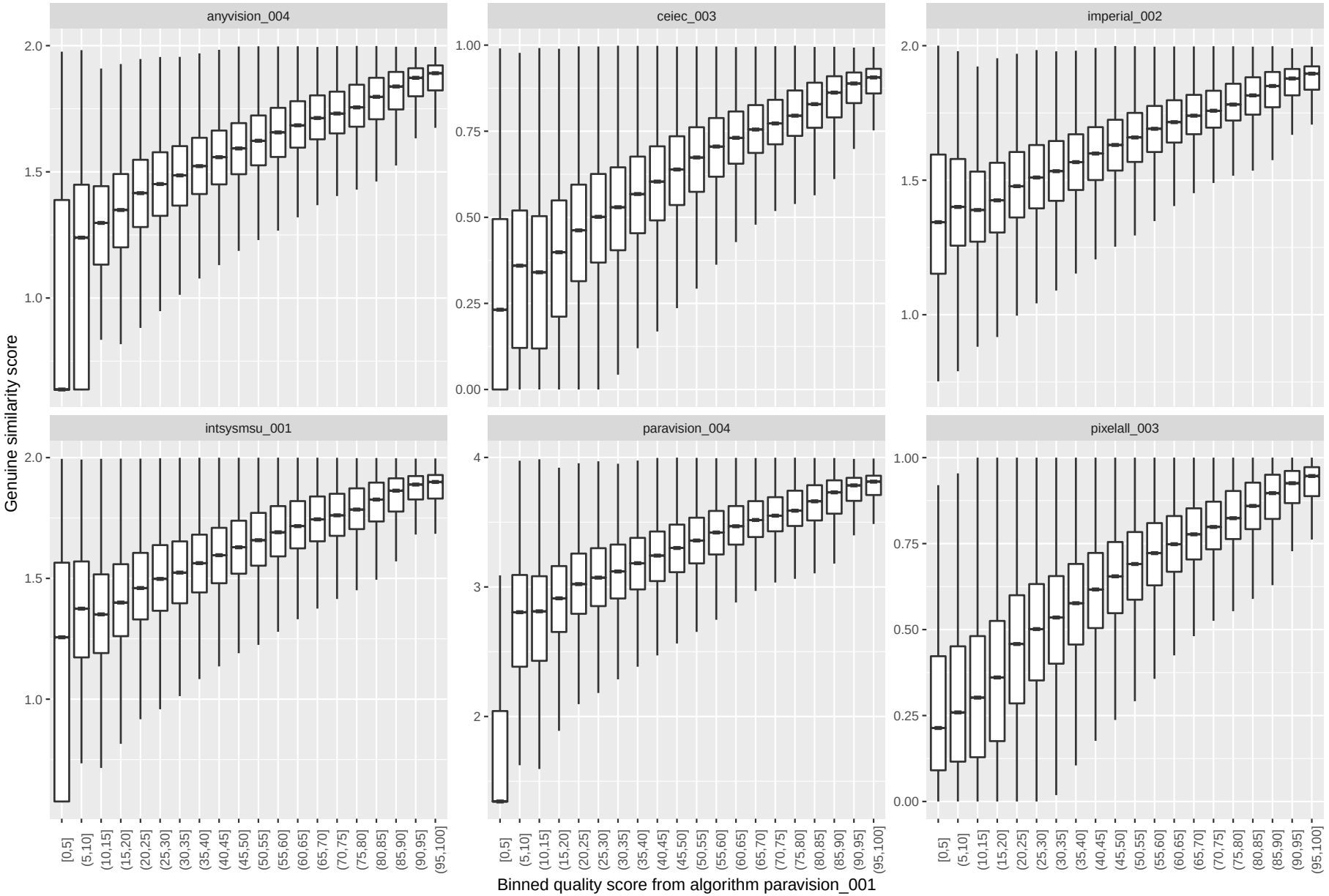


Figure 19: This plot shows boxplots of match scores of various 1:1 face verification algorithms plotted against quality scores. The quality score was generated using the minimum score between the enrollment (wild) and verification (wild) image.

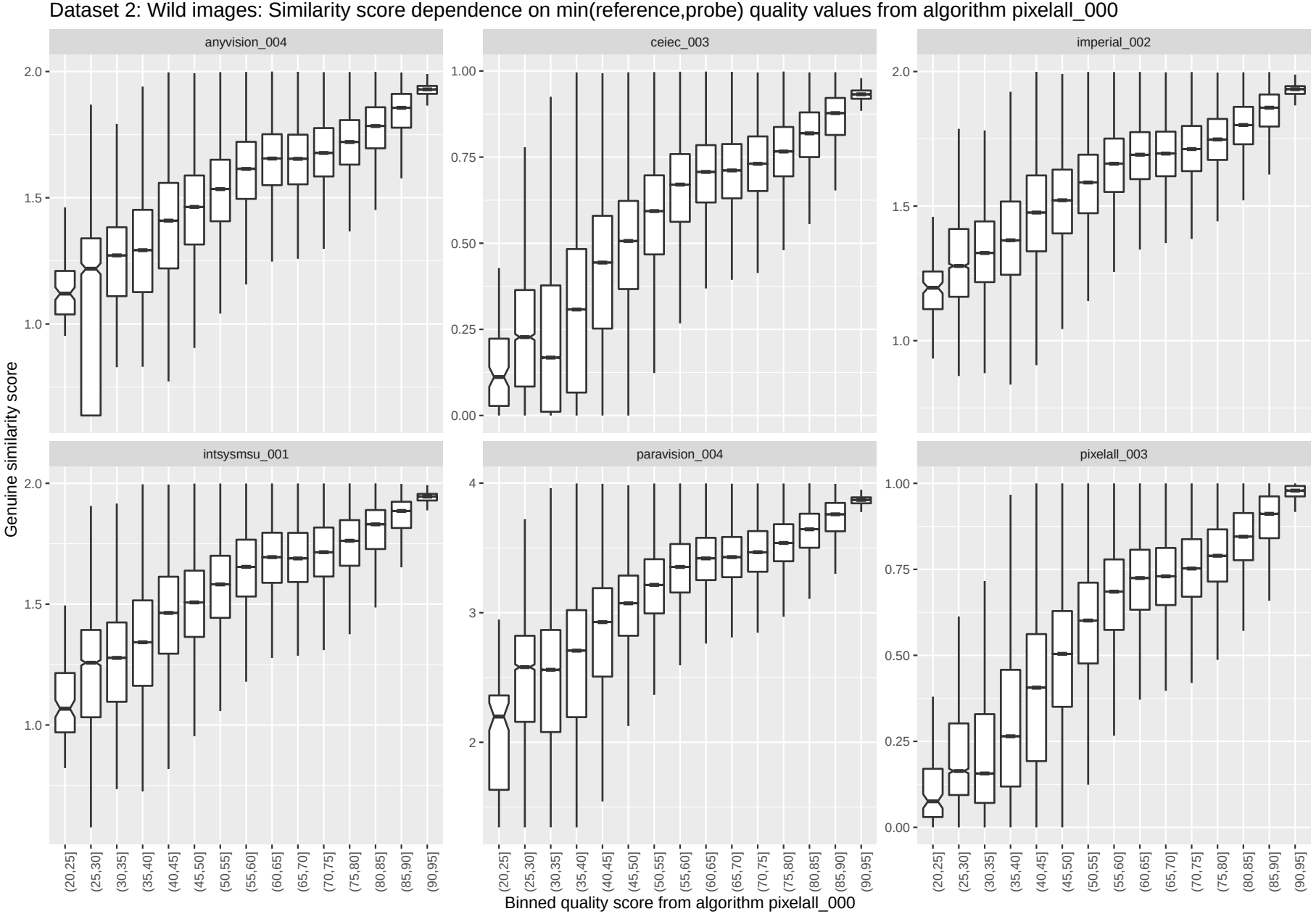


Figure 20: This plot shows boxplots of match scores of various 1:1 face verification algorithms plotted against quality scores. The quality score was generated using the minimum score between the enrollment (wild) and verification (wild) image.

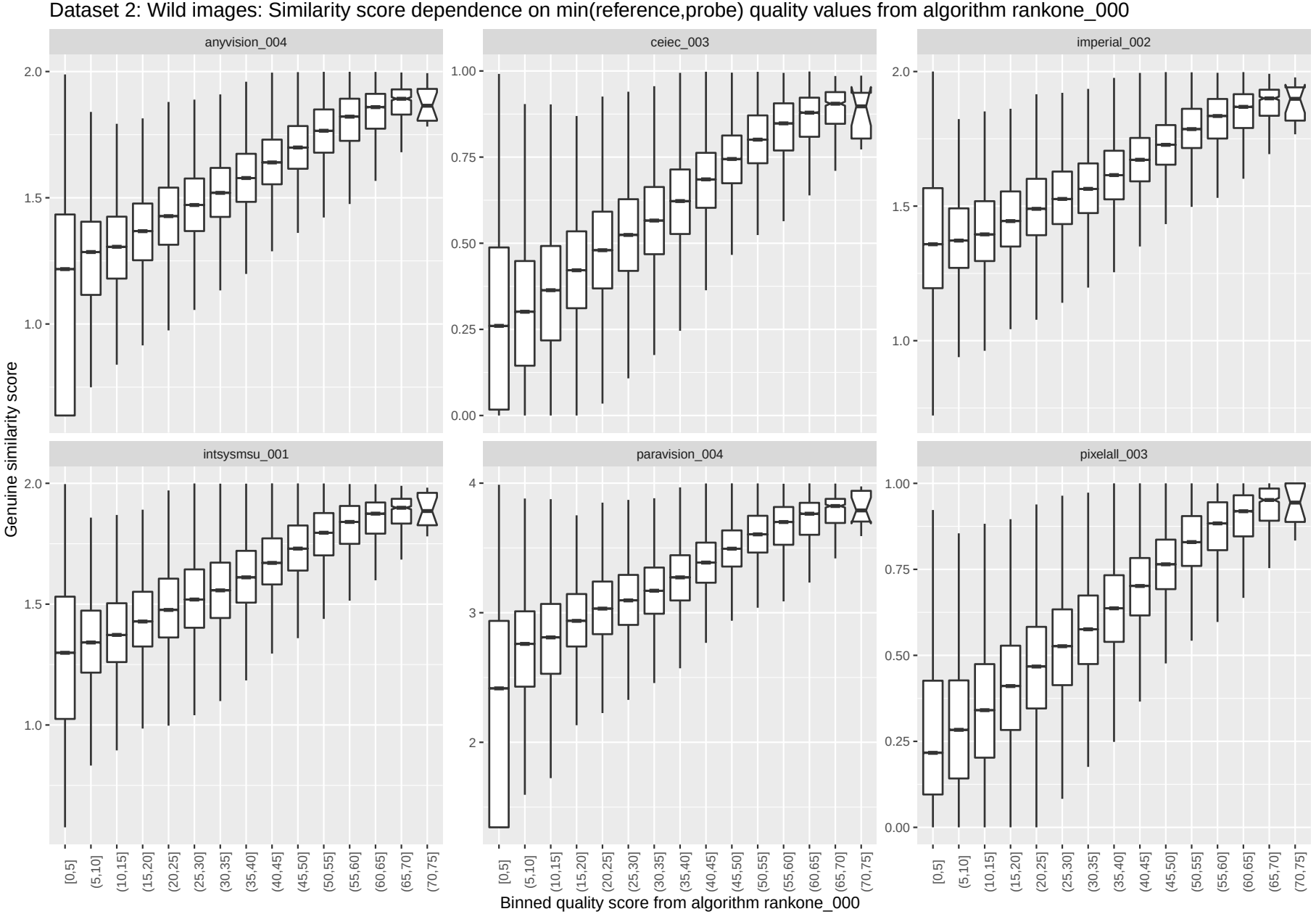


Figure 21: This plot shows boxplots of match scores of various 1:1 face verification algorithms plotted against quality scores. The quality score was generated using the minimum score between the enrollment (wild) and verification (wild) image.

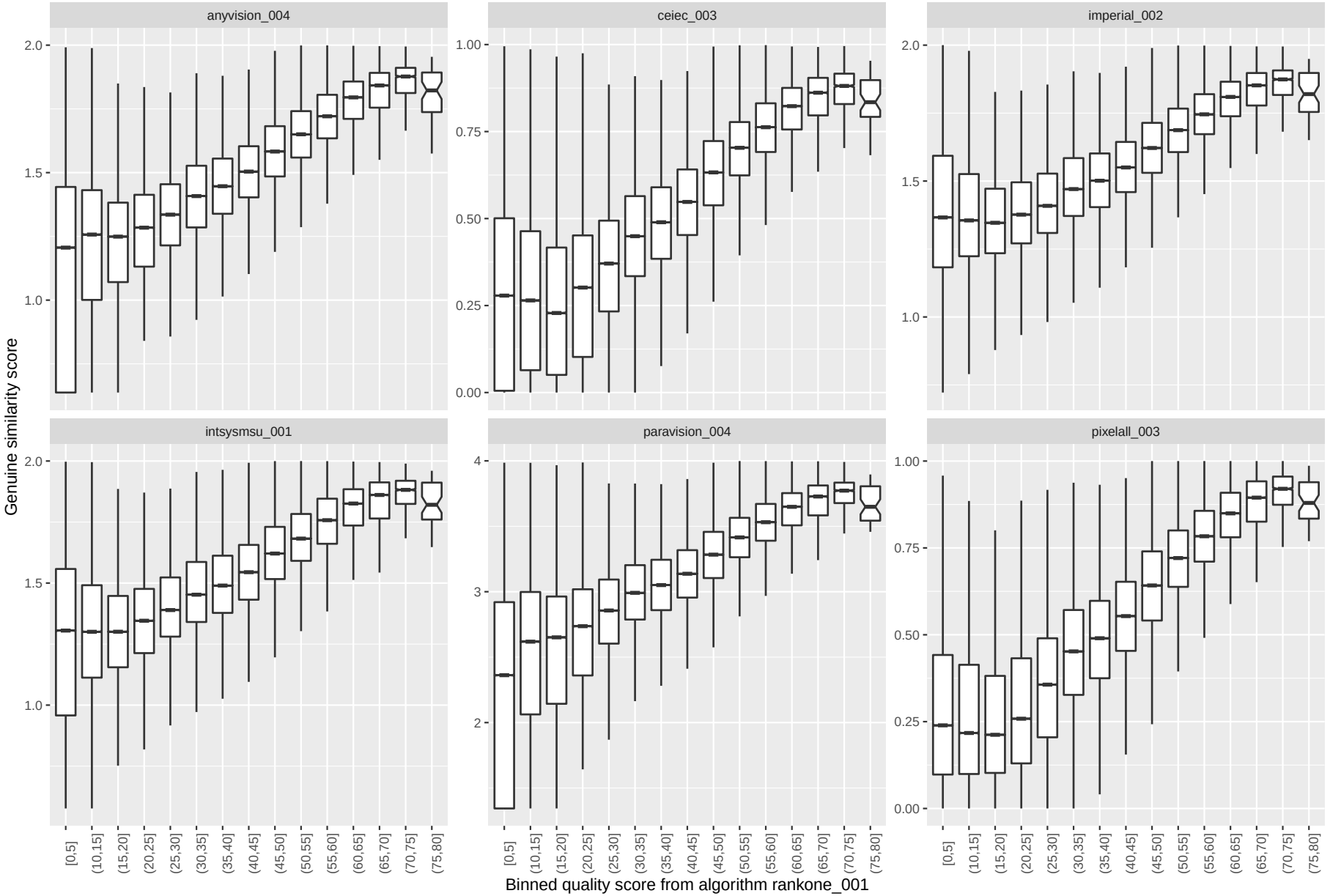


Figure 22: This plot shows boxplots of match scores of various 1:1 face verification algorithms plotted against quality scores. The quality score was generated using the minimum score between the enrollment (wild) and verification (wild) image.

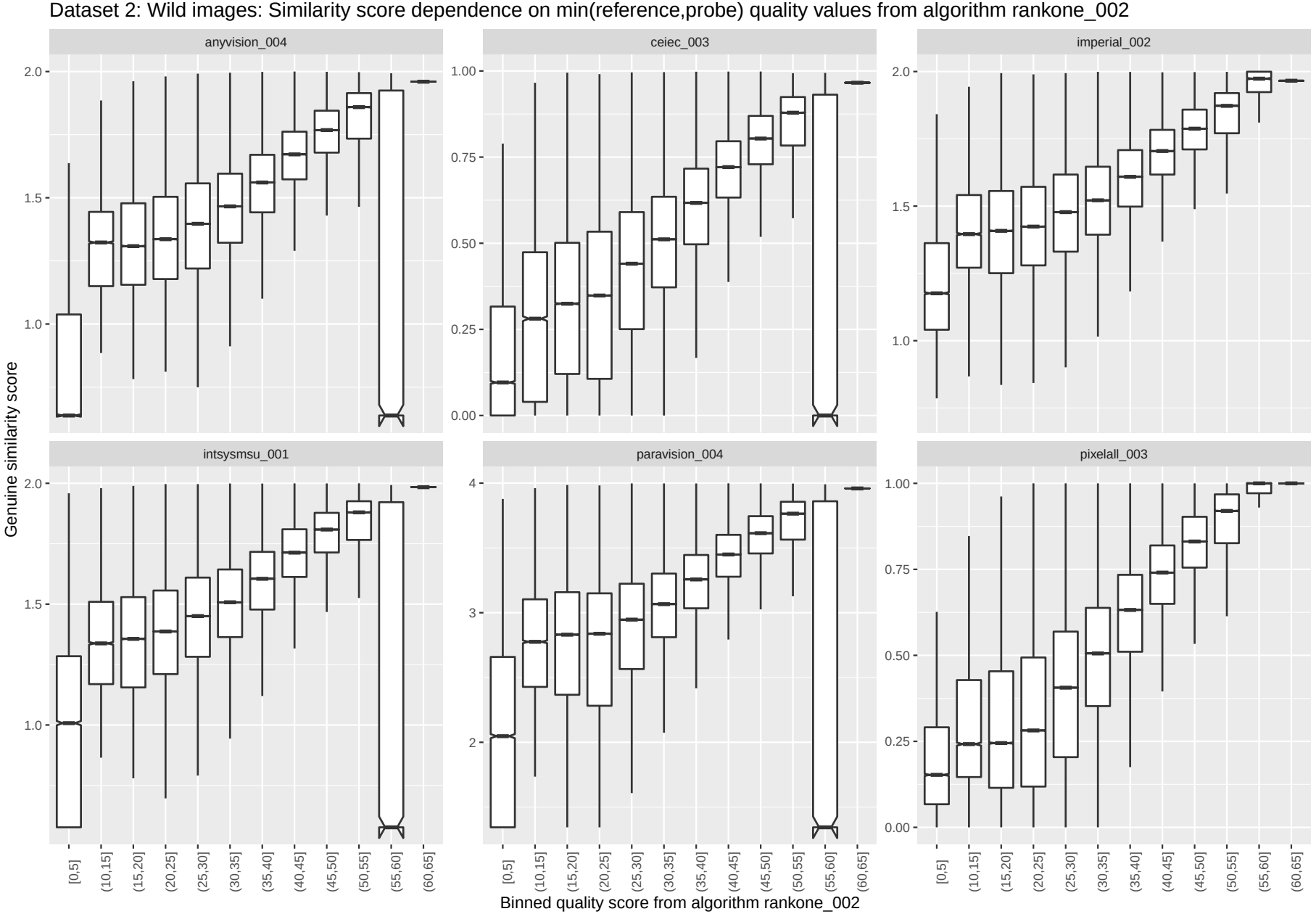


Figure 23: This plot shows boxplots of match scores of various 1:1 face verification algorithms plotted against quality scores. The quality score was generated using the minimum score between the enrollment (wild) and verification (wild) image.

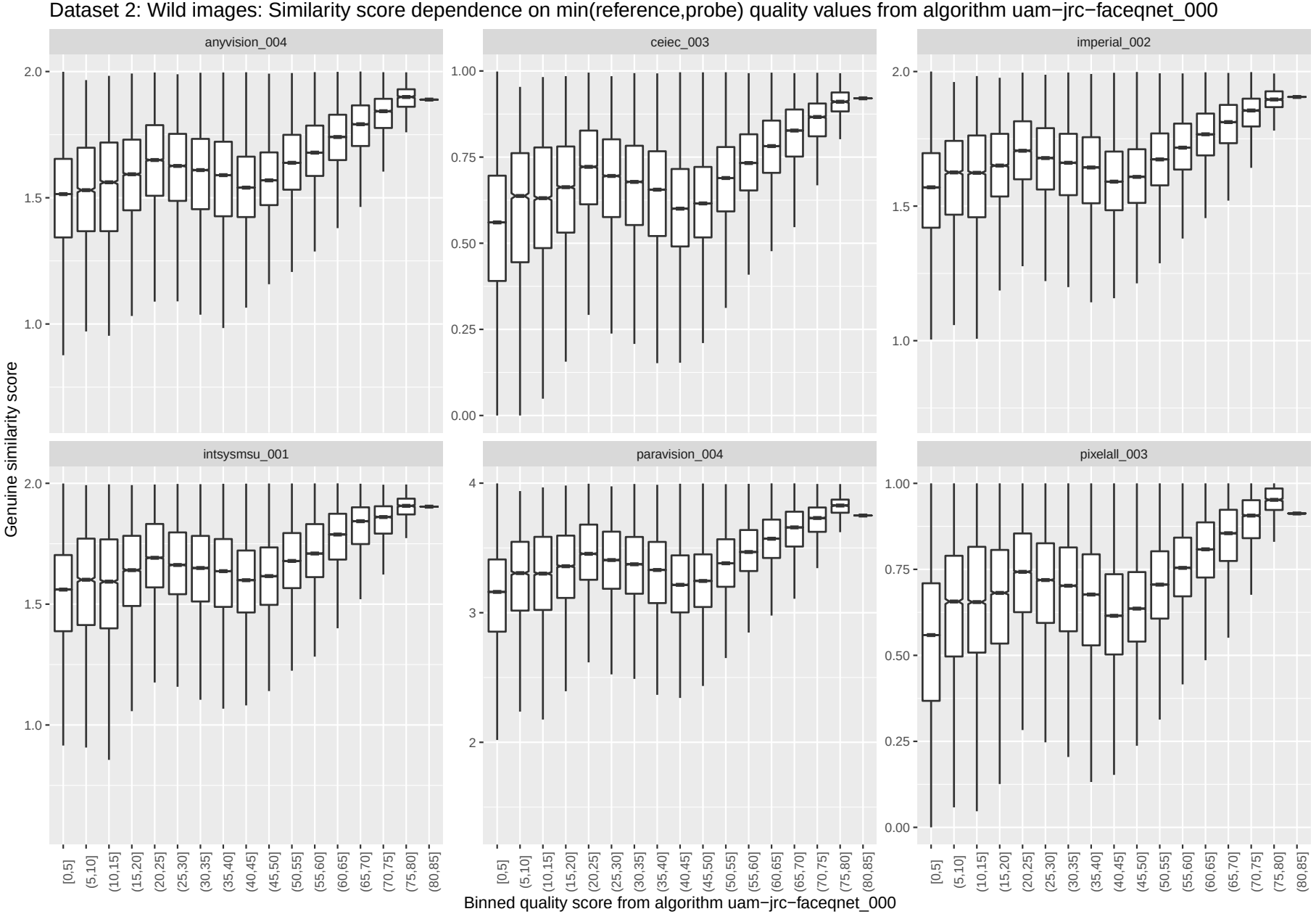


Figure 24: This plot shows boxplots of match scores of various 1:1 face verification algorithms plotted against quality scores. The quality score was generated using the minimum score between the enrollment (wild) and verification (wild) image.

5.3 Calibration

Future work: 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 [7] 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.

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] 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.

[8] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay. FaceQnet: Quality assessment for face recognition based on deep learning. In *Proc. International Conference on Biometrics (ICB)*, June 2019.

[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.