New performance measure of image and video quality assessment algorithms: Subjective Root-Mean-Square Error (SRMSE)

Mikko Nuutinen, Toni Virtanen, Jukka Häkkinen
Institute of Behavioural Sciences, University of Helsinki, Helsinki, Finland

Abstract. Evaluating algorithms used to assess image and video quality requires performance measures. Traditional performance measures (e.g., $PLCC$, $SROCC$ and $RMSE$) compare quality predictions of algorithms to subjective mean opinion scores ($MOS/DMOS$). In this study, we propose a new Subjective Root-Mean-Square Error ($SRMSE$) performance measure for evaluating the accuracy of algorithms used to assess image and video quality. The $SRMSE$ performance measure takes into account dispersion between observers. The other important new property of the $SRMSE$ performance measure is its measurement scale, which is calibrated to units of the number of average observers. The results of the $SRMSE$ performance measure indicate the extent to which the algorithm can replace the subjective experiment (as the number of observers). Furthermore, we have presented the concept of target values, which define the performance level of the ideal algorithm. We have calculated the target values for all sample sets of the CID2013, CVD2014 and LIVE MDIQ databases. The target values and MATLAB implementation of the $SRMSE$ performance measure are available from: `www.helsinki.fi/psychology/groups/visualcognition/`.

Keywords: Quality assessment algorithm, performance measure, subjective evaluation, number of observers.

Address all correspondence to: Mikko Nuutinen, University of Helsinki, Institute of Behavioural Sciences, Visual Cognition Research Group, Siltavuorenpenger 1 A, Helsinki, Finland, FI-00014; E-mail: mikko.s.nuutinen@gmail.com

1 Introduction

The top priority of the field of research related to image and video quality is the creation of a computational model (in the form of an algorithm) that can predict the subjective visual quality of natural images and videos. An established practice is to use publicly available databases (e.g., IVC,[1] LIVE,[2] TID2008,[3] CSIQ,[4] and TID2013[5]) when testing the performance of new algorithms for image or video quality assessment (IQA/VQA).[6,7] These databases include test images or videos that are distorted in different ways and annotated with subjective ratings.

Quantifying and validating the algorithms requires performance measures. The performance measures calculate the accuracy of the quality prediction that the algorithm provides, compared to the mean opinion score (MOS) or differential mean opinion score ($DMOS$) from the subjective evaluations. Root Mean Square Error ($RMSE$), Pearson's Linear Correlation Coefficient ($PLCC$) and Spearman's Rank-Order Correlation Coefficient ($SROCC$) are the three traditional performance measures. The $RMSE$ calculates how closely the predicted scores match to the subjective values. The $PLCC$ calculates the linearity between the predicted scores and the subjective values. The $SROCC$ measures the equality between the ranks based on the predicted scores and the subjective values.

In this study, we propose a new performance measure for the image and video quality assessment algorithms known as Subjective Root-Mean-Square Error ($SRMSE$). $SRMSE$ rectifies two drawbacks of the traditional performance measures. The first shortcoming of the traditional performance measures is that they do not take into account dispersion in the subjective data.[8–10] The traditional performance measures assume that the algorithm should predict the values of the $MOS/DMOS$ as accurately as possible regardless of the level of dispersion. Many factors affect

the degree of dispersion in the data, such as image content and distortion type.[11] Furthermore, observers' opinions can cluster into different groups while increasing the degree of dispersion.[12] The low number of observers[9] or an inaccurate evaluation method[13] also affect the degree of dispersion.

The second shortcoming of the traditional performance measures is the non-informative units of the measuring scales. The units of the $PLCC$, $SROCC$ and $RMSE$ performance measures have not been linked to perceived quality or quality differences. If the $PLCC$ or $SROCC$ value approaches 1, the performance of the algorithm is assumed to be high, but compared to what? For example, the study by[14] indicated that the $PLCC$ and $SROCC$ values of the state-of-the-art algorithms ranged from 0.7 to 0.85 for the TID2008[3] database and from 0.9 to 0.95 for the LIVE[2] database. The $RMSE$ values showed the same trend. With these values, drawing conclusions about the performance level of the algorithms based on human perception is clearly difficult.

The proposed $SRMSE$ is calibrated to the performance of the human observer. The units of the measuring scale are the number of average observers. For example, if the measured value of the $SRMSE$ is 2, the accuracy of the predicted $MOS/DMOS$ is the same as the average of the opinions of two randomly selected observers compared to the ground truth. In other words, the algorithm can replace a subjective evaluation experiment with two observers.

In addition to the $SRMSE$ performance measure, we introduce the concept of target values. A target value is the performance level of the ideal algorithm. It is the level of accuracy achieved with a sufficiently high number of observers (i.e., adding more observers will not change the accuracy of the $MOS/DMOS$). The test material and test method affect target values. If observers have different quality opinions, the differences in quality between the samples are small or the test method is inaccurate, the target values will be more relaxed and vice versa.

The contribution and novelty of this study are related to the $SRMSE$ performance measure and its quality scale, in which the units of the number of observers defines the performance of the algorithm. This new performance measure takes into account the dispersion between observers; if the dispersion is high, the accuracy of the algorithm is relaxed. The second contribution relates to the concept of the target value. We presented the idea of the target value in the conference article[15] and in this study, we have improved the calculation algorithm. The target values are calculated for the CID2013,[16] CVD2014[17] and LIVE Multiply Distorted Image Quality (MDIQ)[18] databases and raw data are available on the project page of this study.

This article is divided into eight sections. After this introduction, Section 2 reviews the performance measures presented in the literature. Section 3 derives the $SRMSE$ performance measure and the calculation of the target values. Sections 4 and 5 present the CID2013 database and the results of the benchmarking study of the IQA algorithms. Section 6 presents the target values for the CID2013, CVD2014 and LIVE MDIQ databases. Section 7 discusses the topics relating an acceptable performance level and offers some recommendations for using $SRMSE$. Section 8 concludes the study.

2 Performance measures of IQA and VQA algorithms

This section reviews the standard performance measures ($RMSE$, $PLCC$ and $SROCC$). Furthermore, we present more advanced measures and methods that can be found from the literature. The section is divided into parts that discuss uncertainty (Section 2.2) and acceptable performance (Section 2.3) according to the nature of the methods presented. The end of this section discusses the novelty and the need for the proposed SRMSE performance measure.

2.1 Standard performance measures

Traditionally, the performance of algorithms is measured by the $RMSE$, Eq. (1), $PLCC$, Eq. (2) and $SROCC$, Eq. (5) performance measures:

$$RMSE = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(X(j) - Y(j))^2} \tag{1}$$

$$PLCC = \frac{\sum_{j=1}^{m}(X(j) - \bar{X}) \times (Y(j) - \bar{Y})}{\sqrt{\sum(X(j) - \bar{X})^2} \times \sqrt{\sum(Y(j) - \bar{Y})^2}} \tag{2}$$

where $X(j)$ is the average subjective score, and $Y(j)$ is the score predicted by the algorithm for sample $j, (j = 1, ..., m)$. Average subjective score $X(j)$ is calculated for sample j with Eq. (3) or (4)

$$MOS(j) = \frac{1}{N}\sum_{n=1}^{N} S_{j,n} \tag{3}$$

$$DMOS(j) = \frac{1}{N}\sum_{n=1}^{N} S_{ref,n} - S_{j,n} \tag{4}$$

where N is the number of observers, $S_{j,n}$ is the quality evaluation score of observer $n, (n = 1, ..., N)$ and $S_{ref,n}$ is the quality evaluation score for (high-quality) reference sample. Reference sample eliminates scoring biases associated with image or video content. $SROCC$ is defined between ranked variables:

$$SROCC = 1 - \frac{6\sum d_j^2}{m(m^2 - 1)} \tag{5}$$

where m is the sample size, $X(j)$ and $Y(j)$ are converted to ranks $x(j)$ and $y(j)$, and $d_j = x_j - y_j$.

The $RMSE$ measures the accuracy of the algorithm to predict subjective score. The $PLCC$ measures the linearity of the predictions. The $SROCC$ measures the performance of the algorithm in terms of sorting the samples in the same order as subjective scores. The correlation measures ($PLCC$ and $SROCC$) have easily interpreted end-points of values (i.e., 0 is the worst, and 1 is the best). A value of 0 is best for the $RMSE$ scale, but the definition for low quality remains unclear.

2.2 Uncertainty

Subjective experiments always have a component of uncertainty. One main shortcoming of the traditional performance measures is that they have not taken uncertainty into account. However, our literature search[4,19–23] identified few methods that took uncertainty into account in at least some way. These methods consider uncertainty from two different viewpoints: uncertainty between observers or the uncertainty of the algorithms.

Standard ITU-T P.1401[22] recommends taking uncertainty between observers into account. It presents the 95% confidence intervals (ci_{95}) calculated with the following equation:

$$ci_{95} = t(0.05, N)\frac{\sigma}{\sqrt{N}} \tag{6}$$

where σ is the standard deviation, N is the number of observers, and $t()$ is the t-value for the given N. Confidence intervals can be calculated for many public databases because the standard

deviations of the $MOS/DMOS$ values are often available.[7] Standard ITU-T P.1401[22] indicates that the ci_{95} value can be used when calculating the outlier ratio, OR:

$$OR = TNO/m \tag{7}$$

where TNO is the total number of outliers, and m is the number of samples. Outliers are defined as points for which the prediction errors exceed the ci_{95} value ($|X(j) - Y(j)| > ci_{95}$).

In addition to the OR, the standard ITU-T P.1401[22,23] introduces the epsilon-insensitive rmse ($rmse^*$) performance measure. The $rmse^*$ performance measure compensates for the error value if the prediction error exceeds the ci_{95} value:

$$error(j) = \max(0, |X(j) - Y(j)| - ci_{95}(j)) \tag{8}$$

and the $rmse^*$ value is calculated with the following equation:

$$rmse^* = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}(error(j))^2} \tag{9}$$

That is, if the prediction error is lower than ci_{95}, the effect of the error is nullified. Furthermore, the study[4] proposed an outlier distance measure d_{out}:

$$d_{out} = \sum_{k \in K_{false}} \min(|Y_k - [X_k + 2\sigma_k]|, |Y_k - [X_k - 2\sigma_k]|) \tag{10}$$

where K_{false} is the set of all predicted ratings outside error bars of $\pm 2\sigma$. That is, the value of d_{out} is the distance from outlier to the closest error bar.

Some have also proposed divergent methods for describing the uncertainty of algorithms. For example, the study[19] presented the resolving power measure, which defines the minimum change in the output value of the algorithm related to a statistically significant change in the value of the subjective average data.

The studies[20,21] presented a concept for collecting specialized image sets for investigating the uncertainty of algorithms. The study[20] presented a strategy for identifying the vulnerabilities of algorithms by using systematically distorted test images. That is, when the image manipulations are known, researchers can investigate whether the modifications affect the predictions of the algorithms in a non-systematic way.

The study[21] presented a method for selecting test image sets with image pairs that are highly likely to identify misclassification errors in algorithms that assess image quality. First, the method uses available IQA algorithms to select a small image set A from the image set B in order to find difficult image pairs that lead to misclassifications. The subjective experiment is then conducted for image set A. Finally, the misclassification performance of the algorithms is calculated with the data of image set A.

2.3 Acceptable performance

The question of an acceptable performance level is important when measuring the performance of a new method or algorithm. It can also be a question of when the performance is acceptable for a specific application. The literature describes some alternative methods and definitions for these questions. Pinson et al.[24] discussed how the PSNR (Peak-Signal-to-Noise-Ratio) algorithm can serve

as a pragmatic minimum performance benchmark. They proposed that an FR (Full-Reference) algorithm must perform with greater accuracy than the PSNR. RR (Reduced-Reference) algorithms must be at least as accurate as the PSNR. Recently, the PSNR has also started to use the level of acceptable performance for new NR (No-Reference) algorithms.[25]

In addition, other established algorithms have been used as a level of acceptable performance. In particular, the SSIM[26] or its multi-scale version (MSSIM)[27] have achieved a de facto status when evaluating the performance of a new algorithm .[14,28–31] The VQM[32] and MOVIE[33] have often served as performance benchmarks for video algorithms.

In addition to de facto benchmarking algorithms, one practice is to compare the performance of the proposed algorithm to the state-of-the-art algorithm. For sufficiently high performance, the prediction accuracy should be higher or at least at the same level as the state-of-the-art algorithm. Alternatively, the proposed algorithm should introduce other assets, such as a lower computational requirement.

Nachlieli and Shaked[34] presented a rank agreement measure (RAM) which compares the accuracy of the algorithm and the subjective evaluations. The comparison is based on the Spearman's rank-order coefficient ($SROCC$) values (see, e.g., Eq. (5)). $SROCC$ values are calculated for algorithm and individual subjective evaluations. If the $SROCC$ of the algorithm is higher than the average $SROCC$ over all subjective evaluations, the performance is considered acceptable. The idea of the RAM approaches the method proposed in this article. However, the units of the RAM measurement scale are unrelated to subjective perception, as in our performance measure.

It should be noted that, occasionally, the performance can be defined as acceptable only if the proposed method and the reference method show a statistically significant difference. The statistically significant difference can be calculated with, for example, the z-test[22] or student's t-test.[35,36] Furthermore, the F-test is used to determine whether the variances in the residuals are identical (i.e., whether the two sample sets are from the same distribution).[31,37]

2.4 Going beyond standard methods

In this study, we stress that the performance measure should have an informative measuring scale and units, should take into account the dispersion of the subjective data and should introduce some levels of acceptable performance. Next, we summarize why previous performance measures do not fulfill these requirements.

According to the review presented in Sections 2.1 to 2.3, the RAM proposed in[34] and d_{out} proposed in[4] as well as OR and $rmse^*$ measures proposed by the ITU-T P.1401 standard[22] take subjective dispersion into account when evaluating the performance of the algorithms. The resolving power measure[19] is also linked to subjective evaluations.

The RAM defines whether the performance of the algorithm is superior to that of the individual subjective evaluators. The accuracy of subjective evaluators is measured using the average $SROCC$ over all evaluators. The scale of the RAM could be more informative. Now, the scale indicates the performance level of one evaluator, but the units of the scale are undefined. For example, from the results data, it is difficult to draw conclusions about the meaning of the performance differences between the algorithms.

The OR performance measure defines the number of inaccurate predictions relative to the sample size. The $rmse^*$ performance measure defines the size of the average error between predictions and subjective average scores while filtering small errors from the sum. One of the problems with

both measures is that they use ci_{95} values to define the limits of the acceptable error. If the number of observers is high, the value of ci_{95} will be very low. That is, according to these performance measures, the algorithm should predict the average subjective score without error if averaged from a large observer group, even when the opinions are distributed in many groups or heavily dispersed. In addition, the scale of both performance measures could be more informative regarding human perception.

The d_{out} performance measure uses the standard deviation (σ) of subjective scores for defining the prediction error. The sum of error distances is calculated from the error bars. However, the scale of the d_{out} could be more informative. Now, it is difficult to draw conclusions about the meaning of the sum. That is, the scale of the measure is not related to the human perception.

Study[19] presented a method for describing minimum change in the subjective values that the algorithm can predict. In principle, the unit is understandably linked to human perception. However, the measure can indicate only that algorithm a identifies more samples than does algorithm b. From the resulting data, concluding what the performance differences between algorithms mean in practice can be difficult.

3 Subjective Root-Mean-Square Error (SRMSE)

This section describes the $SRMSE$ performance measure, which addresses the goals defined in the beginning of Section 2.4. The $SRMSE$ performance measure is used both for measuring the performance of individual algorithms (as units of the number of average observers) and for calculating the target values of sample sets (as units of the root mean square error). The target values can be used to indicate the performance level of the ideal algorithm.

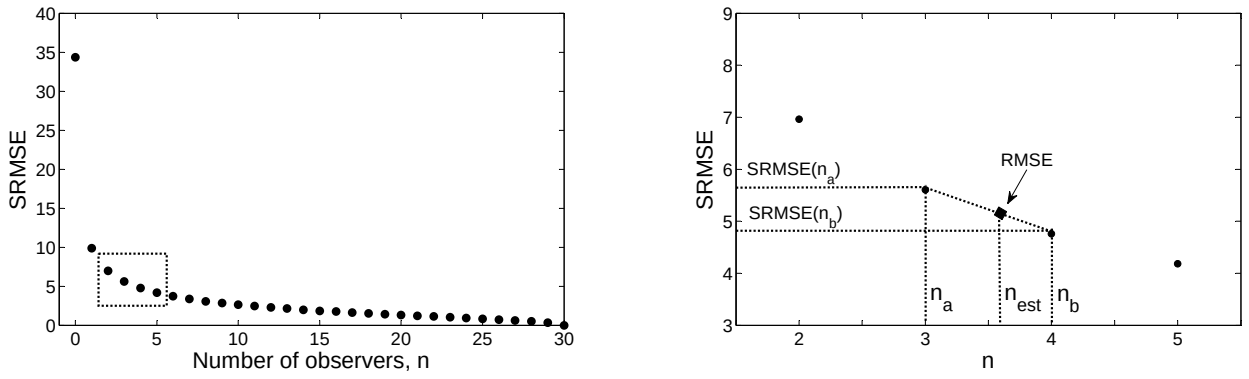
The calculation of the performance and target values is based on the $SRMSE(n)$ functions, which are computed by taking the $RMSE$ between the $MOS/DMOS$ of $n, (n = 1, ..., N)$, random observers and the $MOS/DMOS$ value of N observers when N is the number of all observers. For example, if $n = 3$, the $RMSE$ value between the $MOS/DMOS$ of three randomly selected observers and the $MOS/DMOS$ of N observers is calculated. The $MOS/DMOS$ value of N observers is assumed to be the ground truth for the perceived quality. This assumption requires a sufficiently high number of N (i.e., adding more observers will not change the X or σ).

To calculate the $SRMSE(n)$ functions, the different observer combinations cb, $(cb = 1, ..., K)$ are randomly selected K times (e.g., $K = 1000$) from the group of all N observers for all $n, (n = 1, ..., N)$. The $SRMSE_{cb}(n)$ is calculated with:

$$SRMSE_{cb}(n) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{m} \sum_{j=1}^{m} (X_{cb,j}(n) - X_j)^2} \tag{11}$$

where $X_{cb,j}(n)$ is the $MOS/DMOS$ of n observers for sample $j, (j = 1, ..., m)$ from random observer combination cb and X_j is the $MOS/DMOS$ computed from the group of all N observers. The $SRMSE(n)$ is the average value computed over all observer combinations:

$$SRMSE(n) = \frac{1}{K} \sum_{cb=1}^{K} SRMSE_{cb}(n) \tag{12}$$



(a) (b)

Fig 1: Linear interpolation estimates the number of observers value for the objective algorithm: (a) displays a set of discrete data points $(SRMSE, n)$ and (b) displays the magnified area from (a), indicating an example of linear interpolation estimating the number of average observers for the algorithm

3.1 Performance measure of algorithms

The performance of the algorithms for image and video quality assessment is calculated by linking the $RMSE$ value of the algorithm to the average number of observers n_{est} calculated from the $SRMSE(n)$ functions. The $RMSE$ value of the algorithm is calculated with the formula described in Eq. (1).

Because the performance of algorithm can be lower that of a single average observer, the hypothetical value of $SRMSE(0)$ is calculated by producing K random numbers $X_{cb,j,0}$, such as $(cb = 1, ..., K)$, and the numbers are between $[S_w...S_b]$, where S_w is the worst value, and S_b is the best value in the $MOS/DMOS$ scale.

Figure 1 indicates how the $RMSE$ value of an algorithm is linked to the $SRMSE$ values and how to calculate the performance measure of the average number of observers, n_{est}. First, Eq. (12) serves to define the sample-set-specific discrete data point group. Then, the n_{est} value is estimated for the $RMSE$ by linear interpolation:

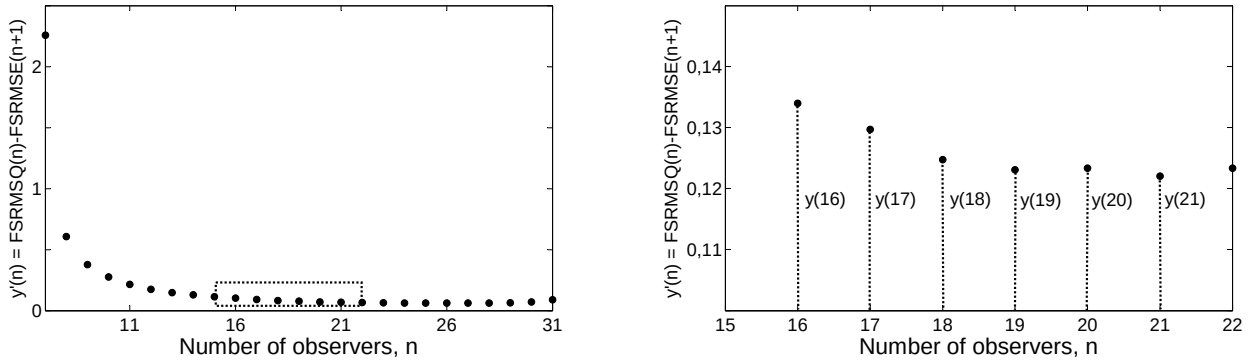
$$n_{est} = n_a + (n_b - n_a)\frac{RMSE - SRMSE(n_a)}{SRMSE(n_b) - SRMSE(n_a)} \tag{13}$$

where $SRMSE(n_a)$ and $SRMSE(n_b)$ are smaller and larger $SRMSE$ values, respectively, of the sample set compared to the $RMSE$ values of the algorithm. The values of n_a and n_b are the data points for the $SRMSE(n_a)$ and $SRMSE(n_b)$ values, respectively. The linear interpolation is illustrated in Fig. 1b.

Linear interpolation is a simple method for estimating missing data points. However, we assume that the accuracy of estimates is sufficient for this application because the problem is simple. The values of $SRMSE(n)$ as a function of n are monotonously decreasing, and the line between two data points can be approximated to a straight line.

3.2 Target values for algorithms

The target values for sample sets are defined by locating the stabilization points of the $SRMSE(n)$ functions. Stabilization means that the $SRMSE(n)$ function no longer decreases beyond that



(a) (b)

Fig 2: Target values for the data set are estimated by finding the stabilization point at which $y'(n) < y(n+1) + th$: (a) indicates the differential values from the $SRMSE(n)$ functions and (b) displays the magnified example from (a), from which $y(19)$ can be defined as the stabilization point

point. In the other words, the number of observers is high enough that adding more observers will not change the $MOS/DMOS$ or σ values.

Calculating the target values includes two phases. First, the moving average filter (F) filters the functions $SRMSE_{cb}(n)$ with so that $FSRMSE_{cb}(n) = \mathrm{F}(SRMSE_{cb}(n))$. In this study, we used the filter $[\frac{1}{8}\frac{1}{4}\frac{1}{4}\frac{1}{4}\frac{1}{4}\frac{1}{8}]$. Filtering is performed with the MATLAB $conv$ function.

In the second phase, the differential value functions $y_{cb}(n), cb = (1, ...K)$ are taken from the filtered functions:

$$\begin{aligned} y_{cb}(n) &= FSRMSE_{cb}(n) - FSRMSE_{cb}(n+1) \\ n &= 1, ..., N-1 \end{aligned} \tag{14}$$

and the final function for calculating the stabilization point is defined as the average of all $y_{cb}(n)$ functions:

$$y'(n) = \frac{1}{K} \sum_{cb=1}^{K} y_{cb}(n) \tag{15}$$

Figures 2a and 2b offer an example of target value derivation for the $y'(n)$ data points. If $y'(n) \leq y'(n+1) + th$ but $y'(n-1) \geq y'(n) + th$ when n is minimized and th is a threshold value, the target value for the sample set is:

$$target\ value = SRMSE(n) \tag{16}$$

The example in Figure 2b shows that $y'(19) \leq y'(20) + th$ and $y'(18) \geq y'(19) + th$ and $target\ value = SRMSE(19)$ if th is a small constant.

The th value is an important parameter that should be defined when finding target values. If the th value is high, the $SRMSE$ value of the stabilization point (target value) will be higher, and the number of observers, n, will be lower, and vice versa. In Section 6, we have presented one selection criterion of the th value.

Fig 3: The images of the CID2013 database were captured from eight different contents by 12 to 14 different cameras

4 Test material

In this study, we used the new $SRMSE$ performance measure to conduct a state-of-the-art benchmarking analysis of IQA algorithms. In this paper, we present the results for the CID2013[16] image database.

4.1 CID2013

In total, the CID2013 database contains 474 images captured by 79 different consumer cameras. The images are distributed in six image sets (I-VI), each captured by 12 to 14 different cameras. The quality levels of the cameras range from low to high. The number of observers was 30, 32, 31, 26, 34, and 34 for image sets I, II, III, IV, V, and VI, respectively. The database contains all subjective data.

Each image set includes images captured from six different contents. Thus, in total, the CID2013 contains 36 sample sets (6 image sets $\times$ 6 image contents). All images of the CID2013 are from the eight different contents. The majority of the contents represent environments in which consumers typically shoot photos. Figure 3 displays sample images of the different contents.

The subjective evaluations were collected using the dynamic reference (DR) method. The DR method creates reference image series from the test images. Before evaluating the test image of a given content on one display, all of the test images of the content in question are presented to the observer as a slide show for reference on the other display. More details on the DR method and the CID2013 database can be found in.[13,16] All experiments were designed and conducted with the VQone MATLAB toolbox.[38]

4.2 Image quality algorithms

We selected 13 No-Reference (NR) IQA algorithms for this study (see Table 1). We selected these algorithms because they were publicly available and are state-of-the-art algorithms or de facto reference algorithms for IQA performance studies.

We used the default settings to implement the algorithms. However, before evaluating the performance of an algorithm, we applied a nonlinear transformation to the predicted scores to bring the predicted (X) and measured (Y) values to the same scale and to account for the nonlinear

Table 1: The quality algorithms for the performance study with the CID2013 database

Metric	Description	Public Implementation
BIQI[39]	Overall quality	http://live.ece.utexas.edu/research/quality/BIQI_release.zip
BLIINDS-II[40]	Overall quality	http://live.ece.utexas.edu/research/quality/BLIINDS2_release.zip
BRISQUE[41]	Overall quality	http://live.ece.utexas.edu/research/quality/BRISQUE_release.zip
CPBD[42]	Distortion-specific sharpness metric	http://ivulab.asu.edu/software
DESIQUE[43]	Overall quality	http://vision.okstate.edu/yi/code/DESIQUE_release.rar
DIIVINE[44]	Overall quality	http://live.ece.utexas.edu/research/quality/DIIVINE_release.zip
FISH[45]	Distortion-specific sharpness metric	http://vision.okstate.edu/phongvu/code/FISH.rar
FISH bb[45]	Distortion-specific sharpness metric	http://vision.okstate.edu/phongvu/code/FISH.rar
LPC[46]	Distortion-specific sharpness metric	https://ece.uwaterloo.ca/~rhassen/LPC-SI/
Martziliano[47]	Distortion-specific sharpness metric	http://ivulab.asu.edu/software
NIQE[41]	Overall quality	http://live.ece.utexas.edu/research/quality/niqe_release.zip
NJQA[48]	Distortion-specific JPEG compression artifact metric	http://vision.okstate.edu/njqa/
S3[49]	Distortion-specific sharpness metric	http://vision.okstate.edu/s3/

relationships between values. This pre-processing is necessary before calculating $RMSE$ values. We used a logistic function with an added linear term:

$$f(X) = \beta_1(\frac{1}{2} - \frac{1}{1+\exp(\beta_2(X-\beta_3))}) + \beta_4 \cdot X + \beta_5 \tag{17}$$

where β_i are the model parameters chosen to minimize the mean square error between the $f(X)$ and the X values.

It should be noted that, in this study, we analyzed Y ($MOS/DMOS$) and X (predicted $MOS/DMOS$) values as content and image set specific way. In the study[16] related to the CID2013 database, IQA algorithms were analyzed using re-alignment data. The re-alignment data were estimated from the content- and image-set-specific average opinions and adjusted to the common scale, resulting in one average opinion score for all single image files. That is, because the $SRMSE$ performance measure requires observer-specific data, we have used the original raw data (observer-specific values after removal of outliers) distributed with the CID2013 instead of the re-alignment data.

5 Performance analysis of the state-of-the-art algorithms

5.1 SRMSE functions

Figure 4 presents the $SRMSE$ values as a function of the average observer n calculated for the image-set-specific contents of the CID2013 database. For the case of the CID2013, we have 36 functions (6 image contents $\times$ 6 image sets). With the aid of these functions, the performance measure values n_{est} of the state-of-the-art algorithms and 36 target values will be calculated.

The form of the functions is nonlinear. First, all functions decrease rapidly and then slowly. That is, the accuracy of the mean opinion compared to the ground truth as a function of n increases

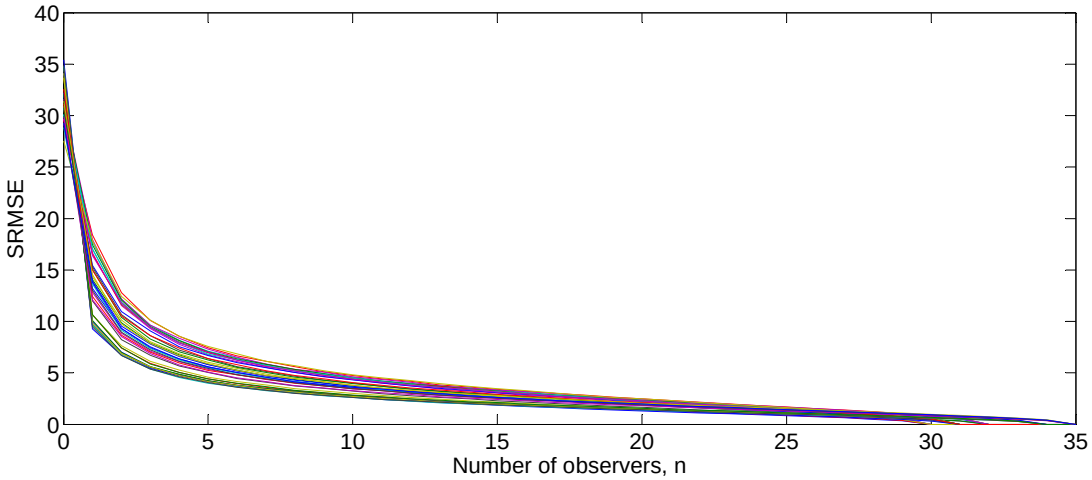


Fig 4: SRMSE functions for all image sets and contents of the CID2013 database

rapidly and then more slowly. From the figure, one can see that the values of the functions depend on the image set and content. If the image content causes greater dispersion among the subjective opinions, the values of $SRMSE(n)$ are higher. For example, based on Figure 4, when $n = 1$, the values of $SRMSE$ range from 9 to 18, depending on the image set and content. In addition, when $n = 5$, the $SRMSE$ values range from 4 to 7.

Figure 5 presents images from two content sources sorted in the order of the MOS values. Only small differences in quality between the images in the top row cause dispersion of the subjective opinions. The high dispersion of the subjective opinions then raises the $SRMSE(n)$ function values. Thus, the values of the $SRMSE(n)$ for the image samples in the top row are high (i.e., $SRMSE(1) = 18.42$). In addition, the quality of the images in the bottom row differ substantially. That is, sorting the images in order of quality is easier (leading to smaller dispersion), and the values of the $SRMSE(n)$ functions for the image samples in the bottom row are lower (i.e., $SRMSE(1) = 9.29$).

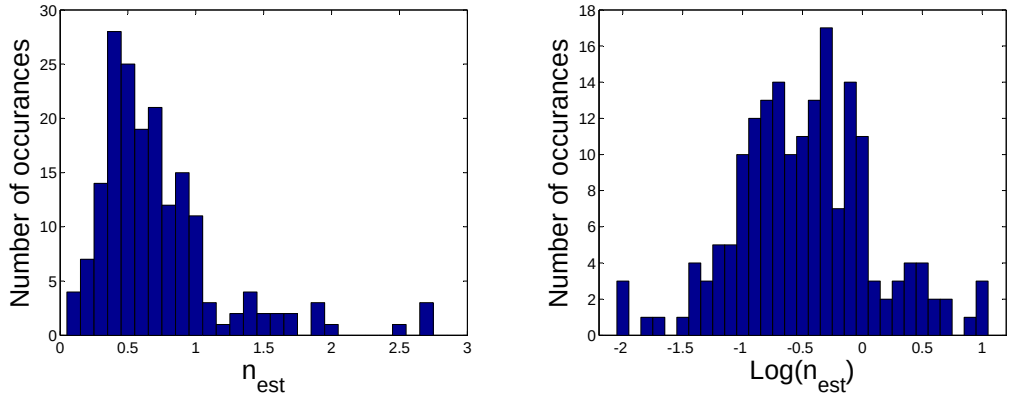
The example in Figure 5 demonstrates the ability of the $SRMSE$ to take into account the uncertainty of observers when calculating the performance of the algorithm. For example, the n_{est} value of 1.0 average observers (the $SRMSE$ performance unit) for the image sets in the top and bottom rows requires that the $RMSE$ values of the algorithm be 18.42 and 9.29, respectively. The same $SRMSE$ performance value is linked to the lower $RMSE$ value (higher accuracy) for the images in the bottom row as opposed to the images in the top row. That is, if the observers are unanimous, the algorithm should also be accurate. This example indicates that the $SRMSE$ performance measure compensates for the uncertainty of observers when calculating the performance of the algorithm.

5.2 Performance of the algorithms: Number of Observers

This section presents the results of the performance study for the state-of-the-art algorithms. The overall performance, $\hat{n}_{est}$, is defined as the average of the image set- and content-specific n_{est} values.

Another way to calculate the overall performance would be to estimate the average $SRMSE(n)$ function over all image set- and content-specific functions. Then, one can estimate the average $\hat{n}_{est}$

Fig 5: If the image set includes images with small quality differences causing higher dispersion among the subjective scores, the $SRMSE(n)$ function values are higher and vice versa (top row: $SRMSE(1) = 18.42$; bottom row: $SRMSE(1) = 9.29$)



(a) (b)

Fig 6: Histogram of the n_{est} values for the CID2013 database (a); Histogram of the $log(n_{est})$ data (b)

value by linking the average $SRMSE(n)$ function and the average $RMSE$ value of the algorithm. However, we recommend calculating the test- and content-specific n_{est} values. Therefore, it is possible to study whether the algorithms show statistically significant differences.

Before conducting the statistical tests, one should check for data normality. Figure 6a shows a histogram of the image set- and content-specific n_{est} values calculated for the algorithms listed in Table 1. The histogram shows that the distribution of the n_{est} values peaks and skews to the right. The kurtosis value of the distribution is 7.98, whereas the kurtosis value of a normal distribution is 3. The skewness value of the distribution is 2.01. If the skewness value is positive, the data are spread out more to the right. The skew of a normal distribution is zero. Also, the Shapiro-Wilks normality test[50] rejects the hypothesis of normality ($p < 0.05$).

According to the values of the normality tests, the data should be converted to normal with a proper modification. Figure 6b presents a histogram of the n_{est} values after logarithmic modification ($\log(n_{est})$). The kurtosis value of the $\log(n_{est})$ values is 3.55, and the skewness value is -0.03. The Shapiro-Wilks normality test indicates that one cannot reject the hypothesis of normality ($p = 0.1218$). According to these values, the data are normal and symmetric and can therefore be used to describe significant differences between the algorithms.

Figure 7 presents the n_{est} values for all algorithms selected for this study. Red dots indicate

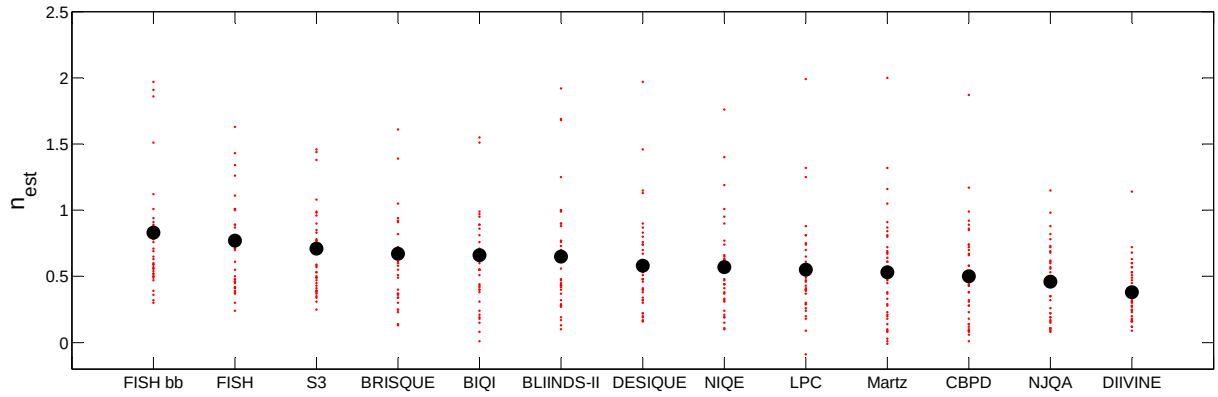


Fig 7: Image set- and content-specific n_{est} values (red points) for the CID2013 database. The black circles indicate the average values of all image sets and contents.

Table 2: The average n_{est}, $PLCC$ and $RMSE$ values of the algorithms for the CID2013 database

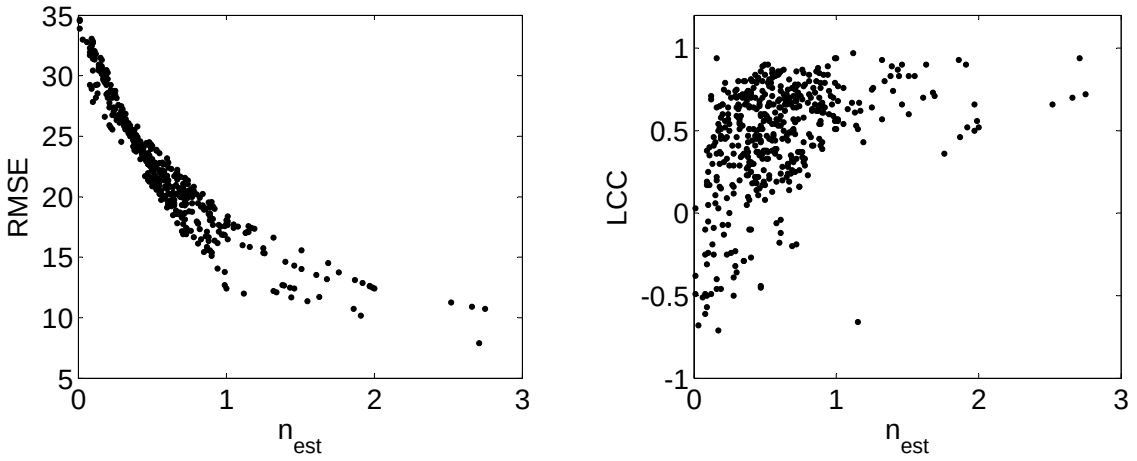
Algorithm	$\hat{n}_{est}$	$PLCC$	$RMSE$
FISH bb	0.83	0.72	19.19
FISH	0.77	0.67	19.97
S3	0.71	0.63	20.63
BRISQUE	0.67	0.50	21.42
BIQI	0.66	0.57	21.89
BLIINDS-II	0.65	0.54	21.70
DESIQUE	0.58	0.46	22.67
NIQE	0.57	0.44	22.83
LPC	0.55	0.48	22.79
Martziliano	0.53	0.29	23.79
CPBD	0.50	0.25	24.20
NJQA	0.46	0.17	24.45
DIIVINE	0.38	0.30	25.52

image set- and content-specific values. The average values $\hat{n}_{est}$ of the image sets and contents are indicated by black circles. These average values are listed in Table 2. The values of $PLCC$ and $RMSE$ can also be found in the table.

Figure 8a presents $RMSE$ values, and Figure 8b presents $PLCC$ values as a function of image set- and content-specific n_{est} values. The figures show that the n_{est} values predict the $RMSE$ values with quite high accuracy. This is an expected result because we calculated the n_{est} values from the connections between the $RMSE$ and $SRMSE$ values, and the distribution in Figure 8a resembles the distribution in Figure 4.

In contrast, the image set- and content-specific n_{est} values do not predict the $PLCC$ values with a high level of accuracy. Figure 8b shows that the $PLCC$ values saturate when the value of n_{est} is higher than 1 average observer. In addition, the $PLCC$ vs. n_{est} values are highly scattered. However, the correlation is rather high between the average $\hat{n}_{est}$ and $PLCC$ values of the algorithms (Table 2). The linear correlation coefficient is 0.92, and the rank-order correlation is 0.93, meaning that the disparity values between the image set- and content-specific $PLCC$ and n_{est} data points presented in Figure 8a are filtered out from the average values presented in Table 2.

According to the results presented in Table 2, the best performing algorithms were FISH bb, FISH and S3. FISH bb predicted the MOS values at a performance level of 0.83 average observers. The accuracy of the FISH bb approached the accuracy of one random observer relative to the



(a) (b)

Fig 8: $PLCC$ values of the state-of-the-art algorithms as a function of the n_{est} values (a); $RMSE$ values as a function of the n_{est} values (b)

ground truth. We claim that the number of average observers is an informative way to report and analyze the performance of image quality application. For example, the $PLCC$ and $SROCC$ values of the FISH bb were 0.72 and 19.19, respectively. The meaning of these numbers is difficult to interpret from the human perception point of view.

While observing the images in Figure 5, according to the $PLCC$ performance measure, the FISH bb predicted the top and bottom row images with accuracies of 0.41 and 0.90, respectively; according to the $SRMSE$ measure, the accuracies were 0.84 and 0.50 average observers, respectively. We can conclude that the $SRMSE$ performance measure indicates that the FISH bb is more accurate for the images in the top row than in the bottom row, if the subjective performance is taken into account. If we use only traditional performance measures, we will miss this important information.

We made a t-test[35] for the n_{est} values after logarithmic modifications. The t-test results are displayed in Table 3. A value of '1' in the table indicates that the row (algorithm) is statistically better than the column (algorithm); a value of '0' indicates that the row is worse than the column, and a value of '-' indicates that the row and column are statistically identical. Table 3 validates our observations from the performance measures: FISH bb, FISH and S3 performed best for the CID2013 database.

6 Target values for the CID2013, CVD2014 and LIVE MDIQ databases

In this section, we analyze the target values computed for the CID2013, CVD2014 and LIVE MDIQ databases. We also discuss the selection of the threshold value th needed for target value computation. The image set- and content-specific target values are available as .csv files from the project page. We hope that the use of target values becomes a common practice when the level of performance for ideal reference is indicated.

We calculated the target values by locating the stabilization points of the functions derived by Eq. (14). The selection of the threshold th greatly affects the location of the stabilization point. If th is higher, the target value ($RMSE$ value) increases, meaning that the performance of the ideal algorithm will be lower. If $th = 0$, there is a risk that the performance of the ideal algorithm will

Table 3: Statistical analysis of algorithm performance: A value of '1' in the cell indicates that the row (algorithm) is statistically better than the column (algorithm). A value of '0' indicates that the row is worse than the column. A value of '-' indicates that the row and column are statistically identical.

Algorithm	FISH bb	FISH	S3	BRISQUE	BIQI	BLIINDS-II	DESIQUE	NIQE	LPC	Martz	CPBD	NJQA	DIIVINE
FISH bb	-	-	-	-	1	-	1	1	1	1	1	1	1
FISH	-	-	-	-	-	-	1	1	1	1	1	1	1
S3	-	-	-	-	-	-	-	-	1	1	1	1	1
BRISQUE	-	-	-	-	-	-	-	-	-	-	1	1	1
BIQI	0	-	-	-	-	-	-	-	-	-	-	1	1
BLIINDS-II	-	-	-	-	-	-	-	-	-	-	-	1	1
DESIQUE	0	0	-	-	-	-	-	-	-	-	-	-	1
NIQE	0	0	-	-	-	-	-	-	-	-	-	-	1
LPC	0	0	0	-	-	-	-	-	-	-	-	-	1
Martziliano	0	0	0	-	-	-	-	-	-	-	-	-	-
CPBD	0	0	0	0	-	-	-	-	-	-	-	-	-
NJQA	0	0	0	0	0	0	-	-	-	-	-	-	-
DIIVINE	0	0	0	0	0	0	0	0	0	-	-	-	-

be so high that even a subjective experiment with an unlimited number of observers could reach that level of accuracy. It is clear that the selection of the th is an important research question, and next, we give one selection criterion for it.

Figures 9a, 9c and 9e present the target values ($RMSE$ values) as a function of the th values for the CID2013, CVD2014 and LIVE MDIQ databases. Image set- and content specific-values are indicated as red dots. The average values for image sets and contents are presented as black circles. Figures 9b, 9d and 9f present the number of observers as a function of the th values. For example, when $th = 0$, the corresponding average number of observers with the CID2013 database is 21.4. When $th = 0.01$, the corresponding average number of observers is 13.7. The corresponding $RMSE$ values are 1.90 and 3.03, which can be defined as the performance levels for the ideal algorithm. It can be noted that, comparing the performance level of the ideal algorithm and the $RMSE$ values presented in Table 2 for the state-of-the-art algorithms, there is room for performance improvements.

We propose that the selection of the th value used to calculate the target value is based on the number of observers recommended for subjective experiments. That is, the target value is selected from the stabilization point corresponding to a specific number of observers. For example, in this study, we found that when $th = 0.01$, the stabilization points for the CID2013 and CVD2014 databases were at the level corresponding to 13.7 and 14.8 average observers. Furthermore, we found that when $th = 0.00$, the stabilization point for the LIVE MDIQ database was at the level corresponding to 13.3 average observers. The standards[51,52] recommend that at least 15 observers participate in the experiment. Furthermore, Winkler[53] determined that, with 10-15 observers, the standard deviation reaches the actual value. Based on this discussion, we recommend using the target values calculated with $th = 0.01$ when determining the reference performance levels of the ideal algorithm for the CID2013 and CVD2014 databases and with $th = 0.00$ for the LIVE MDIQ database.

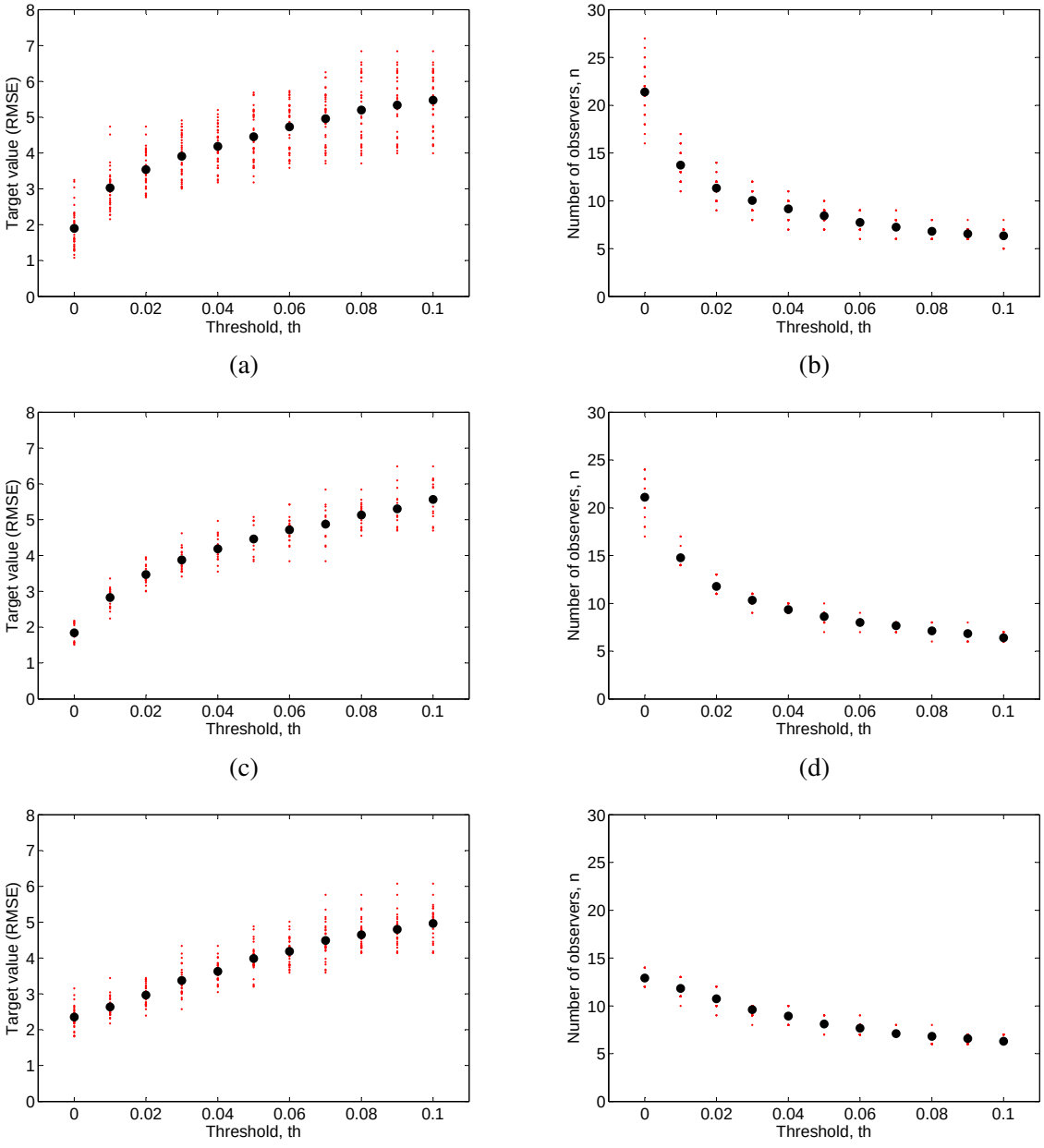


Fig 9: Target values ($RMSE$) as a function of the threshold value for the CID2013 (a), CVD2014 (c) and LIVE MDIQ (e) databases; number of observers related to the target values as a function of the threshold value for the CID2013 (b), CVD2014 (d) and LIVE MDIQ (f) databases

7 Discussions

In this section, we discuss how an acceptable performance level can be defined from the results of the $SRMSE$ performance measure. One traditional definition for an acceptable performance level is that the $PLCC$ and $SROCC$ values of the algorithm are higher than the values of the reference algorithm. The reference can be an algorithm with a de facto status for benchmarking, such as SSIM. It should also be noted that one important requirement from the point of view of scientific publishing is to demonstrate that the performance of the proposed algorithm is higher than that of the state-of-the-art algorithm.

However, the above-mentioned traditional approaches provide no answers to questions about the meaning of the performance levels or differences between algorithms. For example, if the $PLCC$ of reference is 0.75 and the $PLCC$ of the proposed algorithm is 0.80, we can define the performance difference as 0.05 units. However, it is difficult to consider what these numbers actually mean in terms of real-life applications. Furthermore, it is difficult to determine whether the performance of 0.80 units on the $PLCC$ or $SROCC$ scale is adequate for a specific application.

In this study, we propose using the units of the number of average observers to measure the performance of the algorithms. Because the unit is easy to understand and meaningless, we claim that the proposed $SRMSE$ performance measure can serve to define a level of acceptable performance. Furthermore, we consider the question of an acceptable performance level from the application point of view. For example, some applications require an accuracy level that is achievable (at least at this moment) only with a subjective experiment. For this type of application, the acceptable performance level corresponds to the target value (i.e., $th = 0.01$). That is, acceptable performance is a level of accuracy corresponding to the performance achieved by a subjective test with 15 observers. An example application can be the fine-tuning of the image processing of a camera device before market launch. In this case, the acceptable performance level for the implemented algorithm would be the average $RMSE$ value of 3.03, if the validation material is from the CID2013 database.

Some applications can benefit even if the accuracy of the predictions is not as high as the target value indicates. For example, the application of manual image selection can be improved by using an IQA algorithm for image pre-selection. For example, an application for pre-tuning the image processing pipe can be accelerated by using an IQA algorithm for searching the coarse intervals of the tuning parameters. The acceptable performance level can be rather low, i.e., at the level of 1 random observer. Then, the average $RMSE$ value required for the implemented algorithm would be 13.7, if the validation material is the CID2013 database. According to the state-of-the-art analyses in this paper, some best performer algorithm can nearly achieve this performance level.

One limiting factor, for which the $SRMSE$ performance measure could be utilized with all the available data, is that all the publicly available image and video databases do not distribute observer-specific data. For example, the widely used LIVE, TID2008 and CSIQ databases contain only the mean opinions rating and σ values. In the future, it is important for all data to be distributed when new databases are launched, as we have done with our new CID2013 and CVD2014 databases.

8 Conclusions

This study presented a new performance measure for image and video quality assessment algorithms. The unit of the $SRMSE$ is the number of average observers. The unit is informative

compared to the units of the traditional performance measures. The number of average observers is easy to link with the acceptable performance levels of different image quality applications. For example, it can be stated that the accuracy of the implemented algorithm should be higher than one random observer.

In addition, this study presented the concept of target values. The target values define the performance levels for ideal algorithms depending on the level of dispersion between the observers when the sample set is evaluated. If there are large differences between the opinions, the accuracy of the ideal algorithm is lower compared to the situation in which all observers were unanimous. The concept of the ideal algorithm can be used as the reference performance indicator for image and video quality assessment algorithm evaluations and benchmarking.

References

1 A. Ninassi, P. Le Callet, and F. Autrusseau, "Pseudo no reference image quality metric using perceptual data hiding," in *Proc. of SPIE/IS&T Electronic Imaging 2006*, **6057**, 60570G–60570G–12 (2006).

2 H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing* **15**, 3440–3451 (2006).

3 N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "Tid2008-a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics* **10**(4), 30–45 (2009).

4 E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging* **19**(1), 011006 (2010).

5 N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Jay Kuo, "Image database TID2013," *Image Commun.* **30**, 57–77 (2015).

6 R. Streijl, S. Winkler, and D. Hands, "Perceptual quality measurement - towards a more efficient process for validating objective models [standards in a nutshell]," *IEEE Signal Processing Magazine* **27**, 136–140 (2010).

7 S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE Journal of Selected Topics in Signal Processing* **6**, 616–625 (2012).

8 O. Wu, W. Hu, and J. Gao, "Learning to predict the perceived visual quality of photos," in *IEEE International Conference on Computer Vision (ICCV)*, 225–232 (2011).

9 M. Pinson, L. Janowski, R. Pepion, Q. Huynh-Thu, C. Schmidmer, P. Corriveau, A. Younkin, P. Le Callet, M. Barkowsky, and W. Ingram, "The influence of subjects and environment on audiovisual subjective tests: An international study," *IEEE Journal of Selected Topics in Signal Processing* **6**, 640–651 (2012).

10 R. Streijl, S. Winkler, and D. Hands, "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives," *Multimedia Systems* , 1–15 (2014).

11 S. Winkler, "Does inter-subject variability depend on test material?," in *International Workshop on Quality of Multimedia Experience (QoMEX)*, 37–38 (2014).

12 P. Satgunam, R. Woods, P. Bronstad, and E. Peli, "Factors affecting enhanced video quality preferences," *IEEE Transactions on Image Processing* **22**, 5146–5157 (2013).

13 M. Nuutinen, T. Virtanen, T. Leisti, T. Mustonen, J. Radun, and J. Häkkinen, "A new method for evaluating the subjective image quality of photographs: dynamic reference," *Multimedia Tools and Applications* , 1–25 (2014).

14 M. Narwaria, W. Lin, I. McLoughlin, S. Emmanuel, and L.-T. Chia, "Fourier transform-based scalable image quality measure," *IEEE Transactions on Image Processing* **21**, 3364–3377 (2012).

15 T. Virtanen, M. Nuutinen, T. Leisti, J. Radun, and J. Häkkinen, "Alternative performance metrics and target values for the CID2013 database," in *Proc. of SPIE/IS&T Electronic Imaging 2015*, **9396**, 93960Q–93960Q–11 (2015).

16 T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen, and J. Häkkinen, "CID2013: A Database for Evaluating No-Reference Image Quality Assessment Algorithms," *IEEE Transactions on Image Processing* **24**, 390–402 (2015).

17 M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen, "CVD2014: A Database for Evaluating No-Reference Video Quality Assessment Algorithms," *Submitted for review* (2015).

18 D. Jayaraman, A. Mittal, A. Moorthy, and A. Bovik, "Objective quality assessment of multiply distorted images," in *Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, 1693–1697 (2012).

19 M. Brill, J. Lubin, P. Costa, and J. Pearson, "Accuracy and cross-calibration of video-quality metrics: new methods from ATIST1A1," in *IEEE International Conference on Image Processing (ICIP)*, **3**, III–37–III–40 (2002).

20 F. Ciaramello and A. Reibman, "Systematic stress testing of image quality estimators," in *IEEE International Conference on Image Processing (ICIP)*, 3101–3104 (2011).

21 A. Reibman, "A strategy to jointly test image quality estimators subjectively," in *IEEE International Conference on Image Processing (ICIP)*, 1501–1504 (2012).

22 "ITU-T P.1401. Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," (2013).

23 I. Cotanis, "QoE Model Performance Evaluation," *VQEG eLetter* **1**, 6–13 (2014).

24 M. Pinson, N. Staelens, and A. Webster, "The history of video quality model validation," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 458–463 (2013).

25 A. Mittal, M. Saad, and A. C. Bovik, "Assessment of video naturalness using time-frequency statistics," in *IEEE International Conference on Image Processing (ICIP)*, 571–574 (2014).

26 Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing* **13**, 600–612 (2004).

27 Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *Asilomar Conference on Signals, Systems and Computers*, **2**, 1398–1402 (2003).

28 A. Kolaman and O. Yadid-Pecht, "Quaternion structural similarity: A new quality index for color images," *IEEE Transactions on Image Processing* **21**, 1526–1536 (2012).

29 G. Zhai, X. Wu, X. Yang, W. Lin, and W. Zhang, "A psychovisual quality metric in free-energy principle," *IEEE Transactions on Image Processing* **21**, 41–52 (2012).

30 P. Ye and D. Doermann, "No-reference image quality assessment using visual codebooks," *IEEE Transactions on Image Processing* **21**, 3129–3138 (2012).

31 W. Xue, L. Zhang, X. Mou, and A. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing* **23**, 684–695 (2014).

32 M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting* **50**, 312–322 (2004).

33 K. Seshadrinathan and A. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing* **19**, 335–350 (2010).

34 H. Nachlieli and D. Shaked, "Measuring the quality of quality measures," *IEEE Transactions on Image Processing* **20**, 76–87 (2011).

35 R. Wilcox, *Basic Statistics : Understanding Conventional Methods and Modern Insights: Understanding Conventional Methods and Modern Insights*, Oxford University Press, USA (2009).

36 M. Nuutinen, T. Virtanen, and P. Oittinen, "Image feature subsets for predicting the quality of consumer camera images and identifying quality dimensions," *Journal of Electronic Imaging* **23**, 061111 (2014).

37 J. Wu, W. Lin, G. Shi, and A. Liu, "Perceptual quality metric with internal generative mechanism," *IEEE Transactions on Image Processing* **22**, 43–54 (2013).

38 M. Nuutinen, T. Virtanen, O. Rummukainen, and J. Häkkinen, "VQone MATLAB toolbox: A graphical experiment builder for image and video quality evaluations," *Behavior Research Methods* , 1–13 (2015).

39 A. Moorthy and A. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters* **17**, 513–516 (2010).

40 M. Saad, A. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE Transactions on Image Processing* **21**, 3339–3352 (2012).

41 A. Mittal, A. Moorthy, and A. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing* **21**, 4695–4708 (2012).

42 N. Narvekar and L. Karam, "A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection," in *International Workshop on Quality of Multimedia Experience (QoMEx)*, 87–91 (2009).

43 Y. Zhang and D. M. Chandler, "No-reference image quality assessment based on log-derivative statistics of natural scenes," *Journal of Electronic Imaging* **22**(4), 043025 (2013).

44 A. Moorthy and A. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing* **20**, 3350–3364 (2011).

45 P. Vu and D. Chandler, "A fast wavelet-based algorithm for global and local image sharpness estimation," *IEEE Signal Processing Letters* **19**, 423–426 (2012).

46 R. Hassen, Z. Wang, and M. Salama, "Image sharpness assessment based on local phase coherence," *IEEE Transactions on Image Processing* **22**, 2798–2810 (2013).

47 P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: Application to jpeg2000, signal process," *Signal Processing: Image Communication* **19**(2), 163–172 (2004).

48 S. Golestaneh and D. Chandler, "No-reference quality assessment of jpeg images via a quality relevance map," *IEEE Signal Processing Letters* **21**, 155–158 (2014).

49 C. Vu, T. Phan, and D. Chandler, "S3 : A spectral and spatial measure of local perceived sharpness in natural images," *IEEE Transactions on Image Processing* **21**, 934–945 (2012).
50 S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika* **52**(3-4), 591–611 (1965).
51 "ITU-T Rec. P. 910. Subjective video quality assessment methods for multimedia applications," (2008).
52 "ITU-R BT.500. Methodology for the subjective assessment of the quality of television pictures," (2012).
53 S. Winkler, "On the properties of subjective ratings in video quality experiments," in *International Workshop on Quality of Multimedia Experience (QoMEx)*, 139–144 (2009).

Mikko Nuutinen received a M.Sc. (Tech) and a Lic.Sc (Tech) degrees from the Helsinki University of Technology in 2004 and 2007, respectively, and a D.Sc (Tech) degree from Aalto University, Helsinki in 2012. His current research interests are in the areas of predictive analytics, image and video quality assessment algorithms, camera performance measurements, and subjective assessment methods and analysis.

Toni Virtanen received his M.A. (psychology) degree from University of Helsinki, Finland in 2010. He has been working on image quality measurement and subjective evaluation since 2005 and is now pursuing a doctoral degree on related topics in psychology. His main occupation has been developing and conducting subjective image quality experiments in collaboration projects with Nokia, Microsoft and others at the University of Helsinki Visual Cognition research group.

Jukka Häkkinen received PhD in the Institute of Behavioural Sciences, University of Helsinki, Finland. He worked as Principal Scientist at Nokia Research Center, and as Adjunct Professor at Department of Media Technology, Aalto University School of Science. Currently he is Principal Investigator at the Institute of Behavioral Sciences, University of Helsinki, where he leads Visual Cognition Research Group. His interests include visual quality, attention, scene perception, and visual ergonomics of stereoscopic, head-mounted and flexible displays.

List of Figures

List of Tables