Alternative performance and metrics target values for the CID2013 database

T. Virtanen, M. Nuutinen, J. Radun, T. Leisti, J. Häkkinen
Institute of Behavioral Science, University of Helsinki, Helsinki, Finland

ABSTRACT

An established way of validating and testing new image quality assessment (IQA) algorithms have been to compare how well they correlate with subjective data on various image databases. One of the most common measures is to calculate linear correlation coefficient (LCC) and Spearman’s rank order correlation coefficient (SROCC) against the subjective mean opinion score (MOS). Recently, databases with multiply distorted images have emerged [1], [2]. However with multidimensional stimuli, there is more disagreement between observers as the task is more preferential than that of distortion detection. This reduces the statistical differences between image pairs. If the subjects cannot distinguish a difference between some of the image pairs, should we demand any better performance with IQA algorithms? This paper proposes alternative performance measures for the evaluation of IQA’s for the CID2013 database. One proposed alternative performance measure is root-mean-square-error (RMSE) value for the subjective data as a function of the number of observers. The other alternative performance measure is the number of statistical differences between image pairs. This study shows that after 12 subjects the RMSE value saturates around the level of three, meaning that a target RMSE value for an IQA algorithm for CID2013 database should be three. In addition, this study shows that the state-of-the-art IQA algorithms found the better image from the image pairs with a probability of 0.85 when the image pairs with statistically significant differences were taken into account.

Keywords: Image quality assessment algorithms, performance metrics, subjective image quality evaluation

1. INTRODUCTION

One of the main priorities of the image quality research has been the creation of image quality assessment (IQA) algorithms that are capable of predicting the subjective visual quality of natural scenes. For this end, publicly available image databases have been used when validating image quality assessment algorithms [3]–[5]. An established way of testing the new IQA algorithms is to compare how well they correlate with subjective data on various image databases. A proven method have been to calculate linear correlation coefficient (LCC) and Spearman’s rank order correlation coefficient (SROCC) against the subjective mean opinion score (MOS). LCC measures the prediction accuracy and SROCC the relative monotonicity between the predicted values and subjective MOS. One of the reasons why these methods have been selected is that they are the only ones that can be calculated from databases that do not distribute other information than the MOS along with the image files. However, MOS is only a value of central tendency of the observers’ ratings and do not tell what other statistical properties the subjective data might contain. This introduces a risk that the correlation coefficients might not always reveal the actual performance of the IQA algorithms or that the various results from different image databases might not be fully comparable as the distribution behind the subjective data in different database varies.

When the image databases mainly include images with only one type of distortion at a time, such as increasing levels of digital noise or other image attributes, the central tendency can be considered sufficient indicator of observed image quality. Recently, also databases with multiply distorted images have emerged [2], [6]. In these image databases the multiple distortions can mask or enhance each other in the images. With multidimensional stimuli, there can be more disagreement between observers as the task is more preferential than with one dimensional stimulus where the task is closer to distortion detection. One observer might prefer one sample, for example with less noise, while another prefers a different sample with better exposure but more noise. This reduces the statistical differences between image pairs in the subjective data as the standard deviations increase. If the consensus between the observers in sorting the samples in their order of quality is low, the target performance level should be more relaxed as well.
In this paper we propose two alternative performance measures for the newly published multiply distorted real camera image database (CID2013)[2] that can be used to evaluate the performance of IQA algorithms in addition to the established methods LCC and SROCC values. To our knowledge the CID2013 is the first database that includes the complete raw data from the subjective experiments. When the full data is available other descriptive statistics become available. This extra information is especially important when the images have multidimensional properties. The first target value is calculated from the performance measure of subjective Root-Mean-Square-Error (RMSE). The second target value is based on the statistical differences between image pairs in the subjective data. A target value is the point when a performance statistic measure is saturated, e.g. when increase in sample size (number of observers) does not change its value significantly.

2. CID2013 DATABASE

In contrast to previous image databases, the CID2013- Camera Image Database uses retail cameras instead of introducing distortions via post-processing. Retail cameras contain images that can have enhancements and distortions that are multidimensional and more subtle in nature. Retail cameras are often equipped with low quality optics and image sensors; their shooting process can cause motion blur and poor focus, and their low-sensitivity pixels increase noise level. In addition, the conditions that pictures are taken in, such as shooting distance and lighting conditions, affect the quality of raw images. After shooting, the raw images are processed by the image signal processing (ISP) pipe within the camera. The ISP includes operations such as color filter array demosaicking, automatic white balancing, color correction, noise filtering, tone reproduction, gamma correction, edge enhancement, color saturation enhancement, and image compression[7], [8]. The parameters and the order of operations affect the resulting image. Some operations, such as demosaicking, white balancing, and noise filtering, seek to restore the image. Other components, such as edge enhancement and color saturation enhancement, aim to produce a pleasant image.

The database consists of six image sets (I-VI) where on average 30 subjects have evaluated 12-14 devices depicting 8 different scenes, see Figure 1. The subjective evaluation method was a hybrid ACR-Pair Comparison developed for the study and presented in this study[9].
2.1 Test procedure and environment

We used two Eizo ColorEdge CG241W 24” monitors, with a third smaller display underneath for presenting questions (see Picture 3). The data collection and image presentation software was created using the VQone MATLAB toolbox[10], [11]. The room was covered with medium gray curtains to diffuse the ambient illumination. Fluorescent lights (5800K) were positioned behind the monitors and reflected from the back wall covered with grey curtain to create dim and uniform ambient illumination in the room. In dim viewing conditions backlighting has been found to reduce eyestrain and visual fatigue [12]. The light hitting the monitors measured below 2 lx, and the ambient illumination from behind the monitors were 20 lx. The subject’s viewing distance (approximately 80 cm, 2½ picture heights) was controlled by a line hanging from the ceiling, and they were instructed to keep their forehead steady next to the line. Because of the display size, images were scaled to a size of 1600 x 1200 pixels using the bicubic interpolation method resulting in a horizontal size of 30 degrees of visual angle and a resolution of 52 pixels per degree of visual angle. Monitors were calibrated to sRGB having target values of: 80 cd/m2, 6500K and gamma 2.2 using EyeOne Pro calibrator (X-rite co. Grand Rapids, MI, USA) [13]. A relatively low luminance on the monitors was selected as in dim viewing environment bright monitors would cause unnecessary eyestrain and subject fatigue. Gamma curves and chromaticity coordinates of R,G,B for all monitors are shared along with the database as well as provided at www.helsinki.fi/psychology/groups/visualcognition/

2.2 Data statistics

While collecting data for the CID2013 databases, we had Image Sets, which were each estimated by 26-35 observers. Here, we estimate the average standard deviation values as a function of the number of observers, n, to investigate whether or not the number of observers was high enough. The standard deviation of the random observer combination cb, (i = 1,..., 1000), was calculated for image set is ∈ {1,...,6} as an average over the clusters c ∈ {1,...,6} by Equation:

$$
\sigma_{is,n} = \frac{1}{6} \sum_{c=1}^{6} \frac{1}{m} \sum_{j=1}^{m} \left( \frac{1}{n} \sum_{k=1}^{n} (S_{is,c,j,k} - MOS_{is,c,j})^2 \right)
$$

(1)

where $S_{is,c,k}$ is the quality evaluation of observer $k$ ($k = 1,...,n$) for image $i_{c,j}$ when $I_{is,c} = \{i_{c,j} | j = 1,...,m\}$ and

$$
MOS_{is,c,j} = \frac{1}{n_{is}} \sum_{l=1}^{n_{is}} S_{is,c,l,j}
$$

(2)

where $n_{is}$ is the total amount of observers in image set $is$. The observer combinations of different sizes ($n = 1,2,...,n_{is}$) were randomly selected 1000 times from the group of all of the observers. Figure 2 shows the average standard deviation as a function of the number of observers for image sets I - VI. From the figure, it can be seen that standard deviation is being established, when $n > 15$. Additionally, study [14] showed the same type of results. We conclude that the $n$ in all of image sets of the CID2013 was adequate.
3. METHODS

In this section we review the metrics of the subjective RMSE\cite{15}, \cite{16} and the number of statistically different image pairs\cite{17} as alternative performance measures for objective image quality assessment (IQA) algorithms.

3.1 Subjective Root-Mean-Square-Error (RMSE)

We estimated the subjective RMSE values by comparing the mean opinion score (MOS) values of n random observers with the MOS value of all of the observers \cite{15}, \cite{16}. For example, if n = 3, the mean value of three selected observers is compared with the mean of all observers. The different observer combinations are randomly selected from the group containing all observers, and the subjective RMSE as a function of n for image set is \(i\) is calculated by Equation:

\[
RMSE_{is,n} = \sqrt{\frac{1}{6} \sum_{c=1}^{6} \sum_{j=1}^{m} (MOS_{is,c,j,n} - MOS_{is,c,j})^2}
\]  

where \(MOS_{is,c,j,n}\) is the mean opinion score of n observer. The observer combinations of different sizes were randomly selected 1000 times from the group of all of the observers and the subjective RMSE is the average value computed from all combinations. This process was repeated 10 times to ensure adequate sampling and averages of the results were used in further analyses.

The target value (subjective performance) is defined by searching the zero point for the median filtered gradient function \(y(n) = RMSE_{is,n} - RMSE_{is,n+1}\). The performance value of an IQA measure is defined as the number of random observers from the point where the subjective and objective RMSE values are equal.
3.2 Statistical differences between image pairs

With full data we can calculate e.g. the number of statistical differences between every image pair, using Linear Mixed Models. Linear Mixed Models expand the general linear model so that the data can exhibit correlated and non-constant variability. The mixed linear model, therefore, provides the flexibility of modeling not only the means of the data but their variances and covariance’s as well. [17], [18].

We utilized Akaike's Information Criterion (AIC)[18] with to evaluate seven potential covariance structures for, the best model with least complexity for CID2013 data. The AIC tests how well a statistical model fit to a given set of data:

\[
AIC = 2k - 2\ln(L) \tag{4}
\]

where \(k\) is the number of parameters in the model and \(L\) is the maximized value of the likelihood function of the model. The AIC rewards for goodness of fit, but also includes a penalty for the complexity of the model reducing the risk of over fitting. The smaller the value of AIC the better fit the model is:

- Diagonal: 20989.93;
- Compound Symmetry: 21016.08;
- First order Factor Analytic: Constant Diagonal Offset: 21060.99;
- Toeplitz: 20947.06;
- First order Ante-Dependence: 20870.95;
- First-Order Autoregressive: 21226.21; Compound symmetry: Heterogeneous 20671.52.

From the AIC results the best fit for the data is with Heterogeneous Compound Symmetry (HCS) covariance matrix. It has heterogeneous variances and constant correlation between the elements:

\[
\begin{pmatrix}
\sigma_1^2 & \cdots & \sigma_1\sigma_n\rho \\
\vdots & \ddots & \vdots \\
\sigma_n\sigma_1\rho & \cdots & \sigma_n^2
\end{pmatrix}
\tag{5}
\]

Where \(\sigma\) is variance and \(\rho\) is correlation.

When considering the covariance structure using the elements from the data, observers, devices and contents. The HCS takes into account the repeated effect when observers evaluate the combinations same scenes and the devices multiple times. Image quality can also exhibit a strong correlational effect with various scenes and device combinations. High end devices perform in overall better in most of the scenes, while certain scenes can be found more difficult or easy for all the tested devices. Thus the use of HCS covariance structure also makes sense from the data properties point of view.

Testing the statistical difference between the every potential image pair tests increases the risk of Type I error, e.g. the risk of false positive significance. Therefore a correction is needed for the test alpha threshold [19]. One of the simplest corrections is the Bonferroni correction that simply divides the alpha significance threshold of 0.05 with the number of tests made. The down side of the Bonferroni correction is that it makes the limit of statistical significance very conservative, meaning that image pairs with small differences might not be flagged as statistically significant like they might be with more relaxed alpha threshold. However, the benefit is that only image pairs with real substantial difference in quality are flagged as statistically significant, reducing the risk of false positives.

3.3 Number of statistically different image pairs

The statistically significant difference metric examines the percent of image pairs with statistically significant difference from the all possible image pairs in the image sets. [17]

Figure 3 shows how the target and the performance values for an IQA measure is calculated for a set of images \(I_s\). Let \(I_s = \{im_i | i = 1, \ldots, n\}\), when \(n\) is the number of images in the group \(s\). In the figure \(n = 6\). Matrix \(M1\) contains the \(p\) values of the paired comparisons that are calculated from the subjective data. In Matrix \(M2\), the cell value is 1 (-1) if the row image \(im_i\) is statistically significantly better (worse) than the column image \(im_j\). The cell of Matrix \(M3\) is 1 (-1) if the IQA measure predicted image \(im_i\) better (worse) than image \(im_j\). The cell of matrix \(M4\) is 1 if the IQA measure predicted
the better image correctly from image pair \((k, l)\) and if there was a statistically significant difference between the image pair. The performance measure, \(\text{Prob}\), is calculated by Equation:

\[
\text{Prob} = \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{M^4(k,l)}{|M^2(k,l)|} \quad (6)
\]

where the sum of the matrix \(M^4\) cells is divided by the sum of the absolute values of matrix \(M^2\) cells. The proposed measure gives the probability that IQA algorithms predicts the sample pairs in the correct quality order if and only if there is a statistically significant difference between the samples. The target value (subjective performance) for a IQA is calculated by Equation:

\[
\text{Target} = \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{|M^2(k,l)|}{n^2-n} \quad (7)
\]

where \(n\) is the number of images.

<table>
<thead>
<tr>
<th>M1</th>
<th>M2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

\[\text{Prob} = \frac{\sum M^4(i,j)}{\sum|M^2(i,j)|}\]

\[\text{Prob} = \frac{\sum M^4(i,j)}{\sum|M^2(i,j)|}\]

Figure 3: The probability of an IQA measure to find the statistically significant image-pairs is calculated by dividing the number of image pairs found by the measure and the number of image pairs with statistically significant difference.

4. **RESULT: TARGET VALUES**

4.1 **RMSE target values**

Figure 4 presents the Subjective RMSE values as a function of the number of observers. It shows that on average a single randomly chosen observer has a RMSE value between 9-18 depending on the image Cluster and Image set. The figure 4 also shows that the RMSE stabilizes somewhere after 10 observers in the CID2013 database. Table 1 examines this
further by estimating the point where the RMSE stabilizes and no longer decrease significantly with more observers. The estimated Subjective RMSE target value was calculated by taking a differential coefficient from the above RMSE values plotted in figure 4. The “elbow” of that differential coefficient function was traced by taking an average of three sequential values and comparing them to the average of following three sequential values. When the preceding average was smaller than the following average, it could be argued that the RMSE had stabilized. The RMSE values in Table 1, can be interpreted as the best possible outcome any no-reference IQA algorithms needs to achieve to have similar agreement for the quality of the CID2013 images that is present in the subjective data.

Figure 4: Subjective RMSE values as a function of the number of observers from the six image sets of CID2013

<table>
<thead>
<tr>
<th>Image Set</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>12</td>
<td>2.34</td>
<td>11</td>
<td>2.63</td>
<td>12</td>
<td>2.64</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>12</td>
<td>3.18</td>
<td>12</td>
<td>3.76</td>
<td>12</td>
<td>2.90</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>12</td>
<td>2.55</td>
<td>11</td>
<td>2.53</td>
<td>12</td>
<td>2.43</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>13</td>
<td>3.73</td>
<td>12</td>
<td>4.15</td>
<td>13</td>
<td>3.63</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>12</td>
<td>3.69</td>
<td>11</td>
<td>4.32</td>
<td>13</td>
<td>3.65</td>
</tr>
<tr>
<td>Cluster 6</td>
<td>11</td>
<td>3.54</td>
<td>13</td>
<td>3.95</td>
<td>12</td>
<td>3.88</td>
</tr>
<tr>
<td>Cluster 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster 8</td>
<td></td>
<td>11</td>
<td>2.98</td>
<td></td>
<td>12</td>
<td>3.10</td>
</tr>
</tbody>
</table>
Figure 5 compares the overall average of the subjective RMSE value of one hypothetical randomly chosen observer from all the Image sets and image clusters to the RMSE values calculated from various no-reference IQA algorithms [20]–[31] that has been tested with CID2013. The hypothetical single average observer RMSE score was 14, which can be considered a relaxed target value for the IQA algorithms. When an IQA achieves a RMSE value below 14 it can be argued that it can predict the quality from the CID2013 as well as one randomly chosen observer. As none of the tested IQA’s were optimized for the type of images in CID2013, it is understandable that there is still some way before such claim can be made.

Figure 5: RMSE value of one average observer of CID2013 compared to the RMSE values of IQA algorithms.
4.2 Statistically significant difference metric

Table 2 presents the results from the statistically significant difference metric, where the probability to predict the correct order of statistically significant image pairs is given for each image cluster. It needs to be reminded here that as the statistical testing of each image pair used Bonferroni correction, only image pairs with relatively large differences were flagged as significant since the correction made the alpha significance threshold very conservative.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
<th>Cluster 5</th>
<th>Cluster 6</th>
<th>Cluster 7</th>
<th>Cluster 8</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>FISH_bb [18]</td>
<td>0.88</td>
<td>0.82</td>
<td>0.91</td>
<td>0.77</td>
<td>0.87</td>
<td>0.87</td>
<td>0.79</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>FISH [18]</td>
<td>0.85</td>
<td>0.79</td>
<td>0.84</td>
<td>0.82</td>
<td>0.90</td>
<td>0.89</td>
<td>0.81</td>
<td>0.81</td>
<td>0.84</td>
</tr>
<tr>
<td>S3 [19]</td>
<td>0.85</td>
<td>0.76</td>
<td>0.85</td>
<td>0.73</td>
<td>0.80</td>
<td>0.81</td>
<td>0.82</td>
<td>0.78</td>
<td>0.80</td>
</tr>
<tr>
<td>LPC [24]</td>
<td>0.84</td>
<td>0.78</td>
<td>0.93</td>
<td>0.61</td>
<td>0.61</td>
<td>0.56</td>
<td>0.79</td>
<td>0.75</td>
<td>0.74</td>
</tr>
<tr>
<td>CBPD [27]</td>
<td>0.38</td>
<td>0.61</td>
<td>0.56</td>
<td>0.78</td>
<td>0.78</td>
<td>0.76</td>
<td>0.72</td>
<td>0.46</td>
<td>0.63</td>
</tr>
<tr>
<td>NJQA [28]</td>
<td>0.42</td>
<td>0.53</td>
<td>0.38</td>
<td>0.61</td>
<td>0.34</td>
<td>0.42</td>
<td>0.35</td>
<td>0.42</td>
<td>0.44</td>
</tr>
<tr>
<td>Marziliano [26]</td>
<td>0.58</td>
<td>0.27</td>
<td>0.47</td>
<td>0.32</td>
<td>0.26</td>
<td>0.24</td>
<td>0.35</td>
<td>0.44</td>
<td>0.37</td>
</tr>
<tr>
<td>DIVINE [29]</td>
<td>0.67</td>
<td>0.31</td>
<td>0.25</td>
<td>0.21</td>
<td>0.21</td>
<td>0.13</td>
<td>0.16</td>
<td>0.35</td>
<td>0.29</td>
</tr>
<tr>
<td>DESIQUE [23]</td>
<td>0.23</td>
<td>0.33</td>
<td>0.20</td>
<td>0.40</td>
<td>0.22</td>
<td>0.29</td>
<td>0.19</td>
<td>0.30</td>
<td>0.27</td>
</tr>
<tr>
<td>BLINDS-II [21]</td>
<td>0.24</td>
<td>0.30</td>
<td>0.29</td>
<td>0.24</td>
<td>0.21</td>
<td>0.24</td>
<td>0.21</td>
<td>0.43</td>
<td>0.27</td>
</tr>
<tr>
<td>NIQE [25]</td>
<td>0.31</td>
<td>0.29</td>
<td>0.33</td>
<td>0.28</td>
<td>0.22</td>
<td>0.25</td>
<td>0.19</td>
<td>0.23</td>
<td>0.26</td>
</tr>
<tr>
<td>BRISQUE [20]</td>
<td>0.24</td>
<td>0.21</td>
<td>0.20</td>
<td>0.50</td>
<td>0.26</td>
<td>0.26</td>
<td>0.09</td>
<td>0.21</td>
<td>0.25</td>
</tr>
<tr>
<td>BIQI [22]</td>
<td>0.37</td>
<td>0.18</td>
<td>0.14</td>
<td>0.28</td>
<td>0.15</td>
<td>0.13</td>
<td>0.25</td>
<td>0.27</td>
<td>0.22</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

This paper explored two alternative performance metrics for IQA algorithms that are possible to calculate from the subjective evaluations image quality databases. These metrics offer a different view on the performance of IQA’s and could possibly be used in conjunction with the already proven correlation metrics LCC and SROCC. However, to be able to calculate these alternative performance metrics, the full raw data is necessary to be available. As the CID2013 database consists of multiple simultaneous distortions that can mask or expose one another, the subjective evaluation task of image quality becomes more preferential. This increases the variation in the subjective data. If the consensus between the observers in sorting the samples in their order of quality is reduced, the target performance level should be more relaxed as well. One of the suggested methods is to calculate a Subjective RMSE for a hypothetical random average observer. If an IQA algorithms achieves a RMSE value below this limit, it can be claimed it predicts the overall image quality of that database as well as one randomly observer would.

In addition, for these new metrics to be useful tools for performance evaluation, it is important that some common standards for their use are laid down. More research is still needed to find the best way to estimate the Subjective RMSE efficiently and reliably. Also researchers should need to concur in what method is used for calculating the statistical difference in subjective data. In this paper we have opted for MMA with HCS and Bonferroni corrected Post-Hoc testing that could be such a method for statistical testing with subjective preference data.

ACKNOWLEDGEMENTS

This project was supported by funding from Academy of Finland (project no. 267061).
REFERENCES


