https://doi.org/10.2352/CIC.2023.31.1.31

A comprehensive image quality dataset to compare no-reference and with- reference image quality assessment

Nanlin Xu[1], Yuechen Zhu[1], Ming Ronnier Luo[1], and Xinchao Qu[2]*
[1]State Key Laboratory of Extreme Photonics and Intrumentaion, Zhejiang University, Hangzhou, China
[2] Dajiang Innovation Technology Co., Ltd

Abstract

With the prevalence of digital devices, images are now more accessible. A method to judge the image quality of a picture and corresponding datasets are highly desired. However, previous works focused solely on total image quality, without consider image quality separately in terms of color and spatial aspects. The present study aims to fill this gap by evaluating total, color, and spatial image quality together.

The whole experiment was divided into two parts: no-reference (NR) experiment and with-reference (WR) experiment. In the NR part, 30 participants assessed total image quality (tIQ), color image quality (cIQ) and spatial image quality (sIQ) as well as their corresponding weights for color and spatial impact. In the WR part, 30 participants were asked to evaluate the difference in color and total image quality between the original image and rendered image.

Weighted IQ, obtained through linear weighting using ratio, cIQ, and sIQ, demonstrated a high correlation coefficient (0.96) with total IQ. This implies that color and spatial features of image quality can be treated as separate entities.

A no-reference image quality model was proposed to predict IQs whose accuracy of prediction obtained a correlation coefficient value of 0.80.

1. Introduction

Due to the increasing availability of digital products, images are becoming more prevalent in daily life. Image quality assessment task arouse much interest across diverse industries. IQ assessment can be completed either subjectively with the input of observers or objectively using IQ assessment models. IQ assessment model have brought great concern on account of its convenience and high accuracy.

Image quality assessment model (IQM) can be divided into 3 types: full-reference (FR), reduced-reference (RR) and no-reference (NR) methods. As the name implies, the FR methods predict the image quality with a full set of features from a test image and the same set of features from the compared image; Ashirbani Saha [1] predicts the perceptual quality of the test image in terms of an objective score by comparing a test/distortion image and a reference image.

Reduced reference methods predict the image quality with a reduced set of features from a test image and the same set of features from the compared image; Abdul Rehman [2] predicts IQ using partial information – structure similarity (SSIM) index; in TMQI [3], the structural fidelity uses a reference image for comparison while the naturalness is calculated without a reference image.

No-reference methods predict image quality only with features of the test image. Among these methods, NR methods are the most widely used and urgently needed by the industry due to the difficulty of obtaining reference images in practical applications. Bianco trained a neural network model called DeepBIQ using a migration learning approach [4]. Model was pre-trained on the public dataset Image-Net and then migrated to individual database for parameter tuning.

In addition, image evaluation can be divided into two categories: subjective and objective evaluation. Subjective evaluation refers to the process of conducting a psychophysical experiment in which final evaluation scores are obtained by combining statistical methods to synthesize the image quality scores of all observers. The goal of objective evaluation is to fit the subjective evaluation data by analyzing the characteristics of the image, which simulates subjective evaluation.

By analyzing the difference between subjective IQ and the predicted result of objective IQM, features of the image can be judged to be effective or not. Therefore, establishing subjective image quality evaluation datasets is significant in order to build an effective objective IQM.

At present, several image datasets have been developed and applied extensively in image quality modeling, such as the LIVE database of the University of Texas [5], the TID2008 and TID2013 datasets of the Finnish University of Tampere, the CSIQ dataset of the Oklahoma State University [6, 7], the CIDIQ dataset of the Norwegian University of Science and Technology [8] and the KADID dataset [9] of the University of Konstanz in Germany. Recently, a dataset was developed at the Color and Engineering Lab of Zhejiang University consisting of 47 original images and whole 1600 images rendered by over 6 color domain modifications [10].

Most of the image quality datasets mentioned above focus on color and spatial rendering datasets. Nevertheless, the results of these dataset are often linked to overall image quality rather than considered with color and spatial aspects.

In this paper, we establish a dataset that considers the subdivision of image quality and obtains weights for color and spatial aspects for each image. In addition, we proposed a no-reference image quality model consisting of 8 attributes that extract essential information to characterize an image.

2. Experiments

Preparation of experimental images

All images were collected in two ways. First, 9 and 12 original images were picked out from CIDIQ and KADID datasets respectively [8, 9]. Images from CIDIQ were rendered by 5 methods including 2 color rendering and 3 spatial rendering while images from KADID were rendered by 5 color rendering and 3 spatial rendering. What's more, the heavily distorted images were excluded, resulting in the different levels of rendering details being listed in table 1. A total of 208 and 377 images were collected from these two datasets, respectively.

Tabel 1. Details of image preparation, including sources, rendering methods, rendering levels and the number of images.

Datasets	CIDIQ	KADID	NEW
Selected image	9	12	11
Color rendering	SGCK GM (5 levels) Min DE GM (5 levels)	Color saturation (2-3 levels) Brighten (3-4 levels) Darken (3-5 levels) Mean shift (5 levels) Contrast (5 levels)	Vividness (5 levels) Depth (5 levels) Clarity (5 levels) Chroma (5 levels)
Spatial rendering	Gaussian Blur (4 levels) JPEG (4 levels) JPEG2000 (4 levels)	Gaussian blur (4 levels) JPEG2000 (3 levels) JPEG (3 levels) Sharpness (2 levels)	
Final images	208	377	231
Total images	816		

Table 1 lists details of image preparation, including sources, rendering methods, rendering levels and the number of images.

In addition, 11 high-quality images were selected and then rendered using 4 color attribute rendering methods with 5 levels of intensity, including vividness, depth, clarity and chroma of CIELAB color space as given in Eq. (1-4). This method produced a total of 231 images.

$$vividness = \sqrt{L^2 + a^2 + b^2} \quad (1)$$

$$depth = \sqrt{(L-100)^2 + a^2 + b^2} \quad (2)$$

$$clarity = \sqrt{(L-50)^2 + a^2 + b^2} \quad (3)$$

$$chroma = \sqrt{a^2 + b^2} \quad (4)$$

where L, a and b are the CIELAB attributes.

Finally, 816 images were collected for using in subsequent experiments. Figure 1 shows all original images prepared.

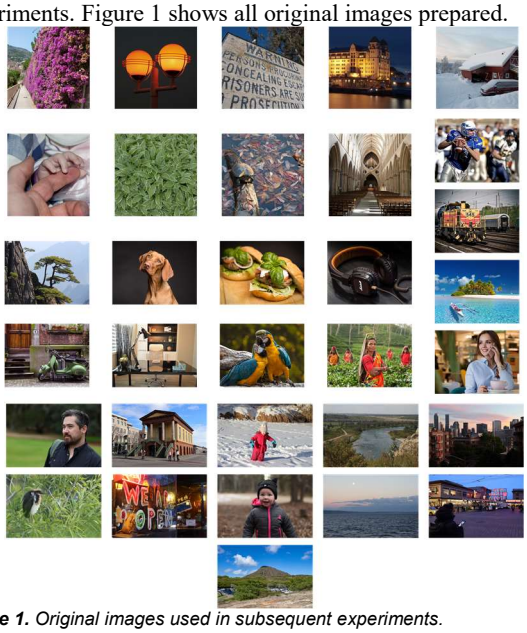

Figure 1. *Original images used in subsequent experiments.*

Experiment setup

Whole experiment includes no-reference (NR) experiment and with-reference (WR) experiment. Both experiments were conducted using an Eizo display with a resolution of 2560 × 1440 pixels. The correlated color temperature (CCT) of the display peak white was set to 6500 K with a luminance of 300 cd/m2. Gama coefficient was set to 2.2 The Gain-Offset-Gamma (GOG) was used to characterize the display. The predictive accuracy of the GOG model was an average of 0.60 CIELAB units over 24 Macbeth Color Checker colors. All the measurements were conducted using a Kosnica Minolta CS2000A tele-spectroradiometer in black surroundings. All experimental images were transformed to the Eizo gamut using GOG model to achieve better display effect. The display was placed at a fixed distance of 60 centimeters from observers.

Prior to formal experiments, observers will undergo some training on image quality including some concepts and shown images. Observers were told the total image quality of image consists of color and spatial parts. Color includes the lightness, hue and saturation aspects, while the spatial part indicates the degree of detail retention in one image.

No-reference experiment

The NR experiment was divided into 2 sessions. In the 1st session, observers were asked to evaluate total image quality (tIQ), as well as the weights of its color impact and spatial impact (summing to 1). In the 2ed session, observers were asked to evaluate color image quality (cIQ) and spatial image quality (sIQ). A six categorical judgment method was used here for the evaluation of IQs. Observers were asked to rate IQ ratings according to the displayed image's perceptual quality using a keyboard whose values ranged from −3, very poor; −2, poor; −1, little poor; 1, little good; 2, good; to 3, excellent. Here, −3 corresponds to the lowest perceptual IQ (very poor), whereas +3 corresponds to the highest (excellent) quality.

To begin with, observers adapted to the experimental conditions for 60 seconds and then provided the evaluation of the image using the experiment software. Figure 2 shows the interface

used to perform the experiment. In the experiment, 10 percent of all images were randomly picked out as a repeated group for observers' performance evaluation. The whole NR experiment took about 180 mins to complete for each observer mixed, with interlaced rest periods. 30 normal observers who performed the Ishihara color vision test between 19 and 29 years of age (mean = 22.2, std = 2.3) participated in the no-reference experiment including 14 males and 16 females. At the end of the NR experiment, a total of 133,800 evaluations were accumulated.

With-reference experiment

In the WR experiment, two images were presented on the window as shown in figure 2, in which one is the reference image (original image) and the other is the corresponding rendered image with randomized positions. Observers were asked to evaluate the color difference and total image difference between the two images. A scale of 5 rating levels was used, from 1, no difference; 2, JND (just noticeable difference); 3, small difference; 4, medium difference; 5, large difference.

Similar to the NR experiment, observers were also required to adapt for 60 seconds at the beginning, and 10 percent of all images were randomly chosen as a repeated group. 30 normal color vision observers who performed the Ishihara color vision test between 19 and 29 years of age (mean = 22.7, std = 2.4) participated in the with-reference experiment including 15 males and 15 females. A total of 53,520 evaluation data were accumulated in the end.

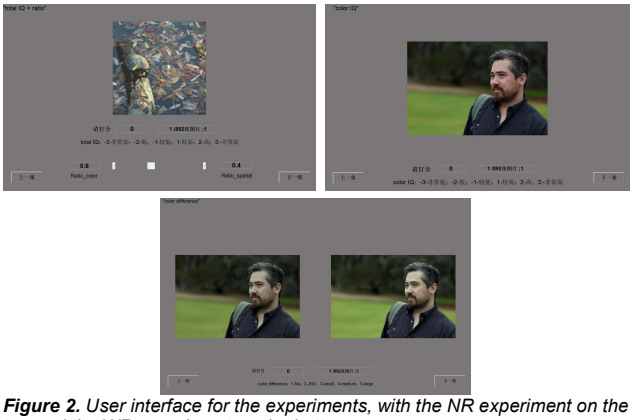

Figure 2. *User interface for the experiments, with the NR experiment on the top and the WR experiment on the bottom.*

Experimental data analysis

Raw data processing

All obtained IQ (tIQ, cIQ, sIQ) data ranging from -3 to +3 were first converted to 1-6. Then the ratings of each image were average to obtain final visual score. The average scores (1-6) were then normalized to a 0-1 scale in which 0 corresponds to the lowest perceptual quality and 1 corresponds to the highest perceptual quality. The data were then used in the training and testing of the image quality model.

Observer variability

Intra- and inter- R were calculated to represent the data consistency of observers using raw data ass

$$\mathrm{R}(\mathrm{x},\mathrm{y}) = \frac{\mathrm{cov}(\mathrm{x},\mathrm{y})}{\mathrm{std}(\mathrm{x})\cdot\mathrm{std}(\mathrm{y})} \quad (5)$$

Where x and y are the refence and batch data respectively and

$$\mathrm{std}(\mathrm{x}) = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x_i}-\bar{\mathrm{x}})^2} \quad (6)$$

$$\mathrm{cov}(\mathrm{x},\mathrm{y}) = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x_i}-\bar{\mathrm{x}})\cdot(\mathrm{y_i}-\bar{\mathrm{y}})}{(\mathrm{n}-1)} \quad (7)$$

n is the number of images.

For a perfect agreement, R should be 1. The mean values of intra-R and inter-R are 0.59 and 0.60 in NR experiment. In addition, the mean values of intra-R and inter-R are 0.62 and 0.73 in WR experiment.

Comparison with original data

As some images used in experiments are from existing datasets (CIDIQ and KADID), Pearson coefficient between present visual score and old data can be calculated to validate the `stability of the experiment results. In the NR experiment, R values are 0.90 and 0.82 for CIDIQ and KADID respectively, while R values are 0.84 and 0.86 for CIDIQ and KADID respectively in the WR experiment. Figure 3 plots the scatters between visual score and old data. Given the number of datasets, the weighted average R values are 0.848 and 0.853 for NR and WR respectively. These high R values indicate the strong stability of whole experimental results.

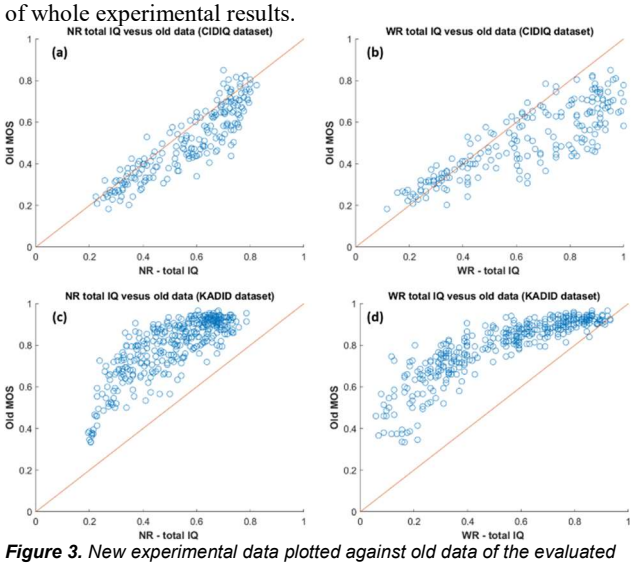

Figure 3. *New experimental data plotted against old data of the evaluated datasets for a) NR on CIDIQ; b) WR on CIDIQ; c) NR on KADID and d) WR on KADID dataset.*

NR and WR comparison

Since the NR and WR experiments were simultaneously conducted, another method to mutually validate data is to calculate total IQ and color IQ difference score using NR experimental score by Eq. (8-9) for comparing results of WR experiment.

$$\Delta tIQ_{\mathrm{NR}} = tIQ - tIQ_{\mathrm{refrence}} \quad (8)$$

$$\Delta cIQ_{\mathrm{NR}} = cIQ - cIQ_{\mathrm{refrence}} \quad (9)$$

Figure 4 plots the scatter between two ΔxIQs. Data shift is caused by different value ranges of NR and WR experiments. Then R value between ΔtIQ_{NR} and ΔtIQ_{wR} is 0.84 and R for ΔcIQ is 0.80, which proves the stability of the whole experiment.

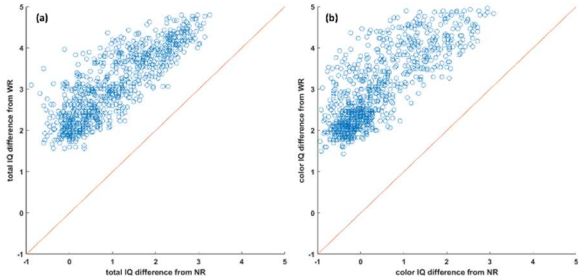

Figure 4. *Scatter plots between ΔxIQ_{NR} and ΔxIQ_{WR}. a) ΔtIQ, b) ΔcIQ. The abbreviations "NR" and "WR" indicate no-reference and with-reference experiments respectively.*

Data analysis with ratio in NR experiment

Images can be divided into three parts in terms of ratio from the NR experiment. 1) ratio-color >= 0.6: color-dominated image; 2) ratio-color <= 0.4: spatial-dominated image;3) 0.4 < ratio-color < 0.6: normal image. Figure 5 shows the histogram of ratio-color distribution, which demonstrated the number of color-dominated images were more than others in the dataset.

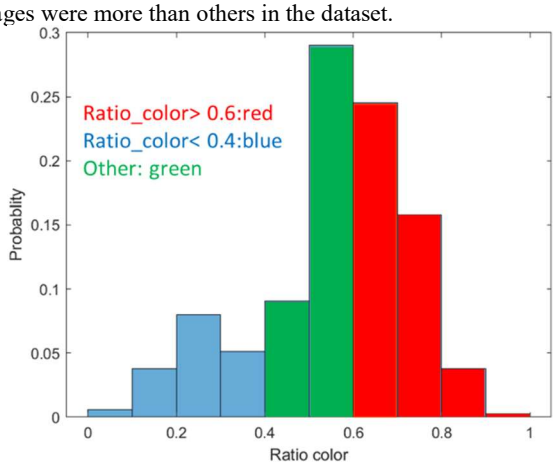

Figure 5. *Histogram of distribution of ratio color. The red spot indicates color-dominated image, the blue spot indicates spatial-dominated image and the green spot indicates normal image.*

Weighted IQ can be obtained through linear weighting with ratios, cIQ and sIQ as Eq. (10).

$$wIQ = cIQ * ratio_{\text{color}} + sIQ * ratio_{\text{spatial}} \quad (10)$$

Table 2 lists R between IQs and figure 6 plots the scatter between tIQ and other IQs. From the table and figure 6 (c), the correlation between cIQ and sIQ is close to 0 at 0.092, which indicated that it is feasible to consider image quality in terms of color and spatial quality. Moreover, the high correlation between the wIQ calculated by combining the ratio and tIQ further confirms the accuracy of the experimental design.

Table 2. Correlation coefficients between the IQs.

R	tIQ	cIQ	sIQ	wIQ
tIQ	1.0	0.720	0.703	0.960
cIQ	0.720	1.0	0.092	0.717
sIQ	0.703	0.092	1.0	0.723
wIQ	0.960	0.717	0.723	1.0

No reference image quality models

To improve the prediction accuracy of the model, the proposed IQM consists solely of 8 attributes listed in Eq. (11) which extracts the key information including chroma (C) of CIELAB color space, chroma contrast 5x5 (CC5), sharpness contrast 9x9 (CS9), chroma ratio (Cr), clarity ratio (Clr), global contrast (GC), local contrast (LC) and sharpness (S) attributes.

Chroma contrast (CC5) was employed to model the contrast changes in the images. The CC5 value was calculated by taking the average of standard deviations of local image regions of sizes 5 × 5, around all pixels of chroma channels. For sharpness (CS9), edge detection was first applied on the lightness channel using the Sobel operator. Then the standard deviations of local image regions were calculated around the detected edge pixels and averaged to obtain a single value for an entire image.

The chroma ratio and clarity ratio values were relative attributes based on the display gamut.

$$IQ = f(C,\ CC5, CS9,\ Cr, Clr, GC, LC, S) \quad (11)$$

$$Clr = \frac{1}{N}\sum \frac{Cl_{\text{t}}}{Cl_{\text{g}}} \quad (12)$$

$$Cr = \frac{1}{N}\sum \frac{C_{\text{t}}}{C_{\text{g}}} \quad (13)$$

For certain attributes, the computational procedure is to increase this attribute of the target pixel until it reaches the gamut boundary [10]. Then the end point on the gamut could be regarded as a reference point showed in figure 7. The ratio of the target pixel and the reference point attribute is considered the relative attribute value. The ratio value of an image for certain attributes

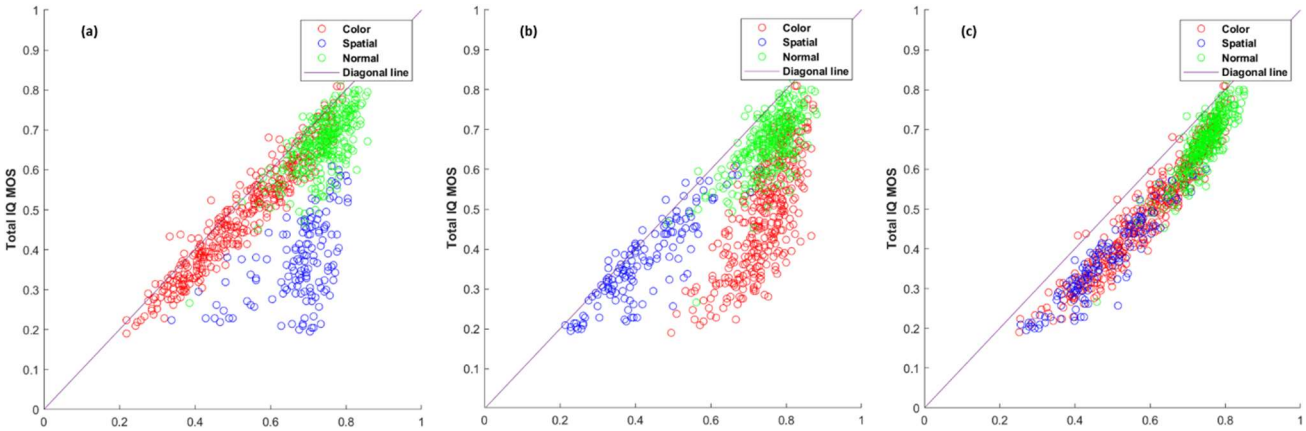

Figure 6. *Scatter plot between tIQ and a) cIQ, b) sIQ, c) wIQ. The red spot indicates color-dominated image, the blue spot indicates spatial-dominated image and the green spot indicates normal image.*

was obtained by averaging the ratio values of all pixels as given by Eq. (12-13)

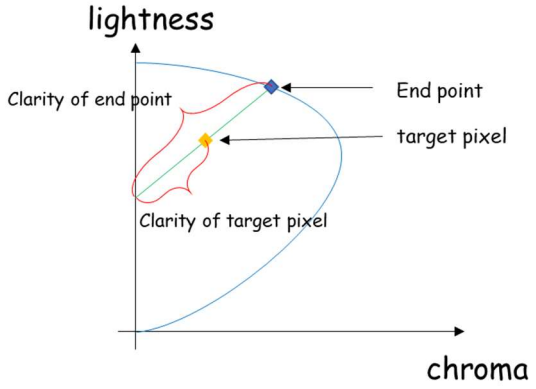

Figure 7. *Calculation procedure of clarity ratio (Clr) value. The blue curve is the gamut boundary in the hue page of the target pixel.*

In addition to the above-mentioned five attributes, three more image attributes were chosen to build the IQ model, including GC, LC and S. These attributes were calculated in the following equations:

$$GC = (\frac{10}{N}\sum_{i=1}^{0.1*N} L_i - \frac{10}{N}\sum_{i=0.9*N}^{N} L_i)/100 \quad (14)$$

$$LC = \frac{1}{N}\sum_{1}^{N} \Delta L \quad (15)$$

$$\Delta L = \frac{1}{24}\sum_{i=1}^{5}\sum_{j=1}^{5}(|\bar{L} - L_{ij}|) \quad (16)$$

Where $\bar{L}^*$ is the lightness of target pixel and L_{ij}^* is the lightness of its 5 X 5 neighboring pixels. Sharpness (S) was calculated as

$$S = \frac{1}{N}\sum_{1}^{N} \Delta E_{ab} \quad (16)$$

Where,

$$\Delta E_{ab} = \frac{1}{24}\sum_{i=1}^{5}\sum_{j=1}^{5} \Delta E_{ab,ij} \quad (17)$$

To evaluate model's performance, attributes values were used to train a support vector machine (SVM) regression.

Results and discussion

In this session, two train-test methods were used to evaluate the performance of IQM. The first method is as follows: For each IQ data in the NR experiment, all images was divided into a train set and a test set by an 80:20 ratio with a random order. In order to ensure the stability of evaluation, this process was repeated 1,000 times. The median value of the correlation coefficient was chosen as the key result of this IQM.

Table 3 lists these results of method 1. It can be found that total IQM achieved an R value of 0.796 similar to 0.773 value of weighted IQM which indicates that the proposed IQM has a good performance in predicting IQ of daily life in a commonly accepted sense. What's more, cIQM achieved the highest R value of 0.84 which implies the proposed IQM is capable of representing the color information of images well.

Table 3: Evaluation performance of IQMs with 8020 training test method

R	tIQM	cIQM	sIQM	wIQM
Min	0.681	0.737	0.502	0.651
Median	0.796	0.841	0.685	0.773
Max	0.882	0.903	0.825	0.864

Another method is that on IQM was trained on one IQ data with all images and then was tested on other 3 IQs using all images to estimate the robustness and generalizability of the model. Table 4 lists R values between the predicted value and IQs.

Table 4. Evaluation performance of IQMs when train on one IQ of all images and test on other IQs of all images.

R	Test on tIQ	Test on cIQ	Test on sIQ	Test on wIQ
Train on tIQ	0.959	0.737	0.627	0.929
Train on cIQ	0.695	0.969	0.063	0.678
Train on sIQ	0.656	0.194	0.847	0.670
Train on wIQ	0.936	0.748	0.639	0.952

The results imply the total IQ model achieved a higher R value of 0.737 when testing on cIQ than 0.627 on sIQ. Furthermore, the performance of wIQM in predicting tIQ obtain R value of 0.936, indicating the feasibility to consider image quality in terms of color and spatial aspects separately.

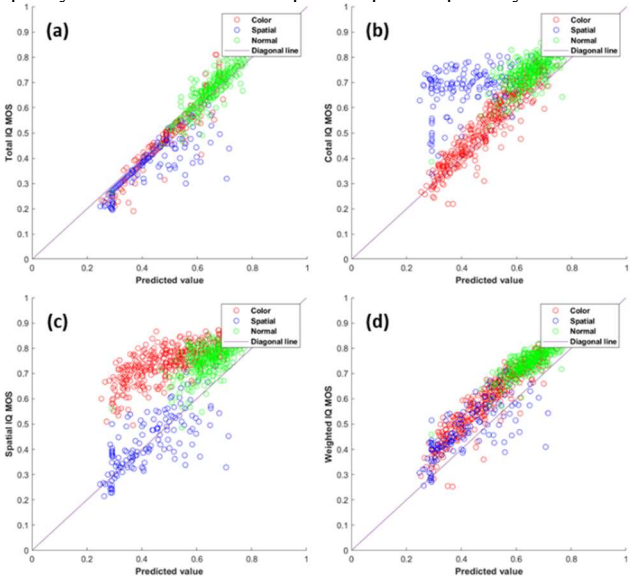

Figure 8. *Scatter plots between the predicted scores with the tIQ model and the different MOS scores. The red spot indicates color-dominated image and the blue spot indicates color-dominated image.*

Figure 8 plots the scatter between the predicted values and IQs. There are more blue spots, not red spots, above diagonal line in figure 8 (b) suggesting that spatial-dominated images of high color IQ with spatial distortion receive a poor total IQ.

Conclusion

This study aimed to establish a comprehensive dataset that separately considers the aspects of image quality in terms of color and spatial features and to propose a no reference image quality model. The experimental dataset contained 816 images, with some selected from existing datasets while the remaining generated via color space. In conclusion, the weighted average Pearson correlation coefficients are 0.848 and 0.853 for NR and WR respectively, between present visual data and old data. These R values indicate strong stability of results obtained via the entire

experimental process. Furthermore, the Pearson correlation coefficient achieved a high value of 0.96 between total IQ and weighted IQ obtained by linear weighting with ratio, which demonstrates the experimental design methodology's accuracy and present an alternative approach to predicting image quality. In addition, a proposed total IQM with an 80:20 train-test method achieved a good R value of 0.80. These results suggest that the model could benefit from further improvement to achieve more precise results.

Acknowledgement

This research was funded by Dajiang Innovation Technology Co., Ltd.

References

[1] Larson, Eric C. and Damon M. Chandler. "Most apparent distortion: full-reference image quality assessment and the role of strategy." J. Electronic Imaging 19 (2010): 011006.

[2] Rehman A, Zhou R. Reduced-reference image quality assessment by structural similarity estimation. IEEE Trans Image Process. 2012 Aug;21(8):3378-89. doi: 10.1109/TIP.2012.2197011. Epub 2012 May 1. PMID: 22562759.

[3] H. Yeganeh, and Z. Wang, "Objective assessment of tone mapping algorithms," in 17th IEEE International Conference on Image Processing (ICIP 2010), pp. 2477–2480.

[4] Bianco S, Celona L, Napoletano P,et al. On the use of deep learning for blind image quality assessment. SIViP 12, 355–362 (2018).

[5] Sheikh H. Live image quality assessment database release 2[J/OL], 2005[2022-04-04].

[6] Ponomarenko N, Lukin V, Zelensky A, et.al. TID2008 – A Database for Evaluation of Full- Reference Visual Quality Assessment Metrics[J]. 14

[7] Ponomarenko N, Jin L, Ieremeiev O, et.al. Image database TID2013: Peculiarities, results and perspectives[J/OL]. Signal Processing: Image Communication, 2015, 30: 57-77.

[8] Liu X, Pedersen M, Hardeberg J Y. CID:IQ – A New Image Quality Database[J]. 2010, 17(6): 503-510.

[9] Lin H, Hosu V, Saupe D. KADID-10k: A Large-scale Artificially Distorted IQA Database[C/OL]//2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX). Berlin, Germany: IEEE, 2019: 1-3[2021-08-29].

[10] Khan MU, Luo MR, Tian D. No-reference image quality metrics for color domain modified images. J Opt Soc Am A Opt Image Sci Vis. 2022 Jun 1;39(6):B65-B77. doi: 10.1364/JOSAA.450595. PMID: 36215544.

Author biography

Nanlin Xu received her BS in Optical Engineering from the Shenzhen University and he is now a master student supervised by Professor Ming Ronnier Luo at Zhejiang University. His research work is on image quality assessment model.