MP2020: Visual quality assessment database for macro photography images

Qingbing Sang1 | Yujie Cao1 | Lixiong Liu2 | Cong Hu1 | Xiaojun Wu1

1 School of Artificial Intelligence and Computer Science, Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi 214122, China
2 School of Computer Science and Technology, Beijing 100081, China

Correspondence
Qingbing Sang, School of IoT Engineering, Jiangnan University, Wuxi 214122, China.
Email: qingbings@jiangnan.edu.cn

Abstract
With the development of mobile phone camera technology, mobile phones can take a large number of macro photography images that previously could only be taken by professional cameras. Therefore, it is of great significance to study the quality of macro photography images. For this reason, a macro photography image visual quality evaluation database is established and it is named as MP2020. The database contains 100 reference images and 800 distorted images of four distortion types, including 200 distorted images of JPEG 2000, 200 distorted images of JPEG, 200 distorted images of white noise, and 200 distorted images of Gaussian blur. The DMOS values in the database were calculated from 48000 data which are provided by 60 subjects. Ten classical image quality assessment algorithms were tested on the MP2020 database. The experimental results show that the existing image quality assessment algorithms, which are widely used, are not applicable to the macro photography images. Therefore, MP2020 would contribute to the improvement of existing algorithms and the development of new algorithms. MP2020 has been uploaded to GitHub for download.

1 INTRODUCTION

Macro photography refers to the use of the optical capabilities of lenses to capture images that are larger or smaller than the actual object. Due to this special shooting technology, the main part of a macro image is clear, and the background part is blurred. Such an image can highlight the theme and increase the aesthetic feeling, which is deeply loved by everyone. In the past, macro photography images could only be taken by technicians with professional cameras. Currently, with the development of mobile camera technology, mobile phones can easily take macro photography images, resulting in an increasing number of macro photography images. However, this type of image is more prone to a variety of distortions in the process of acquisition, processing, transmission and storage than that of ordinary images because of their background blurring characteristics, resulting in image quality degradation and information loss. These distortions seriously affect the practical application of the images and the visual quality, so research on the quality of macro photography images is particularly important.

Image quality assessment (IQA) is a process of analysing image quality. The evaluation methods are divided into subjective image quality assessment and objective image quality assessment [1]. In the current universal image processing applications, human beings are the ultimate recipients of image processing applications, that is, to evaluate images with subjective human will, so subjective image quality assessment is the most effective and reliable method. Corresponding to the subjective evaluation is the objective image quality assessment, whose essence is to simulate the subjective perception of human eyes and give the quality evaluation results with the help of a certain mathematical model. According to the different dependencies of the evaluation system on the reference image, objective image quality assessment can be divided into full-reference (FR) image quality assessment, reduced-reference (RR) image quality assessment and non-reference (NR) image quality assessment [2].

In the past 40 years, image quality assessment (IQA) has developed rapidly. The full reference image quality evaluation algorithm is one of the earliest and most mature algorithms. In the early reference algorithms, the image quality was evaluated by directly calculating the corresponding pixel error between the...
TABLE 1  The number of images per degradation category in each database

|                | LIVE | TID2008 | TID2013 | CSIQ |
|----------------|------|---------|---------|------|
| Reference image| 29   | 25      | 25      | 30   |
| Distorted image| 779  | 1700    | 3000    | 866  |
| MOS/DMOS       | [0, 100] | [0, 9] | [0, 9] | [0, 9] |

FIGURE 1  Example of macro camera and natural images

Figure 1(a) shows a macro camera image, and (b) shows a natural image. As we can see from the figure, this is a special type of image, which is different from the ordinary image in uniform and clear (or blur), with a clear main body and blurred background. Therefore, this kind of image may be a distorted image for most of the existing quality evaluation algorithms. Figure 1(c) and (d) show the differences between the macro photography image and the natural image, using the histogram of oriented gradient (HOG). Currently, some of the existing databases cover this type of image, such as parrots.bmp and monarch.bmp in the live database, but there is no special database for this kind of image. Now, with the mobile phone also being able to take micro photography, the number of this type of image is increasing day by day, so it is very meaningful to study this new image quality. Based on this, we set up an image database dedicated to studying the quality of micro photography images, named MP2020.

2  ESTABLISHMENT OF MP2020

2.1  Reference image and distorted image

The MP2020 database includes 100 reference images and 800 distorted images. The 100 reference images, which are carefully selected from a large number of macro photography images. The selection of the reference image of the database is based on the standard of pixel 255 x 255. If the image is large, the typical part is selected by manual cutting. At the beginning of the experiment, nearly 200 reference images were screened and deleted based on different kinds, and 100 reference images were selected.

This database includes 25 flowers, 20 animals, 16 human figures, 14 insects, 10 vegetation, 9 delicacies, 5 buildings and 1 other type. Some of the images are shown in Figure 2. This selection guarantees the comprehensiveness of the database, makes the experimental pictures more diverse, and reduces visual aesthetic fatigue during artificial scoring.

A total of 800 distorted images are formed by four different types of distortion based on 100 reference images, and two different levels are divided under each type of distortion. The four types of distorted images are processed by MATLAB tools as follows:

1. JPEG compression (200 pieces): Forward discrete cosine transform (FDCT) is used to transform the image in the spatial domain into that in the frequency domain. The weighting function is used to quantize the DCT and then the Huffman variable word length encoder is used to encode the quantization coefficient. In the experiment, the JPEG compression algorithm is used to compress the reference image. Its quality factors are 5 and 25, respectively.

2. JPEG 2000 compression (JP2K, 200 pieces): Discrete wavelet transform (DWT) and the JPEG 2000 compression standard based on wavelet image compression are used to compress the reference image. Its quality factors are 5 and 25, respectively.
3. White noise (WN, 200 pieces): Function `Randn()` is used to generate the normal distribution function. Then, the `WGN(M, N, P)` function is called to generate the Gaussian white noise and generate a matrix of Gaussian white noise of M rows and N columns. P specifies the output intensity in DBW and divides the image into three channels, R, G and B. Add the standard Gaussian white noise of 0.013–0.15 to each channel and scale the image to between 0 and 255.

4. Gaussian blur (GBLUR, 200 pieces): The `Fspecisl()` function is used to obtain the point spread function and then calls the `Imfilter()` function to realize defocusing blur. In this experiment, the standard two-dimensional Gaussian smooth filter with a standard deviation of 0.9–6 is used for processing.

Figure 3 shows a typical picture and its corresponding distorted picture in the MP2020 database, considering factors such as light, colour and whether it has the typical characteristics of macro photography images.

2.2 | Subjective score

The subjective score of the experiment was performed in the normal indoor lighting environment. All images were displayed on an LCD with a screen resolution of 1680 × 1050 pixels. One subject viewed two images displayed on the screen at a distance of approximately 25 cm from the screen; the left image was a high-quality reference image, and the right image was a distorted image. In this experiment, a double stimulus impulse scale (DSIS) was used for subjective scoring; that is, the subjects first looked at the reference image, then at the distorted image, and then compared the two images. By comparing the two images, the subjects rated the distorted images and gave a score between 0 and 100. The lower the score was, the higher the image quality damage.

The number of subjects in this experiment was 60, and the laboratory personnel were scattered to find the subjects. The number of students in this major is 27, the number of non-major students is 14, the number of graduates is 19, the ratio of men and women is 3:2, and the age range is from 21 to 33. Most of the subjects wear glasses, and the personnel are mixed and cover professional and non-professional, ensuring comprehensive experimental data.

To facilitate the scoring of subjects, independent platform software for MP2020 database scoring named subjective quality assessment (SQA) was developed to collect the scoring information of subjects. SQA was developed by using the C# tool in Visual Studio 2017, which mainly includes two functions: training and subjective scoring. This scoring software will be described in detail below.
The software has an independent installation package, which can be installed locally. After the installation, double-click the shortcut generated by the desktop to run. The software installation package can be downloaded from GitHub for self-use. The main part of the main interface of the scoring software is two pictures distributed on the left and right sides of the screen. The image on the left is a reference image, and the other image on the right is an image after artificial distortion processing. The specific interface is shown in Figure 4.

Before the subjects score, they need to click the “Train” button to train to understand the approximate score curvature. The training interface is shown in the Figure 5 above.

The scores printed by the subjects will be saved in the format of .txt. The subjects can click the “Save” button to select the save path and file name.

2.3 Data processing

Different evaluation indexes often have different dimensions and dimension units, which will affect the results of data analysis. In order to eliminate the dimensional impact between indexes, data standardization is needed to solve the comparability between data indexes. After data standardization, all indexes are in the same order of magnitude, which is suitable for comprehensive comparative evaluation. Here are two common normalization methods:

(I) Min-max normalization

Also known as deviation standardization, it is a linear transformation of the original data to map the result value between [0–1]. The conversion function is:

\[ x = \frac{x - \text{min}}{\text{max} - \text{min}} \]  

where max is the maximum value of sample data and min is the minimum value of sample data.

One drawback of this method is that when new data is added, it may lead to changes in max and min, which need to be redefined.

(II) Z-score

This method standardizes the mean and standard deviation of the original data. The processed data conforms to the standard normal distribution, i.e. the mean value is 0, the standard deviation is 1, and the conversion function is:

\[ x = \frac{x - \mu}{\sigma} \]  

where \( \mu \) is the mean value of all sample data and \( \sigma \) is the standard deviation of all sample data.

The data analysis method used in this paper is the min-max normalization. Normalization is to make the features of different dimensions have certain comparability in numerical value.

Generally speaking, the histogram of all subjects’ scores tends to be normal Gaussian distribution after normalization, as shown in Figure 6.

2.4 Statistical analysis of subjective scores

After several months of experimental data collection, all the experimental data are analysed and screened. The data of this experiment are all between 0 and 100. Through the Grubbs
criterion [26], the data with far deviation in the dataset are eliminated to reduce the impact of outliers on the overall data. That is, 60 data points of each picture are divided into 4 groups, each 15 data points is a group, each 15 data points is screened for outliers, and the subjects to which the outliers belong are collected. If there are more than 20 abnormal points, the subject will be recorded as an unreliable subject and will be disqualified.

In the experiment, a boxplot is used to analyse the scoring of distorted pictures. Boxplot is a statistical chart used to display a group of data dispersion data. It is composed of a “box” and “whisker”. There is a straight line in the box that represents the median of the sample. The upper and lower boundaries of the box show 75% and 25% values, respectively. The two “musts” are the maximum and minimum values of the data. Outliers are usually plotted separately and indicated by “+”.

We use the boxplot tool to calculate the average artificial score of distorted pictures. Boxplot is a statistical chart used to display a group of data dispersion data. It is composed of a “box” and “whisker”. There is a straight line in the box that represents the median of the sample. The upper and lower boundaries of the box show 75% and 25% values, respectively. The two “musts” are the maximum and minimum values of the data. Outliers are usually plotted separately and indicated by “+”.

We use the boxplot tool to calculate the average artificial score of distorted pictures under the four distortion types of JP2K, JPEG, WN and GBLUR, including the minimum, the first quartile, the median, the third quartile and the maximum. Each centre blue rectangle spans from the first quartile to the third quartile, as shown in Figure 7. The line segment inside the rectangle (red) represents the median, and the horizontal line segments above and below the rectangle represent the maximum and minimum values of the data. Outliers are usually plotted separately and indicated by “+”.

In Figure 7, we can see that the fraction distribution is low under JP2K distortion, and the fraction span is large under JPEG distortion. The fraction under WN distortion is slightly higher than that under other distortion types, and the fraction under GBLUR distortion is compact, generally low, and has many outliers. On the whole, JP2K, JPEG and WN are better than GBLUR, which may be caused by the macro photography image itself being a kind of image with clear foreground and fuzzy background. Later, we will manually grade and count the GBLUR-type distorted images.

### 3 Analysis of the Experimental Results of Existing Algorithms

#### 3.1 Image quality assessment index

Two commonly used performance metrics were employed to evaluate the competing IQA methods. The first is the Spearman rank-order correlation coefficient (SROCC), which can assess the prediction monotonicity of an IQA method. This metric operates on the ranked data points and ignores the relative distances between data points. The second metric is the Pearson linear correlation coefficient (PLCC) between the MOS and the objective scores after non-linear regression. For the non-linear regression, we used the following mapping function:

\[
\text{Quality}(\alpha) = \beta_1 \times (0.5 - 1/(1 + \exp(\beta_2 \times (\alpha - \beta_3)))) + \beta_4 \times \alpha + \beta_5
\]

where \(\alpha\) is the score obtained from the objective metric, and \(\beta_k\) with \(k = 1, 2, 3, 4, 5\) are parameters. The fitting, i.e. the determination of parameters in [27], is done by the non-linear regression over the dataset.

#### 3.2 Evaluation of the reference image

We select six non-reference IQA algorithms, including BIQI, BRISQUE, SSEQ, CNN, ASIQE and ENIQA, to evaluate 100 reference images. These six non-reference methods have achieved good performance in several open image databases. To evaluate the performance of the algorithms, we use two common indicators, PLCC and SROCC, whose absolute value range is \([0, 1]\). The closer the value is to 1, the closer the performance of the algorithm is to human vision.

Table 2 shows the results of each index obtained by each algorithm after evaluation of 100 reference images in the MP2020 database. It can be seen in Table 2 that the scores of these five unreferenced algorithms on the reference images in the MP2020 database are not ideal, which indicates that the existing

| Algorithm | SROCC | PLCC |
|-----------|-------|------|
| BIQI [12] | 0.0924 | 0.2078 |
| BRISQUE [14] | 0.1219 | 0.1107 |
| SSEQ [16] | 0.1150 | 0.2140 |
| CNN [18] | 0.1027 | 0.0724 |
| ASIQE [20] | 0.0609 | 0.1787 |
| ENIQA [21] | 0.1584 | 0.4402 |
3.3 | Evaluation of the distorted images

Twelve typical quality evaluation algorithms are selected in the experiment, including six full-reference algorithms (PSNR, FSIM, GMSD, RFSIM, PSIM, SUMMER), six shallow-learning unreferenced algorithms (BIQI, BRISQUE, SSEQ, ASIQE, ENIQA), and one deep-learning unreferenced evaluation algorithm (CNN_IQA). The experimental results on the MP2020 database are shown in Tables 3 and 4.

By comparing the IQA algorithms in Tables 3 and 4, it is obvious that these algorithms are not very effective in the evaluation of macro photography images. The accuracy of the full-reference algorithm is higher than that of the non-reference algorithm, mainly because we choose the learning-based non-reference method, and the learning-based method is affected by the generalization performance. The performance of FSIM and GMSD is better than that of the other full-reference algorithms, demonstrating that the two algorithms are excellent. The BIQI and CNN algorithms are relatively better in the non-referenced method, and the CNN method does not show excellent performance, which shows that the generalization performance of the deep-learning method is insufficient.

Among them, we can see that the performance of all algorithms in the GBLUR distortion type is not good. The analysis may be due to the characteristics of the macro photography image itself, which is a kind of image with clear foreground and blurred background. This characteristic of macro photography images will also affect the subjective score of human beings, so the subjective score itself may be inaccurate.

To make the comparison more obvious, we give the experimental results of five kinds of non-reference algorithms in the live database, and the values of SROCC and PLCC are shown in Table 5. Through the comparison of Tables 3, 4 and 5, it can be found that the performance of the algorithm in the macro photography database needs to be further improved.

3.4 | Some suggestions for developing IQA algorithms on MP2020

In view of the particularity of macro photography images, we try to give the following suggestions when developing a quality evaluation algorithm capable of evaluating such images:

1. For the evaluation of macro photography images, the image should be evaluated as a whole instead of block evaluation, and then the average score should be taken. To expand the training samples, some existing IQA algorithms, especially deep learning-based quality evaluation methods, usually carry out block processing, which should not be suitable for micro distance photography images.

2. The attention mechanism should be combined in the evaluation of macro photography images. The basic idea of the attention mechanism in computer vision is to let the system learn to pay attention to the important information instead of the irrelevant information. The attention mechanism focuses computation on certain elements, a set of calculations. It is very similar to human attention to relevant information. Macro photography images are images with...
such characteristics, and people only pay attention to clear foreground objects while ignoring fuzzy backgrounds.
3. The evaluation of macro photography images should be combined with image aesthetic evaluation. When macro photography mainly aims at the aesthetic stimulation formed under the influence of aesthetic factors, such as composition, colour, light, depth of field, and virtual reality. Therefore, the evaluation of macro photography images should be based not only on their own distortion but also on the ‘beauty’ of the image, that is, the combination of aesthetic evaluation.

4 | CONCLUSIONS

To promote research on the quality of macro photography, we have established a micro camera image quality database MP2020, which contains 100 reference images and 800 distorted images. We use 10 existing typical full-reference and non-reference algorithms for evaluation and find that these algorithms are not very suitable for the MP2020 database, so the establishment of this database will improve the performance of existing algorithms and develop new algorithms.

In future work, we will make the following improvements and upgrades to the MP2020 database: (1) expand the number of images and distortion types to meet the requirements of deep learning; (2) improve the evaluation and statistics of GBLUR distortion types to make them more scientific; and (3) develop a general non-reference quality evaluation algorithm that can adapt to both common distortion images and micro distance camera distortion images.

REFERENCES

1. Bo, W., et al.: HVS-based structural similarity for image quality assessment. In: International Conference on Signal Processing. Beijing, China, pp. 1194–1197 (2008).
2. Yong, M., et al.: A novel non-reference image quality assessment algorithm. In: 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). Chengdu, pp. 560–566 (2017).
3. Zhou, W., Alan, C.B.: A universal image quality index. IEEE Signal Process. Lett. 9(3), 81–84 (2002).
4. Zhou, W., et al.: Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. 13(4), 600–612 (2004).
5. Zhou, W., Eero, PS., Alan, C.B.: Multiscale structural similarity for image quality assessment. In: The 37th Asilomar Conference on Signals, Systems & Computers. Pacific Grove, CA, pp. 1398–1402 (2003).
6. Zhou, W., Qiang, L.: Information content weighting for perceptual image quality assessment. IEEE Trans. Image Process. 20(5), 1185–1198 (2011).
7. Lin, Z., et al.: FSIM: A feature similarity index for image quality assessment. IEEE Trans. Image Process. 20(8), 2578–2586 (2011).
8. Lin, Z., Lei, Z., Xuanqin, M.: RFSIM: A feature-based image quality assessment metric using Riesz transforms. In: 2010 IEEE International Conference on Image Processing. Hong Kong, China, pp. 321–324 (2010).
9. Wufeng, X., et al.: Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. IEEE Trans. Image Process. 23(2), 684–695 (2014).
10. Ke, G., et al.: A fast reliable image quality predictor by fusing micro- and macro-structures. IEEE Trans. Ind. Electron. 64(5), 3903–3912 (2017).
11. Dogancan, T., Ghassan, A.: Perceptual image quality assessment through spectral analysis of error representations. Signal Process.: Image Commun. 70, 37–46 (2019).
12. Anush, K.M., Alan, C.B.: A two-step framework for constructing blind image quality indices. IEEE Signal Process. Lett. 17(5), 513–516 (2010).
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
14. Anish, M., Anush, K.M., Alan, C.B.: No-reference image quality assessment in the spatial domain. IEEE Trans. Image Process. 21(12), 4695–4708 (2012).
15. Anish, M., Rajiv, S., Alan, C.B.: Making a ‘Completely Blind’ image quality analyzer. IEEE Signal Process. Lett. 20(3), 209–212 (2013).
16. Lixiong, L., et al.: No-reference image quality assessment based on spatial and spectral entropies. Signal Process.: Image Commun. 29(8), 856–863 (2014).
17. Le, K., Peng, Y., David, D.: Convolutional neural networks for no-reference image quality assessment. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, pp. 1733–1740 (2014).
18. Jongyoo, K., et al.: Deep convolutional neural models for picture-quality prediction: challenges and solutions to data-driven image quality assessment. IEEE Signal Process. Mag. 34(6), 130–141 (2017).
19. Sebastian, B., et al.: Deep neural networks for no-reference and full-reference image quality assessment. IEEE Trans. Image Process. 27(1), 206–219 (2018).
20. Ke, G., et al.: No-reference quality assessment of screen content pictures. IEEE Trans. Image Process. 26(8), 4005–4018 (2017).
21. Xiaqiao, C., et al.: No-reference color image quality assessment: From entropy to perceptual quality. EURASIP J. Image Video Process. 2019, 77 (2019).
22. Hamid, R.S., Muhammad, F.S., Alan, C.B.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Trans. Image Process. 15(11), 3440–3451 (2006).
23. Ponomarenko, N., et al.: TID2008-A database for evaluation of full-reference visual quality assessment metrics. Adv. Mod. Radio Electron. 10(4), 30–45 (2009).
24. Ponomarenko, N., et al.: Color image database TID2013: Peculiarities and preliminary results. In: European Workshop on Visual Information Processing (EUVIP). Paris, France (2013).
25. Aladine, C.: Full reference image quality assessment: Limitation. In: 2014 22nd International Conference on Pattern Recognition. Stockholm, pp. 833–837 (2014).
26. Shirai, M., Kambaya, H., Taniguchi, K.: Data analysis apparatus and data analysis method. (2013). US, US8509227 B2.
27. Tsung-Jung, L., et al.: Visual quality assessment: Recent developments, coding applications and future trends. APSIPA Trans. Signal Inf. Process. 2, e4 (2013).

How to cite this article: Sang, Q., et al.: IET Image Processing, 1–7. (2021) https://doi.org/10.1049/ipt.r2.12198.