Preselection based Subjective Preference Evaluation for the Quality of Underwater Images

Miao Yang[1,2,3], Yixiang Du[1,7], Yue Huang[4], Hantao Liu[5], Zhiqiang Wei[6],
Jintong Hu[1], Ke Hu[1], Zhibin Sheng[1]
[1]School of Electronic Engineering, HuaiHai Institute of Technology, China
[2]Marine Equipment and Technology Institute, Jiangsu University of Science and Technology, China
[3]Qingdao National Laboratory of Marine Science and Technology, China
[4]Key Laboratory of Underwater Acoustic Communication and Marine Information Technology Ministry of Education, China
[5]School of Computer Science and Informatics, Cardiff University, U.K
[6]Ocean University of China School of Information Engineering
[7]Mining and Technology University of China School of Information and Control Engineering

Abstract

Underwater images contain an interactive mixture of distortions due to the physicochemical property of water and the instability of imaging systems, which differ from those in natural images. We cannot obtain the pristine underwater image as the reference applied in the traditional benchmark databases, and the groups of gradual distortions either. In this paper, a novel preselection based preference label evaluation method is proposed to construct a combined subjective test procedure for an extended preference judgment dataset of underwater images. To the best of our knowledge, this is the first subjective evaluation procedure for underwater images, and also a solution for an expanding visual preference benchmark database. We demonstrate the excellent correlation of the proposed subjective evaluation with the traditional image quality assessment. It is also proven that the proposed subjective evaluation procedure could reflect the slight change of image quality and the authentic quality of a picture more accurately better than the traditional methods.

Keywords: *Image quality evaluation, MOS, underwater image, subjective image quality database*

1. Introduction

Vision as a scientific exploration measure is becoming more and more indispensable in marine survey [1], [2]. The images captured in water are usually afflicted with various complex and mixed degradations such as low contrast, blur, non-uniform illumination, non-uniform color casting and noises caused by optical attenuation, adsorption and scattering of the water body [3], [4], which are not necessarily well-modeled as opposed to the synthetic distortions found in existing natural image databases. An automatic quality prediction tool [6], [7] which enable automatically obtain high quality underwater images and marine science artificial intelligence analysis, provide objective criteria for underwater image restoration or enhancement [5] are thus highly desirable goals. Given that the ultimate receivers of images are humans, the most reliable way to understand and predict the effect of distortions is to capture opinions from human subjects [8], [9].

1.1. Traditional subjective image evaluation database

Although the existing subjective image evaluation databases, such as CSIQ database [15], LIVE database [16], TID2008/TID2013 [17], [18], IVC database [19], Toyama database [20], WIQ database [21], [22], Cornell-A57 database [23] and the newest the LIVE In the Wild Image Quality Challenge database [24], play an important role in advancing the field of image quality prediction, the images contained in the databases are of a fixed number of air images with a certain degree of individual distortion artificially synthesized from a small original image dataset and intrinsically different from underwater images. The comparison of existing databases is listed in Table I. There are almost no underwater images included. In addition, observing an original image and its various distortions simulated in one group can lead to over-learning of the distortion. Furthermore, the results obtained in a strictly controlled experimental environment need to be consistent with the subjective perception in real life.

1.2. Underwater images

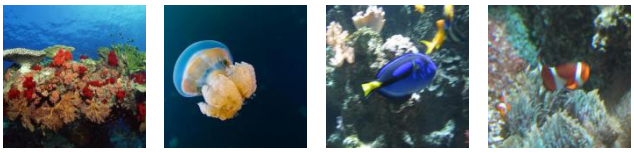
There is no pristine reference for an underwater image [26]. It is also impossible to establish an image database by simulating the original images with different distortions and degrees, or to group underwater images according to the type of distortion as is the case with natural images. In addition to the difficulty of underwater image acquisition, underwater images contain an interactive mixture of

Table I A comparison of IQA databases

Database	Source Images	Distorted Images	Distortion Types	Authenticity of Distortions	Evaluation Method	Underwater Images
CSIQ [15]	30	866	6	Synthetic	Self-design method	0
LIVE IQA [16]	29	779	5	Synthetic	Double stimulus	0
TID2013 [18]	25	3000	24	Synthetic	Pairwise comparison; Single stimulus	0
LIVE Challenge [24]	N/A	1162	Numerous	Authentic	Single stimulus	2

distortions [25] and the qualities of those images are usually relatively similar in visual perception.

Traditional subjective evaluation methods [27]-[31] fail to distinguish the difference between the underwater images. For instance, consider the images shown in Figs. 1(a) and (b), the single-stimulus scores of them in our experiment were all 4.4, but when we put the two images on one screen at the same time, there is a slight blurring distortion on the left comparing to the right one. By using the proposed method, observers scored the two images with 63.1 and 71.4 respectively. In our studies, the Fig. 1(b) has higher chroma contrast than Fig. 1(a), demonstrating the sensitivity of the preference label to subtle mass differences. Figs. 1(c) and (d) are the only two underwater images in the LIVE In the Wild Image Quality Challenge Database [24]. Their mean opinion scores (MOS) are 82.48 and 68.98, respectively (scores are from the LIVE In the Wild Image Quality Challenge Database), although the difference in image qualities are not particularly noticeable. Comparatively, their subjective scores obtained by our subjective test are 61.4 and 54.8, respectively, which is more reasonable. An observer evaluates the distortion (such as Gaussian noise) of an image taken in the air, rather than the authentic mixed degradation caused by the water body factors, often fails to account for the quality of an underwater image.



(a) (b) (c) (d)

Figure 1. (a) - (b) Two images with the same score for the single stimulus five-level quality scoring; (c) - (d) Two underwater images in the LIVE In the Wild Image Quality Challenge database.

1.3. Contributions

A prescreening based preferrence label (PPL) subjective image quality evaluation method is proposed by which we can collect multi-level distorted underwater images widely. Through grouping the underwater image data into pairs, the observers are organized to judge the relative quality, and a preference label is generated. The accumulated label scores of an image is computed, and then the initial underwater image database is established according to the label scores. By gathering the pair labels online, the MOS of the laboratory scored images are updated, and the MOS of a newly added underwater image is obtained.

The approach we described in this paper is on the difficulty of establishing a subjective quality baseline for underwater images with the expanding visual perceptive range of an increased number of underwater images, and summarize our contributions below:

(i)First, a subjective quality evaluation method for underwater images is proposed in which a reference of the original image is unnecessary. Based on a preference label, we settle the difficulty of distinguishing the type and degree of underwater image degradation and avoids the sensitivity of the image content as well as groups the test images according to the distortion types, all of which ensures a consistency of scores among the observers.

(ii)Secondly, an objective image quality prescreening is adopted ahead of subjective evaluation in our method, which ensures the tested underwater images are uniformly sampled within the existing quality range.

(iii)Progressive learning ranking is proposed in extending the established underwater image subjective evaluation database. By combing the laboratory with online voting, we can gather extensive underwater image data in real time. Furthermore, the progressive learning ranking is in line with the psychological process for the broadened subjective image perceptible quality distributions.

(iv)The subjective rating procedure is more efficient and labor-saving, which means that workload on participants is lower and more flexible than other classical subjective image quality evaluations.

The superiority and accuracy of the proposed PPL subjective evaluation method is verified by comparing to the traditional subjective evaluation methods on the established database. Compared with the traditional single-stimulus subjective evaluation method, the PPL subjective underwater image quality evaluation can correctly reflect that the image quality decreases with the increase of the distances to the camera, in different turbidities of water under same imaging conditions. The performance of two common blind IQA algorithms developed using traditional natural image databases and two outstanding underwater IQA indicators were also empirically studied. The experimental results showed that the existing no reference natural image objective quality evaluation methods based on the traditional image databases could not correctly evaluate the underwater

image quality which have interactive blending distortions. This PPL subjective image quality evaluation method that combined of laboratory and online is an effective way to collect a large number of subjective image quality opinions for underwater images.

The rest of this article is arranged as follows. The proposed PPL subjective underwater image quality evaluation method is described in Section 2. In Section 3, how to extend the subjective underwater image quality database based on the PPL method and some strategies (dichotomy, ER random graphs, etc.) is described. Experiments and discussions are given in Section 4. We conclude the paper in Section 5.

2. Underwater image preference label subjective quality evaluation

2.1. Collection of images

We collected nearly 3000 images[1] from laboratory underwater image sequences under controlled imaging environments, online manual collection including underwater photography, near-shore marine aquaculture, pipeline engineering, coastal surveys, deep-sea images and others taken by a variety of underwater optical cameras and light sources. The resolution of these underwater colour images range from 58×83 to 3000×4000, and they have a variety of non-uniform colour degradation, blur, fog effect or non-uniform illuminating, biological disturbance sediment turbidity distortion, and so on. Some of the pictures are shown in Fig. 2.

2.2. Preselection of evaluation pairs

Given N collected underwater images, $N\times(N\text{-}1)/2$ possible image pairs can be generated. According to the International Telecommunication Union [27], [28], 300 images were preselected to generate 44850 priority image pairs. We selected CIELab spatial brightness contrast, hue-variance, and saturation-mean as the criteria for the preselection, which was the conclusion in our previous work [53]. About ten images at each interval were randomly selected. The purpose of this is to make the quality of priority test images not concentrate on a certain interval, so that images quality distribution of the initial data set is as uniform as possible, which is necessary for the extended construction of the subsequent large-scale database.

[1] Part of the collected images comes from public image resources (such as ImageNet, etc.) and some are from the database provided by the cvpr challenge. Some images are authorized, but images taken from the network are difficult to find the real source. If you are the data owner, please contact us, we hope to get the sharing permission and support.

2.3. Image quality assessment experiment

Subjective evaluation phase. All experiments were carried out in the underwater vision laboratory of Huaihai Institute of Technology, which guaranteed the subjective evaluation environment requirements [27], [28]. Forty eight students aged from 20 to 30, including 28 male and 20 female observers who had normal visual acuity and color vision were recruited to evaluate the images. Two images in an image pair were simultaneously displayed on a 27-inch LCD. At the beginning of the test, five "simulation presentations" were broadcast to stabilize the observer's score and the preference labels given in these demos were not recorded in the results. Our playlist was based on a random permutation of 44850 test pairs with a random within-pair order. To avoid the contextual and memory effects, our program would go through the entire playlist to determine if adjacent pairs correspond to the same image. The observer had to vote on each image pair in 3s, otherwise the score of the pair would not be recorded. We also interspersed some image pairs with significant differences in quality to check the observer's attentiveness. For every observer, the number of pairs had to vote on was not limited, each phase of experiment was about 30 minutes (including inspection and demonstration), and fatigue effects were minimized.

Mean opinion score. We unified the image size to 512×512. Suppose there are preference labels $l_{1,2}$ and $l_{2,1}$ for an image pair (I_1, I_2)in P , $P \in \{(I_i, I_j) \mid i, j = 1,2,...,300\}$, if the observer thought I_1 was better in quality than I_2, then $l_{1,2}$=+1 and $l_{2,1}$=-1. On the contrary, we have $l_{1,2}$ =-1, $l_{2,1}$=+1. If the observer did not mark the image pair (I_1, I_2) or believed that the relative quality of the image pair could not be determined, then $l_{1,2}$ and $l_{2,1}$ were both set to 0.

By obtaining the preference label $l_{i,j}$, $i \neq j$ for all image pairs, the cumulative label score S for image i, which is in the range [- 299, 299] can be computed as:

$$S_i = \sum_{j=1}^{300} l_{i,j} \quad i \neq j \qquad (1)$$

The centesimal system score S_{ip} of the image i can be calculated based on the linear mapping.

The observer did not have to evaluate many image pairs in a single session, and only needed to judge the preference for each pair. There was no lag phenomenon between the front and back images or judgment scale mismatch. Consistency in the ratings between the observer groups was ensured because the test images were neither grouped according to the degraded type, nor were graded on a hierarchical basis.

Some of the images we measured along with their centesimal system MOS, S_{ip} are shown in Fig. 3. As shown, the gradient of MOS was in consistent with our visual inspection.

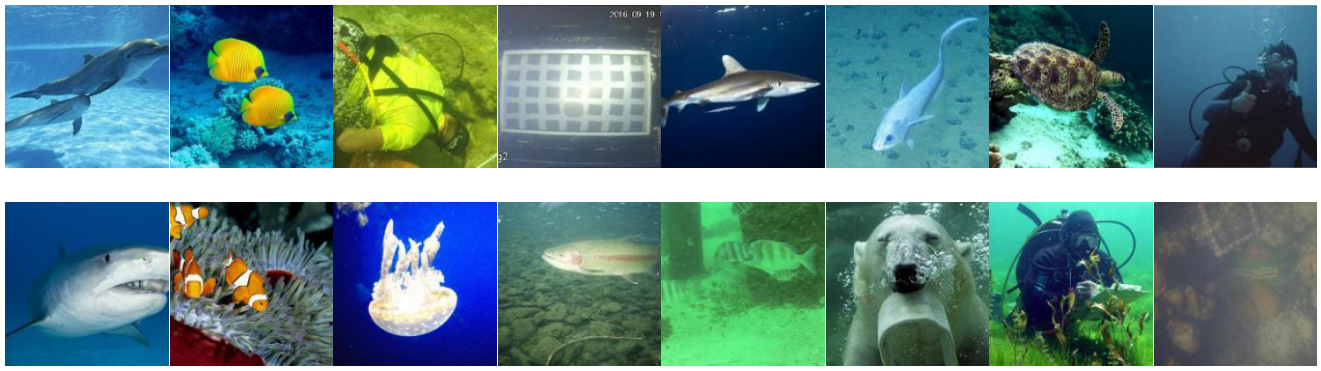


Figure 2. Part of the collected underwater image database.

3. Adding new images

In this section, we introduce the process of inserting a new image into the sorted initial databases. Based on 300 existing laboratory MOS, we use online voting to achieve database extension. A strategy similar to dichotomy is used to compare the new image with the images in the preselected database. For a new image to be added, we process:

(1) Compare it to a random image with the quality score around the central of the current database, allowing to fluctuate with five images.

(2) Rating the added image pairs with no more than $\log_2 N$ times iteratively according to the label given by an observer until the comparison range is reduced to less than ten images.

(3) The new image will be compared to all images in the last range, record the possible location of the new image obtained by the observer in the current database.

(4) Steps (1), (2), and (3) are respectively performed online to collecting the possible locations of the new images from no less than 20 observers.

It is foreseeable that most of the collected positional locations will be concentrated in a smaller interval. We choose the median of these possible positions and remove the 5% scores that is the farthest from the median distance. The remaining scores are averaged and the final result is considered to be the position where the image should be inserted in the existing sequence.

(5) Add one point to all the label scores of images above the new position, and subtract one below it.

(6) Score the new image based on the number of images above, below, and with the same quality.

ER random graphs $G(n, p)$ start from n vertices and draw their edges independently according to a fixed probability p ($0 \leqslant p \leqslant 1$), which was chosen to meet the scenario that in crowd sourcing ranking raters [57]. In the extension of the database by the dichotomy strategy, the 20 observers independently observed $\log_2 N$ image pairs, the edge probability for a new vertice (image) in the image pair ER random graph to the other N vertices (the current image database size) must be greater than $\log N/N$. Therefore, the $G(n, p)$ is almost always connected [60].

The corresponding centesimal system score error $\Delta S_{ip\lambda}$ for the image i with λ pair evaluations is:

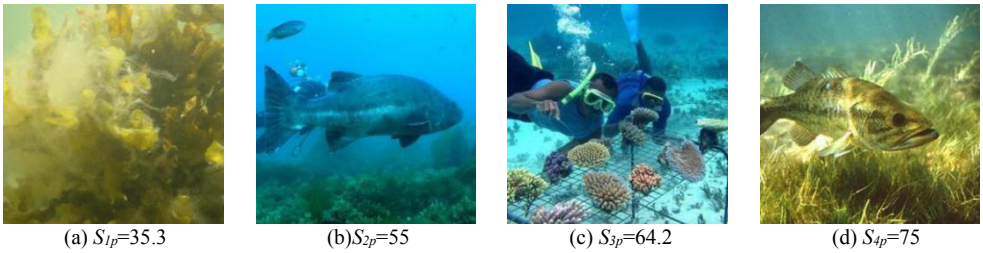
$$\Delta S_{ip\lambda} = S_{ip\lambda} - S_{ip} = \frac{50\Delta S_{i\lambda}}{(N-1)} \tag{2}$$

$$S_{ip\lambda} = (\frac{S_i + \Delta S_{i\lambda}}{2(N-1)} + \frac{1}{2}) \times 100 \tag{3}$$

where N is the number of images in the current database. A label score error $\Delta S_{i\lambda}$ exists if only λ pair evaluations, $1 \leq \lambda \leq \log_2 N$. As the number of the images in database increases, $\Delta S_{ip\lambda}$ decreases. This can be approximated as:

$$\Delta S_{ip\lambda} \approx \frac{2^{(\log_2 N - \lambda)} \times \frac{2(N-1)}{N} \times 50}{(N-1)} = \frac{100}{2^{\lambda}} \tag{4}$$

It can be inferred that an authentic subjective quality score can be achieved with a reasonable number of evaluations (set to λ). For instance, for N=2,000, λ=9, the ΔSipλ is only around 0.2 with Eq. (4). The number of pair evaluations and the images in the database tend to be independent. We can quickly gather a relatively reasonable score of the test image with only a few pair evaluations.



(a) S_{1p}=35.3 (b)S_{2p}=55 (c) S_{3p}=64.2 (d) S_{4p}=75

Figure 3. A few underwater images from the initial evaluation database and their MOS.

4. Experiments and discussion

Through the subjective evaluation experiment described in the previous section, we obtained 134,550 comparisons from 48 observers. We discarded edges that are not part of cycles (*e.g.*, inequalities $\{R_1 > R_2, R_2 > R_3, R_3 > R_1\}$ are not consistent). As introduced previously, we also interspersed some pairs of images with obvious quality differences, and removed the results from individuals with an error rate of more than one-third. We averaged the labels of the same image pair and get 44,850 comparisons out of 300 underwater images.

4.1. Experiments

We explored the performance of the PPL subjective image quality evaluation in groups of experiments to compare it with the traditional subjective test. The accuracy was illustrated through an objective software test and image sequences. We also explored the usefulness of the new initial underwater database by using it to evaluate the quality prediction performance of the blind natural and underwater IQA algorithms.

Comparison with the single stimulus. Under the same environment and with the same observers, a 5-level scoring test was carried out on the initial 300 underwater images, which were divided into six groups (50 in each group) to prevent observer fatigue [27], [28]. The correlation of the MOS collected with the proposed method and the single stimulus method is illustrated in Fig. 4. The correlation reached 0.95. Four groups of underwater images and their MOS produced by PPL and single stimulus subjective image quality evaluation methods are shown in Fig. 5. Each row of images in Fig. 5 had the same score by the single stimulus evaluation, although we could still perceive the subtle difference in quality between them. It can be seen that for images whose quality differences were not obvious, the PPL subjective evaluation method was more susceptible to the subtle quality difference, and the scores gathered by the proposed method therefore, has the ability to distinguish such nuances in scoring.

Changing turbidity of water. We compared the scores of 46 sequence images in the 300 database which were sampled by the preselection automatically. These underwater images were taken at the same area and angle but in water of different turbidities [59], and can be divided into four groups according to the content (called photo1-4 group). The scores of each group are plotted in Fig. 6. The larger the image number is, the lower the turbidity of the water the image was taken in. The Pearson's linear correlation coefficient (PLCC), Spearma's rank ordered correlation coefficient (SROCC) and Kendall's rank ordered correlation coefficient (KROCC) between the MOS obtained by the PPL subjective image quality evaluation and the corresponding image order in the sequence are listed in Table II since these images are numbered according to the gradient of turbidity. The number of images in each sequence is not the same because of the automatic selection.

We can see that the results obtained from the proposed PPL subjective evaluation were linearly related to the turbidity of the water, which correctly reflects the image quality levels. Obvious outliers existed in the groups shown in Figs. 6(b) and (c) and the KROCC of photo2 group is the lowest. The two outlier images in the photo2 group are highlighted in Fig. 7. We can see that the difference between the two adjacent images is extremely ambiguous. The selected images in the photo2 group were more concentrated, as shown in Fig. 6(b).

The photo4 group was the one closest to linearity, as shown in Fig. 6(d). From this, it can be inferred that: 1) The images in the photo4 group were more colourful than photo2, 2) The images randomly selected in the photo4 group were scattered on the image quality gradient, as shown in Fig. 8. There were no consecutive images of similar image qualities.

Table II Correlation of the MOS and the order of the image

	Numbers in each group	PLCC	SROCC	KROCC
Photo1 group	12	0.9877	0.9930	0.9697
Photo2 group	11	0.9880	0.9727	**0.8909**
Photo3 group	14	0.9787	0.9901	0.9503
Photo4 group	9	0.9923	**1.0000**	**1.0000**

Imatest test. The accuracy of the PPL subjective image quality evaluation score with regard to the pristine image quality is discussed by presenting the quality of the ColorChecker 24 X-Rite Chart (21.59×27.94cm) images taken in a tank. The tank was 2.53m long, 1.02m wide, and 1.03m high. The chart images were taken with the OTI-UWC-325/ P/ E colour camera. Images (960×576) were obtained in water of 94.5cm transparency [61] under daytime lighting. The images were taken with increased distances to the camera, part of which is shown in Fig. 9.

The certimential score (S_{ip}) obtained by the PPL subjective image evaluation method, the five-level quality scores obtained by the single-stimulus system and the data output from Imatest for the ColorChecker chart images included in the 300 priority test images are listed in Table III.

Imatest [62], [63] has been the most professional software in qualifying camera imaging by comparing the differences between real images of charts and their testing ones. In Table III, the Mean camera chroma (saturation) is the average camera chroma (colour saturation), which was generally between 100% and 120%. The meaning of this value is to test the differences between the representative 24 colours on the test image and the standard ColorChecker chart. The greater the difference, the larger

this value is. The ΔC*ab chroma corr, ΔC*ab uncorr, and the ΔE*ab are color error metrics in the device-independent CIELAB color space that are used to illustrate the perceived difference between colors by measuring the Euclidean distance between them. The ΔE*ab includes the luminance L*, while ΔC*ab chroma corr and ΔC*ab uncorr compute colors only. It can be seen that the real qualities of the test images were linear with our MOS by comparing the subjective scores of the preference labels with the Imatest test output data. This proves that the PPL subjective evaluation could reflect the slight change of image quality and the authentic quality of a picture more accurately than the traditional methods.

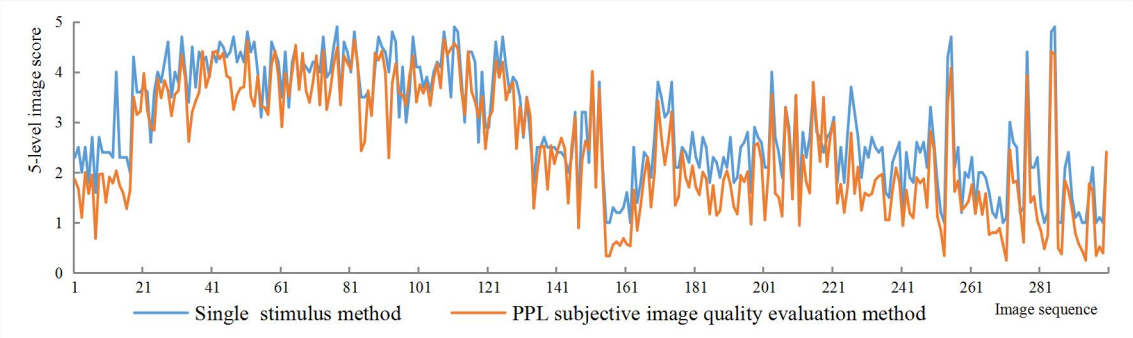
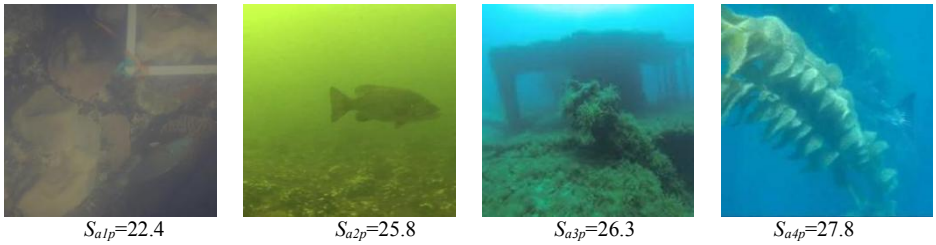


Figure 4. Comparison the scores of the two test methods.



S_{a1p}=22.4 S_{a2p}=25.8 S_{a3p}=26.3 S_{a4p}=27.8

(a) Single stimulus score is 1.8.

S_{b2p}=39.1 S_{b2p}=40.1 S_{b3p}=42.1 S_{b4p}=44.3

(b) Single stimulus score is 2.7.

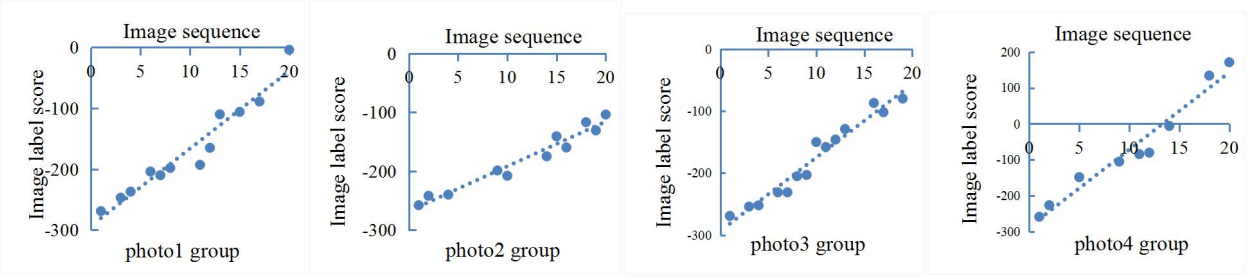
S_{c1p}=58.1 S_{c2p}=59.7 S_{c3p}=60.5 S_{c4p}=61.1

(c) Single stimulus score is 3.5.

S_{d1p}=77.4 S_{d2p}=78.4 S_{d3p}=80.3 S_{d4p}=81.1

(d) Single stimulus score is 4.0

Figure 5. Comparison with the single stimulus methods.



(a) (b) (c) (d)

Figure 6. Label scores of the photo1-4 groups.

Tables III. MOS and software output scores for the ColorChecker chart images.

Distance from camera		60cm	70cm	90cm
Image		Fig. 9 (a)	Fig. 9 (b)	Fig. 9 (c)
Certimential score (S_{ip})		48.16	37.12	25.59
Five-level quality scoring		2.4	2.3	2.3
Mean camera chroma(saturation)		134.10%	136.90%	137.70%
Color errors:	mean	43.2	45.0	47.9
ΔC^*_{ab} chroma corr	max	85.3	89.4	101.0
ΔC^*_{ab} uncorr	mean	52.1	54.9	57.7
	max	24.0	24.0	24.0
ΔE^*_{ab}	mean	55.2	57.6	61.0
	max	93.5	97.5	110.0

Figure 7. Two outlier images in the photo 2 group.

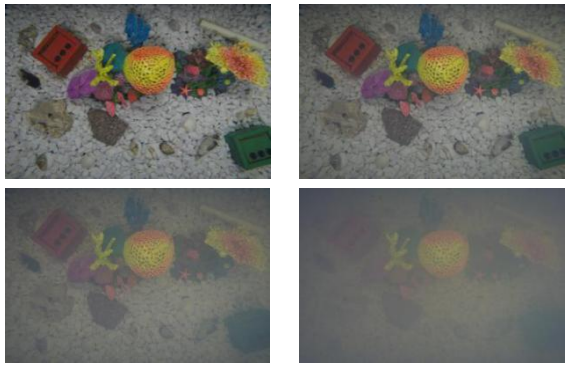


Figure 8. Partial images in the photo4 group.

(a) (b) (c)

Figure 9. Partial images of Imatest test

The Objective Quality Evaluation Methods of Underwater Images. To further highlight the challenges that underwater distortions present in our database pose to the objective image quality evaluation metrics, we also computed the median correlation values when the algorithms were trained on the underwater image database. We computed the median correlation value of two common blind IQA algorithms which were developed using the traditional natural image databases including the Colour Image Quality Index (CIQI) [47] and the Colour Quality Enhancement (CQE) [48], and the two prominent underwater IQA metrics including UIQM [2] and UCIQE [53]. These algorithms were tested on the underwater image database constructed by the proposed dichotomy based PPL subjective image quality evaluation. The data includes 1000 RGB underwater images, with the size averaging from 68×101 to 3000×4000.

We selected 80% of the images in the database as training set and 20% for testing by applying the K-fold decomposition, and repeated the process five times. We computed the median SROCC and PLCC between the predicted and MOS values. A higher value of each of these metrics indicates better performance in terms of the correlation with human opinions. The results were obtained with the same parameters that were originally presented in their work on the constructed underwater image database. The results are reported in Table Ⅳ, where it can be observed that the performance of the UCIQE was significantly better than of the state-of-the-art IQA methods designed based on the natural images and the UIQM when the underwater image database was used for testing. It illustrated the challenges that the authentic

distortions present in underwater images pose to the IQA algorithms based on the traditional image databases.

Table Ⅳ Median SROCC and PLCC on the proposed underwater image database.

	CIQI [47]	CQE [48]	UICM [2]	UCIQE [53]
PLCC	0.5720	0.2019	0.2086	0.7487
SROCC	0.5047	0.1809	0.1142	0.7369

4.2. Discussion

Attention problems and outliers. We checked the scored image pairs in the evaluation process, especially the interspersed images with significant quality differences. Approximately 5% of the observers failed to correctly answer the pros and cons of the two images with significant differences. We have reason to believe that such observers may have given an abnormal choice for other images [64]. For the non-serious individuals, we removed all their records.

Screening of online observers. Although the combination of laboratory and online assessment has advantages in collecting subjective evaluation values, there are still many limitations that require in-depth research. For example, risk of differences in test conditions between a large number of observers, and how to determine whether the observer takes this online assessment seriously. We plan to first provide some sample image pairs when an observer attempts to perform an evaluation on our website. These image pairs have been tested in a lab environment (we will avoid selecting some image pairs that are too close in quality). Observers need to evaluate these image pairs, and we will set a threshold for this screening. The requester with an error rate of more than 1/4 during the test will be rejected. Qualified observers are allowed to do online testing. And image pairs with significant quality differences will be scattered throughout the testing process, as we did during the lab assessments. For those who are not serious, we will delete all the in records. Five image pairs would be randomly presented to each observer twice during an online evaluation. If the observer provides preference labels more than two pairs that are inconsistent in the five same image pairs, we also delete all of their records.

Online progressive learning rank. We will publish the authorized part of the database online and support the addition of more underwater images following the steps described in Section 3. We hope to create a trend for the distribution of underwater mix-distortions and to be able to level uniformly with the increased image number. We hold more than 100,000,000 underwater images given by the Qingdao National Laboratory of Marine Science and Technology, and will first add part of these images to extend the underwater image database in the strict laboratory environment. We will also accept image upload and provide online image quality evaluation services after calibration in the laboratory. The consistence of online evaluation with the laboratory scores is still a question we will explore [68], since the basic infrastructure and procedures of subjective testing for the online evaluation are different from the traditional subjective research conducted in the laboratory. However, we are inclined to believe that based on a huge number of subjective scores conducted in the laboratory and the preference label subjective evaluation, we can compare the ranking order of the sampled underwater images viewed in the laboratory with the online votes to compute the consistency, calibrate the online score, and process the outliers.

5. Summary

A preselection based preferred-pair label subjective quality evaluation method has been proposed in this paper. The proposed method does not require a reference. Compared with the traditional single-stimulus subjective evaluation, better accuracy and a higher correlation were shown by the proposed PPL subjective image quality evaluation. We also designed an underwater image database construction method in which the progressive learning ranking is proposed, and this is a new solution to set up an extending image quality database in real time.

Acknowledge

This work was supported in part by the National Natural Science Foundation under Grant 61601194, the "Six talent peaks" project in Jiangsu Province (No. DZXX-030), the "Double creation talents" science and technology deputy general manager project in Jiangsu Province (2017), Jiangsu colleges High-tech ship collaborative innovation center funded project (HZ20190005), and in part by Lianyungang "521" project and Lianyungang "Haiyan" funded project in China.

References

[1] A. Rosenfeld, “Image Analysis and Computer Vision,” *Computer Vision and Image Understanding.*, vol. 78, no. 2, pp. 222-302, 2000.

[2] K. Panetta, C. Gao, and S. Agaian, “Human-Visual-System-Inspired Underwater Image Quality Measures,” *IEEE Journal of Oceanic Engineering.*, vol. 41, no. 3, pp. 541-551, 2016.

[3] K. Ingrid, “Underwater Imaging and the effect of inherent optical properties on image quality,” Norwegian University of Science and Technology. Master thesis, 2014.

[4] G. Johnsen, Z. Volent, E. Sakshaug, F. Sigernes, and L. H. Pettersson, “Remote sensing in the Barents Sea,” In: E. Sakshaug, G. Johnsen, and K. Kovacs(eds.), *Ecosystem Barents sea*, Trondheim, Norway, Tapir Academic Press, pp. 139-166, 2009.

[5] R. Schettini, S. Corchs, "Underwater image processing: state of the art of restoration and image enhancement methods," *EURASIP Journal on Advances in Signal Processing.*, 2010:746052, Apr. 2010.

[6] H. Lu, *et al.*, "Underwater image descattering and quality assessment," In *ICIP*, pp. 1998-2002, Sep. 2016.

[7] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult," In *ICASSP*, vol. 4, pp. 3313-3316, May. 2002.

[8] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K Cormack, "Study of subjective and objective quality assessment of video," *IEEE transactions on image processing.*, vol. 19, no. 6, pp. 1427-1441, Feb. 2010.

[9] E. Abdou, N. J. Dusaussoy, "Survey of image quality measurements," In *Proceedings of 1986 ACM Fall joint computer conference.*, pp. 71-78, Nov. 1986.

[10] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, *et al*., "Image database tid2013: peculiarities, results and perspectives," *Signal Processing Image Communication.*, vol. 30, pp. 57-77, Jan. 2015.

[11] F. Gao, D. Tao, X. Gao, and X. Li, "Learning to rank for blind image quality assessment," *IEEE Transactions on Neural Networks and Learning Systems.*, vol. 26, no. 10, pp. 2275-2290, Sep. 2013.

[12] N. Ponomarenko, V. Lukin, K. Egiazarian, J. Astola, M. Carli, F. Battisti, "Color image database for evaluation of image quality metrics," In:*IEEE 10th Workshop on Multimedia Signal Processing*, pp. 403-408, Oct. 2008.

[13] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," In *CVPR*, vol. 35, no. 1, pp. 305-312, Jun. 2011.

[14] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L Zhang, and A. C. Bovik, "Deep Convolutional Neural Models for Picture-Quality Prediction:Challenges and Solutions to Data-Driven Image Quality Assessment," *IEEE Signal Processing Magazine.*, vol. 34, no. 6, pp. 130-141, Nov. 2017.

[15] E. C. Larson, D. M. Chandler, "Most Apparent Distortion: Full-Reference Image Quality Assessment and the Role of Strategy," *Journal of Electronic Imaging.*, vol. 19, no. 1, pp. 011006, Jan. 2010.

[16] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE Image Quality Assessment Database Release 2," [Online]. Available: http://live.ece.utexas.edu/research/quality/.

[17] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, F.Battisti, "Tampere image database 2008," VERSION 1.0. [Online]. Available: http://www.ponomarenko.info/tid2008.htm.

[18] B. Vozel, *et al.*, "Color image database TID2013: Peculiarities and preliminary results," *European Workshop on Visual Information Processing IEEE.*, Jun. 2013.

[19] P. L. Callet, F. Autrusseau, "Subjective quality assessment irccyn/ivc database," [Online]. Available: http://www.irccyn.ec-nantes.frivcdb/.

[20] Z. P. Sazzad, Y. Kawayoke, and Y. Horita, "MICT image quality evaluation database," [Online]. Available: http://mict.eng.u-toyama.ac.jp/.

[21] H. J. Zepernick, *et al*., "Wireless imaging quality database," [Online]. Available: http://www.bth.se/tek/rcg.nsf/pages/wiq-db.

[22] U. Engelke, H. J. Zepernick, and T. Kusuma, "Subjective quality assessment for wireless image communication: the wireless imaging quality database," *International Workshop on Video Processing and Quality Metrics for Consumer Electronics.*, Jan. 2010.

[23] D. M. Chandler, S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images," *IEEE Transactions on Image Processing.*, vol. 16, no. 9, pp. 2284-2298, Sep. 2007.

[24] D. Ghadiyaram, A. Bovik, "Massive Online Crowdsourced Study of Subjective and Objective Picture Quality," *IEEE Transactions on Image Processing.*, vol. 25, no. 1, pp. 372-387, Nov. 2015.

[25] J. C. Guo, C. Y. Li, C Guo, *et al.*, "Research Progress in Underwater Image Enhancement and Restoration Methods," *Journal of Image and Graphics*, 2017.

[26] J. C. Guo, C. Y. Li, Y. Zhang, and X. Gu, "Quality assessment method for underwater images," *Journal of Image and Graphics*,. 2017.

[27] *ITU-R Recommendation P.800. Methods for subjective determination of transmission quality*, 1996.

[28] *ITU-R Methodology for the Subjective Assessment of the Quality of Television Pictures*, 2002.

[29] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of Four Subjective Methods for Image Quality Assessment" *Computer Graphics Forum.*, vol. 31, no. 8, pp. 2478-2491, Nov. 2012.

[30] T. Tominaga, *et al*., "Performance comparisons of subjective quality assessment methods for mobile video," *International Workshop on Quality of Multimedia Experience IEEE*., vol. 97, no. 1, pp. 82-87, Jul. 2010.

[31] X. B. Gao, W. Lu, "Quality Assessment Methods for Visual Information," Xi'an: Xidian University Press, 2010.

[32] Y. T. Peng, P. C. Cosman, "Underwater image restoration based on image blurriness and light absorption," *IEEE Trans Image Process.*, vol. 26, no. 4, pp. 1579-1594, Feb. 2017.

[33] C. Mei, H. X. Sheng, *et al*., "Underwater color image enhancement algorithm based on prior dark-channel model," *Chinese Journal of Quantum Electronics.*, 2016.

[34] C. O. Ancuti, C. D. Vleeschouwer, *et al.,* "Color Balance and Fusion for Underwater Image Enhancement," *IEEE Transactions on Image Processing.*, vol. 27, no. 1, pp. 379-393, Oct. 2017.

[35] C. Li , J. Guo, R. Cong, Y. Pang, and B. Wang, "Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior," *IEEE Transactions on Image Processing*., vol. 25, no. 12, pp. 5664-5677, Sep. 2016.

[36] C. O. Ancuti, C. Ancuti, C. D. Vleeschouwer, *et al.*, "Color Balance and Fusion for Underwater Image Enhancement," *IEEE Transactions on Image Processing.*, vol. 27, no. 1, pp. 379-393, Oct. 2017.

[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process*., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[38] C. Gao, K. Panetta, and S. Agaian, "A new color contrast enhancement algorithm for robotic applications," in *Proc. IEEE Conf. Technol. Practical Robot Appl. (TePRA)*, Apr. 2012, pp. 42–47.

[39] Y. Wang, Q. Chen, and B. Zhang, "Image enhancement based on equal area dualistic sub-image histogram equalization method," *IEEE Trans. Consum. Electron*., vol. 45, no. 1, pp. 68–75, Feb. 1999.

[40] M. Kim and M. Chung, "Recursively separated and weighted histogram equalization for brightness preservation and contrast enhancement," *IEEE Trans. Consum. Electron*., vol. 54, no. 3, pp. 1389–1397, Aug. 2008.

[41] S.-D. Chen and A. R. Ramli, "Minimum mean brightness error bi-histogram equalization in contrast enhancement," *IEEE Trans. Consum. Electron*., vol. 49, no. 4, pp. 1310–1319, Nov. 2003.

[42] C. Wang and Z. Ye, "Brightness preserving histogram equalization with maximum entropy: A variational perspective," *IEEE Trans. Consum. Electron*., vol. 51, no. 4, pp. 1326–1334, Nov. 2005.

[43] C. H. Ooi, N. S. P. Kong, and H. Ibrahim, "Bi-histogram equalization with a plateau limit for digital image enhancement," *IEEE Trans. Consum. Electron*., vol. 55, no. 4, pp. 2072–2080, Nov. 2009.

[44] B. Bringier, N. Richard, M.-C. Larabi, and C. Fernandez-Maloigne, "No-reference perceptual quality assessment of colour image," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2006, pp. 1–5.

[45] A. Maalouf and M.-C. Larabi, "A no reference objective color image sharpness metric," in *Proc. Eur. Signal Process. Conf.*

(EUSIPCO), Aug. 2010, pp. 1019–1022.

[46] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," *Proc. SPIE*, vol. 5007, pp. 87–95, Jun. 2003.

[47] Y. Y. Fu, "Color image quality measures and retrieval," Ph.D. dissertation, Dept. Comput. Sci., New Jersey Inst. Technol., Newark, NJ, USA, Jan. 2006.

[48] K. Panetta, C. Gao, and S. Agaian, "No reference color image contrast and quality measures," *IEEE Trans. Consum. Electron*., vol. 59, no. 3, pp. 643–651, Aug. 2013.

[49] Y. Y. Schechner and N. Karpel, "Recovery of underwater visibility and structure by polarization analysis," *IEEE J. Ocean. Eng*, vol. 30, no. 3, pp. 570–587, Jul. 2005.

[50] W. Hou, A. D. Weidemann, D. J. Gray, and G. R. Fournier, "Imagery-derived modulation transfer function and its applications for underwater imaging," *Proc. SPIE*, vol. 6696, pp. 22-29, Sep. 2007.

[51] A. Arnold-Bos, J. P. Malkasse, and G. Kervern, "Towards a model-free denoising of underwater optical images," in *Proc. IEEE Eur. Oceans Conf*., vol. 1. Brest, France, Jun. 2005, pp. 527–532.

[52] M. Arredondo and K. Lebart, "A methodology for the systematic assessment of underwater video processing algorithms," in *Proc. IEEE Eur. Oceans Conf*., vol. 1. Jun. 2005, pp. 362–367.

[53] M. Yang and A. Sowmya, "New image quality evaluation metric for underwater video," *IEEE Signal Process. Lett*., vol. 21, no. 10, pp. 1215–1219, Oct. 2014.

[54] Q. Xu, Q. Huang, and Y. Yao, "Online crowdsourcing subjective image quality assessment," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 359–368.

[55] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, "A crowdsourceable QoE evaluation framework for multimedia content," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 491–500.

[56] Y. Peng, and D. Doermann, "Active Sampling for Subjective Image Quality Assessment," *IEEE Conference on Computer Vision & Pattern Recognition IEEE Computer Society.*, pp. 4249-4256, Jun. 2014.

[57] Q. Xu, T. Jiang, Y. Yao, Q. Huang, B. Yan, and W. Lin, "Random partial paired comparison for subjective video quality assessment via hodgerank," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 393–402.

[58] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye, "Statistical ranking and combinatorial Hodge theory," *Math. Program*., vol. 127, no. 1, pp. 203–244, 2011.

[59] M. Jian, Q. Qi, J. Dong, Y. Yin, W. Zhang, and K. M. Lam, "The OUC-vision large-scale underwater image database," *IEEE International Conference on Multimedia and Expo.*, Jul. 2017, pp. 1297-1302,

[60] P. Erdos and A. Renyi, "On random graphs i. Publ. Math. Debrecen," vol. 6, pp. 290–297, 1959.

[61] R. J. Davies-Colley, "Measuring water clarity with a black disk," *Limnology and oceanography*., vol. 33, no. 4, pp. 616-623, Jul. 1988.

[62] W. C. Hou, "Practical test method for important indicators of urban surveillance system cameras," *Technology Innovation and Application*., pp. 65-66, 2013.

[63] N. Koren, "The Imatest program: Comparing cameras with different amounts of sharpening," *Electronic Imaging International Society for Optics and Photonics*, 2006.

[64] Z. Zhang, J. Zhau, N. Liu, N, X. Gu, and Y. Zhang, "An improved pairwise comparison scaling method for subjective image quality assessment," *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting.*, pp. 1-6, Jul. 2017.

[65] S. S. Stevens, E. C. Poulton, "The estimation of loudness by unpracticed observers," *Journal of Experimental Psychology.*, vol. 51, no. 1, pp. 71, Feb. 1956.

[66] E. Poulton, "Choice of first variables for single and repeated multiple estimates of loudness," *J. Exp. Psychol,* vol. 80, no. 2, pp. 249–253, Jun. 1969.

[67] M. A. Saad, *et al*., "Online subjective testing for consumer-photo quality evaluation," *Journal of Electronic Imaging.*, vol. 25, no. 4 pp. 043009, Jul. 2016.

[68] J. Lin, and Ivan V. Bajić, "A platform for subjective image quality evaluation on mobile devices," Electrical and Computer Engineering. IEEE, May. 2016.