2022 14th International Conference on Quality of Multimedia Experience (QoMEX)

Crowdsourced Quality Assessment of Enhanced Underwater Images – a Pilot Study

Hanhe Lin[1], Hui Men[2], Yijun Yan[3], Jinchang Ren[3], Dietmar Saupe[4]
[1]School of Science and Engineering, University of Dundee, United Kingdom
[2]Allgemeine und Biologische Psychologie, Philipps-Universität Marburg, Germany
[3]National Subsea Centre, Robert Gordon University, United Kingdom
[4]Department of Computer and Information Science, University of Konstanz, Germany

***Abstract*—Underwater image enhancement (UIE) is essential for a high-quality underwater optical imaging system. While a number of UIE algorithms have been proposed in recent years, there is little study on image quality assessment (IQA) of enhanced underwater images. In this paper, we conduct the first crowdsourced subjective IQA study on enhanced underwater images. We chose ten state-of-the-art UIE algorithms and applied them to yield enhanced images from an underwater image benchmark. Their latent quality scales were reconstructed from pair comparison. We demonstrate that the existing IQA metrics are not suitable for assessing the perceived quality of enhanced underwater images. In addition, the overall performance of 10 UIE algorithms on the benchmark is ranked by the newly proposed simulated pair comparison of the methods.**

***Index Terms*—underwater image, image quality assessment, image enhancement, crowdsourcing**

I. Introduction

In order to avoid putting humans into high-risk environments and to reduce emission and cost, underwater robots, e.g., remotely operated underwater vehicles, have been widely used in applications like inspection of underwater structures and ocean research. A high-quality underwater optical imaging system is an essential component in such devices. However, images taken underwater are usually degraded due to the light absorption and scattering in water. To compensate for this deficiency, underwater image enhancement (UIE) is often applied and has been a long standing research topic [1].

While the technology of UIE has made significant progress, there is little work on image quality assessment (IQA) on enhanced underwater images. Guo *et al.* [2] applied 5-point (Bad, Poor, Fair, Good, and Excellent) absolute category ratings (ACR) to rate the quality of five chosen UIE algorithms. However, the obtained subjective ratings may be unreliable since on the one hand, participants may have different interpretations of the ACR scale. On the other hand, UIE algorithms may introduce severe changes in the images as well as additional degradations which makes it harder to define and understand the quality of enhanced underwater images.

We conducted the first crowdsourced subjective IQA study on enhanced underwater images. Instead of using the ACR scale, we applied pair comparison to the enhanced images generated by 10 UIE algorithms in the 3D TURBID benchmark [3]. By reconstructing the quality scales from the pair comparison, we arrived at two results. (1) The existing IQA metrics are not suitable for assessing the perceived quality of enhanced underwater images. (2) The overall performance and ranking of 10 UIE algorithms on the benchmark.

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161 (Project A05).

II. Subjective study

A. 3D TURBID Benchmark

The 3D TURBID benchmark [3] was created in an controlled underwater environment, which is composed of a 1000 liter water tank, two LED lamps, and several planned scenes with 3D objects, e.g., stones and decorations, to simulate a real seabed. For each scene, a photo was first taken in clean water as a reference image. Next the water in the tank was diluted in a controlled way by adding specific particles, and a photo was taken. This procedure was repeated, resulting in 19 increasingly distorted images of the underwater scene.

To sum up, the released 3D TURBID dataset[1] contains four scenes, where one scene is degraded by blue ink, two are degraded by chlorophyll, and the last one by milk, see Fig. 1. Each scene contains one reference image taken with clean water and 19 degraded versions.

B. Selected turbidity level and UIE algorithms

Since the perceptual differences between two successive images in the original 3D TURBID image sequences is quite small, we did not use all the degraded images in our subjective study. Instead, for each scenario, we manually selected four images from 19 distorted images such that the visual quality of the distorted images varies perceptually linearly with the distortion amount. Fig. 1 shows the references, each together with their corresponding four degraded images, having been distorted by one of the three different degradation types.

We collected 10 UIE algorithms, with the source codes made available by their respective authors. These methods are Fusion-based [4], Retinex-based [5], CLAHE [6], UDCP [7], ICM [8], DCP [9], UCM [10], Rayleigh Distribution (RD) [11], RGHS [12], and Water-Net [13]. Among them, the former nine are conventional feature-based approaches,

[1]Available from http://amandaduarte.com.br/turbid/.

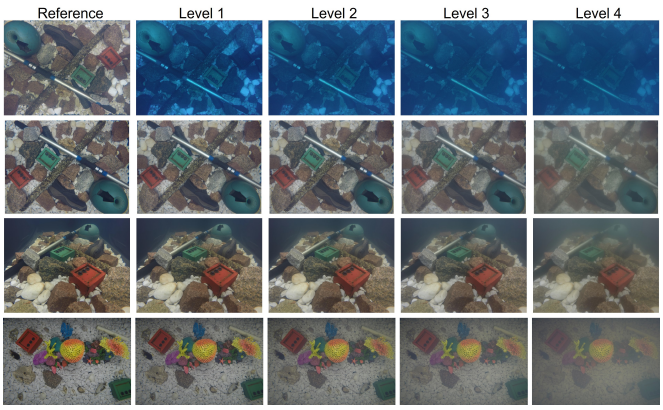

Fig. 1. Selected images from the 3D TURBID dataset for subjective IQA study. The first column contains reference images; the other columns correspond to four applied turbidity levels. The first row shows images degraded by blue ink, the scenes in the second and third row were degraded by chlorophyll, and the fourth row shows the effect of added milk.

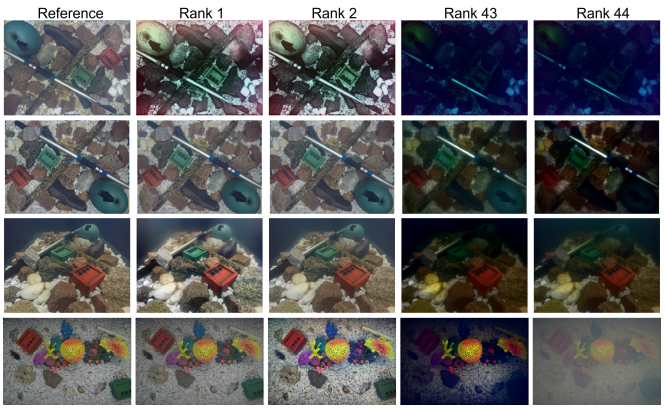

Fig. 2. Images of best two ranks (second and third columns) and worst two ranks (fourth and fifth columns) using reconstructed scores.

and the last one is a deep learning-based approach. For each degraded image, we applied the 10 UIE algorithms to generate 10 enhanced images.

C. Pair sampling

In our study, we have four reference images. Each reference has four degraded images. For each degraded image, apart from the 10 enhanced images, we also include the original degraded image for comparison. Eventually, for each reference image, there are $4 \times (1 + 10) = 44$ images to be compared.

Although ACR is widely used in generic IQA [14] [15], we preferred to apply pair comparison (PC) in our study to address the limitations of ACR, like disturbing observer bias, varying interpretations of the quality categories, and resulting nonlinear perceptual scales. In a typical PC test, two items are presented as pairs and participants are required to choose the preferred one in a forced choice setting.

For pair comparison we had to choose from all possible pairs in 44 images for each of the four reference images, giving a total of $4\binom{44}{2} = 3,784$ pairs. PC is most informative when compared test items have similar qualities. Therefore, we selected a subset of image pairs, based on similar SSIM [16] scores as follows.

1) Calculate for each enhanced image the SSIM w.r.t. its reference image.
2) Sort each of the 4 sets of 44 images derived for one reference image according to decreasing similarity.
3) Define a window size $W < 44$.
4) In each of the 4 sorted image sequences, select all pairs of images with a distance in the sorting of at most W.
5) Randomly swap the orientation of each pair, i.e., swap images in each pair with probability 1/2.

We chose the window size $W = 10$, which reduced the number of pairs from 3,784 to 1,540. This procedure reduced the size and cost of the experiment by omitting only pair comparison for which the responses can be expected to be obvious and, thus, not informative for the scale reconstruction.

D. Subjective underwater IQA study

Our subjective study was conducted on the Amazon Mechanical Turk (AMT) platform, on which *requesters*, e.g., companies or persons, create and submit *human intelligence tasks* (HITs) for *workers*. Workers submit their results w.r.t. HITs and get rewards for completion. Requesters would specify how many workers can submit their results for a HIT, i.e., number of *assignments*. In the study, three images were presented each time, where the images on the left and on the right are those for PC and the image in the middle is the reference image. Workers had 5 seconds to inspect the images. Then, the images disappeared, and workers had 3 seconds to choose the image (left or right) they considered more similar to the reference (middle). To reduce stress, a ternary choice (i.e., "left", "right", and "not sure") was used. If a worker did not make a choice within the total of 8 seconds, the answer was treated as "skipped".

In our study, each HIT contained 20 PC questions, where one of them was a test question (i.e., the proper answer was clear and already known). The rest were study questions. The test questions were generated by randomly sampling suitable pairs of original degraded images. In such a pair, the image with a lower distortion level has higher quality, is closer to the reference image, and thus is the ground truth answer that is expected to be given by an attentive crowd worker. After randomly partitioning 1,540 pairs into $1540/19 \approx 82$ HITs, we collected 9 assignments for each HIT. In total, we collected 13,860 responses for study questions.

To ensure the quality of the crowdsourcing study, we filtered out unreliable assignments. If in an assignment four or more questions were skipped or the hidden test question was answered incorrectly, the assignment was rejected. This reduced the number of ratings to 10,638.

III. Data Analysis and Results

Using the collected responses of pair comparison, we reconstructed quality scale values for each image based on Thurstone's model using the code provided by [17]. This was done separately for each of the four scenes, respectively sets of 44 images each. Since in this study, we had adopted pair

TABLE I
SRCC BETWEEN RECONSTRUCTED SUBJECTIVE SCORES AND OBJECTIVE SCORES OF STATE-OF-THE-ART OBJECTIVE IQA METHODS.

Method	Scene 1	Scene 2	Scene 3	Scene 4	Average
SSIM [16]	0.893	0.900	0.842	0.688	0.831
PSNR	0.830	0.782	0.785	0.436	0.708
VSI [19]	0.741	0.619	0.619	0.623	0.651
SCQI [20]	0.573	0.470	0.547	0.639	0.557
UCIQE [22]	0.096	0.426	0.304	0.173	0.250
FDUM [23]	0.229	0.125	0.079	0.156	0.147
UIQM [21]	0.129	0.093	0.226	0.084	0.133

comparison with a ternary choice, responses of type "not sure" were interpreted as two votes, one for "left" and one for "right", both weighted by 1/2, the same as in [18].

We ranked the enhanced images (for each reference) according to the obtained quality scale values. Fig. 2 presents the images of the top two and the bottom two ranks next to the corresponding reference image.

We evaluated the Spearman's rank correlation coefficient (SRCC) between the reconstructed subjective scores and objective scores of a few state-of-the-art objective IQA methods, including four full-reference (FR) IQA methods (i.e., PSNR, SSIM [16], VSI [19], SCQI [20]) and three no-reference (NR) underwater IQA methods, namely UIQM [21], UCIQE [22], and FDUM [23].

Although it has been demonstrated that generic FR-IQA methods perform quite well on authentic natural images, they have rather low correlations with the evaluations made by human observers for the enhanced underwater images, as shown in Table I. Among them, SSIM performed the best.

The NR-IQA methods performed even worse. Even though they were designed specifically for underwater imagery, this apparently could not compensate for the missing reference image that is available for FR-IQA methods.

Following the IQA for a benchmark of original and enhanced underwater images, the estimation of the success of the involved underwater image enhancement algorithms may be desired. In our small pilot study we did not compare all 10 participating enhancement methods for all underwater images in order to limit the number of comparisons. Thus, a simple statistic of which method won the most comparisons against its competitors was not possible. To solve this problem, we propose a new approach, based on the reconstructions of image qualities from incomplete pair comparison. This method is also applicable for larger studies with more source images and more enhancement methods for which it would be too costly to compare all methods for all images. Basically, we propose to use Thurstonian reconstruction of latent scores of the methods, based on simulated pair comparison of enhancement methods applied to the source images.

More precisely, we have $4 \times 4 = 16$ distorted underwater images I_j, $j = 1, .., 16$, and 10 enhanced versions each, $M_k(I_j)$, $k = 0, 1, ..., 10$, including the base case $k = 0$, indicating that no enhancement is applied. For each enhanced underwater image $M_k(I_j)$, we have one reconstructed quality score, s_{kj}. We used pair comparison between methods

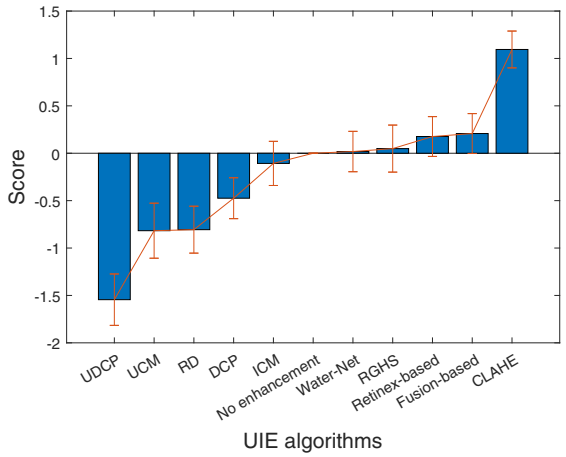

Fig. 3. Performance scores of 10 UIE algorithms. The scores were calibrated against the base case of performing no image enhancement, which was assigned a zero score.

Fig. 4. Comparison of the best and worst UIE methods, CLAHE and UDCP, for one of the original images. In the reconstructed quality scores, CLAHE beats UDCP 16 times (i.e., for all degraded images).

$M_k, M_l, k \neq l$ for all images I_j. M_k won over M_l for image I_j, iff $s_{kj} > s_{lj}$. This yielded $16 \times 10 \times 11/2 = 880$ comparisons, enough for the reconstruction of scores for the 11 methods. Corresponding confidence intervals were obtained by bootstrapping the original PCs to generate new sets of image quality scores s_{jk}. From each of these sets a new method score reconstruction was generated.

Fig. 3 shows that only half of the ten methods actually improved the visual image quality. The best method in this pilot study with 16 underwater images is CLAHE, a method based on based on histogram equalization. Fig. 4 shows an example with the best and the worst enhancement result.

IV. CONCLUSION

In this paper, we conducted pair comparison by crowdsourcing for assessing qualities of underwater images that were processed by ten modern image enhancement methods. After reconstruction, we demonstrated that the objective quality scores of existing IQA methods have a rather low correlation with the reconstructed scale values. From the reconstructed scale values we also simulated PC of the ten enhancement methods themselves and ranked the enhancement methods.

Our pilot study proves the feasibility of conducting underwater IQA via crowdsourcing. Future work should create a large-scale underwater IQA dataset by conducting a large study on authentic images and develop a more efficient objective underwater IQA metric.

REFERENCES

[1] M. Yang, J. Hu, C. Li, G. Rohde, Y. Du, and K. Hu, "An in-depth survey of underwater image enhancement and restoration," *IEEE Access*, vol. 7, pp. 123 638–123 657, 2019.

[2] P. Guo, L. He, S. Liu, D. Zeng, and H. Liu, "Underwater image quality assessment: Subjective and objective methods," *IEEE Transactions on Multimedia*, vol. 24, pp. 1980–1989, 2022.

[3] A. Duarte, F. Codevilla, J. D. O. Gaya, and S. S. Botelho, "A dataset to evaluate underwater image restoration methods," in *OCEANS 2016-Shanghai*, 2016, pp. 1–6.

[4] C. Ancuti, C. O. Ancuti, T. Haber, and P. Bekaert, "Enhancing underwater images and videos by fusion," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 81–88.

[5] X. Fu, P. Zhuang, Y. Huang, Y. Liao, X.-P. Zhang, and X. Ding, "A retinex-based enhancing approach for single underwater image," in *IEEE International Conference on Image Processing*, 2014, pp. 4572–4576.

[6] K. Zuiderveld, "Contrast limited adaptive histogram equalization," *Graphics Gems*, pp. 474–485, 1994.

[7] P. L. Drews, E. R. Nascimento, S. S. Botelho, and M. F. M. Campos, "Underwater depth estimation and image restoration based on single images," *IEEE Computer Graphics and Applications*, vol. 36, no. 2, pp. 24–35, 2016.

[8] K. Iqbal, R. A. Salam, A. Osman, and A. Z. Talib, "Underwater image enhancement using an integrated colour model." *IAENG International Journal of Computer Science*, vol. 34, no. 2, 2007.

[9] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341–2353, 2010.

[10] K. Iqbal, M. Odetayo, A. James, R. A. Salam, and A. Z. H. Talib, "Enhancing the low quality images using unsupervised colour correction method," in *IEEE International Conference on Systems, Man and Cybernetics*, 2010, pp. 1703–1709.

[11] A. S. A. Ghani and N. A. M. Isa, "Underwater image quality enhancement through composition of dual-intensity images and Rayleigh-stretching," *SpringerPlus*, vol. 3, no. 1, pp. 1–14, 2014.

[12] D. Huang, Y. Wang, W. Song, J. Sequeira, and S. Mavromatis, "Shallow-water image enhancement using relative global histogram stretching based on adaptive parameter acquisition," in *International Conference on Multimedia Modeling*. Springer, 2018, pp. 453–465
.

[13] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE Transactions on Image Processing*, vol. 29, pp. 4376–4389, 2019.

[14] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[15] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[17] J. Li, R. Mantiuk, J. Wang, S. Ling, and P. Le Callet, "Hybrid-MST: A hybrid active sampling strategy for pairwise preference aggregation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 3475–3485.

[18] H. Men, H. Lin, M. Jenadeleh, and D. Saupe, "Subjective image quality assessment with boosted triplet comparisons," *IEEE Access*, vol. 9, pp. 138 939–138 975, 2021.

[19] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270–4281, 2014.

[20] S.-H. Bae and M. Kim, "A novel image quality assessment with globally and locally consilient visual quality perception," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2392–2406, 2016.

[21] 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, 2015.

[22] M. Yang and A. Sowmya, "An underwater color image quality evaluation metric," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 6062–6071, 2015.

[23] N. Yang, Q. Zhong, K. Li, R. Cong, Y. Zhao, and S. Kwong, "A reference-free underwater image quality assessment metric in frequency domain," *Signal Processing: Image Communication*, vol. 94, p. 116218, 2021.