# A COMPREHENSIVE EVALUATION OF FULL REFERENCE IMAGE QUALITY ASSESSMENT ALGORITHMS

*Lin Zhang*[a], *Lei Zhang*[b*], *Xuanqin Mou*[c], *and David Zhang*[b]
[a] School of Software Engineering, Tongji University, Shanghai, China
[b] Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong
[c] Institute of Image Processing and Recognition, Xi'an Jiaotong University, China

## ABSTRACT

Recent years have witnessed a growing interest in developing objective image quality assessment (IQA) algorithms that can measure the image quality consistently with subjective evaluations. For the full reference (FR) IQA problem, great progress has been made in the past decade. On the other hand, several new large scale image datasets have been released for evaluating FR IQA methods in recent years. Meanwhile, no work has been reported to evaluate and compare the performance of state-of-the-art and representative FR IQA methods on all the available datasets. In this paper, we aim to fulfill this task by reporting the performance of eleven selected FR IQA algorithms on all the seven public IQA image datasets. Our evaluation results and the associated discussions will be very helpful for relevant researchers to have a clearer understanding about the status of modern FR IQA indices. Evaluation results presented in this paper are also online available at http://sse.tongji.edu.cn/linzhang/IQA/IQA.htm.

***Index Terms—*** Image quality assessment, SSIM, FSIM.

## 1. INTRODUCTION

Rapid proliferation of digital imaging and communication technologies has rendered the image quality assessment (IQA) an important issue in numerous applications. To this end, the scientific community has developed numerous automatic IQA methods in the past decades. According to the availability of a reference image, objective IQA indices can be classified as full reference (FR), no reference (NR) and reduced reference (RR) methods [1]. In this paper, our discussion is confined to FR methods.

It has been widely acknowledged that the conventional IQA indices, such as the peak signal-to-noise ratio (PSNR), which operate directly on the intensity of the image, do not correlate well with the subjective fidelity ratings. Thus, many efforts have been made on designing sophisticated computational IQA models and great progress has been achieved in this area in the past decade.

Due to historical reasons, most of the representative IQA indices were designed based on and evaluated only by the Laboratory for Image and Video Engineering (LIVE) dataset [2]. On seeing that LIVE only comprises limited distortion types (5 types of distortions in total), in order to provide more comprehensive test beds, several large scale image datasets for evaluating FR IQA algorithms have been established in recent years, such as the Tampere Image Database 2008 (TID2008) [3] and the Categorical Image Quality Database (CSIQ) [4]. In order to figure out clearly the status of the current FR IQA research, it is necessary to have a thorough evaluation of those state-of-the-art FR IQA metrics on all the available datasets. However, to the best of our knowledge, so far there is no work reported to conduct such a comprehensive evaluation.

In this paper, we try to fulfill this task by providing comprehensive evaluations of eleven selected representative FR IQA indices on all the seven publicly available datasets. For each selected IQA index, its prediction capability and running speed are both evaluated. The results provided will benefit a lot the interesting researchers to have a clearer understanding about the current research and development status of FR IQA methods. In addition, the elicited discussions based on the evaluation results are expected to inspire new thoughts for designing new IQA algorithms.

The rest of this paper is organized as follows. Section 2 briefly reviews the FR IQA indices selected for evaluation. Section 3 introduces the benchmark IQA image datasets and the performance metrics adopted. Section 4 presents in detail the evaluation results and the associated discussions. Finally, Section 5 concludes the paper.

## 2. FULL REFERENCE IQA ALGORITHMS

In the past decade, various FR IQA indices were proposed. From them, we select ten salient ones plus PSNR for evaluation in this paper. Generally speaking, the selected IQA indices are widely cited in the literature and have been reported to have reliable performance by researchers. In addition, for all of these selected IQA indices, their authors have released the source codes so that the results can be easily reproduced. The 11 FR IQA indices used in our

* Corresponding author. Email: cslzhang@comp.polyu.edu.hk.

evaluation include PSNR, the noise quality measure (NQM) index [5], the universal quality index (UQI) [6], the structural similarity (SSIM) index [1], the multi-scale SSIM (MS-SSIM) index [7], the information fidelity criterion (IFC) index [8], the visual information fidelity (VIF) index [9], the visual signal to noise ratio (VSNR) index [10], the information content weighted SSIM (IW-SSIM) index [11], the Riesz transforms based feature similarity (RFSIM) index [12], and the feature similarity (FSIM) index [13]. We briefly review them in the following.

NQM [5] and VSNR [10] are two representative Human Visual System (HVS)-based IQA models. They emphasize the importance of HVS's sensitivity to different visual signals, such as the luminance, the contrast, and the frequency content. SSIM [1] can be considered as a milestone of the development of FR IQA models. It was extended from its predecessor, the UQI index [6], and it is based on the hypothesis that HVS is highly adapted to extract the structural information from the visual scene. In their later work, Wang *et al*. proposed a multi-scale extension of SSIM, namely MS-SSIM [7], and it has been corroborated that MS-SSIM could produce better results than its single scale counterpart. In [11], Wang and Li improved the original MS-SSIM to IW-SSIM by introducing an information-content weighting (IW) based quality score pooling strategy. In IW-SSIM, local weights are assigned to SSIM map based on the total amount of information that can be extracted from the examined local patches in the reference and the distorted images. In [9], Sheikh *et al*. proposed the VIF index, which was an extension of its former version, i.e., the IFC index [8]. In VIF, Sheikh *et al*. treated the FR IQA problem as an information fidelity problem and the fidelity were quantified by the amount of information shared between the reference image and the distorted image. In [12], Zhang *et al*. proposed the RFSIM index. In RFSIM, $1^{st}$-order and $2^{nd}$-order Riesz transforms are used to characterize image's local structures and the Canny edge detector is employed to generate the mask for quality score pooling. The FSIM index proposed in [13] employs two features to compute the local similarity map, the phase congruency and the gradient magnitude. At the quality score pooling stage of FSIM, phase congruency map is utilized again as a weighting function since it can roughly reflect how perceptually important a local patch is to the HVS.

## 3. DATASETS AND PERFORMANCE METRICS

To the best of our knowledge, there are seven publicly available image datasets for evaluating FR IQA indices. They are TID2008 [3], CSIQ [4], LIVE [2], Image and Video Communication Database (IVC) [14], Media Information and Communication Technology Database (MICT) [15], Wireless Imaging Quality Database (WIQ) [16], and Cornell-A57 Database (A57) [10]. The important information of these seven datasets, in terms of the number of reference images, the number of distorted images, the number of quality distortion types, the image format (grayscale or colorful) and the number of subjects, is summarized in Table 1. Totally, there are 3832 distorted images with all these datasets.

Table 1: Benchmark image datasets for IQA

| Dataset | Ref. Images No. | Distorted Images No. | Distortion Types No. | Image Format | Subjects No. |
|---|---|---|---|---|---|
| TID2008 | 25 | 1700 | 17 | color | 838 |
| CSIQ | 30 | 866 | 6 | color | 35 |
| LIVE | 29 | 779 | 5 | color | 161 |
| IVC | 10 | 185 | 4 | color | 15 |
| MICT | 14 | 168 | 2 | color | 16 |
| WIQ | 7 | 80 | 5 | gray | 60 |
| A57 | 3 | 54 | 6 | gray | 7 |

Four commonly utilized performance metrics [11] to evaluate IQA indices are employed to evaluate the selected FR IQA indices in this paper, including Spearman rank-order correlation coefficient (SROCC), Kendall rank-order correlation coefficient (KROCC), Pearson linear correlation coefficient (PLCC), and Root Mean Squared Error (RMSE).

## 4. EVALUATION RESULTS AND DISCUSSIONS

### 4.1. Evaluation of prediction performance

In this section, the prediction performance measured by SROCC, KROCC, PLCC, and RMSE of each selected IQA index on each benchmark dataset is given. The results are listed in Table 2. For each performance measure, the two IQA indices producing the best results are highlighted in boldface. From the results listed Table 2, it can be seen that an IQA index may have quite different prediction performance on different datasets. Thus, in order to provide an evaluation of the overall performance of the FR IQA indices, in Table 3 we present their weighted-average SROCC, KROCC and PLCC results over seven datasets and the weight assigned to each dataset linearly depends on the number of distorted images contained in that dataset. The overall performance ranking of the evaluated IQA indices based on three different performance metrics, SROCC, KROCC, and PLCC, is presented in Table 4.

### 4.2. Evaluation of running speed

We also evaluated the running speed of each selected IQA index. Experiments were performed on a Dell Inspiron 530s PC embedded with an Intel E6550 processor and 2GB RAM. The software platform was Matlab R2009a. The average time cost consumed by each IQA index for measuring the similarity of a pair of 384×512 color images (taken from TID2008) is listed in Table 5.

Table 2: Performance comparison of 11 IQA indices on 7 benchmark datasets

| | | PSNR | NQM | UQI | SSIM | MS-SSIM | IFC | VIF | VSNR | IW-SSIM | RFSIM | FSIM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TID 2008 | SROCC | 0.5531 | 0.6243 | 0.5851 | 0.7749 | 0.8542 | 0.5675 | 0.7491 | 0.7046 | 0.8559 | **0.8680** | **0.8805** |
| | KROCC | 0.4027 | 0.4608 | 0.4255 | 0.5768 | 0.6568 | 0.4236 | 0.5860 | 0.5340 | 0.6636 | **0.6780** | **0.6946** |
| | PLCC | 0.5734 | 0.6142 | 0.6643 | 0.7732 | 0.8451 | 0.7340 | 0.8084 | 0.6820 | 0.8579 | **0.8645** | **0.8738** |
| | RMSE | 1.0994 | 1.0590 | 1.0031 | 0.8511 | 0.7173 | 0.9113 | 0.7899 | 0.9815 | 0.6895 | **0.6746** | **0.6525** |
| CSIQ | SROCC | 0.8058 | 0.7402 | 0.8098 | 0.8756 | 0.9133 | 0.7671 | 0.9195 | 0.8106 | 0.9213 | **0.9295** | **0.9242** |
| | KROCC | 0.6084 | 0.5638 | 0.6188 | 0.6907 | 0.7393 | 0.5897 | 0.7537 | 0.6247 | 0.7529 | **0.7645** | **0.7567** |
| | PLCC | 0.8000 | 0.7433 | 0.8312 | 0.8613 | 0.8991 | 0.8384 | **0.9277** | 0.8002 | 0.9144 | **0.9179** | 0.9120 |
| | RMSE | 0.1575 | 0.1756 | 0.1460 | 0.1334 | 0.1149 | 0.1431 | **0.0980** | 0.1575 | 0.1063 | **0.1042** | 0.1077 |
| LIVE | SROCC | 0.8756 | 0.9086 | 0.8941 | 0.9479 | 0.9513 | 0.9259 | **0.9636** | 0.9274 | 0.9567 | 0.9401 | **0.9634** |
| | KROCC | 0.6865 | 0.7413 | 0.7100 | 0.7963 | 0.8045 | 0.7579 | **0.8282** | 0.7616 | 0.8175 | 0.7816 | **0.8337** |
| | PLCC | 0.8723 | 0.9122 | 0.8987 | 0.9449 | 0.9489 | 0.9268 | **0.9604** | 0.9231 | 0.9522 | 0.9354 | **0.9597** |
| | RMSE | 13.3597 | 11.1926 | 11.9823 | 8.9455 | 8.6188 | 10.2643 | **7.6137** | 10.505 | 8.3473 | 9.6642 | **7.6780** |
| IVC | SROCC | 0.6884 | 0.8347 | 0.8244 | 0.9018 | 0.8980 | 0.8993 | 0.8964 | 0.7983 | **0.9125** | 0.8192 | **0.9262** |
| | KROCC | 0.5218 | 0.6342 | 0.6252 | 0.7223 | 0.7203 | 0.7202 | 0.7158 | 0.6036 | **0.7339** | 0.6452 | **0.7564** |
| | PLCC | 0.7196 | 0.8498 | 0.8302 | 0.9119 | 0.9108 | 0.9093 | 0.9028 | 0.8032 | **0.9231** | 0.8361 | **0.9376** |
| | RMSE | 0.8460 | 0.6421 | 0.6792 | 0.4999 | 0.5029 | 0.5069 | 0.5239 | 0.7258 | **0.4686** | 0.6684 | **0.4236** |
| MICT | SROCC | 0.6132 | 0.8911 | 0.7028 | 0.8794 | 0.8874 | 0.8354 | **0.9077** | 0.8614 | **0.9202** | 0.7731 | 0.9059 |
| | KROCC | 0.4443 | 0.7129 | 0.5227 | 0.6939 | 0.7029 | 0.6370 | **0.7315** | 0.6762 | **0.7537** | 0.5752 | 0.7302 |
| | PLCC | 0.6429 | 0.8955 | 0.7164 | 0.8887 | 0.8927 | 0.8403 | **0.9138** | 0.8710 | **0.9248** | 0.7783 | 0.9078 |
| | RMSE | 0.9585 | 0.5569 | 0.8731 | 0.5738 | 0.5640 | 0.6784 | **0.5084** | 0.6147 | **0.4761** | 0.7857 | 0.5248 |
| WIQ | SROCC | 0.6257 | 0.7644 | 0.6084 | 0.7261 | 0.7495 | 0.7159 | 0.6918 | 0.6558 | **0.7865** | 0.7368 | **0.8006** |
| | KROCC | 0.4626 | 0.5803 | 0.4360 | 0.5569 | 0.5740 | 0.5290 | 0.5246 | 0.4873 | **0.6038** | 0.5493 | **0.6215** |
| | PLCC | 0.7939 | 0.8170 | 0.6974 | 0.7980 | 0.8095 | 0.7678 | 0.7605 | 0.7736 | **0.8329** | 0.8103 | **0.8546** |
| | RMSE | 14.138 | 13.209 | 16.416 | 13.805 | 13.449 | 14.675 | 14.873 | 14.515 | **12.677** | 13.424 | **11.895** |
| A57 | SROCC | 0.6189 | 0.7981 | 0.4260 | 0.8066 | 0.8414 | 0.3185 | 0.6223 | **0.9355** | 0.8709 | 0.8215 | **0.9181** |
| | KROCC | 0.4309 | 0.5932 | 0.3330 | 0.6058 | 0.6478 | 0.2378 | 0.4589 | **0.8031** | 0.6842 | 0.6324 | **0.7639** |
| | PLCC | 0.7073 | 0.8271 | 0.6356 | 0.8017 | 0.8603 | 0.5772 | 0.6915 | **0.9502** | 0.9034 | 0.8475 | **0.9393** |
| | RMSE | 0.1737 | 0.1381 | 0.1897 | 0.1469 | 0.1253 | 0.2007 | 0.1784 | **0.0766** | 0.1054 | 0.1305 | **0.0844** |

Table 3: Overall performance of IQA indices over 7 datasets

| IQA Index | SROCC | KROCC | PLCC |
|---|---|---|---|
| PSNR | 0.6874 | 0.5161 | 0.7020 |
| NQM | 0.7355 | 0.5649 | 0.7349 |
| UQI | 0.7137 | 0.5398 | 0.7602 |
| SSIM | 0.8430 | 0.6593 | 0.8407 |
| MS-SSIM | 0.8885 | 0.7087 | 0.8831 |
| IFC | 0.7128 | 0.5524 | 0.8084 |
| VIF | 0.8423 | 0.6827 | 0.8728 |
| VSNR | 0.7875 | 0.6132 | 0.7776 |
| IW-SSIM | **0.8955** | **0.7215** | **0.8960** |
| RFSIM | 0.8866 | 0.7092 | 0.8845 |
| FSIM | **0.9094** | **0.7409** | **0.9050** |

Table 4: Overall performance ranking of IQA indices

| IQA Index | SROCC | KROCC | PLCC |
|---|---|---|---|
| PSNR | 11 | 11 | 11 |
| NQM | 8 | 8 | 10 |
| UQI | 9 | 10 | 9 |
| SSIM | 5 | 6 | 6 |
| MS-SSIM | 3 | 4 | 4 |
| IFC | 10 | 9 | 7 |
| VIF | 6 | 5 | 5 |
| VSNR | 7 | 7 | 8 |
| IW-SSIM | 2 | 2 | 2 |
| RFSIM | 4 | 3 | 3 |
| FSIM | 1 | 1 | 1 |

Table 5: Time cost of each IQA index

| IQA Index | Time (milliseconds) |
|---|---|
| PSNR | 14.3 |
| NQM | 545.2 |
| UQI | 105.8 |
| SSIM | 45.2 |
| MS-SSIM | 141.7 |
| IFC | 3352.9 |
| VIF | 3399.9 |
| VSNR | 382.8 |
| IW-SSIM | 870.6 |
| RFSIM | 219.4 |
| FSIM | 705.3 |

### 4.3. Discussions

Based on the evaluation results, we can have the following findings. First of all, all the IQA indices deliberately designed to have a quality prediction capability consistent with human perceptions can get better performance than PSNR, which is a widely used pixel-based quality index.

Secondly, objective scores predicted by three recently proposed indices, FSIM [13], IW-SSIM [11], and RFSIM [12], are highly consistent with the subjective evaluations. They can provide statistically better prediction accuracy than the other methods evaluated. Especially, no matter which criterion is used, FSIM can always achieve the best overall performance. The superiority of FSIM, IW-SSIM,

and RFSIM can be partially attributed to the quality score pooling strategies they adopt. These three indices all use a spatially varying weighting function to indicate the different significance of different local image regions at the quality score pooling stage. It is believed that with a more proper score pooling strategy, better performance of an FR IQA index can be obtained. Thus, how to devise a more powerful score pooling strategy is a promising direction for the FR IQA research. Furthermore, it needs to point out that NQM and VSNR are two typical pure HVS-based IQA methods, trying to systematically model relevant psychophysical and physiological properties of HVS. Nevertheless, since how to model the components and their interactions of HVS is itself still a challenging topic, NQM and VSNR cannot perform as good as other signal-driven methods in most cases. However, this does not mean that HVS knowledge can be disregarded. On the contrary, though they do not depend on fundamental vision modeling, most of the signal-driven methods actually make use of the HVS knowledge implicitly.

Thirdly, for some IQA indices, they may work well on some datasets but fail to provide good results on other datasets. For example, though VIF can get very pleasing results on LIVE, it performs quite poor on TID2008, WIQ, and A57. Similarly, VSNR can get very good results on A57 while it performs much poorer than most of the other indices on the rest datasets. One possible reason is that VIF (VSNR) may be over-tuned on LIVE (A57). On the other hand, it also indicates that different IQA datasets have quite different characteristics and different capabilities in assessing IQA indices. Though the latest TID2008 dataset is much larger than the previously published datasets in terms of the number of images, the number of quality distortion types, and the number of subjects, it still has some deficiencies to some extent. For example, multiply distorted images are not considered in TID2008. Hence, constructing an even more comprehensive dataset is of tremendous significance to the IQA research and it still requires further efforts by researchers in this field.

At last, with respect to the running speed, VIF and IFC perform much poorer than the others. With our experimental settings, they both need more than 3 seconds for computing the similarity between a pair of images. Their low speed should be attributed to the multi-orientation and multi-scale wavelet decomposition required by them. FSIM, IW-SSIM, and RFSIM, methods that have pleasing prediction performance have moderate computation complexity. Among all the modern IQA indices evaluated, SSIM runs the fastest. Additionally, it can also achieve acceptable prediction performance. That is why SSIM has been so widely used after its birth. Thus, how to devise an IQA index that can have FSIM-like prediction performance and SSIM-like computation simplicity still needs further efforts.

## 5. CONCLUSION

In this paper, we extensively evaluated 11 selected FR IQA indices on all the seven publicly available IQA image datasets. Their prediction performance and the running speed were reported. Such evaluations can facilitate the IQA research by presenting researchers clearly the current status of modern FR IQA methods. Furthermore, constructive discussions based on the evaluation results were presented, aiming to inspire new insights for the further IQA research.


## ACKNOWLEDGEMENT

This research is supported by the Fundamental Research Funds for the Central Universities (2100219033), and the HK RGC GRF Grant (PolyU 5375/09E).


## 6. REFERENCES


[1] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. IP*, vol. 13, pp. 600-612, 2004.

[2] H.R. Sheikh, M.F. Sabir, and A.C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. IP*, vol. 15, pp. 3440-3451, 2006.

[3] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008 - A database for evaluation of full-reference visual quality assessment metrics," *Adv. Modern Radioelectron.*, vol. 10, pp. 30-45, 2009.

[4] E.C. Larson and D.M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *J. Electr. Imaging*, vol. 19, pp. 001006:1-21, 2010.

[5] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, and A.C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. IP*, vol. 9, pp. 636-650, 2000.

[6] Z. Wang and A.C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, pp. 81-84, 2002.

[7] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multi-scale structural similarity for image quality assessment," *ACSSC'03*, pp. 1398-1402, 2003.

[8] H.R. Sheikh, A.C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. IP*, vol. 14, pp. 2117-2128, 2005.

[9] H.R. Sheikh and A.C. Bovik, "Image information and visual quality," *IEEE Trans. IP*, vol. 15, pp. 430-444, 2006.

[10] D.M. Chandler and S.S. Hemami, "VSNR: a wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. IP*, vol. 16, pp. 2284-2298, 2007.

[11] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. IP*, vol. 20, pp. 1185-1198, 2011.

[12] L. Zhang, L. Zhang, and X. Mou, "RFSIM: a feature based image quality assessment metric using Riesz transforms," *ICIP'10*, pp. 321-324, 2010.

[13] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: a feature similarity index for image quality assessment," *IEEE Trans. IP*, vol. 20, pp. 2378-2386, 2011.

[14] Subjective quality assessment IRCCyN/IVC database, http://www2.irccyn.ec-nantes.fr/ivcdb

[15] MICT Image Quality Evaluation Database, http://mict.eng.u-toyama.ac.jp/mictdb.html

[16] U. Engelke, M. Kusuma, H.J. Zepernick, and M. Caldera, "Reduced-reference metric design for objective perceptual quality assessment in wireless imaging," *Signal Process.: Image Communication*, vol. 24, pp. 525-547, 2009.