RANKING CONSISTENT RATE: NEW EVALUATION CRITERION ON PAIRWISE SUBJECTIVE EXPERIMENTS

Yeji Shen, Tingting Jiang

National Engineering Laboratory for Video Technology
Cooperative Medianet Innovation Center
School of Electronics Engineering and Computer Science, Peking University

ABSTRACT

Subjective experimental results are widely used as the ground truth in objective Image Quality Assessment (IQA). Specifically, Pairwise Comparison method has superiority over Mean Opinion Scores (MOS), but there is a problem when measuring the consistency between subjective pairwise comparisons and objective quality predictions. In this paper, we first analyze the existing problem of current evaluation method for the consistency between the pairwise comparisons given by human subjects and the ranking results given by objective IQA algorithms. Then we propose a new direct evaluation method, Ranking Consistent Rate, to solve this problem. Moreover, through our method, we can check the self-consistency of datasets based on pairwise comparisons and evaluate the performance of an IQA algorithm more accurately.

Index Terms— Subjective Image Quality Assessment, Paired Comparison, Mean Opinion Score

1. INTRODUCTION

As digital images are universally produced, presented and perceived in our everyday life, it is important to know the nature of human's perception of images. Image Quality Assessment (IQA) aims at automatic evaluation of image's quality, which represents human's overall perception of images. Currently, the evaluation of quality is usually given as *Quality of Experience (QoE)* [1], indicating the degree of a user's satisfaction.

In applications of IQA, subjective experiments are used as the ground truth of evaluation. Therefore, correlation between the subjective experimental results and the objective predictions given by an IQA algorithm is measured to indicate the consistency between the subjective and the objective results. The performance of the IQA model can also be shown in this correlation. Typically, SROCC is the most commonly used method[2, 3, 4].

In subjective experiments, MOS (Mean Opinion Score) and PC (Pairwise Comparison) are two most commonly used methods. In MOS framework, subjects are to report scores ranging from 1 to 5, representing their opinions of "Poor", "Bad", "Fair", "Good" and "Excellent", respectively. Then, mean score among all subjects is the ground truth of the quality score of the image. In contrast, in PC experiments, subjects are to give direct answer on which one in a pair is better rather than a score. Besides, in [5], the quality of an image is represented in the form of a vector, which is quite different from both MOS and PC experiments.

Compared with MOS, PC has superiority in subjective consistency and workload. MOS rating system suffers from *Scale Usage Heterogeneity* problem [6], which means different interpretation of one rating scale for different subjects. This problem is avoided in PC where subjects have no need to make the absolute decisions on objects. This feature provides PC with higher subjective consistency. Moreover, according to [1], PC is considered to has lower workload for subjects.

Although increasing popularity can be observed on PC experiments [1, 4, 7, 8], some inevitable problems will arise when objective rankings given by an IQA algorithm is judged by PC experimental results. In order to deal with this inconsistency between the subjective and the objective, pairs are usually converted to a global ranking [1], or MOS-like scores in TID2008 [7]. There are some research on the methodology of global ranking [9, 10], including famous BTL model [11]. As Figure 1 indicates, there are some cases where there exists a subset of images of similar quality in PC experiments.

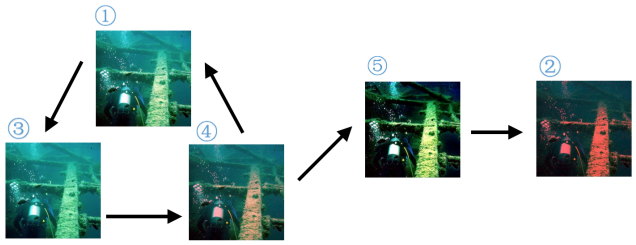
In this paper, we propose a new method, Ranking Consistent Rate (RCR), for the evaluation of the consistency between the subjective pairs and objective rankings given by IQA algorithms. This method has two benefits: one is that it can overcome the inconsistency problem of traditional method, second is that it can also evaluate the self-consistency of subjective pairwise experimental results through Intrinsic Contradiction Rate (ICR), which is a direct application of RCR. We conduct PC experiments on images selected from TID2013 [8]. Our method is performed on this dataset and PKU-EAQA [4]. Our experimental results of RCR can offer a better evaluation of an IQA algorithm with higher consistency with subjective data, while our ICR results shows that our dataset has better self-consistency than PKU-EAQA.

2. PROBLEM AND METHOD

In this section, we firstly describe the traditional evaluating method of consistency between the subjective and the objective. Then, we analyze the problem of this traditional method. After that, we propose a new method to solve this problem. At last, we discuss another usage of our proposed method.

2.1. Traditional Evaluation of Consistency

Consistency between the subjective pairwise experimental result and the objective ranking can be defined as a metric function $f(P, R)$, where P is the result of subjective pairwise experiments, see Figure 1. And R is the ranking result of an objective IQA algorithm. By convention, we assume P to be a matrix, where $P_{i,j}$ is the number of votes that i^{th} image is better than j^{th} image, and R to be a vector, where R_i is the rank of i^{th} image. Smaller rank number means better quality.



(a) A group of images with ambiguous subset.

	1	2	3	4	5
1	0	60	33	41	52
2	0	0	0	2	5
3	27	60	0	37	53
4	19	58	23	0	46
5	8	55	7	14	0

(b) Corresponding P matrix.

Fig. 1. From (a), we can see image (1), (3) and (4) are ambiguous in quality, where the arrow from image i to image j means image i is considered to be better than image j. This ambiguity can be confirmed in P, where $P_{i,j}$ is the number of votes that i^{th} image is better than j^{th} image. From (b), we can see close voting numbers among image (1), (3) and (4).

Typically, an IQA algorithm will produce the quality score of an image as output. We can compute all the scores of an image group and sort the scores into a ranking. Then, this objective ranking can be compared to subjective ranking if subjective experiments are did in MOS framework where similar sorting can be performed. In MOS, it is natural to use coefficients of correlation, like SROCC, to evaluate the consistency of the objective ranking and the subjective experiments:

$$SROCC = 1 - \frac{6\sum_i^n (x_i - y_i)^2}{n(n^2-1)}, \tag{1}$$

where $\mathbf{x}$ and $\mathbf{y}$ are two rankings and n is the number of images. However, objective ranking cannot be directly compared to pairwise subjective experiments. Traditional solution to this problem is to convert subjective pairs into rankings before computing the coefficients of correlation using some global ranking models [9, 10]. Assume the global ranking model as a function g. For any P, $R = g(P)$ is the most convincing ranking based on given P.

Therefore, we taking SROCC as an example, traditional evaluation of consistency can be expressed as:

$$f_{old}(P, R) = SROCC(g(P), R). \tag{2}$$

2.2. Problems of Traditional Evaluation

Because it is hard to quantify how good an evaluation method is, we use several typical examples to illustrate the problems of traditional SROCC-based method.

In Figure 1, we can see that image (1), image (3) and image (4) are quite similar, but significantly different with the other two. These three images are intrinsically ambiguous even for human subjects. From matrix P in Figure 1, we can further find that the numbers of votes for pair $(1, 3)$, $(3, 4)$ and $(1, 4)$ are close, supporting our observation. In this case, any rankings of quality scores make sense among image (1), (3) and (4), which means any rankings are also equally improper. Therefore, producing a global ranking and computing SROCC will improperly ignore the ambiguity within certain subsets and potentially exaggerate the mistake among the ambiguous subset.

This phenomenon is originated from the fact that most current IQA algorithms have an assumption that there exists a perfect global ranking which can be extracted from PC results. However, this may not be true all the time. We can also take the visual JND (Just Noticeable Difference) effect as an analogy. Human beings are naturally not sure about the exact ranking of image qualities when they are within the range of JND effect.

2.3. Ranking Consistent Rate

In order to solve the previous problem on intrinsic contradiction of subjective pairs, we propose an new method of evaluation of consistency, Ranking Consistent Rate:

$$RCR(P, R) = \frac{\sum_{i,j} 1\{R_i < R_j\} P_{i,j}}{\sum_{i,j} P_{i,j}}, \tag{3}$$

where P is the voting results in pairwise subjective experiments and R is an objective ranking. We can see that RCR is defined as the ratio of consistent pairs in P with respect to the ranking R.

Instead of computing the coefficients of correlation after conversion from pairs to global rankings, we directly utilize

the correspondence information of subjective pairs P and objective ranking R. Through RCR as an index of the evaluation of consistency, we can achieve a more reliable and more robust evaluation of objective IQA algorithms.

Finally, we will have our new metric function f for the evaluation of consistency between the subjective and the objective:

$$f_{new}(P, R) = RCR(P, R). \tag{4}$$

2.4. Application of RCR

When the ranking given to RCR is subjective rather than objective, the result of RCR will represent the self-consistency of subjective data. First, we can then define Ground Truth Ranking(GTR) as the most convincing ranking:

$$GTR(P) = \underset{R}{argmax}\, RCR(P, R), \tag{5}$$

and corresponding maximum RCR is defined as Intrinsic Contradiction Rate(ICR):

$$ICR(P) = 1 - RCR(P, GTR(P)) \tag{6}$$

ICR is an indicator of the self-consistency of the dataset. If ICR of a group of images is too large, we can assert that ambiguous images dominate this image group, which should be considered invalid accordingly.

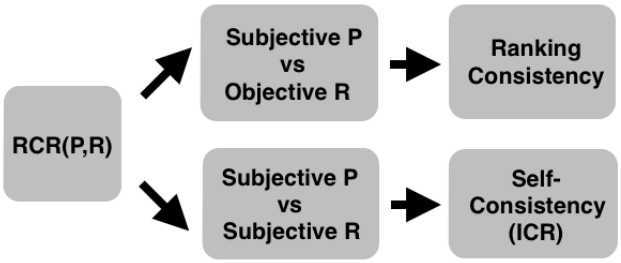


Fig. 2. Diagram for two utilities of RCR: one as the evaluation criterion for objective ranking, the other as an indicator of self-consistency for pairwise data.

3. EXPERIMENTAL RESULTS

In this section, we perform our proposed RCR method on two datasets, PKU-EAQA and our dataset. First, the datasets used are briefly introduced. Then, we show the results of RCR method and SROCC-based method on two datasets with respect to five objective IQA algorithms, followed by examples illustrating the problems of SROCC-based method. Finally, we compute ICR on two datasets as an indicator of their self-consistency.

3.1. Datasets

PKU-EAQA [4] is a dataset on Enhanced Image Quality Assessment comprised of a total of 1500 images. In PKU-EAQA, each of 300 reference images of 3 different scenes, including 100 haze images, 100 water images and 100 low-light images, is processed by 5 corresponding enhancement algorithms, e.g. de-hazing for haze images, de-blurring for images taken in water and lightening for dim (low-light) images. Therefore, 500 enhanced images are produced in each scene. PC experiments were provided among each group of 5 enhanced images based on the same reference image. 30 comparisons were made in each pair.

In our dataset, 100 distorted images derived from 4 reference images from TID2013 [8] are presented, where each reference image has 5 distortion types and 5 distortion levels. Distortion types are Gaussian noise (GN), Gaussian blur (GB), JPEG lossy compression (JPEG), JPEG2000 lossy compression (JPEG2K) and non-traditional distortion (NT). PC experiments were did among every pair of distorted images with the same reference. 8 to 10 comparisons were made in each pair.

3.2. Results

We perform Chen's Ranking SVM based algorithm [4], Mittal's BRISQUE [12], Moorthy's DIIVINE, Saad's BLIINDS-II and Wang's SSIM [2] on PKU-EAQA. Their results are evaluated by both traditional SROCC-based method as defined in Eq.2 where g is GTR in Eq.5, and RCR method as defined in Eq.4. Detailed results are shown in Table 1 and 2. All the images were randomly

	haze	water	night
Chen's algorithm	0.701	0.714	0.798
Mittal's BRISQUE	0.462	0.394	0.633
Moorthy's DIIVINE	0.649	0.516	0.524
Saad's BLIINDS-II	0.603	0.613	0.388
Wang's SSIM	0.204	0.400	0.728

Table 1. Results of RCR in PKU-EAQA

	haze	water	night
Chen's algorithm	0.635	0.785	0.909
Mittal's BRISQUE	0.421	0.491	0.452
Moorthy's DIIVINE	0.627	0.377	0.510
Saad's BLIINDS-II	0.537	0.529	0.283
Wang's SSIM	-0.105	0.325	0.875

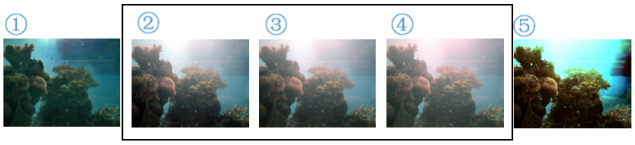
Table 2. Results of SROCC in PKU-EAQA

In Figure 3, we can see two rankings [1] given by Chen's algorithm with the same SROCC but quite different RCR. Though

[1] The index of images are re-arranged for the convenience of presentation. Next example is the same case.

both rankings are exactly the same as the converted global ranking $g(P) = [1, 2, 3, 4, 5]$, we believe that ranking (a) is better than ranking (b). This is because the difference of the images in (a) is very clear. So we know its consistency with subjective experiments is relatively high. In contrast, the middle 3 images in (b) are similar in quality, reducing its reliability and leading to a low RCR.

(a) $SROCC = 1, RCR = 0.918, g = [1, 2, 3, 4, 5]$



(b) $SROCC = 1, RCR = 0.717, g = [1, 2, 3, 4, 5]$

	1	2	3	4	5
1	0	52	48	58	58
2	8	0	52	56	60
3	12	8	0	54	57
4	2	4	6	0	56
5	2	0	3	4	0

(c) P matrix for (a)

	1	2	3	4	5
1	0	39	46	41	56
2	21	0	41	42	52
3	14	19	0	35	43
4	19	18	25	0	40
5	4	8	17	20	0

(d) P matrix for (b)

Fig. 3. Example of two groups of images with the same SROCC but very different RCR

In Figure 4, there are two rankings of images with very close RCR but different SROCC. We can see the first 3 images in (a) are similar in quality and the middle 3 images in (b) are also similar. Though the RCR indicates that the two rankings have close degree of consistency, ranking (a) makes an inverse prediction (compared with ground truth) on the first 3 images, while ranking (b) "luckily" makes a correct prediction on those ambiguous subsets.

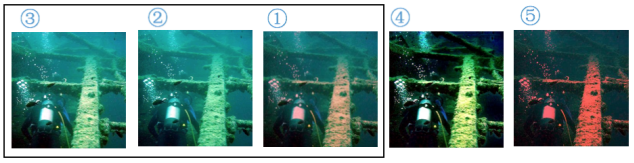
From Figure 3 and Figure 4, we can see that SROCC is not a good evaluation criterion of consistency, because its results are not consistent with human's perception.

3.3. Results of ICR

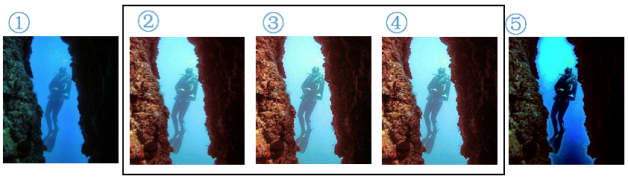
From Figure 3.3 (a), we can see that our dataset has lower ICR than PKU-EAQA, and thus has higher self-consistency. This can also be confirmed by P matrix in (b)

4. CONCLUSION

In PC experiments, traditional SROCC-based method of the evaluation of the consistency between subjective pairs and objective rankings suffers from ambiguous subset in term of image quality. Our proposed method, RCR can overcome this problem, and provide us with a more reliable evaluation on the performance of an objective IQA algorithms more accurately. Moreover, we can use ICR, which is an application of RCR, to check the self-consistency of a PC-based dataset.



(a) $SROCC = 0.6, RCR = 0.755, g = [1, 2, 3, 4, 5]$



(b) $SROCC = 1, RCR = 0.742, g = [1, 2, 3, 4, 5]$

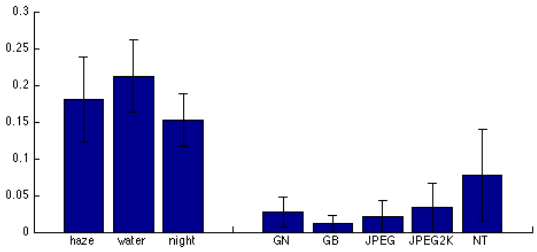
	1	2	3	4	5
1	0	23	19	46	58
2	37	0	27	53	60
3	41	33	0	52	60
4	14	7	8	0	55
5	2	0	0	5	0

(c) P matrix for (a)

	1	2	3	4	5
1	0	48	46	49	54
2	12	0	34	40	48
3	14	26	0	36	44
4	11	20	24	0	46
5	6	12	16	14	0

(d) P matrix for (b)

Fig. 4. Example of two groups of images with close RCR but very different SROCC



(a) Results of IRC.

	1	2	3	4	5
1	0	5	11	10	9
2	2	0	7	9	8
3	0	1	0	10	9
4	0	0	0	0	10
5	0	0	0	0	0

(b) A P matrix for GB of reference image (1).

Fig. 5. In (a), the left three bars are results of PKU-EAQA and the right five bars are results of our dataset. In (b), we can see that most pairs conform to the global ranking $g = [1, 2, 3, 4, 5]$ which is also consistent with its distortion level.

5. ACKNOWLEDGMENT

This work was partially supported by National Basic Research Program of China (973 Program) under contract 2015CB351803 and the Natural Science Foundation of China under contracts 61572042, 61390514, 61421062, 61210005, 61527084, as well as the grant from Microsoft Research-Asia.

6. REFERENCES

[1] Kuan-Ta Chen, Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei, "A crowdsourceable QoE evaluation framework for multimedia content," in *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 2009, pp. 491–500.

[2] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

[3] Hamid Rahim Sheikh, Alan Conrad Bovik, and Gustavo De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *Image Processing, IEEE Transactions on*, vol. 14, no. 12, pp. 2117–2128, 2005.

[4] Zhengying Chen, Tingting Jiang, and Yonghong Tian, "Quality assessment for comparing image enhancement algorithms," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 3003–3010.

[5] Ou Wu, Weiming Hu, and Jun Gao, "Learning to predict the perceived visual quality of photos," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 225–232.

[6] Peter E Rossi, Zvi Gilula, and Greg M Allenby, "Overcoming scale usage heterogeneity: A bayesian hierarchical approach," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 20–31, 2001.

[7] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, M Carli, and F Battisti, "Tid2008-a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009.

[8] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al., "Image database tid2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015.

[9] Qianqian Xu, Qingming Huang, and Yuan Yao, "Online crowdsourcing subjective image quality assessment," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 359–368.

[10] Qianqian Xu, Jiechao Xiong, Qingming Huang, and Yuan Yao, "Robust evaluation for quality of experience in crowdsourcing," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 43–52.

[11] Ralph Allan Bradley and Milton E Terry, "Rank analysis of incomplete block designs the method of paired comparisons," *Biometrika*, vol. 39, no. 3-4, pp. 324–345, 1952.

[12] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *Image Processing, IEEE Transactions on*, vol. 21, no. 12, pp. 4695–4708, 2012.