IEEE International Conference on Image Processing, Beijing, China, Sep. 2017.

QUALITY ASSESSMENT OF IMAGES UNDERGOING MULTIPLE DISTORTION STAGES

Shahrukh Athar, Abdul Rehman and Zhou Wang

Dept. of Electrical & Computer Engineering, University of Waterloo, Waterloo, ON, Canada
Email: {shahrukh.athar, abdul.rehman, zhou.wang}@uwaterloo.ca

ABSTRACT

In practical media distribution systems, visual content often undergoes multiple stages of quality degradations along the delivery chain between the source and destination. By contrast, current image quality assessment (IQA) models are typically validated on image databases with a single distortion stage. In this work, we construct two large-scale image databases that are composed of more than 2 million images undergoing multiple stages of distortions and examine how state-of-the-art IQA algorithms behave over distortion stages. Our results suggest that the performance of existing IQA models degrades rapidly with distortion stages, especially when the distortion types of different stages vary. We also find that full-reference and no-reference frameworks, though both readily applicable, have major drawbacks at predicting the quality of images at middle distortion stages. However, when the quality level of the previous stage is accessible, significantly improved quality prediction performance may be achieved. This study points out a new avenue of degraded-reference IQA research that is both practically desirable and technically challenging.

Index Terms— image quality assessment, multiple distortion stages, degraded-reference, performance evaluation

1. INTRODUCTION

Objective Image Quality Assessment (IQA) methods aim to predict the quality of images perceived by human eyes. Depending upon the accessibility to the pristine reference content, they are traditionally classified into *full-reference* (FR), *reduced-reference* (RR) and *no-reference* (NR) or *blind* IQA methods [1, 2], as illustrated in Figure 1. In the literature, IQA algorithms are usually tested and at times trained on image databases of different distortion types, but typically, each distorted image has undergone a single stage of distortion. This is in clear contrast to real-world visual content distribution scenarios, as illustrated in Figure 2, where visual content may have undergone multiple stages of distortions. For example, an image or video maybe contaminated by noise during acquisition due to limitations of camera sensors, exposure conditions, and lighting conditions, etc. Subsequently, most consumer cameras and camcorders, including mobile phone cameras, store captured content using lossy compression standards such as JPEG and H.264. Furthermore, when these images and videos are uploaded to a social networking or video sharing website, they usually undergo another round of lossy compression depending upon the operations of the hosting website. The question is, do IQA algorithms developed under the classical FR, RR and NR frameworks suffice to provide useful quality predictions in the real world scenarios? A pioneering work in this direction is the *corrupted-reference* (CR) IQA framework laid out in the context of an image restoration problem [3, 4]. The quality of the restored image with respect to an absent pristine reference image is estimated by using a noise contaminated corrupted reference image. Another important work is the LIVE Multiply Distorted database [5], which consists of images with two distortion stages. The distortion combinations include 1) Blur followed by JPEG compression and 2) Blur followed by Noise contamination. However, in these early works, whether the performance of IQA models sustains over multiple stages, and how to make the best use of available information at mid-stage distortions, has not been deeply investigated.

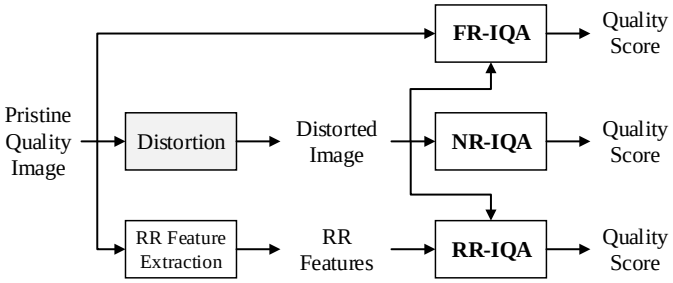


Fig. 1: General framework of FR, RR and NR IQA.

In this research, we make the first attempt aiming for a systematic understanding on the practical issue of quality assessment of images undergoing multiple stages of distortions. We are interested in knowing how the performance of objective IQA models changes over distortion stages, what the options are to carry out IQA tasks at a mid-stage, how well existing IQA frameworks and models fit the problem, and what

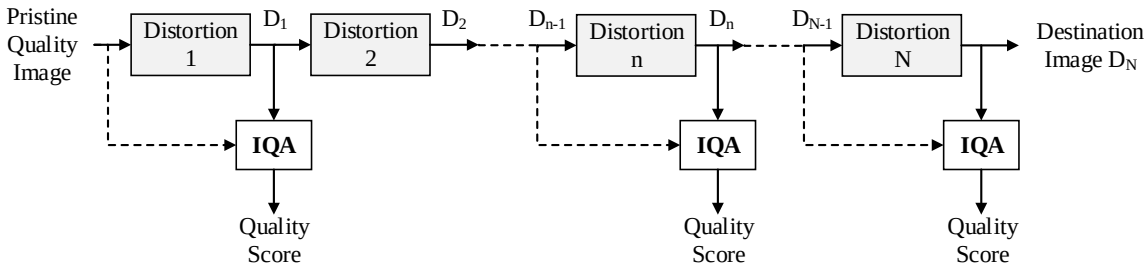


Fig. 2: Quality tracking in visual content distribution with multiple stages of distortions.

new research needs to be done to meet the practical needs.

2. PERFORMANCE VARIATIONS OF IQA MODELS AT MULTIPLE STAGES OF DISTORTIONS

Among FR, RR and NR options, NR IQA algorithms do not require access to pristine reference images and thus are easily applied at each distortion stage. Therefore, we employ state-of-the-art NR IQA models to study IQA performance variations across distortion stages. We have selected representative and best performing NR IQA algorithms for our analysis, including the opinion-aware algorithms BRISQUE [6] and CORNIA [7], opinion-unaware distortion-unaware algorithms LPSI [8] and NIQE [9], and distortion-aware algorithm WANG02 [10]. All these NR IQA algorithms are tested and/or trained on image databases that are composed of a single distortion stage [11, 12, 13].

We created a new image dataset called the *IVC-MD5* database, which is composed of 70 pristine reference images, five distortion stages and three distortion combinations. Table 1 shows the type and number of images at each stage of the three distortion combinations. The three distortion types are Gaussian white noise contamination, JPEG and JPEG2000 compression. Eight distortion levels are determined for each distortion type to roughly evenly cover from poor to excellent quality of images. Images at a particular stage of a distortion combination are then obtained by distorting each image of the preceding stage at three randomly selected distortion levels from the available set of eight distortion levels.

Since it is difficult to perform reliable subjective tests on tens of thousands of images, we use the FR IQA algorithm MS-SSIM [14, 15] for benchmarking purposes as it has been found to provide accurate and robust quality predictions [16, 17, 18]. The FR objective quality score of a distorted image is computed with respect to its respective pristine reference. The NR IQA algorithms mentioned earlier were used to obtain the quality scores of the distorted images at each stage of the database and Spearman's Rank Correlation Coefficient (SRCC) was computed to compare the results against the benchmark data (other correlation metrics were also computed and similar results were obtained but not shown due to space limit). The NR IQA algorithm WANG02 [10] was applied to distortion combination 1 only since it is designed specifically for JPEG compression.

Table 1: Composition of the *IVC-MD5* Database.

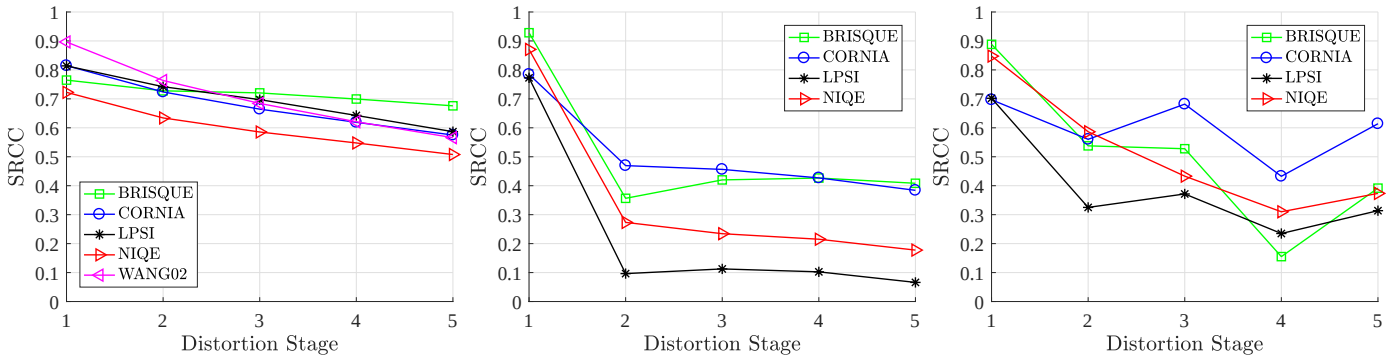
Number of Pristine Images in Database				70
Distortion Stage	**Distortion Combination**			**Number of Images**
	1	**2**	**3**	
1	JPEG	Noise	Noise	210
2	JPEG	JPEG	JPEG2000	630
3	JPEG	JPEG	JPEG	1890
4	JPEG	JPEG	JPEG2000	5670
5	JPEG	JPEG	JPEG	17010

The results are provided in Figure 3. It can observed that for distortion combination 1, which consists of 5 stages of JPEG compression, the performance of NR IQA algorithms degrades consistently with distortion stages. For distortion combination 2, which consists of Gaussian white noise at Stage-1 followed by four stages of JPEG compression, there is a substantial drop in performance between Stages 1 and 2, which indicates that IQA models suffer greatly from multiple distortion stages of mixed distortion types. For distortion combination 3, which is Gaussian white noise followed by alternating JPEG2000 and JPEG compression, the performance of IQA models drops substantially at the beginning but could oscillate afterwards.

Overall, this study reveals the complex nature of the problem. The IQA difficulty level increases with distortion stages, especially when the distortion types vary. In general, the performance of top-performing NR IQA models drops from Stage 2 to a level that would hinder their practical use.

3. IQA AT MID-STAGE

In Section 2, we study how objective IQA measures behave over multiple stages of distortions. In this section, we focus on the practical problem of objective quality assessment at



(a) Distortion Combination-1 (b) Distortion Combination-2 (c) Distortion Combination-3

Fig. 3: Performance of NR IQA algorithms for different distortion combinations of the *IVC-MD5* database in terms of SRCC.

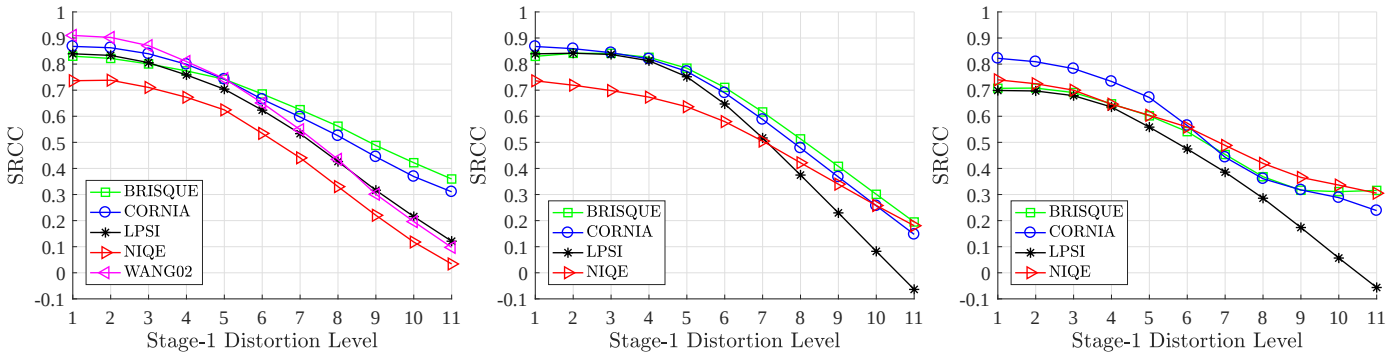
Table 2: Composition of the *IVC-MD-Te* Database.

Number of Pristine Images in Database				3570
Distortion Stage	**Distortion Combination**			**Number of Images**
	1	**2**	**3**	
1	JPEG	Noise	Noise	39270
2	JPEG	JPEG	JPEG2000	667590

a middle stage of a chain of image distortions, and investigate different options to perform quality predictions. For this purpose we created another large-scale image dataset called the *IVC-MD-Te* database, which is composed of 3570 pristine reference images, two distortion stages and three distortion combinations which we shall refer to as *JPEG-JPEG*, *Noise-JPEG* and *Noise-JP2K*. Table 2 shows the type and number of images at each stage of the three distortion combinations. Stage-1 has eleven distortion levels and Stage-2 has seventeen distortion levels such that these stages cover a wide quality range of images. This gives us a two-distortion-stage database. Although this is a simplification of practical problems (of more than 2 distortion stages), we find that by exploiting the quality prediction options after Stage-2 distortion, it is sufficient to reveal the fundamental problems and challenges of visual quality assessment at the middle stages.

At the mid-stage, we exploit three scenarios for IQA. In *Scenario-1*, we assume that access is available only to the distorted content after Distortion Stage n, i.e., to image D_n in Figure 2, and therefore only NR IQA algorithms can be used. NR IQA algorithms mentioned in Section 2 were used to evaluate the quality of the Stage-2 distorted images of the *IVC-MD-Te* database. MS-SSIM [14] scores of the Stage-2 distorted images were obtained by comparing them with their respective pristine references in order to provide benchmark scores to evaluate the performance of the NR IQA algorithms. The Pearson Linear Correlation Coefficient (PLCC) and SRCC were used to evaluate *prediction accuracy* and *prediction monotonicity* respectively [19]. Non-linear mapping of the NR IQA scores to the MS-SSIM score range was performed before evaluating PLCC [19]. The performance comparison of the NR IQA predicted quality scores and MS-SSIM benchmark scores is given in Table 3 for the entire *IVC-MD-Te* database. The opinion-aware NR IQA algorithm CORNIA [7] performs better than the other algorithms for the combinations of *Noise-JPEG* and *Noise-JP2K* while the distortion-aware NR IQA algorithm WANG02 [10] performs better for the *JPEG-JPEG* combination. However, it is evident that all NR IQA algorithms perform in an unsatisfactory manner for all distortion combinations and that their performance deteriorates more significantly for images that have been afflicted with different distortion types. The performance of NR IQA predicted quality scores at the 11 individual Stage-1 distortion levels is presented in Figure 4. Stage-1 distortion increases from left to right, with Level-1 having minimum distortion and Level-11 having maximum distortion. It is important to note that contemporary NR IQA research corresponds to Stage-1 distortion Level-1 only. It can be seen from Figure 4 that the performance of all NR IQA algorithms for all three distortion combinations degrades significantly with increasing Stage-1 distortion. An NR IQA algorithm that performs ideally on images that have undergone multiple stages of distortion should fulfill two conditions: (1) It should lead to high PLCC and SRCC values and (2) it should have consistent performance across all Stage-1 distortion levels. None of the NR IQA algorithms fulfills these performance requirements.

In *Scenario-2*, we assume that in addition to image D_n, access is also available to image D_{n-1}, which we shall refer to as a *degraded-reference* image. Such a scenario often occurs in practice, for example, when one has access to both the input and output of an image/video transcoder. Here, in addition to NR IQA algorithms, FR IQA algorithms can also be used to evaluate the quality of image D_n relative to image D_{n-1}. We evaluate the MS-SSIM quality scores of Stage-2 images of the *IVC-MD-Te* database relative to their *degraded references* and the performance comparison of these scores with respect to the benchmark MS-SSIM scores is presented



(a) JPEG-JPEG (b) Noise-JPEG (c) Noise-JPEG2000

Fig. 4: Performance evaluation of NR IQA predicted scores at individual Stage-1 distortion levels of the *IVC-MD-Te* database.

Table 3: Performance evaluation of *Scenario-1*.

Distortion Combination	NR Method	PLCC	SRCC
JPEG-JPEG	BRISQUE	0.7238	0.7163
	CORNIA	0.7131	0.7170
	LPSI	0.6800	0.6719
	NIQE	0.5795	0.5605
	WANG02	**0.7444**	**0.7399**
Noise-JPEG	BRISQUE	0.5770	0.5599
	CORNIA	**0.6263**	**0.6018**
	LPSI	0.4502	0.3920
	NIQE	0.4523	0.4161
Noise-JPEG2000	BRISQUE	0.5811	0.5717
	CORNIA	**0.6252**	**0.6358**
	LPSI	0.4187	0.3945
	NIQE	0.5735	0.5674

Table 4: Performance evaluation of *Scenario-2*.

Distortion Combination	PLCC	SRCC
JPEG-JPEG	0.8199	0.7443
Noise-JPEG	0.7080	0.6902
Noise-JPEG2000	0.6975	0.6780

in Table 4. A comparison of Tables 3 and 4 suggests that this approach outperforms the NR IQA case for all three distortion combinations of the *IVC-MD-Te* database.

In *Scenario-3*, we assume that in addition to images D_n and D_{n-1}, prior knowledge about the quality of image D_{n-1} is also available. Ideally, one would desire the FR quality score of image D_{n-1} with respect to the pristine reference image. Practically, it can also be the estimated quality score from the previous stage that is relayed to the current stage. Therefore, this scenario requires the quality scores to be evaluated at each stage and transmitted along the delivery chain. For the *IVC-MD-Te* database, we used the FR quality scores of the degraded reference images relative to the pristine references and the FR quality scores of the final distorted images relative to the degraded references as the inputs to a Support Vector Regression (SVR) [20, 21] model to predict the quality scores of the final distorted images relative to the pristine references. The results are presented in Table 5, which largely elevate the performance as compared to Scenarios 1 and 2, demonstrating the great potential of this approach. However, it should be noted that the deployment of this approach requires all distortion stages to adopt the same framework and relay the current evaluation results to the next stage. Moreover, the performance gain against Scenarios 1 and 2 may degrade with the number of distortion stages.

Table 5: Performance evaluation of *Scenario-3*.

Distortion Combination	PLCC	SRCC
JPEG-JPEG	0.9931	0.9944
Noise-JPEG	0.9554	0.9523
Noise-JPEG2000	0.9408	0.9383

4. CONCLUSION

In this work, we attempt to tackle the challenging practical problem of IQA of images undergoing multiple stages of distortions. We constructed two new image databases and use them to investigate the performance of IQA models as a function of distortion stages. We also study three different application scenarios for IQA at a mid-stage along a chain of distortions. Our results suggest that traditional FR and NR IQA frameworks and models fail to sustain their performance with multiple distortion stages. A promising approach is to relay the IQA results along the distortion chain and use it as side information to help with the IQA at the next stage. Our results demonstrate great potentials of this approach. This study indicates a new avenue of degraded-reference IQA research, which is calling for novel methodologies and algorithms.

5. REFERENCES

[1] Z. Wang and A. C. Bovik, "Modern image quality assessment," *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 2, no. 1, pp. 1–156, 2006.

[2] Z. Wang and A. C. Bovik, "Reduced- and no-reference image quality assessment," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 29–40, 2011.

[3] W. Cheng and K. Hirakawa, "Corrupted reference image quality assessment," in *2012 IEEE International Conference on Image Processing (ICIP),*. IEEE, 2012, pp. 1485–1488.

[4] C. Zhang and K. Hirakawa, "Blind full reference quality assessment of Poisson image denoising," in *2014 IEEE International Conference on Image Processing (ICIP),*. IEEE, 2014, pp. 2719–2723.

[5] D. Jayaraman, A. Mittal, A. K. Moorthy and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR),*. IEEE, 2012, pp. 1693–1697.

[6] A. Mittal, A. K. Moorthy and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[7] P. Ye, J. Kumar, L. Kang and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),*. IEEE, 2012, pp. 1098–1105.

[8] Q. Wu, Z. Wang and H. Li, "A highly efficient method for blind image quality assessment," in *2015 IEEE International Conference on Image Processing (ICIP),*. IEEE, 2015, pp. 339–343.

[9] A. Mittal, R. Soundararajan and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[10] Z. Wang, H. R. Sheikh and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *2002 IEEE International Conference on Image Processing (ICIP),*. IEEE, 2002, vol. 1, pp. I477–I480.

[11] H. R. Sheikh, Z. Wang, L. Cormack and A. C. Bovik, "LIVE image quality assessment database release 2," 2005.

[12] E. C. Larson and 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. 011006–1–011006–21, 2010.

[13] 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," *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009.

[14] Z. Wang, E. P. Simoncelli and A. C Bovik, "Multiscale structural similarity for image quality assessment," in *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers (ASILOMAR),*. IEEE, 2003, vol. 2, pp. 1398–1402.

[15] 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.

[16] 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.

[17] K. Ma, Q. Wu, Z. Wang, Z. Duanmu, H. Yong, H. Li and L. Zhang, "Group MAD Competition-A New Methodology to Compare Objective Image Quality Models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1664–1673.

[18] M. Pettersson, R. Sjoberg, P. Wennersten, K. Andersson, J. Strom and J. Enhorn, "MS-SSIM as an additional mandatory metric to PSNR for future video coding," in *Joint Video Exploration Team (JVET) of ITU-T, Document: JVET-F0064 v2, Source: Ericsson*, 2017.

[19] Video Quality Experts Group and others, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment," Mar. 2000.

[20] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer science & business media, 2013.

[21] C-W Hsu, C-C Chang and C-J Lin, "A Practical Guide to Support Vector Classification," 2003.