Subjective Quality Assessment of Animation Images

Guanghui Yue [*1], Chunping Hou [1], Ke Gu [2,3]

[1] *School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China,*
[2] *Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China*
[3] *Beijing Key Laboratory of Computational Intelligence and Intelligent System*
[*] *Corresponding author: yueguanghui@tju.edu.cn*

***Abstract*—In the past few decades, many attempts have been maken to evaluate the image quality assessment (IQA) of natural scene images. However, the IQA research of animation images (AIs) has been highly overlooked. In this article, we carry out in-depth study on perceptual quality assessment of AIs. As the lack of a public and diverse testing database currently, this paper builds a large-scale *Animation Images Quality Assessment Database* (AIQAD). This database totally includes 1050 distorted images derived from 30 source images by corrupting seven distortion types with multiple distortion levels. Then, a subjective experiment, which is the basic and accurate quality evaluation measurement, is conducted to obtain the mean opinion score (MOS) for each image. Furthermore, we also investigate the feasibility of utilizing existing mainstream full reference (FR) IQA metrics to solve the IQA problem of AIs. Experimental results demonstrate that existing mainstream FR IQA metrics merely achieve fair performance on the proposed database.**

***Index Terms*—Animation Images (AIs), image quality assessment (IQA), subjective assessment**

I. Introduction

Compared to camera-recorded image or video, animation is full of artificial scene composed of various anthropomorphic characters. In an animation, a dog even cloud may open its mouth to speak. Animation makes audiences really believe in those characters, whose misfortunes and adventures make people cry and laugh [1]. The development of animation has gone through a long period of time, from the early murals to present maturely computer-rendered animation. Early animation is mainly drawn with limited color manually. With the rapid development of computer, animation begins to be produced by rendering technology. Nowadays, animation has grown to become a mainstream artistic tool and been successfully applied to many areas, such as motion film, television, advertising, game and remote education [2]. So far, numerous works have been reported about animation [3], [4].

However, animation, like other multimedia categories, is inevitably processed due to the limitation of bandwidth, memory and other system resources, leading to occurrence of blurring, blocking, ringing and other distortions. For example, contrast change may be induced by different brightness or contrast of screens. Compression artifacts (e.g., blocking, ringing and quantization noises) may be caused by different compression standards, such as JPEG, JPEG2000 and HEVC. Blurring appears on images along with hand-shake or out-of-focus of camera, when we capture animation to share it by smart phones or other cameras in hands. With these distortions, the quality of animation is seriously affected. Therefore, it is essential and urgent to engage the quality assessment of AIs. Currently, IQA has become a basic and important research topic and can be basically classified into subjective assessment and objective assessment. The results of subjective assessment are usually regarded as the groundtruth of objective assessment. Since subjective assessment requires participants, the experimental results authentically reflect the perceptual quality. On the contrary, objective assessment overcomes the subjective assessment's shortcomings (e.g., cumbersome, slow, and expensive) and evaluates the image quality by building mathematical model without participants. Although many objective IQA methods have been proposed for quality assessment of natural scene images [5], [6], whether these IQA methods can be applicable to AIs is still an open question.

In this work, we aim to carry out in-depth study on perceptual quality assessment of AIs from subjective aspects. For this purpose, we build the first large-scale *Animation Image Quality Assessment Database* (AIQAD), which contains a total of 1050 images derived from 30 source images. Then, the database is built by assigning seven distortion types with multiple distortion levels on the source images. Given that the perceptual qualities of images corrupted by the same degree of distortion may be different among different source images, the contents of source images are varied. To be specific, our database contains both simple image with limited color and structure and complex image analogous to natural scene images. Finally, a subjective evaluation is conducted to obtain the mean opinion score (MOS) for each image.

The remainder of this article is organized as follows. Section II describes the database in detail. Section III gives a scheme of subjective evaluation and summaries the subjective experiment results. An investigation about whether existing mainstream FR metrics suit for quality assessment of AIs is described in Section IV. Finally, we remark the conclusion and future work in Section IV.

II. The Animation Image Quality Assessment Database

A. Source image

The entire database was generated from a set of source images that reflect adequate diversity in image content. Totally,

978-1-5386-0462-5/17/$31.00 ©2017 IEEE.

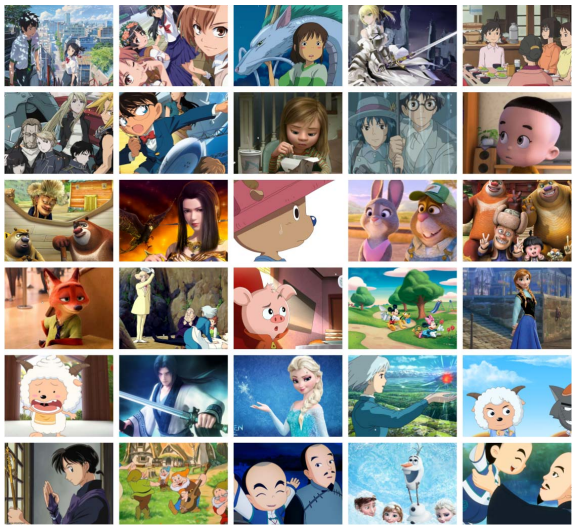

Fig. 1. The contents of source images used in the database

30 color AIs were collected from famous cartoons. Fig. 1 shows all the source images [1]. As can be seen, these images are very different in content. To be specific, part of them have simple structure and limited color. Whereas, the rest of them, which are rendered by computer for simulating the natural scene, are full of color variations. All these images are captured via screenshot tool with the resolution of 1024 $\times$ 768 [2].

B. Distorted image

Seven commonly encountered distortion types (i.e., Gaussian blur, white noise, Contrast change, Mean shift, Motion blur, JPEG2000 compression and JPEG compression) are applied on all the source image to generate distorted images. All distortion types except Contrast change and Mean shift contains 5 distortion levels, while Contrast change (Mean shift) contains 4 (6) distortion levels. The distortion types and their arrangments are as follows.

- Gaussian blur: For each color channel (i.e., R, G and B), a Gaussian kernel (with standard deviation σ_G and window 11 $\times$ 11) was used to generate Gaussian blur. In this article, σ_G is set as 0.25, 0.5, 1, 1.75, 2.5. Note that, each of R, G and B channel was blurred by the same kernel.
- White noise: To generate the white noise, we employ the Matlab *imnoise* command with standard normal probability density function of variance σ_N^2 (0.0003, 0.001, 0.003, 0.01, 0.03) on each of three color channel, R, G and B.
- Contrast change: We utilize the same cubic and logistic functions to obtain the contrast change images [6]. The parameters are the same with those used in [7]. For each function, we only set two kinds of parameter and totally generate 2 distortion levels.
- Mean shift: The mean-shifted image is obtained by $y = x + k$, where x is the source image, k is the shifted levels quantized as $\{\pm 20, \pm 40, \pm 60\}$.
- Motion blur: We use the Matlab *fspecial* and *imfilter* commands to create Motion blurred images by setting the motion angle as 0 and motion length parameter as 3, 6, 9, 12 and 15.
- JPEG2000 compression: We employ the Matlab *imwrite* command to create JPEG2000 compressed images by setting the *CompressionRatio* parameter as 1600, 800, 480, 300 and 120.
- JPEG compression: We employ the Matlab *imwrite* command to create JPEG compressed images by setting the Q parameter as 2, 6, 10, 15 and 25.

After processed by the distortions above, an image can generate 35 distorted images. Fig. 2 gives a group of source images and its associated distorted images. Eventually, total 1050 distorted images are derived from 30 source images [3]. These distortions reflect a broad range of image impairments, from smoothing to structured distortions, image-dependent distortions, and random noise. The level of distortion was varied from imperceptible levels to high levels of impairment.

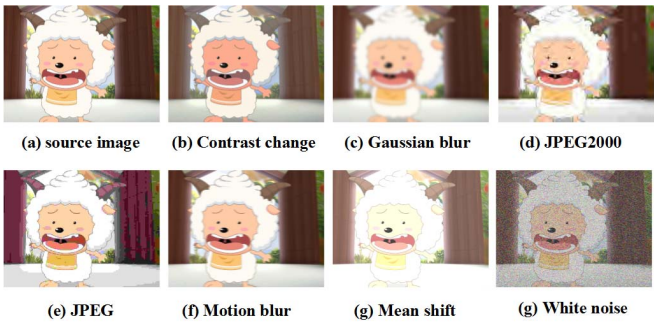

Fig. 2. The source image and its corrupted versions.

III. Subjective Quality Evaluation Experiment

A. Testing Methodology

We conducted the subjective viewing test with a Double-stimulus (DS) method in accordance with ITU-R BT.500 [8]. The subjective evaluation experiment was conducted in a laboratory environment. A 23-inch monitor with resolution of 1920 $\times$ 1080 pixels was used to display the AIs. The viewing distance was three times the image height. Twenty-one subjects, who are no-experts in the field of image processing and quality assessment, participated in the subjective evaluation. All the subjects had normal or corrected-to-normal vision with no previous history of neurological or psychiatric diseases and were free of any medication.

Before the start of testing stage, subjects have to go through the training stage in which some examples with representative distortion types and levels are presented. In order to prevent

[1] All images are copyright of their rightful owners, and the authors do not claim ownership. No copyright infringement is intended. The database is to be used strictly for non-profit educational purposes.

[2] Actually, we use the browser shortcut key for screenshot.

[3] We will release this database online soon.

subjects from remembering these images, all these examples are not included in the testing stage. The testing stage would not be started until the subject can adroitly complete the task. During testing, AIs were randomly presented to the viewers with no duplication. A double stimulus setup was used. Every image in the database was viewed by each subject 6 seconds, over five sessions of an hour each, separated by roughly 24 hours. A ten-point grading scale was used to assign the perceptual quality (from 0 to 9, 0: bad, 9: excellent). Every 5 minutes the participant was stopped to release the accumulated visual fatigue.

B. Subjective Data Processing

After the subjective experiment, the recorded data needs to be analyzed and the abnormal data should to be eliminated. We strictly follow the observers screening procedures recommended by ITU-R BT.500 [8] to exclude outliers. As a result, most subjects complete the task carefully and only one of their data is eliminated. Then, the MOS of each image can be calculated as:

$$s_i = \sum_{i=1}^{N} r_{ij} \tag{1}$$

where i is the image index; j is the subject index; and N is the number of valid subjects. Generally speaking, the quality scales of the distorted AIs in the database should exhibit good separation of perceptual quality and span the entire range of visual quality (from distortion imperceptible to severely annoying). Fig. 3 shows the histogram of the MOS values (rang from 0 to 9) of all distorted images in the database. It can be observed that the MOS values of images range from low to high, and have a good spread at different levels.

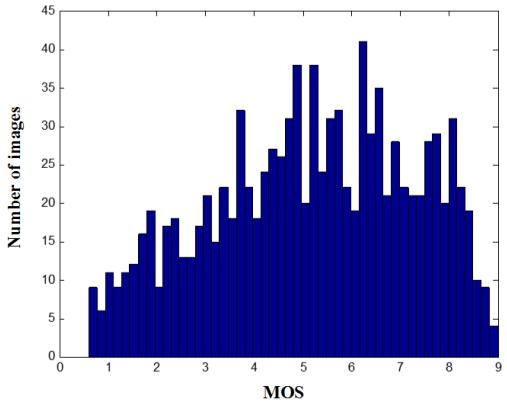

Fig. 3. Histogram of MOS values of images in the AIQAD.

Apart from excluding outliers, we are also interested in exploring each subject's performance. The performance of individual subject can be evaluated by computing the correlation coefficients between subjective ratings and MOS values for each image set, then averaging the correlation coefficients of all image sets. For this purpose, one commonly used performance measures (i.e., Pearson linear correlation coefficient (PLCC)) is employed as the evaluation criterion. With experience, values of PLCC close to one mean higher prediction accuracy. Obviously, the higher value of PLCC corresponds to the better performance and higher variation denotes that the image content has large effect on the perceptual quality. The standard deviation (std) and mean of the results are depicted in Fig. 4. The average performance across all individual subjects is also given in the rightmost columns of Fig. 4. It can intuitively find that the subjects perform quite consistently with relatively low variations for different image contents.

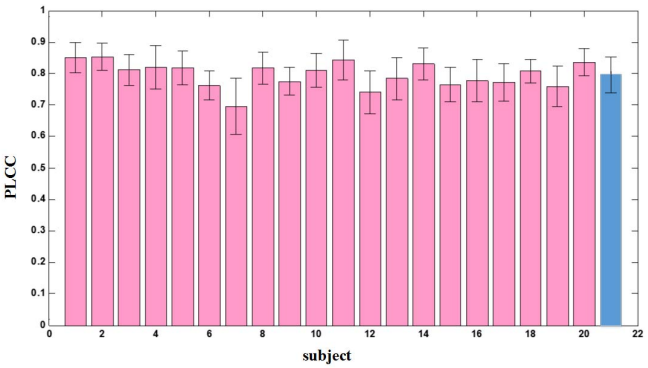

Fig. 4. PLCC between individual subject rating and MOS. Rightmost column: performance of an average subject.

IV. Objective Metrics Comparison

In this section, we will investigate the performance of existing mainstream full reference (FR) IQA methods to evaluate the visual quality of the images in the AIQAD database.

A. Experiment setup

The AIQAD database is employed as testing bed for performance evaluation and comparison. Eight popular FR IQA metrics for natural scene image and three state-of-the-art FR IQA metrics for screen content image are chosen to test on the database. The selected FR metrics are SSIM [9], MS-SSIM [10], FSIM [11], GMSD [12], GMSM [12], MAD [13], ADD-GSIM [14], IGM [15], SQMS [16], IW_SCC [17] and GSS [18]. All these metrics are implemented using the codes published on their websites or released via email. As suggested by video quality experts group (VQEG), four commonly used performance criteria (i.e., PLCC, Spearman rank correlation coefficient (SRCC), Kendall's rank correlation coefficient (KRCC) and Root mean-squared error (RMSE)) are employed to evaluate and compare the performance between different metrics. With experience, higher values of PLCC, SRCC and KRCC, unlike that of RMSE, indicate a superior performance of the tested metric. Note that, according to VQEG, a logistic regression procedure is required, before the calculation of PLCC and RMSE, for removing the nonlinearity of objective quality predictions.

B. Experimental results and analysis

The experimental results are given in Table I. From Table I, we can get some meaningful information. First, both FR

metrics for natural scene images and screen content images achieves fair performance on the proposed AIQAD database. Second, compared to metrics for natural scene images, those for screen content images (e.g., SQMS, IW_SCC and GSS) surprisingly obtain worse results. The reason for the above phenomena may be attributed to the following reasons. First of all, some AIs, compared to natural scene images, are produced manually with limited color variations. With this property, some commonly encountered distortions (such as JPEG and Contrast change) may more easily affect the perceptual quality. Therefore, the metrics with general assessment ability of these distortion types are not strictly suit for quality assessment for such AIs. However, as the computer-rendered images are similar to natural scene image, these metrics are be competent to assessment for such AIs. Combining the two aspects, existing mainstream metrics for natural scene images have fair assessment ability (e.g., MS-SSIM obtains PLCC value of 0.7423). Secondly, screen content images are usually composed of text and pictures [19], [20]. Therefore, metrics for screen content images should consider both textual regions and pictorial regions. Since textual region has distinct characteristics compared with pictorial region and AIs are full of pictures, it is acceptable to face the result that IQA metrics of screen content images are not suitable for AIs.

TABLE I
OBJECTIVE METRICS' COMPARISON ON THE PROPOSED DATABASE.

Metrics	Criteria			
	PLCC	SRCC	KRCC	RMSE
SSIM [9]	0.5830	0.5729	0.4069	1.6811
MS-SSIM [10]	0.7423	0.7370	0.5421	1.3866
FSIM [11]	0.8036	0.7993	0.6008	1.2315
GMSD [12]	0.7448	0.7272	0.5267	1.3808
GMSM [12]	0.6850	0.6659	0.4737	1.5075
IGM [15]	0.7704	0.7431	0.5457	1.3192
ADD-GSIM [14]	0.7959	0.7788	0.5780	1.2527
MAD [13]	0.7862	0.7737	0.5742	1.2788
SQMS [16]	0.6043	0.5475	0.3793	1.6487
IW_SCC [17]	0.6338	0.6192	0.4331	1.6006
GSS [18]	0.5370	0.4707	0.3306	1.7456

V. CONCLUSION AND FUTURE WORK

Distortions caused during image processing and transmission highly affect the quality of experience. In this paper, we have investigated into the problem of quality assessment of AIs. We first set up a database of AIs (named AIQAD) to promote the IQA study for this issue. The AIQAD database includes seven kinds of distortion type with multiple distortion levels. In order to clarify whether or not existing mainstream FR IQA metrics are complete to solve the IQA problem of AIs, we evaluate and validate the performance of these metrics on the proposed animated database by conducting an experiment. Experimental results showed that these IQA metrics merely achieve fair performance and far from the ideal quality model of AIs. Therefore, metrics particularly designed for AIs are urgently needed. In future work, we will put more efforts into analyzing the characteristics of AIs and establish an objective quality evaluation model that is in accordance with the human visual system.

VI. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China under Grant 61520106002, 61471262.

REFERENCES

[1] F. Thomas, O. Johnston, and W. Rawls, *Disney animation: The illusion of life*. Abbeville Press New York, 1981, vol. 4.
[2] R. Parent, *Computer animation: algorithms and techniques*. Newnes, 2012.
[3] J. Wang, S. M. Drucker, M. Agrawala, and M. F. Cohen, "The cartoon animation filter," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 1169–1173, 2006.
[4] J. Yu and D. Tao, *Modern machine learning techniques and their applications in cartoon animation research*. John Wiley & Sons, 2013, vol. 4.
[5] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297–312, 2011.
[6] G. Yue, C. Hou, K. Gu, S. Mao, and W. Zhang, "Biologically inspired blind quality assessment of tone-mapped images," *IEEE Transactions on Industrial Electronics*, 2017, DOI: 10.1109/TIE.2017.2739708.
[7] K. Gu, G. Zhai, W. Lin, and M. Liu, "The analysis of image contrast: From quality assessment to automatic enhancement," *IEEE Transactions on Cybernetics*, vol. 46, no. 1, pp. 284–297, 2016.
[8] I.-R. R. BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," 2002.
[9] 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.
[10] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2. IEEE, 2003, pp. 1398–1402.
[11] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.
[12] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, 2014.
[13] 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. 011 006–011 006, 2010.
[14] K. Gu, S. Wang, G. Zhai, W. Lin, X. Yang, and W. Zhang, "Analysis of distortion distribution for pooling in image quality prediction," *IEEE Transactions on Broadcasting*, vol. 62, no. 2, pp. 446–456, 2016.
[15] J. Wu, W. Lin, G. Shi, and A. Liu, "Perceptual quality metric with internal generative mechanism," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 43–54, 2013.
[16] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, and W. Zhang, "Saliency-guided quality assessment of screen content images," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1098–1110, 2016.
[17] S. Wang, K. Gu, K. Zeng, Z. Wang, and W. Lin, "Objective quality assessment and perceptual compression of screen content images," *IEEE Computer Graphics and Applications*, 2016, DOI: 10.1109/MCG.2016.46.
[18] Z. Ni, L. Ma, H. Zeng, C. Cai, and K.-K. Ma, "Gradient direction for screen content image quality assessment," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1394–1398, 2016.
[19] H. Yang, Y. Fang, and W. Lin, "Perceptual quality assessment of screen content images," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4408–4421, 2015.
[20] S. Wang, K. Gu, X. Zhang, W. Lin, L. Zhang, S. Ma, and W. Gao, "Subjective and objective quality assessment of compressed screen content images," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 6, no. 4, pp. 532–543, 2016.