2023 15th International Conference on Quality of Multimedia Experience (QoMEX)

Dataset of Subjective Assessment for Visually Near-Lossless Image Coding based on Just Noticeable Difference

Soichiro Honda
Nagoya Institute of Technology
Nagoya, Japan
0009-0008-2849-5910

Yoshihiro Maeda
Tokyo University of Science
Tokyo, Japan
0000-0001-6919-637X

Norishige Fukushima
Nagoya Institute of Technology
Nagoya, Japan
0000-0001-8320-6407

***Abstract*—Image compression is essential in image processing, and image quality assessment (IQA) is important in determining the image compression level. This study aims to construct a dataset for evaluating image quality at low compression in this coding degradation, i.e., for high-quality images. Typical IQA databases are selected for general-purpose image degradation, not high-quality images. If one tries to evaluate high-quality images, multi-level evaluations are difficult to construct successfully. In addition, the evaluated encoder is not a de facto standard encoding algorithm. Therefore, this study constructs a dataset for subjective evaluation of visually near-lossless level image compression quality based on binary level evaluation of the just noticeable difference (JND). Experimental results showed that the new dataset was validated for correlation by various IQAs. It was also shown that more than the compression quality covered by the conventional dataset is needed for the binary evaluation of high-quality images. The dataset is available at: https://norishigefukushima.github.io/iqanearlossless/.**

***Index Terms*—just noticeable difference, visually near-lossless image coding, subjective image quality assessment**

I. Introduction

Lossy image coding is essential for image delivery and storage, and its image quality assessment (IQA) is important in determining the compression level. Although their subjective judgments should determine the quality perceived by people, evaluating the quality on a case-by-case basis is costly. For this reason, image quality is usually evaluated automatically by creating a dataset of IQA and using objective evaluation indices according to the dataset.

Representative IQA databases include LIVE [1], CSIQ [2], TID2008/2013 [3], [4], and KADID-10K [5]. These datasets evaluate universal image distortions, including coding distortion as an example of distortion. They are not intended to evaluate high-quality coded images, as the primary purpose of the datasets is to create an IQA index for universal distortion. Distortions are evaluated as a mean opinion score (MOS) of 5, 10, or 100 levels, and the encoding distortions include JPEG [6] and JPEG2000 [7]. However, the visual differences due to encoding distortion must be clear to evaluate at multiple levels; thus, the compression levels will be high, and their values will be beyond the practical range. In addition, JPEG2000 is not a de facto standard encoding.

When the visual difference of degradation is small and difficult to judge, a subjective evaluation of the just noticeable difference (JND) is selected, a binary evaluation method of similarity or dissimilarity. Typical JND evaluations of coding degradation use an adjustment or bi-selection method [12] to evaluate similarity or dissimilarity by changing quality parameters. KonJND-1k [10] uses the adjustment method. The method requires a skilled subject to make a proper evaluation. Otherwise, subjects tend to select a point where degradation is more clearly visible to quickly determine the quality parameter of the JND point, resulting in the selection of a low-quality parameter. MCL-JCI [8] and Shen et al. [9] use the bi-selection method, which is more stable. These datasets use 64/55-inch TV with long distances; thus, the characteristics differ from the other dataset. Moreover, what is being measured is a relative quality parameter (QP) of the JND point when looking at continuously changing quality parameters, and no experiment has been constructed for making IQA metrics.

Unlike typical JND-based coding degradation evaluation, learned perceptual image patch similarity (LPIPS) [11] uses JND for building IQA metrics. LPIPS uses JNDs to evaluate image quality, which is difficult to determine whether the quality degradation is close to the original image. While this data is created for deep learning and serves the purpose of the LPIPS paper well, the data is insufficient to create a classical IQA metric. First, 64×64 patches are used to collect much subjective data, so the evaluation is not made on an image-by-image basis. Therefore, the influence of the surroundings should be addressed. In addition, to ensure patch diversity, there is only one evaluator for each patch. When a single patch is extracted, it cannot be used to statistically determine whether the patches are identical, i.e., a binary decision. Furthermore, the data is collected through crowdsourcing to ensure diversity. Since the data is not under a controlled laboratory environment, it is difficult to judge based on a single image alone with a few data since it is greatly affected by the

This work was supported by JSPS KAKENHI (21H03465, 21K17768).

979-8-3503-1173-0/23/$31.00 ©2023 IEEE

TABLE I: Charactoristics of Datasets. *The Internet-based IQA. †big-screen TV-based IQA. ‡Patch-based IQA.

	Dataset	Quality parameters	Score	Content	Resolution of references	Test environment
IQA (JPEG)	LIVE [1]	{82, 35, 23, 14, 3, 1}	MOS	29	768×512	lab
	*TID2013 [4]	{80, 60, 23, 8, 4}	MOS	25	512×384	lab or crowdsourcing
	*KADID [5]	{43, 36, 24, 7, 4}	MOS	81	512×384	crowdsourcing
JND	†MCL-JCI [8]	continuity	QP	50	1920×1080	lab
	†Shen et al. [9]	continuity	QP	39	1920×1080	lab
	*KonJND-1k [10]	continuity	QP	1008	640×480	crowdsourcing
	*‡LPIPS [11]	random	0 or 1	9.6k	64×64	crowdsourcing
Proposed	JPEG	{90, 80, 70, 60, 50, 35, 20}	Ratio	10	512×512 and 1024×1024	lab
	WebP	{90, 80, 70, 60, 50, 35, 20}	Ratio	10	512×512 and 1024×1024	lab
	HEIF	{55, 50, 45, 40, 35, 30, 25}	Ratio	10	512×512 and 1024×1024	lab

(a) 604.14 (b) 488.26 (c) 445.84 (d) 327.09 (e) 207.09

(f) 97.46 (g) 93.16 (h) 84.51 (i) 70.06 (j) 67.78

Fig. 1: Test images; (a)∼(e) high-frequency images, (f)∼(j) high-frequency images. Noted values are the average of 3×3 patch's variances for each image.

display and the surrounding environment.

Therefore, we construct a new dataset for visually near-lossless compression with standard encoders based on JND in this study, named *Meikoudai image distortion dataset (MIDD)*. A distorted image is evaluated by 30 subjects under laboratory control. Also, the data is evaluated by the typical IQA metrics: peak signal-noise-ratio (PSNR), structural similarity (SSIM) [13], and gradient magnitude similarity deviation (GMSD) [14]. The contributions of our dataset are as follows.

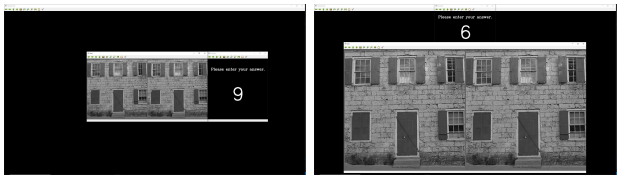
- identification ratios based on JND instead of MOS.
- de facto encoders: JPEG, WebP with/without deblocking filter, and HEIF.
- different resolutions (512×512, 1024×1024) par image to represent dpi difference.

Table I summarizes the characteristics of the datasets.

II. MIDD Dataset Construction

The JND-based subjective evaluation was conducted to measure the degree of various coding deterioration. Participants were asked to identify the differences between the original and compressed images. Ten grayscale images are used as test images selected from the Kodak images. The original images are 768×512 or 512×768, but were cropped to 512×512 (See Fig. 1). The images can be divided into two categories: high-frequency (a,b,c,d,e) and low-frequency (f,g,h,i,j).

These images were degraded by four types of compression: JPEG [6], WebP [15] with and without deblocking filtering, and HEIF [16]. For each degradation, we have 50 images. Also, we upscale these images by the nearest neighbor method from 512×512 to 1024×1024 to simulate a halved dpi resolution. Note that the image is upscaled after compression. In total, we have $4 \times 50 \times 2 = 400$ degraded images. Each compression's quality parameters differ for 512×512 and 1024×1024. Each image was judged by 30 participants who were not image-processing researchers in their 10s and 20s. Participants were recruited primarily through bulletin boards at Nagoya Institute of Technology and did not include students within the laboratory conducting the study. Totally, we have $400 \times 30 = 12,000$ judgments.



(a) 512×512 (b) 1024×1024

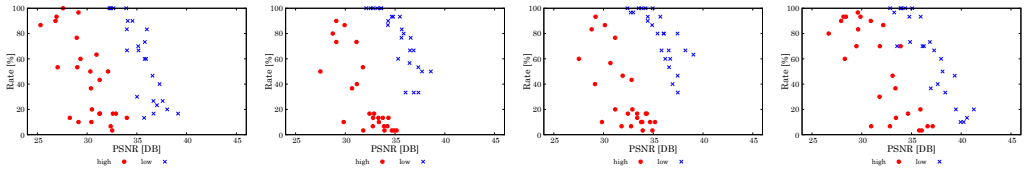
Fig. 2: Screenshot during the experiment of two size cases.

Next, we show the experimental protocol. After simultaneously showing the original and degraded images side-by-side, participants were asked to compare the uncompressed images to the compressed ones. The experimental interface is shown in Fig. 2. If participants judged two images to be the same, they were instructed to input "1"; if they judged them to be different, they were instructed to input "0". The time for making each judgment is up to 12 seconds. If the participants felt that the images were the same, they were required to spend at least 6 seconds making the judgment before moving on to the following image. If 12 seconds had elapsed, the participant judged that the images were identical. This procedure was repeated for 200 images of 1024×1024 and then for the last 200 images of 512×512. The display was EIZO CS 270 (27inch 2560×1440 / IPS), and the viewing distance was 0.5 m.

Furthermore, we took care to minimize any potential impact of the experimental environment on the results. To reduce the effect of afterimages from the previous image display, we set the interval between showing image pairs to 500 ms. To reduce the effect of habituation during the experiment, we informed the participants of the points in the images where block noise and blurring might occur due to the encoding.

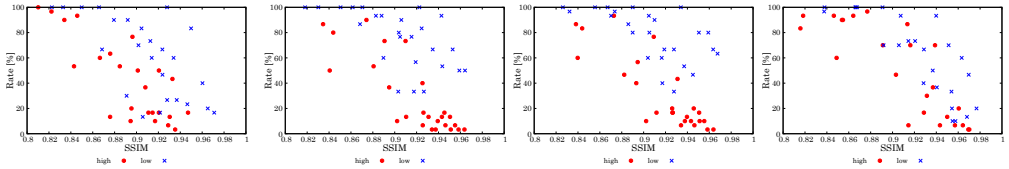
III. Experimental results

Figures 3 to 6 show assessment results for three IQA methods: PSNR, SSIM, and GMSD. The y-axis is the identification rate, and the x-axis is the objective IQA value. A scatter



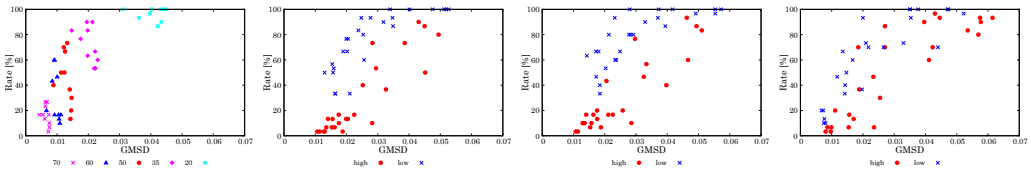
(a) JPEG (b) WebP (off) (c) WebP (on) (d) HEIF

Fig. 3: PSNR (512×512)



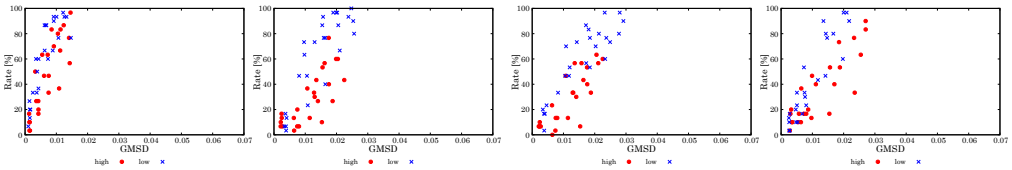
(a) JPEG (b) WebP (off) (c) WebP (on) (d) HEIF

Fig. 4: SSIM (512×512)



(a) JPEG (b) WebP (off) (c) WebP (on) (d) HEIF

Fig. 5: GMSD (512×512)



(a) JPEG (b) WebP (off) (c) WebP (on) (d) HEIF

Fig. 6: GMSD (1024×1024)

plot indicates a degraded image's identification rate and IQA index. The identification rate is the average of the binary JND decisions from 30 subjects. Also, the high label shows the high-frequency image; the other is the low-frequency image. Only in Fig. 5, the scatter plots are colored by quality factors. Note that high-resolution cases for PSNR and SSIM are omitted due to the page limits.

In PSNR (Fig. 3), they are divided into high/low-frequency groups. It is difficult to evaluate visually near-lossless images with PSNR because it is affected by the frequency of the input image. In SSIM (Fig. 4), the group difference is smaller than PSNR but remains. GMSD (Fig. 5) has a higher correlation than PSNR and SSIM. The difference between groups has decreased in the GMSD's $\times 2$ resolution case.

In Fig. 5, quality factors of conventional IQA dataset [1], [4], [5] cannot cover the range of the plots due to the limited QP ranges (See Tab. I). Points of QP=20 or QP=70 are edge clusters; thus, middle QP points are missing.

Next, we compute Spearman's rank correlation coefficient (Tab. II). The PSNR correlations are very weak; JPEG and WebP have almost no correlation, and HEIF correlations are also weak. The correlation improves slightly for SSIM and is highest for GMSD. Note that WebP with deblocking filtering has an effect when the display size is large and the opposite

TABLE II: Spearman's rank correlation coefficient.

512×512	JPEG	WebP (off)	WebP (on)	HEIF
PSNR	-0.1839	-0.0454	0.0565	-0.4583
SSIM	-0.5716	-0.6677	-0.6271	-0.7825
GMSD	**0.8084**	**0.7815**	**0.7683**	**0.8072**
1024×1024	JPEG	WebP (off)	WebP (on)	HEIF
PSNR	-0.4090	-0.3483	-0.3238	-0.6167
SSIM	-0.6855	-0.7322	-0.7534	**-0.8403**
GMSD	**0.8107**	**0.7935**	**0.8378**	0.8307

effect when the display size is small. Deblocking filter is designed to smooth out blocky artifacts caused by compression, which improves visual quality; however, smoothing the image can also remove textures. The presence or absence of textures is a crucial factor for the judgment of this experiment.

IV. Conclusion

This paper provided a JND-based subjective evaluation dataset for high-quality compression, named *MIDD*. JPEG, WebP (with and without deblocking filter), and HEIF were evaluated as coding degradation, and two types of dpi were evaluated. The dataset is evaluated by using PSNR, SSIM, and GMSD. The current limitation of this dataset is only grayscale image dataset.

REFERENCES

[1] 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.

[2] E. Cooper Larson and D. Micheal Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, p. 011006, 01 2010.

[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," *Advances of Modern Radioelectronics*, vol. 10, pp. 30–45, 01 2009.

[4] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "Image database tid2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, p. 57 – 77, 2015.

[5] H. Lin, V. Hosu, and D. Saupe, "Kadid-10k: A large-scale artificially distorted iqa database," in *Proceedings of International Conference on Quality of Multimedia Experience (QoMEX)*, 2019, pp. 1–3.

[6] G. K. Wallace, "The jpeg still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. 18–34, 1992.

[7] M. Rabbani, "JPEG2000: Image Compression Fundamentals, Standards and Practice," *Journal of Electronic Imaging*, vol. 11, no. 2, p. 286, 2002.

[8] L. Jin, J. Y. Lin, S. Hu, H. Wang, P. Wang, I. Katsavounidis, A. Aaron, and C.-C. J. Kuo, "Statistical study on perceived jpeg image quality via mcl-jci dataset construction and analysis," in *Proceedings of IS&T International Symposium on Electronic Imaging: Image Quality and System Performance XIII*, 2016, pp. 1–9.

[9] X. Shen, Z. Ni, W. Yang, X. Zhang, S. Wang, and S. Kwong, "Just noticeable distortion profile inference: A patch-level structural visibility learning approach," *IEEE Transactions on Image Processing*, vol. 30, pp. 26–38, 2021.

[10] H. Lin, G. Chen, M. Jenadeleh, V. Hosu, U.-D. Reips, R. Hamzaoui, and D. Saupe, "Large-scale crowdsourced subjective assessment of picturewise just noticeable difference," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5859–5873, 2022.

[11] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

[12] J. Y. Lin, L. Jin, S. Hu, I. Katsavounidis, Z. Li, A. Aaron, and C.-C. J. Kuo, "Experimental design and analysis of jnd test on coded image/video," in *Proceedings of SPIE: Applications of Digital Image Processing XXXVIII*, International Society for Optics and Photonics. SPIE, 2015, p. 95990Z.

[13] 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.

[14] 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.

[15] WebP homepage, http://code.google.com/speed/webp/, (15 Feb 2023).

[16] M. M. Hannuksela, J. Lainema, and V. K. M. Vadakital, "The high efficiency image file format standard [standards in a nutshell]," *IEEE Signal Processing Magazine*, vol. 32, no. 4, pp. 150–156, 2015.