Compression Artifacts Image Patch database for Perceptual Quality Assessment

1st Tung Pham Thanh
Faculty of Basic Sciences
University of Fire fighting and Prevention
Hanoi, Vietnam
tung@vinafire.com.vn

2nd Chau Ma Thi
Faculty of Information Technology
The VNU University of Engineering and Technology
Hanoi, Vietnam
ma.thi.chau@gmail.com

3rd Tuan Nguyen Manh
Faculty of Basic Sciences
University of Fire fighting and Prevention
Hanoi, Vietnam
nguyenmanhtuan.ppa@gmail.com

4th Linh Le Dinh
Faculty of Information Technology
The VNU University of Engineering and Technology
Hanoi, Vietnam
ledinhlinh910@gmail.com

5th Ha Le Thanh
Faculty of Information Technology
The VNU University of Engineering and Technology
Hanoi, Vietnam
lthavnu@gmail.com

***Abstract*—Ground truth is one of the most important component for training, testing, and benchmarking algorithms for objective quality assessment. In this paper, we propose an image patch quality database with compression artifacts. We create a new database of image patches with High Efficiency Video Coding (HEVC) compression artifacts. Then, the subjective test is conducted in a controlled environment to obtain the ground truth of image patch quality, where we collect differential mean opinion scores (DMOS) from a larger amount of observers. Finally, the rank order correlation factors between DMOS and a set of popular image quality metrics are calculated and presented. The proposed database is expected for learning patch based IQA model for block size in video rate-distortion optimization.**

***Index Terms*—Image quality assessment, coding distortion, Image-Patch Quality Assessment, Image/Video with Compression Artifacts**

I. Introduction

In modern applications of digital image and video processing, visual quality is a basic and important requirement. All these applications require quality, in the first order, visual quality metrics are able to adequately characterize images. Many good full-reference convolutional neural network (CNN) based metrics for which a reference image (or frame sequence) that take into account specific features of human visual system (HVS) have been already proposed. To characterize adequateness and performance of these metrics, several publicly available quality image databases are used, including the databases LIVE [1], CSIQ [2], TID2008 [3], TID2013 [4] etc. The largest subjective database is TID2013, which has 3,000 images, but still not enough to train a good deep neural network.

All available image quality benchmark databases are only suitable for evaluating the quality of images as a whole. These databases are not fundamental investigating which parts of the testing image contribute to the testing results or the score for a particular patch of image. But the problem is that the perceptual quality of each image patch is different at the same level of noise. Fig. 1 shows that the distortions around the houses and on the sky regions (solid square) are easily observable while those on textural regions (dash dot square) are less noticeable. In recent work [5], we propose a quality assessment approach database for image patch with the desire to create a new perception-based metric to apply for each region. Experimental results show that compressed image quality decreases depending on the visual features of image.

Based on these observations, we propose a large experimental database to evaluate the quality that human perceive for each image patch. Firstly, we randomly create 61,600 image patch pairs in both sizes ($128 \times 128, 64 \times 64$) generated from 40 video test sequences with HEVC compression artifacts. Secondly, we select the double stimulus categorical rating (DSIS) method to rate the subjective quality score for each patch pair. Additionally, the software of our designed subjective test is made up following Rec.ITU-R BT.500-13 standard [6]. Then, the experiment is conducted in a testing process to obtain the ground truth of image quality where we collect over 697,000 opinion scores. Together with, the testing

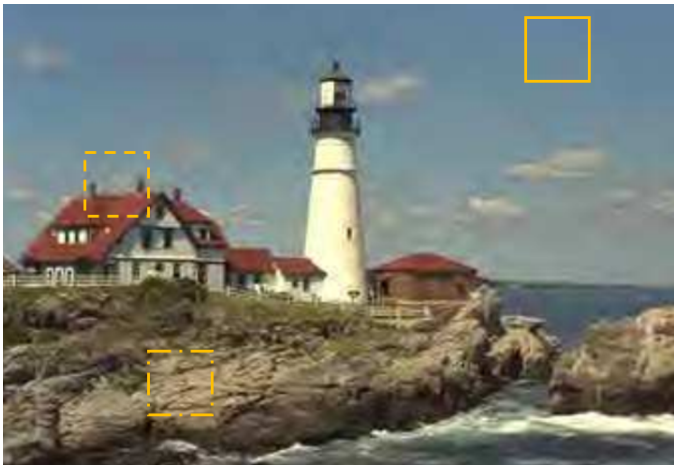


Fig. 1: Example of distorted image

data are pre-processed to remove noises and outliers. Finally, the proposed database contains 40,286 image patch pairs, differential mean opinion score (DMOS), standard deviation, number and time observation. This new database is tested with a number of state-of-the-art IQA models and as the result, the proposed database can help researchers in image compression community to select the best IQA method to conduct the perceptual based image optimization.

The following Section II describes the procedure of database creation. Then, extensive experimental results and discussion are presented in Section III, and conclusions are given in Section IV.

II. Database building process

A. Test material

In video encoding, noise types are added to the original video by H265/HEVC compression, then, testing images normally are cropped from extracted frames in the video test sequence. The goal of our study is, however, to create a testing image database for local image perception. So, we randomly select several patches from each pair of reference and testing images and the process of building database (Fig 2) are depicted as following. There are 40 original source videos of both high-definition (1280×720) and full high-definition (1920×1080). For each video sequence, depending on the length of such video, we select a different number of original frames as reference images. The reference frames are selected evenly throughout the video sequence to diversify the content. On the other hand, the video sequence is compressed with different quantization parameters (QPs) in range from 2 to 50. Testing images are extracted from the compressed video sequence in a similar method with reference frames of original video sequences. As a result, we obtain different groups with 246400 images, each group includes one reference image and some testing images. After that, for each group, we select 200 pair of 128×128 patches when randomly choose positions to crop as well as quantization parameters in selecting testing image to make pairs of reference and testing patches. We also crop pairs of 64×64 center patches from the original pair of 128×128 original patches to evaluate in the experiments. As a result, we obtain 246400 images, 61600 pairs of 64×64 patches and 61600 pairs of 128×128 patches. All patches are annotated with their position. Those images and pairs of patches make HMII (Human Machine Interaction Image) database.

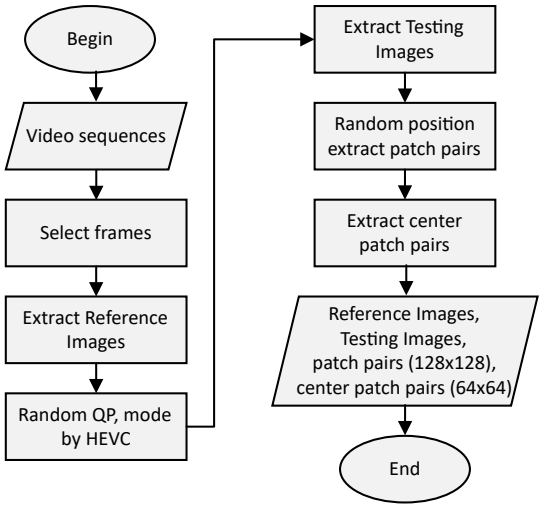


Fig. 2: Process of database building.

B. Testing methodology

For the purpose of subjective testing methodology, the International Telecommunication Union set the ITU-R BT.500-11 standard [6]. In such standard, there are several popular subjective methodologies for testing such as "Single stimulus categorical rating", "Double stimulus categorical rating", "Ordering by force-choice pairwise comparison" and "Pairwise similarity judgments". Double stimulus categorical rating is chosen in our experiment because it matches the full reference quality assessment and observation time is enough to evaluate the score of small patch. In this study, more than 2000 people who are undergraduates, graduates, researchers, and lecturers of University of Fire Fighting and Prevention are employed. Five - score - scale is used corresponding to five levels of assessment: "excellent", "good", "fair", "poor" and "bad".

C. Testing software

As mentioned above, double stimulus categorical rating is chosen in this testing. The image quality assessment method and system stated in [6] is good choice, however, it is only suitable for assessing quality of image as a whole. On the other hand, it cannot be directly applied for our testing experiments. Therefore, we design a new testing software (Fig 3) which is used for patch image testing. Instead of the whole images, pair of reference and testing patches (Fig 4) is read from HMII database and then the testing pair of patches is displayed and disappeared in a standard time. The time between the appearance of the testing patch pairs lasts at least five seconds so that the observer can analyze comprehensively and makes decision of evaluation.

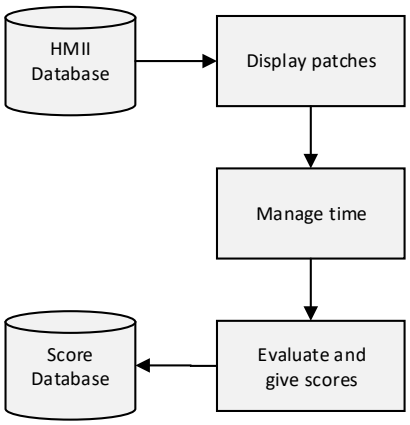


Fig. 3: Testing software.

D. Testing process

In this experiment, observers can only concentrate and assess the local image patch instead of the whole image. After taking vision test, each observer is required to observe pairs of testing patches on the testing software and to provide evaluation (Fig 5). Each pair quality is assessed following the procedure of [4] . Observer watches the original image patch within the time T1 (minimum five seconds) then clicks on the observing image patch to observe the compressed image within the time T2. This cycle of observation is repeated at least twice per pair of patches. Finally, observer gives points at five different levels as mentioned above.

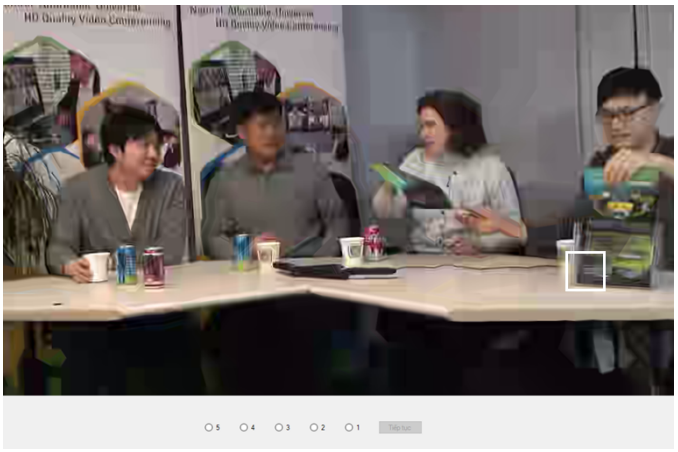


Fig. 4: Tests on testing software

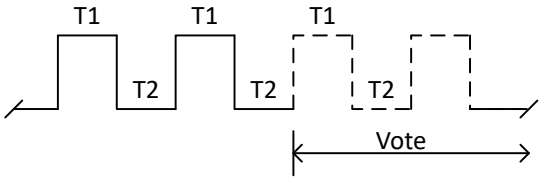


Fig. 5: Presentation structure of test material

E. Data preprocessing

Outlier rejection for image quality assessment is important for many image processing systems such as those for acquisition, compression, restoration, enhancement, reproduction etc. It has been proven to be efficient in application such as network intrusions and credit card fraud detection. Because of a large number of participated observers in the subjective testing experiment, the reliability of scores raw data is not guaranteed. This section is concerned with detecting outliers in the raw data. As a result, we would have a reliable data of scores and cleaned HMII database that can be applied to a further application.

In our experiment, 2189 different subjects rate 61600 image patch pairs of which each p^{th} is observed by S_p (up to 20) subjects. The differential mean opinion score (DMOS) of each patch pair is calculated by:

$$\overline{y_p} = \frac{1}{S_p} \sum_{s=p}^{S_p} y_{p,s}, \tag{1}$$

where $y_{p,s}$ is the differential opinion score of a subjective rating by subject s for patch pair p^{th}. Let Y_o denote the raw data and each image patch pair (R_p, D_p) is evaluated by at least 15 observers as folow:

$$Y_o = \{((R_p, D_p), \overline{y_p}) | S_p \geq 15\}. \tag{2}$$

The raw score database is not entirely good because some observers evaluate carelessly. To remove outlier in this data, we use z-score. The z-score of a subjective rating for patch pair p^{th} is calculated by the following formula:

$$Z_{p,s} = \frac{y_{p,s} - \mu_p}{\sigma_p}, \tag{3}$$

where μ_p is the mean and σ_p is the standard deviation of rating pairs. The properties of data before applying outlier rejection includes 697179 subjective ratings to 40708 image patches from 2189 different subjects. The figure below (Fig.6) shows that distribution of z-score is the standard normal distribution side-by-side. According to empirical rule, 95%, 98.7% and 99.7% of the values lie within 2, 2.5 and 3σ, respectively. After applying this rule we have result shown in table I.

We choose 2σ following the outlier rejection process suggested by [6]. The result is a reduction of 422 image patch pair scores in database. Fig. 7 shows the standard deviation of subjective ratings before and after outlier rejection. Most samples having deviations greater than 0.5 have been removed. The filtered data as follows:

$$Y_f = \{((R_p, D_p), \overline{y_p}) \in Y_o | Z_{p,s} \geq 2\sigma_p; S_p \geq 15\}, \tag{4}$$

Finally, each pair of patches is evaluated by the average clean scores of at least 15 observers (with 5 levels). These includes $N = |Y_f| = 40286$ pairs of quality annotated image patches that are subject to different distortion levels of compression. Differential mean opinion score (DMOS) for

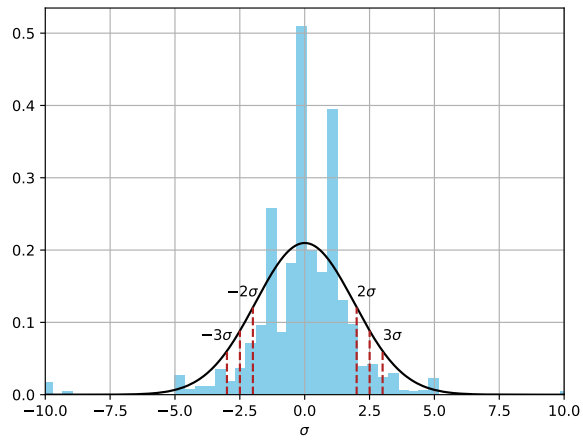


Fig. 6: The standard normal distribution of Z-score

this dataset were computed for each pair, which is in the range 1 to 5. So, final HMII comprises 40286 clean pairs and their DMOS. We compare characteristics of the proposed database with the other ten general purpose IQA databases in Table II. In summary, our database is the first patch based database and has the largest number of scores.

TABLE I: Outlier rejection results

Properties	2σ	2.5σ	3σ
Number of outliers	33199	8631	1991
Number of image patch pairs	422	136	37
Number subjects	21	7	3
Percentage outlier	5.0%	1.3%	0.3%

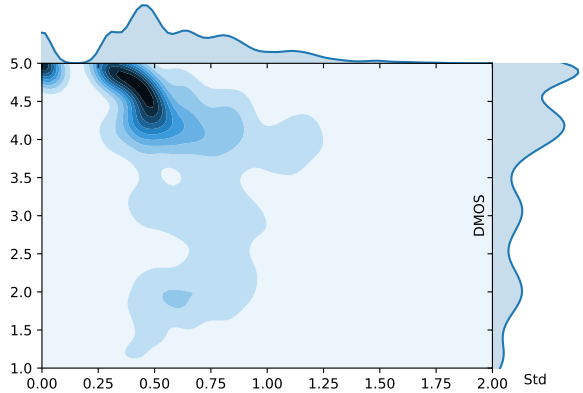


Fig. 7: Standard deviation of subjective ratings

III. Evaluation and Discussion

A. Subjective Ratings

The main aspects considered in the analysis of the subjective ratings here are the distributions of the mean (DMOS) and standard deviation σ of the subjective ratings, as these are indication of the quality range of the test material and the precision of the results. Fig. 7 shows the standard deviations of proposed database mostly in range 0 to 0.5. Those pairs have an approximate DMOS score at 5, the standard deviation is almost zero because observers are easy to make the same decision. While the $DMOS$ decreases in range 3.5 to 4.5 when standard deviations increase along but still in acceptable range.

Fig 8 illustrates the statistical relationship between $DMOS$ and compression levels (quantization parameters) for all patch pairs. It shows that in the range of 1-20, image quality has no significant change. However, it is easily detected quality changes by observer within this range. It also indicates that $DMOS$ decreases when qp increases. A typical relationship between $DMOS$ and distortion level qp, generally exhibits a skew-symmetric sigmoid form. The Pearson Linear Correlation Coefficient (PLCC) and Spearman's Rank order Correlation Coefficient (SRCC) are used as a measure of the accuracy fit of them respectively -0.807 and -0.8438. The result shows a partial but not complete relation because quality change of compressed image depending on the visual features of patches, such as, edge density, average brightness, variance. We compare subjective testing with testing pairs of patches that at the same level of compression, Fig. 9 lists score of the pairs. There is a difference in quality among complex texture area (1, 3), edge texture area (5, 6) and smooth texture area (2). The complex texture samples (1) have higher scores in compare with other. It means that the quality measurements based on signal-fidelity such as RMSE, PSNR are not homologous as those of human perceptual.

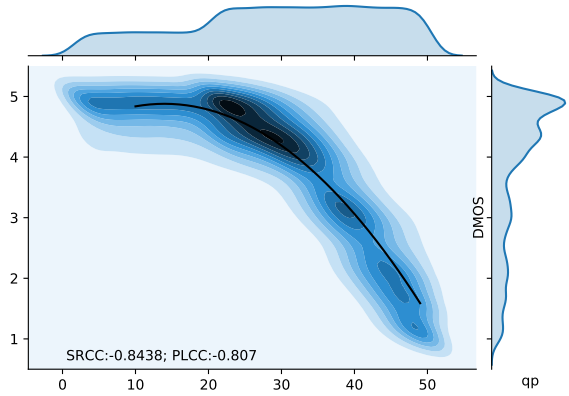


Fig. 8: Data after reject outlier

B. HMII Benchmark Analysis

1) Evaluation Method:

- *Purpose*: the purpose of the experiment is to evaluate how well an objective metric agrees with subjective preferences of subjects in HMII database evaluation.
- *Evaluation Metrics*: To evaluate the performances of the IQA algorithms, we use two standard measures including Spearman's rank order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC).
- *Experiment Setup*: We implement seven state-of-the-art algorithms PSNR, UQI, VSI, SSIM, RFSIM, FSIM and

TABLE II: Comparison characteristics of subjective Image Databases

Database	Year	Type	Data	Scores	SRC	LOD	NOD	Subjects	Ratings	Resolution	Method	PSNR
CSIQ [2]	2010	Full	DMOS+σ	866	30	5	6	25	5-7	512×512	Custom	
LIVE(I) [1]	2006	Full	DMOS+σ	779	29	7-8	5		20-29	768×512	ACR	88%
TID2008 [3]	2008	Full	Raw	1700	25	4	17	838	17	512×384	ACR	55%
TID2013 [4]	2013	Full	MOS+σ	3000	25	4	10	971	33	512×384	ACR	
PDAP-HDDS [7]	2018	Full	MOS	12000	250	4	24	38	30	FHD	ACR	54%
HMII (Proposed)	**2018**	**Patch**	**MOS**	**40286**	**308**	**49**	**1**	**2189**	**15-20**	**HD, FHD**	**DSIS**	**65-67%**
	Scores		Number of images or videos with subjective ratings.									
	SRC		Number of source (reference) images.									
	LOD		Levels of distortions.									
	NOD		Number of different types of distortions.									
	PSNR		Approximate correlation between PSNR and MOS.									

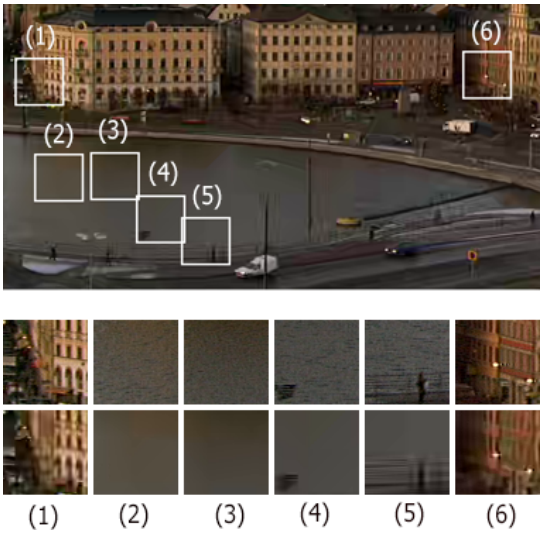


Fig. 9: Patch scores of a testing image in order are: 2.3, 1.5, 2.0, 2.2, 1.7, 1.9; RMSE: 10.9, 6.4, 5.7, 6.8, 8.4, 10.8; PSNR: 27.3, 31.9, 32.9, 31.4, 29.5, 27.4

SRSIM to predict object scores for the entire HMII database. In addition, two new methods DIQaM-FR and WaDIQaM-FR[8] are also used to predict such scores. We evaluate preference consistency using the classic correlation coefficients SRCC and PLCC.

2) Experiment results: As shown in Table III. The SRCC and PLCC are the average values for the distorted images of the same reference image, and the top two correlation coefficient values are highlighted. We can see that the PSNR and UQI are poorly correlated with human perceptual quality, and even contrary to subjective results. This defective performance of PSNR is also mentioned in the work of Zhang *et al.*[9] about Fine-Grained Quality Assessment. Although UQI combines VSI and HVS features and achieves more consistent results than PSNR in global image assessment, it is poorly correlated with human perceptual quality in fine-grained patch quality assessment. As a whole, FSIM achieves top two performances for all the cases and the SSIM achieves better performance with PLCC while SRSIM performs better with SRCC. In summary, perceptual quality metrics give better results than quality measurement based on signal-fidelity. For the two correlation coefficients, above IQA methods show quite similar characteristics, while two new methods based on machine learning fails in proposed database. Because each structure of a machine learning problem is only suitable for its own database. From the results of this experiment, it can be seen that the larger size of the patch seems to be more accurate when assessing image-patch quality by IQA algorithms.

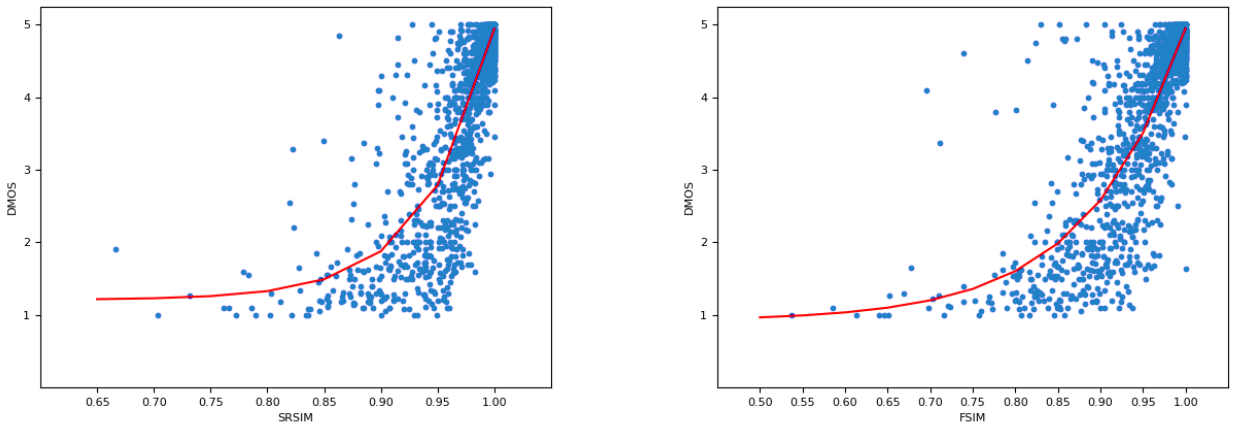
TABLE III: PLCC and SRCC for different IQA algorithms

IQA ALGORITHM	**HMII** 64 × 64		**HMII** 128 × 128	
	SRCC	**PLCC**	**SRCC**	**PLCC**
PSNR	0.7056	0.6596	0.7233	0.6723
UQI [10]	0.0233	0.0233	0.0129	0.0124
VSI [11]	0.7659	0.7659	0.7680	0.7861
SSIM [12]	**0.7878**	0.7714	**0.7989**	0.7894
RFSIM [13]	0.7747	0.7574	0.7891	0.7596
FSIM [14]	**0.7941**	**0.7997**	**0.8241**	**0.8154**
SRSIM [15]	0.7776	**0.8030**	0.7188	**0.8035**
DIQaM-FR [8]	0.5525	0.5521	0.6075	0.6057
WaDIQaM-FR [8]	0.5648	0.6738	0.6750	0.7661

Fig 10 shows the scatter distributions of subjective DMOS versus the predicted scores obtained by the FSIM and SRSIM on the proposed database. From the plots, we can see that these IQA algorithms tend to predict higher score for patches. SRSIM and FSIM frequently predict score which is higher than 0.9 (in range of [0;1]) for the image with DMOS is greater than 2 (in range of [1;5]). Although FSIM achieves highest performance with the two correlation coefficients, SRSIM achieves more consistent results with subjective results on the diagrams. These results prove that some existing IQA models perform poorly in distinguishing the fine-grained distortion levels, which are feasible to determine by human visual system. Therefore, these metrics may not be suitable for perceptual-based image compression because the distortion differences between various coding modes are usually marginal. Moreover, the fine-grained image-patch quality assessment is demanded and should be evaluated on the HMII databases. This means the researchers in the IQA field have more work to do. We expect more advanced and better IQA methods to be developed to conquer this database.

IV. Conclusion

In this work, we present a new subject quality rating database considering image patch quality assessment method for image with compression artifacts. To the best of our



(a) SRSIM (b) FSIM

Fig. 10: Objective Score by top 2 IQA on HMII

knowledge, this new database is the first image patch based among existing general purpose image quality databases with human opinion scores. We use video distortion artifacts with random quantization parameters and positions to generate the testing image patches. Therefore, 40286 image patch pairs are included in the HMII database with DMOS obtaining from 663980 opinions collected by more than 2000 observer. Finally, we test nine well-known image quality metrics on this database. Another advantage of this database is that it can provide a larger dataset to make the realization of deep learning neural networks on IQA possible.

REFERENCES

[1] L. C. H. R. Sheikh, Z. Wang and A. C. Bovik. (2006) Live image quality assessment database release 2. [Online]. Available: http://live.ece.utexas.edu/research/quality/subjective.htm

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

[3] N. P. et al. (2008) Tid2008 – a database for evaluation of full-reference visual quality assessment metrics. [Online]. Available: http://www.ponomarenko.info/tid2008.htm

[4] ——. (2013) Tampere image database 2013. [Online]. Available: http://www.ponomarenko.info/tid2008.htm

[5] T. T. Pham, T. D. Dinh, V. X. Hoang, T. Vu Huu, and T. H. Le, "Distortion model based on perceptual of local image content," *In 4th International Conference on Consumer Electronics Asia*, 06 2019.

[6] I. R. Assembly and I. T. Union, *Methodology for the Subjective Assessment of the Quality of Television Pictures*. International Telecommunication Union, 2003.

[7] T. Liu, H. Liu, S. Pei, and K. Liu, "A high-definition diversity-scene database for image quality assessment," *IEEE Access*, vol. 6, pp. 45 427–45 438, 2018.

[8] S. Bosse, D. Maniry, K. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, Jan 2018.

[9] X. Zhang, W. Lin, S. Wang, J. Liu, S. Ma, and W. Gao, "Fine-grained quality assessment for compressed images," *IEEE Transactions on Image Processing*, vol. PP, pp. 1–1, 10 2018.

[10] Zhou Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, March 2002.

[11] L. Zhang, Y. Shen, and H. Li, "Vsi: A visual saliency-induced index for perceptual image quality assessment," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 23, 08 2014.

[12] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "Ssim-motivated rate-distortion optimization for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 4, pp. 516–529, April 2012.

[13] L. Zhang, L. Zhang, and X. Mou, "Rfsim: A feature based image quality assessment metric using riesz transforms," in *2010 IEEE International Conference on Image Processing*, Sep. 2010, pp. 321–324.

[14] 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, Aug 2011.

[15] L. Zhang and H. Li, "Sr-sim: A fast and high performance iqa index based on spectral residual," in *2012 19th IEEE International Conference on Image Processing*, Sep. 2012, pp. 1473–1476.