MCL-JCV: A JND-BASED H.264/AVC VIDEO QUALITY ASSESSMENT DATASET

Haiqiang Wang*, Weihao Gan*, Sudeng Hu*, Joe Yuchieh Lin*, Lina Jin*,
Longguang Song*, Ping Wang*, Ioannis Katsavounidis†, Anne Aaron † and C.-C. Jay Kuo *
* University of Southern California, Los Angeles, California, USA
† Netflix, Los Gatos, California, USA

ABSTRACT

A compressed video quality assessment dataset based on the just noticeable difference (JND) model, called MCL-JCV, is recently constructed and released. In this work, we explain its design objectives, selected video content and subject test procedures. Then, we conduct statistical analysis on collected JND data. We compute the difference between every two adjacent JND points and propose an outlier detection algorithm to remove unreliable data. We also show that each JND difference group can be well approximated by a normal distribution so that we can adopt the Gaussian mixture model (GMM) to characterize the distribution of multiple JND points. Finally, it is demonstrated by experimental results that the proposed JND analysis performed in the difference domain, called the D-method, achieves a lower BIC (Bayesian information criteria) value than the previously proposed G-method.

Index Terms— Video Quality Assessment, Just Noticeable Difference, Gaussian Mixture Model, Outlier Detection

1. INTRODUCTION

Modern video coding standards such as H.264/AVC [1] and High Efficiency Video Coding (HEVC) [2] have greatly improved video compression efficiency by removing spatial, temporal and statistical redundancies effectively. However, the mean-squared-errors (MSE) distortion measure used in these standards has been criticized for not being well correlated with human visual experience. To address this problem, there has been a large amount of efforts in developing new visual quality metrics [3, 4]. All of these proposed distortion/quality metrics, such as SSIM [5], FSIM [6], EVQA [7], are in form of parametric curves (or fused parametric curves), where model parameters can be determined by regression. They provide a continuous-scale quality function. Although there is an inspiring study in replacing the PSNR measure with the SSIM measure in video coding [8], we do not see a decisive advantage of the new quality metric in offering better perceptual quality at the same bit rate.

Actually, humans cannot perceive small variation in pixel differences. This has been clearly illustrated in the psychovisual study of just-noticeable difference (JND) [9, 10]. The traditional rate-distortion (R-D) function does not take the nonlinear human perception process into account. In the context of image/video coding, recent subjective tests in [11] show that humans can only perceive discrete-scale quality levels over a wide range of coding bitrates. To be more precise, the quality of H.264/AVC video perceived by humans is a stair function of the quantization parameter (QP), where the jumps between adjacent quality levels are called the JND points. The discrete nature of perceived video quality should somehow be exploited in perceptual video coding.

Subjective tests for video coding are typically conducted by very few experts called gold eyes for the worst-case analysis. However, the worst-case analysis does not reflect the statistical behavior of the group-based quality of experience (QoE). For given visual content, the number of perceived distortion levels and JND positions depend on each individual. When the subjective test is conducted with respect to a viewer group, it is meaningful to study their QoE statistically to yield an aggregated function.

Being motivated by the above two observations, we build an H.264/AVC coded video quality dataset consisting of 30 video clips of a wide content variety, and each of them are viewed by 50 subjects. The JND points of each subject with respect to each video clip are recorded and analyzed. This dataset, called MCL-JCV (Media Communications Lab JND-based Coded Video), is available to the public [12]. For the JND data analysis, we compute the difference between every two adjacent JND points and propose an outlier detection algorithm to remove unreliable data. We also show that each JND difference group can be well approximated by a normal distribution so that we can adopt the Gaussian mixture model (GMM) to characterize the distribution of multiple JND points.

The rest of this paper is organized as follows. The data collection procedure for the MCL-JCV dataset is described in Sec. 2. Statistic analysis of the difference of two consecutive JND points is performed in Sec. 3. Statistic modeling of multiple JND points using the GMM is examined in Sec. 4. Experimental results in Sec. 5 show that the JND analysis in the difference domain, called the D-method, achieves a lower Bayesian information criteria (BIC) value than the G-method proposed in [13]. Finally, concluding remarks and future work are given in Sec. 6.

2. MCL-JCV DATASET CONSTRUCTION

Humans cannot perceive quality difference between two video sequences of the same content if their coding QP values are very close to each other although their PSNR (or MSE) values are still different. The MCL-JCV dataset is designed to measure this phenomenon for each test subject. Clearly, this phenomenon depends on the visual content as well as the test individual.

As compared with image quality assessment datasets [14], existing video quality assessment datasets are rather limited in three aspects: the number of test sequences, the diversity of video contents, and the number of participating subjects. For example, the LIVE video quality dataset [15] contains only 10 source sequences assessed by 38 subjects.

Table 1: Diversity of video characteristics in the MCL-JCV dataset.

Genres		Semantics		Features	
Cartoon	4	People	13	Fast Motion	9
Sports	3	Water	3	Camera Motion	11
Indoor	4	Salience	9	Dark scene	6

Fig. 1: Representative thumbnail images of 30 source sequences in the MCL-JCV dataset.

To address these limitations, we collected 30 source (uncompressed) video sequences of resolution 1902×1080 in building the MCL-JCV dataset. Their representative thumbnail images are shown in Fig. 1. They cover a wide diversity of video characteristics. The number of MCL-JCV source video clips in several semantic categories are shown in Table 1. Each source video clip has a duration of 5 seconds but different frame rate. They are encoded by H.264/AVC, and the quality of coded video is determined by the quantization parameter (QP), which takes an integer value ranging from 1 to 51. By including the source video clips, we have $30 \times 52 = 1560$ video sequences in the MCL-JCV dataset.

The 51 degraded clips encoded from the same source form one test set. Each test set was evaluated by 50 subjects. They were seated in a controlled environment. The viewing distance was 2 meters (1.6 times the monitor height) from the center of the monitor to the seat. The video sequences were displayed on a $65''$ TV with native resolution of 3840×2160. A subject compared the quality of two sequences displayed one after another, and determined whether these two sequences are noticeably different or not by choosing 'yes' or 'no'. More than 150 volunteers randomly recruited in the campus participated in the subjective test, and each subject evaluated 10 sets of the sequences. The subjects were trained to be familiar with coding artifacts before the test starts. It takes about 40 minutes to finish a test section. The sequences were randomly selected and the same sequence did not appear more than once for the same subject.

The first JND point of a sequence for a test individual is the transitional point between the best and the second best quality levels obtained in the test as shown in Fig. 2. Generally speaking, the ith JND point is the transitional point that separates the ith and the $(i+1)$th quality levels. The JND points are searched in a sequential manner.

To begin with, we use the video sequence with $QP = 1$ as the anchor sequence, which is fixed throughout the search of the first JND point. That is, we would like to find the sequence that is not only noticeably different from the anchor sequence but also with the smallest QP value. To speed up the search process, we adopt the bisection method in the QP domain, where $QP = 2, 3, \cdots, 51$. For more details about the bisection search process, we refer readers to [16]. Once the ith JND point is determined, we use it as the new anchor, and search among sequences with QP values larger than the anchor QP for the $(i+1)$th JND point. The process repeats until we reach the largest QP (i.e., $QP = 51$.)

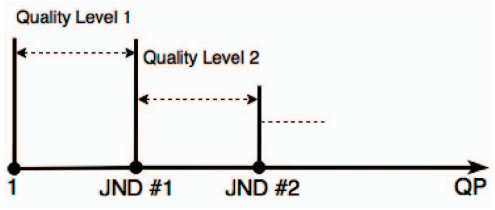


Fig. 2: Illustration of multiple quality levels with JND points as their transition points.

3. GAUSSIAN TEST AND OUTLIER REMOVAL FOR JND DIFFERENCES

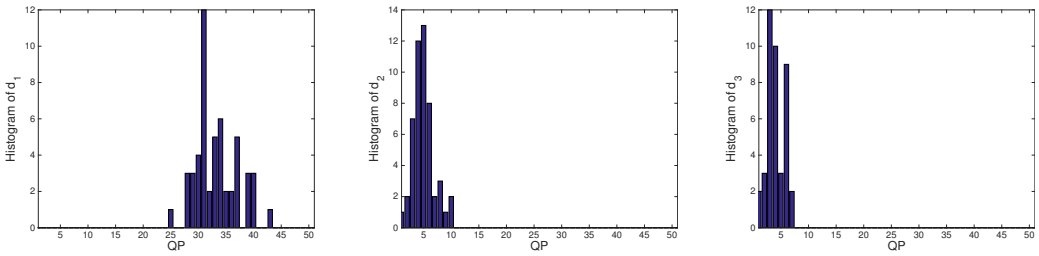
3.1. Normality Test

For each video sequence, we obtain a set of JND data samples for each subject. Due to the change of the anchor position, it is more meaningful to analyze the JND difference value as discussed below. We use $\boldsymbol{x} = \{x_1, x_2, \cdots, x_N\}$ to denote the set of JND points for a subject against a video sequence and denote the difference between consecutive JND points by

$$d_n = x_n - x_{n-1},$$

where $x_0 = 1$ is the reference of the first JND. We would like to study the distribution test of d_n among all 50 subjects for the same content. Without loss of generality, we show the distributions of d_1, d_2 and d_3 from 50 samples of sequence # 7 in Fig. 3, and use the Jarque-Bera test [17] to validate the normality for each of them. The Jarque-Bera test is a two-sided goodness-of-fit test for normality of observations with unknown parameters. Its test statistic is defined as

$$JB = \frac{n}{6}(s^2 + \frac{(k-3)^2}{4}), \quad (1)$$



(a) (b) (c)

Fig. 3: Distributions of JND differences with 50 samples of video No 7, where (a), (b), and (c) are the histograms of d_1, d_2, and d_3, respectively.

where n is the sample size, s is the sample skewness, and k is the sample kurtosis. The test rejects null hypothesis if the statistic JB in Eq. 1 is larger than the precomputed critical value at a given significance level.

Table 2: Normality test results on JND differences.

	d_1	d_2	d_3	d_4	d_5
Passed	22	18	22	9	2
Total	30	30	29	15	2
Percentage	73.3%	60%	75.8%	60%	100%

The normality test results on the MCL-JCV dataset are given in Table 2. Recall that the MCL-JCV contains 30 sequences. If the sample number is much less than 50 for a given sequence, we remove it from the test. Thus, the total test cases for d_3, d_4 and d_5 are less than 30. We see that the JND difference can be adequately approximated by a Gaussian distribution with a probability of 60%-70%. Thus, we can express it as

$$d_n \sim \mathcal{N}(\mu_n, \sigma_n^2), \tag{2}$$

where μ_n is the mean and σ_n is the standard deviation. Under the independent assumption of d_n and the fact

$$x_n = 1 + \sum_{i=1}^{n} d_n, \quad n = 1, 2, \cdots,$$

we can express the distribution of x_n in form of

$$x_n \sim \mathcal{N}(\mu_{X,n}, \sigma_{X,n}). \tag{3}$$

where $\mu_{X,n} = 1 + \sum_{i=1}^{n} \mu_i$ and $\sigma_{X,n}^2 = \sum_{i=1}^{n} \sigma_i^2$.

3.2. Outlier Detection and Removal

It is important to detect and remove outliers in any statistical procedure. In this work, we conduct the outlier detectioin in the JND difference domain for each sequence independently. To proceed, we use

$$\boldsymbol{d}^m = (d_1^m, d_2^m, \cdots, d_N^m),$$

to denote the JND difference sample for subject m and find the median vector among 50 test subjects

$$\tilde{\boldsymbol{d}} = (\tilde{d}_1, \tilde{d}_2, \cdots, \tilde{d}_N),$$

where $\tilde{d}_i = median(d_i^1, d_i^2, \cdots, d_i^M)$, and M is the subject number who observed the ith JND during the subjective test.

We compute the sample correlation coefficient to determine the closeness of an individual sample vector and the median vector as:

$$r = \frac{\sum_{i=1}^{N}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})}{\sqrt{\sum_{i=1}^{N}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^2}\sqrt{\sum_{i=1}^{N}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^2}}, \tag{4}$$

where $\boldsymbol{x} = \boldsymbol{d}^m$ and $\boldsymbol{y} = \tilde{\boldsymbol{d}}$. If the correlation is too low, we treat the sample vector as an outlying sample. Here, we choose a threshold of 0.9 and identify samples with $r < 0.9$ outliers. There are 0, 1, 2, and 3 outliers for 17, 7, 5, and 1 sequences, respectively.

4. JOINT JND MODELING AND PROCESSING

The JND differences were approximated by the Gaussian distribution independently. However, the distribution of JND is of more interest. Based on the above discussion, we can integrate multiple JND points together with a GMM of N components. That is, the JND distribution over quantization parameter x can be expressed as

$$f(x) = \sum_{i=1}^{N} \pi_i \mathcal{N}(\mu_{X,i}, \sigma_{X,i}^2), \tag{5}$$

where π_i is the mixture weight that satisfies $\sum_{i=1}^{N} \pi_i = 1$. Our goal is to optimize the following set of parameters iteratively

$$\Theta = \{\pi_i, \mu_{X,i}, \sigma_{X,i}^2, \quad i = 1, \cdots, N\}, \tag{6}$$

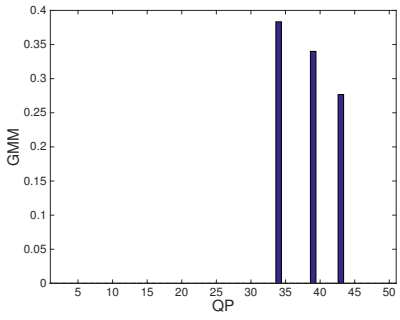
so that it will fit the collected JND data as close as possible. This can be done by the Expectation Maximization (EM) algorithm [18]. It is well known that the EM algorithm is sensitive to initial values of parameters. We adopt the prior derived in Eq. (3) and the uniform weight $\pi_i = 1/N$ as the initialization.

Once a GMM is derived, we would like to consider the mean quality of experience (QoE) of the viewer group, and draw the stair quality function (SQF) accordingly. This can be done as follows. First, the posterior distribution of the ith component can be written as

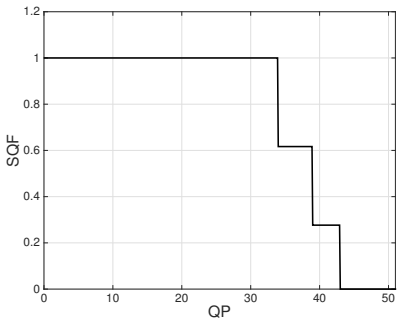
$$H_i = \frac{\pi_i \mathcal{N}(\mu_{X,i}, \sigma_{X,i}^2)}{\sum_{i=1}^{N} \pi_i \mathcal{N}(\mu_{X,i}, \sigma_{X,i}^2)}, \tag{7}$$

Then, we can approximate the JND distribution to the sum of N spikes:

$$JND(x) = \sum_{i=1}^{N} H_i \delta(x - \mu_{X,i}), \tag{8}$$



(a)



(b)

Fig. 4: An example of the relationship between GMM and SQF: (a) locations and heights of three posterior Gaussian components and (b) the corresponding SQF.

where $\delta(\cdot)$ is the Dirac delta function. The SQF is obtained by integrating the JND function from the largest QP to the smallest QP. Mathematically, this is equivalent to

$$SQF(x) = 1 - \int_0^x JND(t)\,\mathrm{d}t, \tag{9}$$

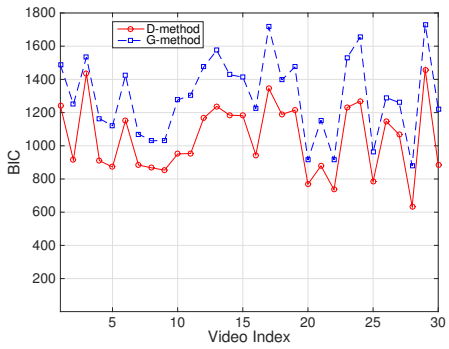
which is a monotonically non-increasing piecewise-constant stair function of x (i.e. QP). One example to illustrate the relationship between the simplified JND distribution and the SQF is given in Fig. 4.

5. EXPERIMENTAL RESULTS AND DISCUSSION

Two processing techniques were proposed in the literature to analyze the JND distribution for a compressed image quality dataset called MCL-JCI [16]. They are the K-method [11] based on a k-mean clustering algorithm and the G-method [13] by applying the GMM to raw JND data directly. The G-method outperforms the K-method in terms of a lower Bayesian information criterion (BIC) value [19]. The BIC value offers a relative estimate of information loss in modeling samples with a statistical model. It is commonly used in model selection among a finite set of models. Mathematically, it is defined as

$$BIC = -2 \cdot ln(\hat{L}) + k \cdot ln(n), \tag{10}$$

where $\hat{L} = p(x|\Theta)$ is the maximized value of the likelihood function of the model, ln is the natural log, k is the number of free parameters in the model, and n is the number of samples. A better fit will derive the negative log likelihood term lower and a smaller k will derive the model complexity term lower for fixed n. The BIC value strikes a balance between goodness of fit and model complexity.



(a)

Fig. 5: The BIC comparison between the D-method and the G-method applied to the 30 video sequences in the MCL-JVC dataset.

The method proposed in this work is called the D-method since it processes the JND difference data first to get the prior for the EM algorithm. We apply both the G-method and the D-method to the MCL-JCV dataset and compare the BIC values for 30 video sequences in Fig. 5. As shown in the figure, the BIC value of the D-method is always lower than that of the G-method, which means the proposed D-method offers a better processing method than the G-method.

The main difference between the G-method and the D-method lies in the different initializations in the EM algorithm. In the G-method, local peaks in the histogram of all samples are adopted as the initial component means and their initial variances are set to unity. This processing has several shortcomings. First, it fails to take the difference between JND points into consideration. Actually, the first JND points have a wide spread and some of them may overlap with the 2nd JND group. Second, the JND points are divided into 3 groups with an ad hoc rule, and the number of components in each group is optimized based on the BIC value individually. It does not address global optimality. In contrast, The D-method attempts to model the ith JND differences by checking its normality and uses derived means and variances to initialize the EM algorithm. There is no artificial group partitioning. Furthermore, an outlier detection scheme is conducted to remove inconsistent samples. This explains the superior BIC performance of the D-method over the G-method.

6. CONCLUSION AND FUTURE WORK

The design objectives, selected video content and subject test procedures of the MCL-JCV dataset were explained, an outlier detection scheme was presented to provide more robust data samples, and a new JND data processing technique was proposed and justified. The new processing technique, called the D-method, always achieves a lower BIC than the G-method proposed in [13]. All 30 source and coded video sequences will be available to the public with measured raw JND data for each test subject. Thus, it allows users to do their own processing. The SQF derived by the D-method will also be released. In the near future, we will adopt a machine learning approach to predict the SQF of new video content not in the training dataset.

7. REFERENCES

[1] Laurent Aimar, Loren Merritt, Eric Petit, Min Chen, Justin Clay, Mns Rullgrd, Christian Heine, and Alex Izvorski, "x264-a free h264/avc encoder," *Online (last accessed on: 04/01/07): http://www. videolan. org/developers/x264. html*, 2005.

[2] Gary J Sullivan, J-R Ohm, Woo-Jin Han, and Thomas Wiegand, "Overview of the high efficiency video coding (hevc) standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1649–1668, 2012.

[3] W. Lin and C.-C. Jay Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297–312, 2011.

[4] A. Aaron, Zhi Li, M. Manohara, J.Y. Lin, E.C.-H. Wu, and C.-C.J. Kuo, "Challenges in cloud based ingest and encoding for high quality streaming media," in *Image Processing (ICIP), 2015 IEEE International Conference on*, Sept 2015, pp. 1732–1736.

[5] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, April 2004.

[6] Lin Zhang, D. Zhang, Xuanqin Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *Image Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2378–2386, Aug 2011.

[7] J.Y. Lin, Chi-Hao Wu, I. Katsavounidis, Zhi Li, A. Aaron, and C.-C.J. Kuo, "EVQA: An ensemble-learning-based video quality assessment index," in *Multimedia Expo Workshops (ICMEW), 2015 IEEE International Conference on*, June 2015, pp. 1–6.

[8] Tiesong Zhao, Kai Zeng, A. Rehman, and Zhou Wang, "On the use of SSIM in HEVC," in *Signals, Systems and Computers, 2013 Asilomar Conference on*, Nov 2013, pp. 1107–1111.

[9] Gordon E Legge and John M Foley, "Contrast masking in human vision," *JOSA*, vol. 70, no. 12, pp. 1458–1471, 1980.

[10] Zhenyu Wei and K.N. Ngan, "Spatio-temporal just noticeable distortion profile for grey scale image/video in DCT domain," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 19, no. 3, pp. 337–346, March 2009.

[11] Joe Yuchieh Lin, Lina Jin, Sudeng Hu, Ioannis Katsavounidis, Zhi Li, Anne Aaron, and C.-C. Jay Kuo, "Experimental design and analysis of JND test on coded image/video," in *SPIE Optical Engineering+ Applications*. International Society for Optics and Photonics, Sep 2015, pp. 95990Z–95990Z.

[12] "MCL-JCV dataset," `http://mcl.usc.edu/mcl-jcv-dataset/`, Accessed: 2016-02-30.

[13] Sudeng Hu, Haiqiang Wang, and C-C Jay Kuo, "A GMM-based stair quality model for human perceived JPEG images," *arXiv preprint arXiv:1511.03398*, 2015.

[14] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "Color image database TID2013: Peculiarities and preliminary results," in *Visual Information Processing (EUVIP), 2013 4th European Workshop on*, June 2013, pp. 106–111.

[15] K. Seshadrinathan, R. Soundararajan, A.C. Bovik, and L.K. Cormack, "Study of subjective and objective quality assessment of video," *Image Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1427–1441, June 2010.

[16] Lina Jin, Joe Yuchieh Lin, Sudeng Hu, Haiqiang Wang, Ping Wang, Ioannis Katasvounidis, Anne Aaron, and C.-C. Jay Kuo, "Statistical study on perceived JPEG image quality via MCL-JCI dataset construction and analysis," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2016.

[17] Carlos M Jarque and Anil K Bera, "A test for normality of observations and regression residuals," *International Statistical Review/Revue Internationale de Statistique*, pp. 163–172, 1987.

[18] Arthur P Dempster, Nan M Laird, and Donald B Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[19] David F Findley, "Counterexamples to parsimony and BIC," *Annals of the Institute of Statistical Mathematics*, vol. 43, no. 3, pp. 505–514, 1991.