Subjective test methodology optimization and prediction framework for Just Noticeable Difference and Satisfied User Ratio for compressed HD video

Jingwen Zhu, Anne-Flore Perrin, Patrick Le Callet

Subjective test methodology optimization and prediction framework for Just Noticeable Difference and Satisfied User Ratio for compressed HD video

Jingwen Zhu *, Anne-Flore Perrin*, Patrick Le Callet*
*Nantes Université, Ecole Centrale Nantes, CAPACITÉS SAS, CNRS, LS2N, UMR 6004, F-44000 Nantes, France,

***Abstract*—Just Noticeable Difference (JND) and Satisfied User Ratio (SUR) has been widely investigated for compressed image and video to use the least resources (*e.g., storage and bandwidth*) without damaging the Quality of Experience (QoE) for end users. However, the current JND subjective test methodologies are extremely time consuming due to the large range of encoding parameters. Besides, the state-of-the-arts SUR/JND prediction models get non-negligible prediction error due to the limited masking effect features. To this end, we first proposed a pre-processing method to reduce the JND subjective test time by using dynamic range of encoding parameters and collected a new Video-Wise JND (VW-JND) datasets for HD videos: HD-VJND. Afterwards, based on the collected datasets, we proposed a SUR prediction framework by extracting 3 types of features 1) masking effect features; 2) bitstreams features; 3) content features. Feature selection is applied to extracted features before regression. Besides, we also compared the direct and indirect SUR value predictions methods. Experiment results shows that our proposed optimization can reduce 7.14% of the subjective experiment time compared to the widely used Robust Binary Search (RBS). Furthermore, the proposed SUR and JND prediction frameworks outperform the SOTA model in HD-VJND datasets.**

***Index Terms*—Video Quality Assessment, Just Noticeable Difference, Satisfied User Ratio, HD videos**

I. Introduction

Human Visual System (HVS) cannot perceive detailed differences between the pristine images/videos and their compressed versions with small distortion. In the context of image/video compression, Just Noticeable Difference (JND) is the minimum distortion level from which human eye begins to perceive difference between the reference image/video and the compressed ones when increasing the level of compression. Specifically, when the reference is the no-compression version, *e.g.,* sources (SRCs) video, the JND is called *1st* JND, and if we take the *1st* JND as the new reference/anchor, the next JND is then the *2nd* JND, etc. The proxy of JND for compressed image/video is usually the encoding parameters of codec, *e.g.,* quantization parameter (QP) in H.264. Acquisition of JND has been widely investigated both subjectively [1]–[5] and objectively [6]–[13] for image/video compression because it helps save resources *e.g.,* storage and bandwidth, without damaging the quality for end-users.

It is well known that JND depends on 3 factors: (1) visualization setting,(*e.g.,* display, environment...) (2) subjects and (3) image/video contents [14]. In this paper, the first two factors are fixed and we will focus on the impact of difference of contents on the *1st* JND of High-Definition (HD) videos.

For the same content, the JND varies with observers because of their various visual sensitivity. To tackle this variation, Satisfied User Ratio (SUR) curve has been widely used [5]–[7], [12], [13] to measure JNDs for a group of observers. SUR curve is defined as the Complementary Cumulative Distribution Function (CCDF) of the distribution of JNDs for a group of observers [14]. Therefore, SUR curve is monotonic non-increasing, it indicates the relationship between the distortion level d and the percentage of observers $p\%$ who are satisfied. Specifically, for a given distortion level d, its corresponding value $p\%$ in the SUR curve means that there are $p\%$ observers in the group who do not perceive any difference between the reference and distorted videos with distortion level smaller than d. For a given threshold p, the corresponding distortion level d in the SUR curve is defined as $p\%$SUR. $p\%$SUR is denoted as $p\%$JND in some previous works [5], [12], [13]. In this work, we use $p\%$SUR to avoid misunderstanding.

Subjective study: Subjective JND datasets are the base on JND studies. Table I is the comparison of existing VW-JND. One of the biggest challenges in JND subjective testing is that the JND search process is extremely time consuming because of the large range of encoding parameters, for example in VideoSet [5], every SRC is encoded into 51 Processed Video Sequences (PVSs) using H.264 with QP from 1 to 51. The list of these PVSs is named as JND candidate playlist (JCP). Instead of comparing the reference with each PVS from QP = 1 until JND is found, *Wang et al.* [5] proposed a Robust Binary Search (RBS) procedure which is inspired by the widely used binary search algorithm in computer science. During the Classic Binary Search (CBS), the observer will be asked at the first time if they can perceive the difference between the video with QP = 0 (reference) and QP = 25 (middle of the original interval of JCP). If "YES", the interval QP = [26, 51] will be excluded in the next comparison; if "NO", the interval QP = [0, 24] will be excluded. However, this will bring problems when the observer makes unconfident decision for the previous comparison. The RBS is a modified version of the CBS which only eliminate one quarter of the original interval instead of removing the half of it, *e.g.,* interval QP = [39, 51] instead of QP = [26, 51] in the previous example.

However, the JND subjective experiment time is still non-negligible even with the help of RBS. In this work, we

TABLE I
COMPARISON OF VW-JND DATASETS.

Name	Number of contents	A/B[1]	Codec	Encoding parameter range	Step size
MCL-JVC	26	50/120	H.264	QP (1-51)	1
VideoSet	220	30/800	H.264	QP (1-51)	1
HD-VJND (**ours**)	180	20/20	HEVC	CRF (**dynamic**)	0.25

[1] A: number of the observers for each content; B: number of observers during the entire subjective test; when A is not equals to B, it means that different contents are evaluated by different observer groups.

proposed a method to reduce the subjective test time and collected a new VW-JND datasets named as HD-VJND for HEVC with dynamic range of Constant Rate Factor (CRF).

Objective study: It is not practical and extremely time consuming to conduct JND subjective tests for every new content. Therefore, it is important to develop objective JND/SUR prediction models based on the subjective JND datasets. *Wang et al.* [10] proposed a model using Support Vector Regression (SVR) to predict SUR curve and accordingly 75%SUR based on masking effect features [15], [16] extracted from SRCs and quality degradation features computed from all PVSs on VideoSet. By considering the spatial and temporal information features via deep learning, *Zhang et al.* [13] improved the SUR prediction accuracy using Video Wise Spatial-Temporal SUR. Instead of using encoding parameter QP as proxy of SUR curve, *Zhang et al.* [12] used bitrates as the proxy of SUR curve and made predictions with Gaussian Processes Regression (GPR) by extracting masking effect features, re-compression features and basic attribute features.

However, all the above mentioned JND/SUR prediction models are based on the assumption that the distribution of individual JND for a group of observers follows Gaussian distribution, which may not be the optimal mathematical modeling. Furthermore, when predicting SUR curve, they predict each point of the curve by using features from every PVSs, which is not practical and has high computational complexity because each video needs firstly to be compressed into many PVSs. A previous work [14] investigated these problems by firstly comparing several different mathematical modelings of SUR curve and secondly computing the SUR curve by predicting the modeling parameters only based on the features of SRC. Nevertheless, the prediction errors of SUR curve and 75%SUR are still non-negligible due to the limitation of masking effect features. In this work, we propose a SUR prediction framework (see Fig. 2) that extract many different types of features followed by features selection and regression.

II. METHODOLOGY

A. HD-VJND datasets

We collected a new VW-JND datasets named as HD-VJND datasets, as shown in the last row of Table I. 180 contents are selected from 606 video clips provided by research partner based on the content selection strategy proposed in [17] such that the selected contents are with wide-range of difference. The duration of each video is 10 seconds. 20 observers with correct visual acuity involved in the subjective JND test. The displays are 55-inch calibrated 4K TVs and the viewing distance was set as $3H$ for HD contents as recommended in ITU-R BT.2022 [18], where H is the height of the screened video. As mentioned in Section I, the JND search procedure is still time consuming with the RBS process. The JND search time depends on the length of JCP (*e.g.,* length of JCP = 51 in VideoSet per content). Meanwhile, it is well known that from a certain level of compression (*e.g., QP = 40*), it is almost certain that anyone with correct visual acuity can perceive difference between the reference and the PVS. Therefore, we proposed a method to optimize the JND subjective test time by reducing the length of JCP with the help of a pre-processing using the mapping function from VMAF to JND proposed in [19]. The mapping function for HD videos is shown in Fig. 1. It can be observed that the higher the VMAF difference (ΔVMAF) between two videos (same content, different encoding recipes), the more likely it is for humans to perceive differences between them in terms of quality.

The idea is to remove the low quality PVSs that human eyes can perceive "for sure" differences to reduce the numbers of comparison before finding the JND. For a given threshold $thr\%$, the corresponding value of ΔVMAF in the mapping function is denoted as $V_{thr\%}$. The reference for the *1st* JND is SRC, therefore $\Delta\text{VMAF} = \text{VMAF}(SRC) - \text{VMAF}(PVS)$ and $\text{VMAF}(SRC) = 100$. The PVS whose ΔVMAF is larger than $V_{thr\%}$ will be removed from the JCP. Eq.(1) stipulates the condition to eliminate the PVS to save subjective test time.

$$\text{VMAF}(PVS) < 100 - V_{thr\%} \tag{1}$$

As illustrated in Table II, we compared the maximum number

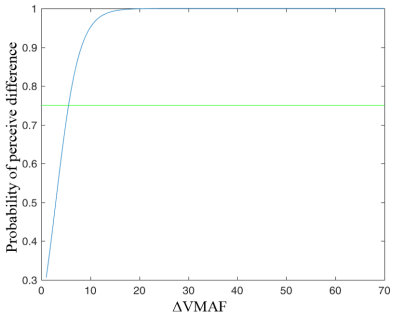

Fig. 1. Mapping function between ΔVMAF and Probability of perceive difference between 2 videos with different encoding recipes for HD videos

of comparison. Estimated time of one comparison is 20 seconds, it includes the time for one observer to watch the reference and distorted videos and to answer the question of whether they can perceive differences. Mean of $len(JCP)$ is the average length of the JCP across the entire datasets. With the decrease of the $thr\%$ in the mapping function, $V_{thr\%}$ will reduce and the number of PVSs eliminated will increase according to Eq.(1), thus the average length for JCP will decrease. However, it is possible that the JND video will be eliminated during this procedure. The last columns in Table II indicate the number of videos whose JND are excluded during the pre-processing. It can be observed that only when the threshold is close to 1, we can ensure not to remove any JND in VideoSet. For CBS, it is well known that the maximum

number of comparison is

$$\log_2\left(len(JCP)\right) = \log_{(1/2)}\left(1/len(JCP)\right), \quad (2)$$

However, the interval to keep for RBS is 3/4 instead of 1/2 in each iteration, thus the maximum number of comparison is calculated by $\log_{(3/4)}\left(1/len(JCP)\right)$. It can be concluded that our proposed method can reduce by 7.14% the subjective test duration without removing any mandatory information.

TABLE II
BENCHMARK BETWEEN OUR SOLUTIONS AND CLASSIC RBS JND SEARCH STRATEGIES IN TERMS OF TESTING TIME IN VIDEOSET (1080P)

JND search method		Mean of *len*(JCP)	Max comparison	Duration (s)	JND excluded
baseline [5]		52	14	280	0
thr. =	99%	36.99	13	260	0
	95%	29.54	12	240	45
	85%	27.21	11	220	113
	75%	24.85	11	220	159

Therefore, we used this pre-processing for HD-VJND collection. To our best knowledge, this is the first VW-JND datasets using HEVC as codec and CRF as JND proxy.

B. SUR prediction framework

There are two steps (see Fig. 2) for the entire pipeline: (1)Modeling; (2)Prediction. Modeling includes computing the empirical SUR curve ($\mathrm{SUR_{emp}}$) from the JND distribution of the group-users and finding the best mathematics model to fit the $\mathrm{SUR_{emp}}$. The fitted SUR curve is denoted as $\mathrm{SUR_{analy}}$. For more details about the modeling, we refer readers to [14]. After generating ground truth from modeling, we use SRC as input to extract features, select features and make predictions. The prediction framework is detailed as follows:

1) Feature extraction: Three types of features: (1) masking effect features, (2) bitstream features and (3) content features are extracted as illustrated in Fig. 2.

Masking effect Features measures randomness/regularity temporally (temporal randomness (TR) [16]) and spatially (spatial randomness (SR) [15]). When the randomness is high, it will be difficult for human to perceive difference, it masks the distortion for HVS. Masking effect features were used in [10], [14] to predict JND.

As shown in Fig. 2, SRC is segmented into small video patches both spatially and temporally to extract features from the eye fixation level. The dimensions of video patches are set the same as [10]. SR and TR are calculated on each small video patche to obtain feature matrices F_{SR} and F_{TR}. The statistic histogram (Eq. (3)) with number of bins equals to 20 is applied as pooling method to reduce the feature dimension.

$$\overrightarrow{SR} = Hist_{20}(F_{SR}), \quad \overrightarrow{TR} = Hist_{20}(F_{TR}) \quad (3)$$

Bitstream Features are widely used for light-weight quality estimation [20]. Before extracting bitstream features in Table III, SRCs are first compressed into a near lossless PVS with CRF = 5. The bitstream features are extracted using videoparse [21] without decoding pixel information.

The temporal and spatial pooling function are defined as:

$$F_{time} = \{Mean, Std, Max, Skew, Kurt\} \quad (4)$$

$$F_{space} = \{Mean, Std\} \quad (5)$$

where $Mean$ is the average value, Std indicates standard deviation, Max denotes maximum, $Skew$ represents skewness, and $Kurt$ is the kurtosis. The dimension of features equals to the product of the dimension of F_{time} and F_{space} (*e.g.,* for motion features, we first compute the $Mean$ and Std of motions intra frame (spatially); afterwards, $Mean$, Std, Max, $Skew$ and $Kurt$ are calculated based on the two previous-computed spatial value ($Mean$ and Std for each frame) inter frame (temporally) respectively.)

TABLE III
BITSTREAM FEATURES SUMMARY

Features	dimension
Average framerate	1
Bitrate	1
$\mathrm{Ratio(non-I)} = \frac{Nb(\mathrm{non-I\ \ frame})}{Nb(\mathrm{all\ \ frame})}$	1
Max(Framerate)	1
F_{time}(non - I frame size)	5
$F_{time}\{F_{space}$(horizontal motion)$\}$	5*2 = 10
$F_{time}\{F_{space}$(vertical motion)$\}$	5*2 = 10
$F_{time}\{F_{space}$(motion)$\}$	5*2 = 10
F_{time}(Temporal Complexity [21] per frame)	5
F_{time}(*Spatial* Complexity [21] per frame)	5

[1] Nb: number;
[2] F_{time} and F_{space} are the temporal (Eq. (4)) and spatial (Eq. (5)) pooling function.

Content Features include 7 types of features: Spatial Information(SI) [22], Temporal Information (TI) [22], Chrominance Information (CI) [23], Contrast Information(CTI) [23], Spatial Perceptual Information (SPI) [23], Colorfulness (CF) [24] and Grey Level Co-occurrence Matrix(GLCM) [25]. As illustrated in Fig. 2, they are extracted directly from the pixel level of the SRC [17]. The temporal pooling function for content features is the same with bitstream features (Eq. (4)), and the spatial pooling function is defind as:

$$F_{space} = \{Mean, Std, Max, Skew, Kurt\}. \quad (6)$$

The co-occurrence matrix (CM) is computed based on image patches. For each small patche, we calculated 6 features as shown in Eq. (7):

$$F_{patch} = \{contrast, dissimilarity, homogeneity, ASM, energy, correlation\}, \quad (7)$$

where $contrast = \sum_{i,j=0}^{l-1} CM_{i,j}(i-j)^2$, the dissimilarity $diss = \sum_{i,j=0}^{l-1} CM_{i,j}\,|i-j|$, the homogeneigy $homo = \sum_{i,j=0}^{l-1} \frac{CM_{i,j}}{1+(i-j)^2}$, Angular Second Moment: $ASM = \sum_{i,j=0}^{l-1} CM_{i,j}^2$, $energy = \sqrt{ASM}$ and $correlation = \sum_{i,j=0}^{l-1} CM_{i,j}\left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{\sigma_i^2\sigma_j^2}}\right]$, in which l is the level of luminance of original image patch (l=255 for 8 bit image), i, j are the horizontal and vertical index of CM respectively; μ, σ are the mean and variance of CM.

2) Features selection: As shown in Fig. 2, all the extracted features are concatenated into one vector. The exhibited vector has dimension of 399. We then used Forward-Sequential Feature Selection (F-SFS) [26]. It is a greedy procedure. More specifically, we initially find the feature that maximizes a cross-validated score when an estimator is trained on this single feature. Once that first feature is selected, we repeat the

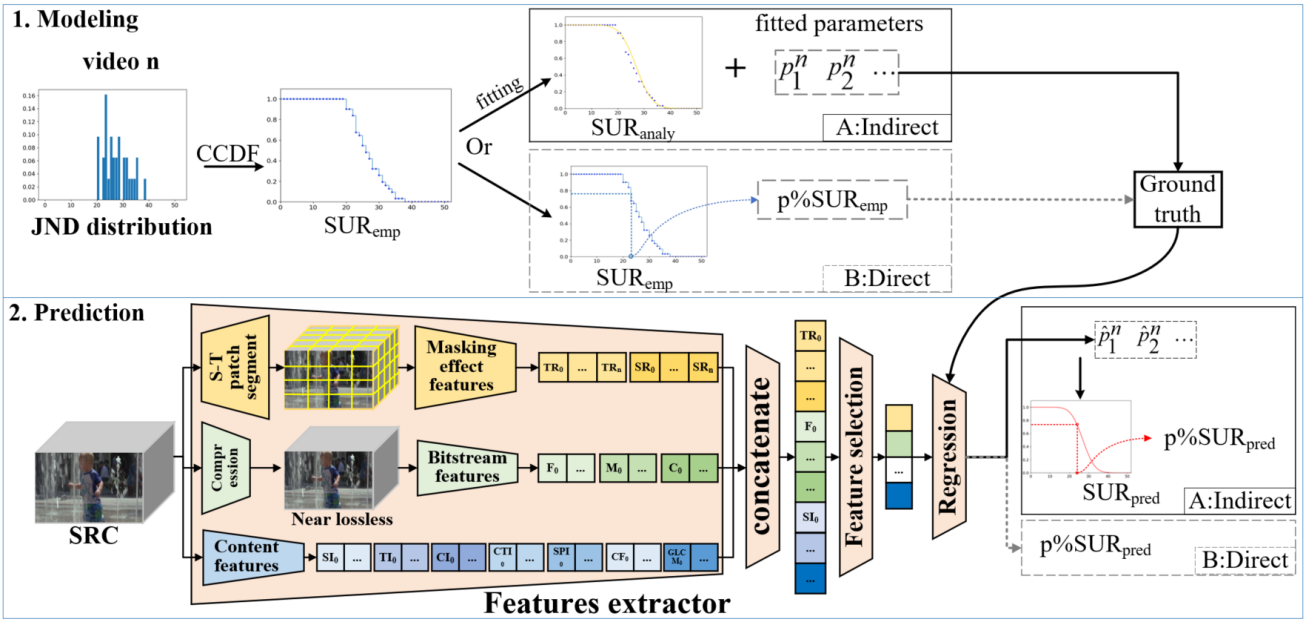

Fig. 2. Illustration of the pipeline of SUR and JND (1) modeling and (2) prediction framework

TABLE IV
CONTENT FEATURES SUMMARY.

Features	dimension
$SI^{+} = F_{time}\{F_{space}\{\text{Sobel}[Y_n(i,j)]\}\}$	5*5 =25
$TI^{+} = F_{time}\{F_{space}[M_n(i,j)]\}$ where $M_n(i,j) = Y_n(i,j) - Y_{n-1}(i,j)$	5*5 =25
$CI_{Cb} = F_{time}\{F_{space}[Cb_n(i,j)]\}$ $CI_{Cr} = F_{time}\{W_R \times F_{space}[Cr_n(i,j)]\}$ where $W_R = 1.5$	5*5+5*5=50
$CTI = F_{time}\{F_{space}[Y_n(i,j)]\}$	5*5 = 25
$SPI_{SI13} = F_{time}\{F_{space}[R_n(i,j)]\}$ where $R_n(i,j) = \sqrt{H_n(i,j)^2 + V_n(i,j)^2}$, $SPI_{HV13} = F_{time}\left\{\frac{mean[HV(i,j)]}{mean[\overline{HV}(i,j)]}\right\}$, $rg = R_n(i,j) - G_n(i,j)$, $yb = \frac{1}{2}(R_n(i,j) + G_n(i,j)) - B_n(i,j)$	5*5+5 = 30
$CF = F_{time}\{CF_n\}$ where $CF_n = \sigma_{rgyb} + 0.3\mu_{rgyb}$ $\sigma_{rgyb} = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2}$, $\mu_{rgyb} = \sqrt{\mu_{rg}^2 + \mu_{yb}^2}$	5
$GLCM = F_{time}\{F_{space}[F_{patch}(CM)]\}$ where CM is the co-occurrence matrix[1]	5*(5*6)=150

[1] https://scikit-image.org/docs/0.7.0/api/skimage.feature.texture.html
[2] F_{time} and F_{space} are the temporal (Eq. (4)) and spatial (Eq. (6)) pooling function; F_{patch} is the functions to compute the texture features of the spatial patch with size $n = 64 * 64$.
[3] Y, Cr and Cb are the luminance and two chroma components; R, G and B are the red, green and blue channels.

procedure by adding the new feature that maximizes the cross-validated score to the set of selected features. The procedure stops when the desired number N of selected features is reached. Grid search was adapted to determine N.

3) Regression: The selected features will be fed into a SVR for prediction. As shown in Fig. 2, there are two ways to predict p%SUR: (A) indirect p%SUR prediction through SUR modeling, (B) direct p%SUR prediction without modeling. For the indirect mode, analytical SUR curve ($\text{SUR}_{\text{analy}}$) and its parameters are determined by fitting the empirical SUR curve (SUR_{emp}). We first obtain the SUR_{pred} curve by predicting the fitted parameters from the features. p%SUR can be computed from the SUR_{pred} curve. If one is only interested in the value of p%SUR (*e.g.,* the demand of a streaming service provider is to satisfy 75% clients), but not the SUR curve, the direct prediction without modeling can be adapted as illustrated in the dotted box in Fig. 2.

III. EXPERIMENTS AND RESULTS

The logistic function showed best prediction performance in [14], hence it is employed in the indirect prediction model. Before feeding the extracted features to the SVR, all the features are normalized by applying z-score transformation. The estimator for F-SFS is the SVR and the metric of features selection is the Mean Square Error (MSE). The optimal number of selected features is 55, detailed in Table V. It can be observed that bitstream features has highest selection rate, which means the bitstream features such as motions are significant for SUR value prediction.

TABLE V
NUMBER OF FEATURES SELECTED PER CATEGORY

Features type	Selected/original	Ratio of selection
Masking effect	9/40	0.2250
Bitstream	15/49	**0.3061**
Content features	31/310	0.1000

Each model is evaluated in HD-VJND datasets with 5-fold cross validation with random split (fixed random state). Hyper parameters of SVR are determined by grid search (kernel='rbf', C=0.1, epsilon=0.0001, gamma='scale'). Difference between **P**redicted and **A**nalytical SUR curve ($\Delta\text{SUR}_{|\text{P}-\text{A}|}$) is the Mean Average Error (MAE) between them. $\Delta\text{SUR}_{|\text{P}-\text{A}|}$ indicates the error between the fitted analytical SUR curve and the predicted one. Difference between **P**redicted and **E**mpirical SUR curve ($\Delta\text{SUR}_{|\text{P}-\text{E}|}$) is evaluated as well. $\Delta 75\%\text{SUR}$ is evaluated in the same way. The results are shown in Table VI. For the model who predict directly the p%SUR

(Direct mode), the $\Delta\mathrm{SUR}_{|\mathrm{P-A}|}$ and $\Delta\mathrm{SUR}_{|\mathrm{P-E}|}$ don't exist. Similarly, we cannot compute $\Delta 75\%\mathrm{SUR}_{|\mathrm{P-A}|}$ because the $\mathrm{SUR}_{\mathrm{analy}}$ doesn't exist without modeling.

Experiment results show that our proposed models (both direct and indirect mode) outperforms (reduced $\Delta\mathrm{SUR}_{|\mathrm{P-E}|}$ by 40%) the baseline model [14]. The direct p%SUR prediction mode without modeling has the smallest prediction error in terms of $\Delta 75\%\mathrm{SUR}$. However, the indirect model provides us more information (the SUR curve and the 75%SUR value) compared to the direct model that outputs only the 75%SUR value. Furthermore, it could be observed in Table VI that the indirect model has smaller standard deviation than direct p%SUR prediction model which indicates that the indirect model helps stabilize the variation of the prediction error.

TABLE VI
BENCHMARK OF AVERAGE AND VARIANCE OF PREDICTION ERROR IN HD-VJND DATASETS

Model		ΔSUR		$\Delta 75\%$SUR	
		$\lvert\mathrm{P-A}\rvert$	$\lvert\mathrm{P-E}\rvert$	$\lvert\mathrm{P-A}\rvert$	$\lvert\mathrm{P-E}\rvert$
Baseline [14]	mean	0.1121	0.1146	1.3251	1.2559
Indirect		**0.0916**	**0.0789**	**0.8285**	0.8575
Direct					**0.7489**
Baseline [14]	Var.	0.0513	0.0671	1.1921	1.1635
Indirect		**0.0298**	**0.0406**	**0.7689**	**0.8382**
Direct					0.9222

IV. CONCLUSION

In this paper, we proposed a pre-processing method for JND subjective test by using the mapping function between VMAF and JND. This method helps us to determine the dynamic range of encoding parameters which helps to reduce 7.14% of subjective test time. Furthermore, we proposed a SUR/JND prediction framework including feature extraction/selection and regression through 3 types of features. Experiment results show that our proposed framework outperforms the SOTA both in SUR curve and 75%SUR value prediction.

ACKNOWLEDGMENT

This work was supported in part by Amazon Prime Video. We thank Dr. Sriram Sethuraman and Dr. Kumar Rahul for helpful discussions and insights into this topic.

REFERENCES

[1] 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," *Electronic Imaging*, vol. 2016, no. 13, pp. 1–9, 2016.

[2] C. Fan, Y. Zhang, H. Zhang, R. Hamzaoui, and Q. Jiang, "Picture-level just noticeable difference for symmetrically and asymmetrically compressed stereoscopic images: Subjective quality assessment study and datasets," *Journal of Visual Communication and Image Representation*, vol. 62, pp. 140–151, 2019.

[3] X. Liu, Z. Chen, X. Wang, J. Jiang, and S. Kowng, "Jnd-pano: Database for just noticeable difference of jpeg compressed panoramic images," in *Pacific Rim Conference on Multimedia*. Springer, 2018, pp. 458–468.

[4] H. Wang, W. Gan, S. Hu, J. Y. Lin, L. Jin, L. Song, P. Wang, I. Katsavounidis, A. Aaron, and C.-C. J. Kuo, "Mcl-jcv: a jnd-based h. 264/avc video quality assessment dataset," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 1509–1513.

[5] H. Wang, I. Katsavounidis, J. Zhou, J. Park, S. Lei, X. Zhou, M.-O. Pun, X. Jin, R. Wang, X. Wang *et al.*, "Videoset: A large-scale compressed video quality dataset based on jnd measurement," *Journal of Visual Communication and Image Representation*, vol. 46, pp. 292–302, 2017.

[6] C. Fan, H. Lin, V. Hosu, Y. Zhang, Q. Jiang, R. Hamzaoui, and D. Saupe, "Sur-net: Predicting the satisfied user ratio curve for image compression with deep learning," in *2019 eleventh international conference on quality of multimedia experience (QoMEX)*. IEEE, 2019, pp. 1–6.

[7] H. Lin, V. Hosu, C. Fan, Y. Zhang, Y. Mu, R. Hamzaoui, and D. Saupe, "Sur-featnet: Predicting the satisfied user ratio curve for image compression with deep feature learning," *Quality and User Experience*, vol. 5, no. 1, pp. 1–23, 2020.

[8] 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, 2020.

[9] H. Liu, Y. Zhang, H. Zhang, C. Fan, S. Kwong, C.-C. J. Kuo, and X. Fan, "Deep learning-based picture-wise just noticeable distortion prediction model for image compression," *IEEE Transactions on Image Processing*, vol. 29, pp. 641–656, 2019.

[10] H. Wang, I. Katsavounidis, Q. Huang, X. Zhou, and C.-C. J. Kuo, "Prediction of satisfied user ratio for compressed video," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6747–6751.

[11] H. Wang, X. Zhang, C. Yang, and C.-C. J. Kuo, "Analysis and prediction of jnd-based video quality model," in *2018 Picture Coding Symposium (PCS)*. IEEE, 2018, pp. 278–282.

[12] X. Zhang, C. Yang, H. Wang, W. Xu, and C.-C. J. Kuo, "Satisfied-user-ratio modeling for compressed video," *IEEE Transactions on Image Processing*, vol. 29, pp. 3777–3789, 2020.

[13] Y. Zhang, H. Liu, Y. Yang, X. Fan, S. Kwong, and C. J. Kuo, "Deep learning based just noticeable difference and perceptual quality prediction models for compressed video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1197–1212, 2021.

[14] J. Zhu, P. L. Callet, A.-F. Perrin, S. Sethuraman, and K. Rahul, "On the benefit of parameter-driven approaches for the modeling and the prediction of satisfied user ratio for compressed video," *arXiv preprint arXiv:2206.09854*, 2022.

[15] S. Hu, L. Jin, H. Wang, Y. Zhang, S. Kwong, and C.-C. J. Kuo, "Compressed image quality metric based on perceptually weighted distortion," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5594–5608, 2015.

[16] S. Hu, L. Jin, H. Wang, Y. Zhang, S. Kwong, and C. J. Kuo, "Objective video quality assessment based on perceptually weighted mean squared error," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 9, pp. 1844–1855, 2016.

[17] S. Ling, Y. Baveye, P. Le Callet, J. Skinner, and I. Katsavounidis, "Towards perceptually-optimized compression of user generated content (ugc): Prediction of ugc rate-distortion category," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[18] I. BT, "General viewing conditions for subjective assessment of quality of sdtv and hdtv television pictures on flat panel displays," *International Telecommunication Union*, 2012.

[19] J. Zhu, S. Ling, Y. Baveye, and P. L. Callet, "A framework to map vmaf with the probability of just noticeable difference between video encoding recipes," *arXiv preprint arXiv:2205.07565*, 2022.

[20] A. Raake, M.-N. Garcia, W. Robitza, P. List, S. Göring, and B. Feiten, "Scalable video quality model for itu-t p. 1203 (aka p. nats) for bitstream-based monitoring of http adaptive streaming," in *Proc. QoMEX*, 2017.

[21] R. R. Ramachandra Rao, S. Göring, W. Robitza, A. Raake, B. Feiten, P. List, and U. Wüstenhagen, "Bitstream-based model standard for 4k/uhd: Itu-t p.1204.3 – model details, evaluation, analysis and open source implementation," in *QoMEX*, Athlone, Ireland, May 2020.

[22] P. ITU-T RECOMMENDATION, "Subjective video quality assessment methods for multimedia applications," 1999.

[23] S. Wolf and M. Pinson, "Video quality measurement techniques," *2002.*, 2002.

[24] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *Human vision and electronic imaging VIII*, vol. 5007. SPIE, 2003, pp. 87–95.

[25] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[26] F. J. Ferri, P. Pudil, M. Hatef, and J. Kittler, "Comparative study of techniques for large-scale feature selection," in *Machine Intelligence and Pattern Recognition*. Elsevier, 1994, vol. 16, pp. 403–413.