Study on Subjective Quality Assessment of Screen Content Images

Sheng Shi, Xiang Zhang, Shiqi Wang, Ruiqin Xiong and Siwei Ma
Institute of Digital Media, Peking University, Beijing 100871, China
Cooperative Medianet Innovation Center, Shanghai, China
Email: {shshi, x_zhang, sqwang, rqxiong, swma}@pku.edu.cn

***Abstract*—With the coming age of big data, the cloud technology, referred to as the computations or applications through the Internet, is dramatically developed. The screen content has become one of the most common data form due to the extraordinary advance of network communication technology, and the JCT-VC has started to develop new standard focusing on improving the efficiency of screen content based on High Efficiency Video Coding (HEVC). Nevertheless, the research on the quality assessment of screen content is still quite limited at the current stage. In this paper, we present a study on subjective quality assessment of the Screen Content Images (SCIs) and investigate whether the existing objective Image Quality Assessment (IQA) methods can effectively evaluate the quality of distorted SCIs. We construct a new Screen Content Database (SCD) including 24 source SCIs and 492 compressed ones with two codecs including HEVC as well as the HEVC extension. The Single Comparison (SC) method is employed for the subjective viewing to guarantee the reliability of the results. In our experiment, the correlations of eight popular IQA methods with the obtained Mean Opinion Score (MOS) values are evaluated. The result indicates that visual information fidelity method can achieve highest consistency with human visual perception.**

I. Introduction

Advances on communication and transmission techniques over Internet have been witnessed under the background of the big data. The amount of screen content is dramatically increasing and various applications are rapidly developed such as remote education, virtual screen, remote gaming, etc. Therefore, it is crucial to study the new strategy of representation, compression and transmission specifically for screen content. The newest video coding standard, High Efficiency Video Coding (HEVC) [1], has achieved more than 50% coding gain at the same subjective image quality compared to H.264/AVC. However, the coding performance on screen contents can be further improved by adopting some advanced techniques [2], and the Screen Content Compression (SCC) [3] is being developed as an extension of HEVC.

Different from the natural images which are captured by camera devices, the screen images have some distinct characteristics as follows:

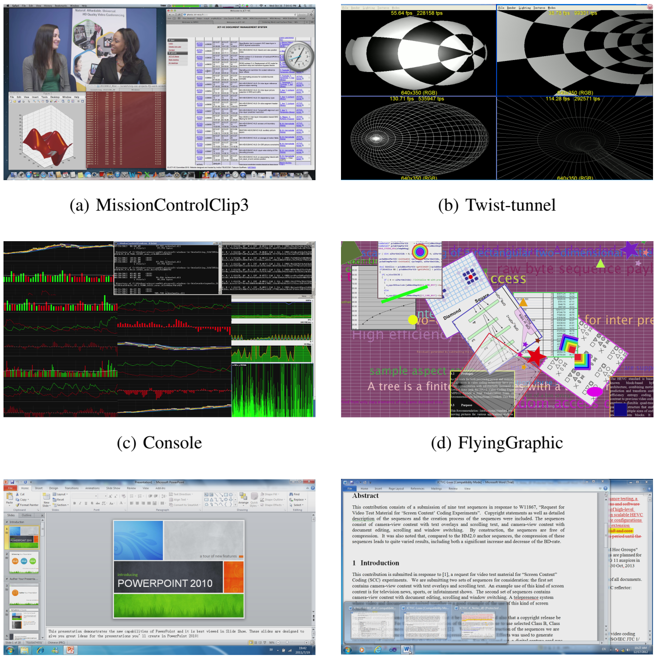
- Screen images are directly generated by computers without the image acquisition process. Therefore, the computer generated regions are totally noise-free.
- Screen images always contain multiple types of content, such as texts, graphics, buttons, icons and natural images.
- Typical screen contents, such as web pages, PowerPoint and text pages, are with higher contrast, sharper edges and larger area of smooth regions.

As various distortions such as noise, blurring, contrast change and compression artifacts may be involved during compression, transmitting and processing, it is important to accurately predict the visual quality of Screen Content Images (SCIs). Some popular models have been proposed for natural image quality assessment. A simple and widely used fidelity measure is the Peak Signal-to-Noise Ratio (PSNR), or the mean squared error (MSE). It is attractive due to its simplicity and mathematical convenience, but not always consistency with the quality perceived by Human Visual System (HVS). Thus modern IQA models always take advantages of the HVS features for obtaining more accurate metrics. Structure similarity (SSIM) [4] and its variant Multi-scale SSIM [5], information weighted SSIM [6] schemes have shown good correlation with visual perception. Some SSIM related works [7], [8] have been proposed in improving the performance of video coding. Feature similarity (FSIM) [9] is developed based on the fact that HVS understands an image mainly according to its low-level features. Specifically, phase congruency (PC) and image gradient magnitude (GM) are extracted as the two features in FSIM. Although many objective IQA methods have been proposed to evaluate quality of natural images [10], whether these objective IQA methods can be directly applied to SCIs is still not fully investigated, especially in the context of HEVC compression. Therefore, it is highly desired to investigate subjective metrics for the quality evaluation of SCIs.

In this paper, we conduct an in-depth study on subjective quality assessment of SCIs. A new Screen Content Database (SCD) is established, which includes 24 source SCIs and their 492 compressed versions. The Single Comparison (SC) method [11] is adopted to obtain the subjective values. Several state-of-the-art objective IQA models are compared with the subjective scores on this database. The SCD including both images and scores is fully available in the website [12].

The remaining of this paper is organized as follows. In Section II, the subjective assessment methodology of screen content images is described. In Section III, we process and analyze the data for the database. In Section IV, some objective quality metrics are introduced and evaluated on the built database, and the correlation between objective and subjective

978-1-4799-7783-3/15/$31.00 ©2015 IEEE



(a) MissionControlClip3

(b) Twist-tunnel

(c) Console

(d) FlyingGraphic

(e) SlideShow

(f) WordEditing

Fig. 1. Samples of the reference SCIs for testing.

scores are presented. Finally, Section V concludes this paper.

II. Subjective Assessment Methodology of Screen Content Images

To investigate quality evaluation of SCIs, a new database (i.e. SCD) is constructed, which includes 24 source SCIs and their 492 compressed versions. Subjective quality study of these SCIs is conducted to obtain the subjective quality scores.

A. Screen Content Database

The reference SCIs are selected with various combination styles of texts, graphics and pictures. Twenty-four SCIs are collected from web pages, slices, PDF files and digital magazines through screen snapshot. Fig. 1 illustrates some samples of the reference SCIs. In order to make database more comprehensive, we selected SCIs with three different resolutions: 2560×1440, 1920×1080 and 1280×720.

We apply the following two distortion types to the reference images. These distortion types are commonly generated in compression of SCIs:

- *High Efficiency Video Coding (HEVC)*: HEVC adopts 4:2:0 color format for compression. Adaptive quad-tree coding structure based on the concept of coding tree unit (CTU) is employed, and three new concepts named coding unit (CU), prediction unit (PU) and transform unit (TU) are developed to specify the basic processing unit of coding, prediction and transform.
- *Screen Content Compression (SCC)*: SCC is an extension of HEVC which supports 4:4:4 format compression. It can improve the coding performance on screen contents by employing advanced techniques, such as intra block copy, transform skipping and base color representation.

For the two distortion types, we vary the distortion level to generate images from noise imperceptible level to high annoying level. Specifically, considering the reference images are simple, 264 distorted images are generated by using SCC intra coding to compress the reference images with quality factors ranging from 30 to 50 at 11 levels. For HEVC, we downsample the UV components of the reference images from 4:4:4 format to the color distorted images with 4:2:0 format. Similarity, we use HEVC intra coding to compress the color distorted images with quality factors ranging from 30 to 50 at 11 levels. Finally, 528 distorted images are derived from the 24 reference SCIs.

B. Subjective Testing Methodology

Several subjective testing methodologies for assessing image quality have been defined by ITU-R BT500-11 [11], including Single-Stimulus (SS), Double-Stimulus Impairment Scale (DSIS) and Paired Comparison (PC). These methods can be roughly categorized into two types: the single stimulus and double stimulus approaches. The former asks the subjective viewers to rate the quality of just one distorted image while the later asks the subjective viewers rate the quality or change in quality between two images (reference and distorted). Each subjective test methodology has its own advantages as discussed in [13]. The single stimulus approach can ensure a faster and more efficient subjective testing process, compared with the double stimulus one. Thus we employ the single stimulus in this paper.

1) Display Configuration: The desktop used to perform the experiments has 16 GB RAM and 64-bit Windows operation system which is placed in a laboratory room with normal indoor light. In the experiment, All the viewers are asked to sit at a viewing distance of about 2-2.5 screen heights.

2) Single Stimulus: In order to avoid contextual and memory effects in viewers quality judgements, the playlist is generated by a random permutation of 528 distorted images, and the every consecutive images are not generated from the same reference image. The different resolutions of the images and the introduced shape distortions may lead subjective viewers to have difficulties in judging the perceptual quality of the image and providing a precise subjective value. Therefore, in order to get more precise subjective values, we employ the 10-category discrete scale to obtain the subjective opinions to build the image subjective quality database.

In the subjective testing, we use MATLAB to develop the user interface as shown in Fig. 2. The distorted images are loaded into the memory before displaying. In order to help human subjects to perform the quality evaluate, the quality scales are labeled in the bottom of the interface. The quality scales ranging from the lowest to the highest perceptual quality index are labeled as “Bad”, “Poor”, “Fair”, “Good” and “Excellent”. During the subjective testing, the subjective values are recorded in numerical values, where “Bad” corresponds to 1 and 2 and the “Excellent” corresponds to 9 and 10. The

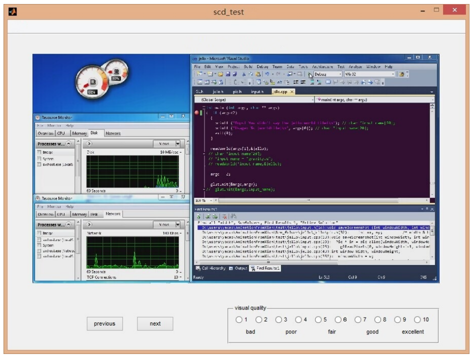


Fig. 2. Screenshot of the subjective study interface.

human subjects select the appropriate quality index according to their own opinions. The subject was allowed to take 10 seconds to evaluate the quality of every image.

Before starting the testing, the objective of this subjective test and instructions are briefly introduced to each human subject. Subsequently, a training session will be presented to the human subject. After the training process, each subject has the idea on how to provide their opinions on the distorted image quality. In the subjective testing, all the subjects with or without corrective glasses have normal vision and have passed the color blindness test. We choose 20 subjects to provide their personal ratings on the perceptual quality of each image.

III. Data Processing and Analysis

A. Subjective Agreement

Examining the similarity of choices between participants is necessary before we process the subjective ratings to build the database. For a large proportion of the images in the database, most of the participants should have similar agreements on the perceptual quality. The image is not suitable for including into the database if the subjective results demonstrate diversely among the human subjects.

In this paper, we employ the method mentioned in [14] to analyze the subject agreement. We can get the 25th and 75th percentiles of subjective ratings after sorting the subjective scores for each image. Then the central 50% of subject ratings will lie in these range. The outlier coefficient (OC) which is used to quantify the subjective agreement of the database is represented as:

$$OC = \frac{Num_{outlier}}{Num_{total}} \tag{1}$$

where Num_{total} denotes the total number of the distorted images in the database, and $Num_{outlier}$ denotes the number of the images, which are regarded as the outlier. The image is recognized as the outlier image if the interval between the 25^{th} and 75^{th} percentiles of subjective ratings is larger than 2. The reason is that viewers may have different opinions on the image quality, but they should at least have the similar judgment. For one image, different viewers may interpret the same image as "Bad" or "Poor", which are neighboring values but not interpret with greatly differences, such as "Bad" or "Good". Therefore, we believe that the participants have an agreement of the screen image quality if the central 50% subjective ratings are constrained within the interval of 2. For the built database, 36 out of 528 are recognized as the outlier images, which implies OC=6.82%. That is to say, the participants have similar agreements on the 93.18% of the images in the database. Consequently, these 492 images can be used to build the database and further employed for evaluation of the quality metrics.

Furthermore, we check the subject agreement between every two subjective ratings. The normalized cross correlation (NCC) and the Euclidean distance (EUD) between every two vectors x and y is represented as follows,

$$NNC = \frac{x^t \cdot y}{\|x\| \, \|y\|} \tag{2}$$

$$EUD = \frac{\|x - y\|_2^2}{k} \tag{3}$$

where k denotes the dimension of the subjective rating vector, and $\|\cdot\|_2^2$ defines the norm of the vector. Suppose that the subject rating values given by each viewer compose a vector. As there are 20 subjective ratings, $C\binom{20}{2} = 190$ NCC and EUD values are obtained. The average NCC value is 0.9851 and the EUD value is 0.0407. The NCC value is close to 1 when the angle difference between every two subjective rating vectors is very small. Similarly, when the magnitude difference between every two subjective rating vectors is relatively small, the EUD value is close to 0. Therefore, these two NCC and EUD values demonstrate that the subjects have achieved a great agreement on the perceptual qualities of the distorted images.

B. Screening of the Observers

The subject agreement on the distorted image quality has been examined in the previous section. However, in order to obtain the final MOS and standard deviation value for each image, we employ the subject rejection process suggested in [11]. Let S_{ij} denotes the subjective rating by the subject i to the image j. The S_{ij} values are firstly converted to Z-scores persession [15]:

$$\mu_{\mathrm{i}} = \frac{1}{N_i} \sum_{j=1}^{N_i} S_{ij} \tag{4}$$

$$\sigma_{\mathrm{i}} = \sqrt{\frac{1}{N_i - 1} \sum_{j=1}^{N_i} (S_{ij} - \mu_i)^2} \tag{5}$$

$$Z_{ij} = \frac{S_{ij} - \mu_i}{\sigma_{\mathrm{i}}} \tag{6}$$

where N_i is the number of the distorted images seen by the subject i. So Z-scores which accounts for the differences in subject preferences for reference images, and the different human subjects are obtained.

After converting the subjective ratings into Z-scores, we discard scores from unreliable subjects by using the subject rejection procedure specified in the ITU-R BT 500-11 [11].

Firstly, we compute the kurtosis β_i of the scores to determine whether the scores assigned by a subject are normally distributed. The kurtosis β_i of the scores is presented as:

$$\beta_i = \frac{m_4}{(m_2)^2} \qquad with \qquad m_\Delta = \frac{\sum_{j=1}^{N_i}(S_{ij}-u_i)^\Delta}{N_i} \tag{7}$$

The scores are regarded to be normally distributed when the kurtosis value β_i falls between 2 and 4. The subject rejection procedure is depicted in Alg. 1. By performing the procedure, no one of the 20 subjects should be rejected.

Algorithm 1: Subject rejection process.

```
foreach subject $i$ do
    $P_i = 0$;
    $Q_i = 0$;
end
if $2 \leq \beta_j \leq 4$ (normally distributed) then
    if $S_{ij} \geq u_i + 2\sigma_i$ then
        $P_i = P_i + 1$;
    end
    if $S_{ij} \leq u_i - 2\sigma_i$ then
        $Q_i = Q_i + 1$;
    end
end
else
    if $S_{ij} \geq u_i + \sqrt{20}\sigma_i$ then
        $P_i = P_i + 1$;
    end
    if $S_{ij} \leq u_i - \sqrt{20}\sigma_i$ then
        $Q_i = Q_i + 1$;
    end
end
if $(P_\mathrm{i}+\mathrm{Q_i})/N_\mathrm{i} > 0.05$ & $|(P_\mathrm{i}-\mathrm{Q_i})/(P_\mathrm{i}+\mathrm{Q_i})| < 0.3$ then
    reject the subject $i$.
end
```

Finally, the MOS value is computed using the mean of the scores:

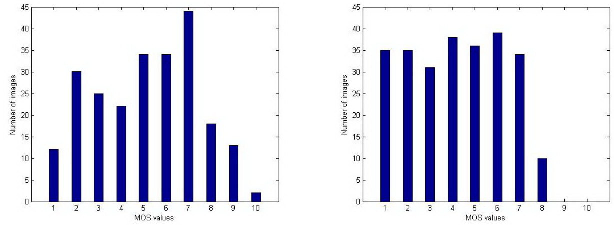
$$MOS_j = \frac{1}{M_k}\sum_{i=1}^{M_k} S_{ij} \tag{8}$$

where M_k is the number of subjects after the subject rejection. In order to have a clear observation, we round the MOS values to integers. The histogram of the processed MOS values is shown in Fig. 3. Observing the perceptual qualities of the images range from low to high values, we find the subjective quality scores for SCC is better than that for HEVC. That is to say, the distorted images compressed by SCC have better performance than that by HEVC at the same distortion level.

IV. Experimental Results

A. Traditional IQAs for Natural Images

We apply the following 8 objective IQA metrics on the SCD: PSNR, SSIM [4], MS-SSIM [5], IW-SSIM [6], FSIM [9], VIF [16], GSM [17] and VSI [18] to investigate efficiency of the existing IQA methods to evaluate quality of SCIs. Most of these IQA methods are implemented using the codes available [19]. Some metrics are briefly introduced as follows:



(a)

(b)

Fig. 3. Histogram of the MOS values for (a) SCC and (b) HEVC.

- **PSNR**: Peak signal-to-noise ratio (PSNR) or MSE is most commonly used to measure the quality of signals including images/videos.
- **SSIM**, **MS-SSIM** and **IW-SSIM**: The structural similarity (SSIM) index is a method for measuring the similarity between two images. It is designed to improve on traditional methods like PSNR and MSE, which have proved to be inconsistent with human eye perception. SSIM considers image degradation as perceived change in structural information. Structural information is the idea that the pixels have strong inter-dependencies especially when they are spatially close. Multi-SSIM (MS-SSIM) index is an extension of SSIM. It performs much better than its single-scale counterpart. Information Content Weighted SSIM (IW-SSIM) index is an extension of MS-SSIM by using spatially varied weights.
- **FSIM**: The well-known SSIM index brings IQA from pixel based stage to structure based stage. Feature-similarity (FSIM) index is based on the fact that human visual system (HVS) understands an image mainly according to its low-level features. Specifically, the phase congruency (PC) and the image gradient magnitude (GM) play important roles in characterizing the image local quality.
- **VIF**: IQA algorithms generally interpret image quality as fidelity or similarity with a “reference” or “perfect” image in some perceptual space. Visual Information Fidelity (VIF) index is presented by Sheikh and Bovik in [16]. The paper approaches the IQA problem as an information fidelity problem. Specifically, it proposes to quantify the loss of image information to the distortion process and explore the relationship between image information and visual quality.
- **VSI**: Visual saliency (VS) has been widely studied by psychologists, neurobiologists, and computer scientists during the last decade to investigate which areas of an image will attract the most attention of the human visual system. Intuitively, VS is closely related to IQA in that suprathreshold distortions can largely affect VS maps of images. The Visual Saliency Induced Index for Perceptual Image Quality Assessment is represented in [18].

TABLE I
PERFORMANCE OF EIGHT IQA MODELS IN TERMS OF PLCC, MAE, RMS, SRCC, KRCC METRICS FOR THE SCC COMPRESSED IMAGES.

	PLCC	MAE	RMS	SRCC	KRCC
PSNR	0.8666	0.8862	1.1298	0.871	0.6933
SSIM	0.868	0.851	1.1241	0.8696	0.708
MS-SSIM	0.8784	0.8287	1.0819	0.8864	0.7247
FSIM	0.8899	0.807	1.0328	0.8968	0.7288
IW-SSIM	0.8853	0.8177	1.0529	0.8951	0.7322
VIF	**0.9152**	**0.6979**	**0.9123**	**0.9229**	**0.7743**
VSI	0.8952	0.7648	1.0089	0.903	0.7477
GSM	0.8795	0.8329	1.0772	0.8842	0.7229

TABLE II
PERFORMANCE OF EIGHT IQA MODELS IN TERMS OF PLCC, MAE, RMS, SRCC, KRCC METRICS FOR THE HEVC COMPRESSED IMAGES.

	PLCC	MAE	RMS	SRCC	KRCC
PSNR	0.8529	0.8651	1.0823	0.8549	0.6617
SSIM	0.8588	0.8426	1.062	0.8605	0.6743
MS-SSIM	0.8901	0.748	0.945	0.8929	0.7118
FSIM	0.8884	0.7241	0.9516	0.8888	0.7143
IW-SSIM	0.8987	0.7197	0.9091	0.8992	0.7195
VIF	**0.9268**	**0.5864**	**0.7786**	**0.9265**	**0.7662**
VSI	0.8564	0.8451	1.0704	0.8541	0.6628
GSM	0.7276	1.1447	1.4222	0.7217	0.5385

B. Correlation between Objective and Subjective Scores

We evaluate the correlation between the predicted scores delivered by the 8 IQA metrics and the MOS values in terms of Pearson Linear Correlation Coefficient (PLCC), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Spearman rank-order correlation coefficient (SROCC) and Kendall rank order correlation coefficient (KROCC). Strong correlation means that the IQA methods are highly consistent with human visual perception on assessing the quality of SCIs. We report the results in Tables I and Table II, respectively.

From the Table I and Table II, we find that correlations between the VIF and MOS values of the distorted images generated by SCC intra coding and HEVC intra coding is fairly strong, which indicates that VIF can well measure the artifacts caused by the traditional compression methods. This might be attributed to the reason that HVS is more careful about the information content when viewing an image, which makes the information content based scheme more efficient.

V. CONCLUSION

In this paper, the new image database SCD, which includes 24 source SCIs and 492 compressed ones undergone two types of codec including traditional HEVC as well as SCC, has been constructed to investigate the subjective quality of SCIs. The Single Comparison method is employed for the subjective viewing to guarantee the reliability of the results. According to the correlation analysis of the 8 IQA scores and the MOS values, we find the VIF method can achieve high consistency with human visual perception when judging the SCIs. The database has been published and can be downloaded from [12].

ACKNOWLEDGE

This work was supported in part by the Major State Basic Research Development Program of China (2015CB351800) and in part by the National Science Foundation (61322106 and 61210005).

REFERENCES

[1] G. J. Sullivan, J. Ohm, W.-J. Han, and T. 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.

[2] S. Wang, J. Fu, Y. Lu, S. Li, and W. Gao, "Content-aware layered compound video compression," in *Circuits and Systems (ISCAS), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 145–148.

[3] W. Zhu, W. Ding, J. Xu, Y. Shi, and B. Yin, "Screen content coding based on HEVC framework," *Multimedia, IEEE Transactions on*, vol. 16, no. 5, pp. 1316–1326, Aug 2014.

[4] Z. 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, 2004.

[5] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2. IEEE, 2003, pp. 1398–1402.

[6] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *Image Processing, IEEE Transactions on*, vol. 20, no. 5, pp. 1185–1198, 2011.

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

[8] ——, "Perceptual video coding based on SSIM-inspired divisive normalization," *Image Processing, IEEE Transactions on*, vol. 22, no. 4, pp. 1418–1429, April 2013.

[9] L. Zhang, D. Zhang, and X. Mou, "FSIM: a feature similarity index for image quality assessment," *Image Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2378–2386, 2011.

[10] D. M. Chandler, "Seven challenges in image quality assessment: past, present, and future research," *ISRN Signal Processing*, vol. 2013, 2013.

[11] I. Recommendation, "500-11 methodology for the subjective assessment of the quality of television pictures," *International Telecommunication Union, Geneva, Switzerland*, vol. 4, p. 2, 2002.

[12] S. Shi and etc., "Screen Content Database." [Online]. Available: https://drive.google.com/file/d/0B0CvNYJSEmzOOG9nN0lRX3VNdFk/view?usp=sharing

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

[14] L. Ma, W. Lin, C. Deng, and K. N. Ngan, "Image retargeting quality assessment: A study of subjective scores and objective metrics," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 6, no. 6, pp. 626–639, 2012.

[15] A. M. van Dijk, J.-B. Martens, and A. B. Watson, "Quality asessment of coded images using numerical category scaling," in *Advanced Networks and Services*. International Society for Optics and Photonics, 1995, pp. 90–101.

[16] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *Image Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 430–444, 2006.

[17] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *Image Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 1500–1512, 2012.

[18] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency induced index for perceptual image quality assessment," 2014.

[19] L. Zhang, "Reseach on image quality assessment." [Online]. Available: http://sse.tongji.edu.cn/linzhang/IQA/IQA.htm