PERCEPTUAL SCREEN CONTENT IMAGE QUALITY ASSESSMENT AND COMPRESSION

Shiqi Wang[1], *Ke Gu*[1,2], *Kai Zeng*[1], *Zhou Wang*[1], *Weisi Lin*[3],

[1]Dept. of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada
[2]Institute of Image Comm. Info. Processing, Shanghai Jiao Tong University, Shanghai, China
[3]School of Computer Engineering, Nanyang Technological University, Singapore

ABSTRACT

Compression of screen content has recently emerged as an active research topic due to the increasing demand in many applications such as wireless display and virtual desktop infrastructure. Screen content images (SCIs) exhibit different statistical properties in textual and pictorial regions, and the human visual system (HVS) also behaves differently when viewing the textual and pictorial regions in terms of the extent of visual field. Here we propose a perceptual SCI quality assessment approach that incorporates visual field adaptation and information content weighting. Furthermore, we propose a perceptual coding scheme in an attempt to optimize the HEVC Screen Content Coding encoder. Experimental results show that the proposed quality assessment method not only better predicts the perceptual quality of SCIs, but also leads to an effective way to optimize screen content coding schemes.

Index Terms— Screen Quality Assessment, Screen Content Compression, Information Content, SSIM Index

1. INTRODUCTION

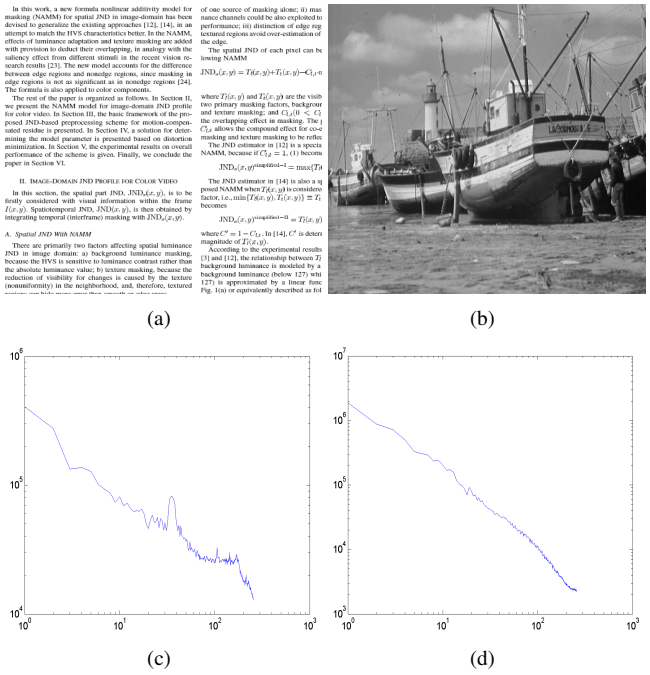
With the rapid development of Internet technology and cloud computing, there has been an increasing desire to enable clients with mobile devices to enjoy and utilize the computationally intensive and graphically rich services by transmitting the remote screen to the clients. In these scenarios, the time variant interface can be regarded as a screen content image (SCI), which is a mixture of pictorial regions and computer generated textual content [1]. The quality of the SCI directly determines the interactivity performance and the user experience of a remote system. Therefore, developing accurate SCI quality measures is an urgent need, as it can further serve as a benchmark for monitoring, adjusting and optimizing the quality of remote computing systems.

Recently, much work has been done to develop objective quality assessment measures which can automatically predict perceived image quality. Popular methods include the Structural Similarity (SSIM) index [2], visual information fidelity (VIF) [3], information content weighted SSIM (IW-SSIM) [4] and feature-similarity (FSIM) [5], etc. However, most of them are designed and validated on natural images, which do not always share the same properties of screen content. Typically, the discontinuous-tone computer generated screen image is featured by sharp edges and thin lines with few colors [6], while natural images usually have continuous-tone, smoother edges, thicker lines and more colors. In view of the importance of SCI quality assessment, in [7] a database of distorted SCIs with subjective quality ranking was created, which includes distortion types such as Gaussian noise, Gaussian blur, motion blur, contrast changing, JPEG, JPEG2000 and layer segmentation based coding. The correlation between the scores of subjective and objective measures demonstrate that there is still large room to improve for SCI quality assessment [7].

As widely hypothesized in computational vision science, the major task of the human visual system (HVS) when viewing an image is to act as an optimal information extractor, or an efficient coder [8]. This motivated us to evaluate the quality of screen image with local information content. Another psychology finding regarding the perception of screen images is that the extent of the visual field used to extract useful information is much larger in pictorial portions than in textual content [9]. These observations inspired us to predict the screen image quality with adaptive window size and local information weighting. Furthermore, to demonstrate the application of this method, we incorporate the proposed quality measure into an HEVC screen content codec to improve its coding efficiency.

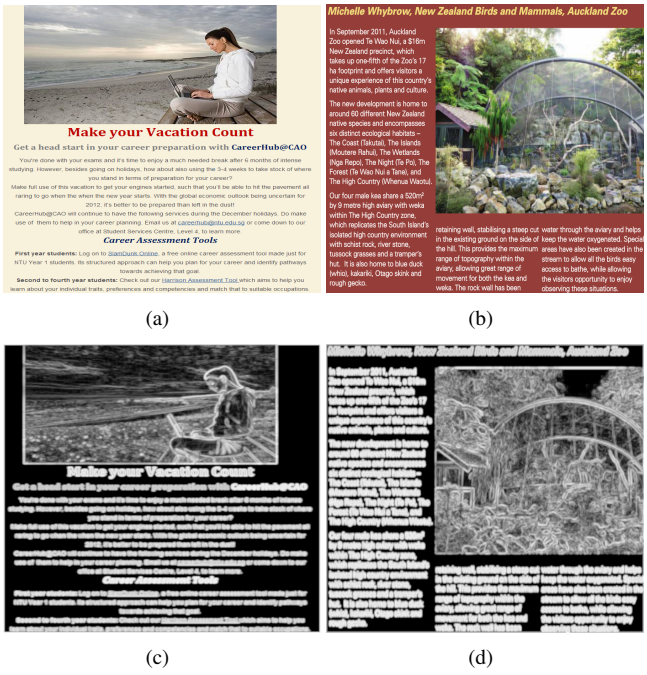
2. SCREEN CONTENT IMAGE QUALITY ASSESSMENT

It has been discovered that the amplitude spectrum of natural images falls with the spatial frequency approximately proportional to the $1/f_s^p$ law [10], where f_s is the spatial frequency and p is an image dependent constant. To examine this, we decompose typical natural and textual images using Fourier transform and then compute the frequency energy, as demonstrated in Fig. 1. It is observed that the falloffs for natural images are approximately straight lines in log-log scale, which is consistent with the $1/f_s^p$ relationship. However, for textual images there is a peak at high frequency, which is some-



(a) (b) (c) (d)

Fig. 1. Frequency energy falloffs of textual and natural images in log-log scale. (a) Textual image; (b) Natural image; (c) Frequency energy of textual image; (d) Frequency energy of natural image.



(a) (b) (c) (d)

Fig. 2. Screen content images and the corresponding local information content maps.

what "unnatural". Therefore, the statistical properties of textual images differ from natural images, which motivated us to distinguish them in the design of quality assessment method.

As in [11], we estimate the information received by the HVS by assuming a local Gaussian source and additive Gaussian channel model [3]. Specifically, when perceiving an image, the input signal is locally modeled as a Gaussian source that is transmitted through a Gaussian noise channel to the receiver. The mutual information between the input and the received signal is the amount of the perceived information content. In spatial domain, this can be quantified by

$$\omega = \log_2\left(1 + \frac{\sigma_p^2}{\sigma_n^2}\right), \tag{1}$$

where σ_p is the local variance within a local window $\mathbf{x}$, and σ_n is a constant parameter accounting for the noise level in the visual channel. The local information maps computed using (1), together with the corresponding original images are shown in Fig. 2, which provide a useful indicator about how perceptual information is distributed over space and how the distributions are different in textual and pictorial regions.

In [9], the authors compared the eye movements when people view textual and pictorial content, and it is observed that the perceptual span in reading textual content is clearly smaller than that in either scene perception or visual search. This motivated us to adapt the window size when accessing the local quality of textual and pictorial content.

The local quality of SCIs is predicted based on SSIM [2], which has been demonstrated to be an effective quality measure that achieves a good compromise between quality evaluation accuracy and computation efficiency. Given two local image patches $\mathbf{x}$ and $\mathbf{y}$ extracted from the original and distorted images, respectively, the SSIM index between them is evaluated as

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \frac{(2\mu_{\mathrm{x}}\mu_{\mathrm{y}} + \mathrm{C}_1)(2\sigma_{\mathrm{xy}} + \mathrm{C}_2)}{(\mu_{\mathrm{x}}^2 + \mu_{\mathrm{y}}^2 + \mathrm{C}_1)(\sigma_{\mathrm{x}}^2 + \sigma_{\mathrm{y}}^2 + \mathrm{C}_2)}, \tag{2}$$

where μ_x, σ_x and σ_{xy} are the mean, standard deviation and cross correlation within a local window with size $l \times l$, respectively. C_1 and C_2 are positive constants used to avoid instability when the means and variances are close to zero.

In the literature, there are various ways to classify textural and pictorial content, for example by using gradient-based methods [12] or text detection approaches [13]. In this work, to improve efficiency, instead of dividing the image into large segments of textual and pictorial regions, we propose a block-classification approach by making use of the information content map as shown in Fig. 2. More specifically, since textual regions that contain abundant high contrast edges typically have higher local information content than pictorial regions, we classify each 4×4 block by applying a threshold on the mean of the information content in the block. The overall quality of the textual and pictorial regions Ω_T and Ω_P, denoted by Q_T and Q_P, respectively, are computed by information

content-based local quality-weighted pooling:

$$Q_T = \frac{\sum_{i\in\Omega_T} \mathrm{SSIM_i}\cdot\omega_i^\alpha}{\sum_{i\in\Omega_T}\omega_i^\alpha}$$
$$Q_P = \frac{\sum_{j\in\Omega_P} \mathrm{SSIM_j}\cdot\omega_j^\alpha}{\sum_{j\in\Omega_P}\omega_j^\alpha}, \quad (3)$$

where the parameter α is a constant used to adjust the strength of weighting and is set to 0.3. Since textual content is perceived with smaller extend of visual field than pictorial regions, the local SSIM map is calculated by employing different sizes of Gaussian windows and standard deviations (std) ($l_t = 5, std = 0.5$ for textual, and $l_p = 17, std = 2.5$ for pictorial). The local information ω_i and ω_j are calculated with their respective windows, within which the SSIM indices are computed.

The final quality of the SCI is estimated by a weighted average of Q_T and Q_P,

$$Q_S = \frac{Q_T\cdot E(\omega_T^\alpha) + Q_P\cdot E(\omega_P^\alpha)}{E(\omega_T^\alpha)+E(\omega_P^\alpha)}, \quad (4)$$

where $E(\omega_T^\alpha)$ and $E(\omega_P^\alpha)$ denote the expectation of the local information for the textual and pictorial content portions, respectively. These quantities indicate the relative importance of the textual and pictorial blocks of the screen image, which are computed with an uniform patch size that lies in the median between l_t and l_p ($l = 11, std = 1.5$). The whole quality assessment method is summarized in Algorithm 1, which adopts the adaptive window size and local information weighting to predict the perceptual quality of SCIs.

We verify the quality assessment method by comparing its performance with PSNR, SSIM [2], MS-SSIM [14], IW-SSIM[4], GSIM [9], FSIM [5], GMSD [15], VSI [16] and VIF [3] using the newly proposed SIQAD database [7]. The database is composed of 980 screen images created by corrupting 20 source images with 7 distortion types at 7 distortion levels. As illustrated in Table 1, the performance of the overall database and compression distortions (JPEG, JPEG2000) are reported, where four frequently used performance measures (PLCC, SRCC, KRCC and RMSE), as suggested by the video quality experts group (VQEG) [17], are evaluated. Overall, the proposed quality measure significantly improves the prediction accuracy and monotonicity.

3. PERCEPTUAL SCREEN CONTENT IMAGE COMPRESSION

The proposed screen content coding scheme follows divisive normalization based perceptual video coding approach [18, 19], in which the DCT transform coefficient of a residual block C_k is normalized with a positive normalization factor f to transform the DCT coefficients into a perceptually uniform domain, which is expressed as $C(k)' = C(k)/f$. As

Algorithm 1 Summary of SCI quality assessment method

Input: SCI
Output: Q_S

1: Compute the local information at each pixel location using 11×11 Gaussian window with $std = 1.5$.
2: For each 4x4 block, the mean of information content is used to classify the block type into textual or pictorial.
3: For textual blocks, compute Q_T using 5×5 Gaussian window with $std = 0.5$.
4: For pictorial blocks, compute Q_P using 17×17 Gaussian window with $std = 2.5$.
5: Compute final Q_S using (4).
6: **return** Q_S.

such, the quantization process of the normalized residuals for a given predefined Q_s can be formulated as

$$\begin{aligned} Q(k) &= \mathrm{sign}\{C(k)'\}\mathrm{round}\{\frac{|C(k)'|}{Q_s}+p\} \\ &= \mathrm{sign}\{C(k)\}\mathrm{round}\{\frac{|C(k)|}{Q_s\cdot f}+p\}, \end{aligned} \quad (5)$$

where p denotes the rounding offset in the quantization.

This implies that the quantization parameters for each coding unit can be adaptively adjusted according to the divisive normalization process. Following this process, the rate distortion optimization is performed as

$$D = \sum_{k=0}^{N-1}(C(k)'-R(k)')^2 = \sum_{k=0}^{N-1}\frac{(C(k)-R(k))^2}{f^2}, \quad (6)$$

where N is the block size and $R(k)$ is the reconstruction coefficient. As the divisive normalization is performed to transform the DCT coefficients into perceptually uniform space, the Lagrangian multiplier λ in rate distortion optimization is set as the predefined value in the encoder.

To derive the normalization factor, the local information and energy of DCT domain coefficients are extracted from the original image. In particular, the normalization factor for the $i\text{-}th$ textual block is given by,

$$f = \frac{\sqrt{\frac{\sum_{k=1}^{N-1}(X_i(k)^2+Y_i(k)^2)}{N-1}+C_2}\cdot g\cdot E(\omega_T^\alpha)}{E(\sqrt{\frac{\sum_{k=1}^{N-1}(X(k)^2+Y(k)^2)}{N-1}+C_2})\cdot(E(\omega_T^\alpha)+E(\omega_P^\alpha))}, \quad (7)$$

where X and Y represents the DCT coefficients of the original and distorted blocks, and g denotes the relative importance of the local block in terms of the local information content:

$$g = \frac{\sum_{j=1}^{N}\omega_j^\alpha}{N\cdot E(\omega_T^\alpha)}. \quad (8)$$

We incorporate the proposed SCI quality measure in the encoder optimization process, where the divisive normalization factor for each block is derived by analyzing the

Table 1. Performance evaluation of the IQA models on SIQAD.

Quality Measures	All				JPEG&JPEG2000			
	PLCC	SRCC	KRCC	RMSE	PLCC	SRCC	KRCC	RMSE
PSNR	0.5788	0.5539	0.4144	11.569	0.7271	0.7118	0.5157	7.111
SSIM	0.7445	0.7433	0.5455	9.471	0.7586	0.7532	0.5586	6.693
MS-SSIM	0.6047	0.5968	0.4408	11.299	0.7539	0.7522	0.5545	6.749
IW-SSIM	0.6408	0.6414	0.4834	10.892	0.7627	0.7635	0.5662	6.644
GSIM	0.5515	0.5311	0.3894	11.834	0.6656	0.6667	0.4770	7.667
FSIM	0.5741	0.5647	0.4088	11.616	0.6738	0.6711	0.4803	7.590
GMSD	0.7277	0.7168	0.5342	9.731	0.7793	0.7764	0.5774	6.437
VSI	0.5403	0.5199	0.3712	11.938	0.7306	0.7249	0.5285	7.014
VIF	0.8026	0.7857	0.5879	8.464	0.7786	0.7725	0.5742	6.446
Q_S	**0.8573**	**0.8456**	**0.6550**	**7.303**	**0.7987**	**0.7955**	**0.6035**	**6.181**

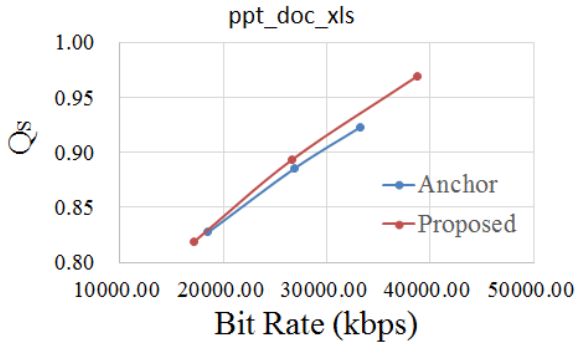
input SCI. The newly developed HEVC extension codec for screen content (HM15.0+RExt-8.0+SCM-2.0rc1) [20] is employed. The test images are in YUV4:4:4 format from both the SIQAD database (news and sports) and HEVC test sequences (the first frame of each sequence is used). The test images include common scenarios in screen image processing, such as web browsing, office software editing and video-conferencing. The R-D performance in terms of the proposed quality measure is demonstrated in Table 2 and Fig. 3. It is observed that significant bit rate saving is achieved, which further demonstrates the effectiveness of the proposed quality measure in potential applications such as encoder optimization.

Table 2. RD performance for different SCIs.

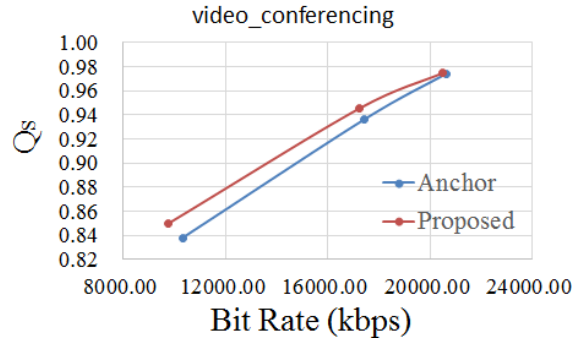
Image	*BD-Rate*
ppt_doc_xls	-5.1%
sc_programming	-4.7%
video_conferencing	-7.2%
sc_web_browsing	-2.0%
wordEditing	-5.6%
twist_tunnel	-0.7%
news	-1.8%
sports	-4.9%

4. CONCLUSIONS

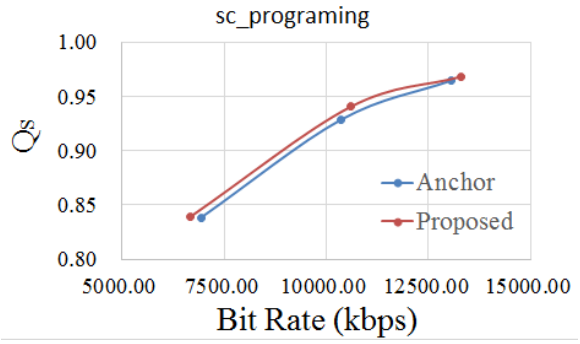
We propose a perceptual screen content quality assessment method and then employ it to optimize the encoding process of screen content compression. The quality assessment method differentiates textual and pictorial blocks, and applies different window sizes to compute the local visual quality. SSIM-based quality assessment method is employed with adaptive window size selection and information content weighting. Experimental results show the superior performance of the proposed method in predicting the quality of screen content images, and also demonstrate its potential in improving the performance of screen content compression.



(a)



(b)



(c)

Fig. 3. Rate-Distortion performance comparison.

5. REFERENCES

[1] Yan Lu, Shipeng Li, and Huifeng Shen, "Virtualized screen: A third element for cloud-mobile convergence," *MultiMedia, IEEE*, vol. 18, no. 2, pp. 4–11, 2011.

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

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

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

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

[6] Tao Lin, Peijun Zhang, Shuhui Wang, Kailun Zhou, and Xianyi Chen, "Mixed chroma sampling-rate high efficiency video coding for full-chroma screen content," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 23, no. 1, pp. 173–185, 2013.

[7] Huan Yang, Yuming Fang, Weisi Lin, and Zhou Wang, "Subjective quality assessment of screen content images," *Proc. IEEE Int. Workshop on Quality of Multimedia Experience*, 2014.

[8] Eero P Simoncelli and Bruno A Olshausen, "Natural image statistics and neural representation," *Annual review of neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.

[9] Monica S Castelhano and Keith Rayner, "Eye movements during reading, visual search, and scene perception: An overview," *Cognitive and cultural influences on eye movements*, pp. 175–195, 2008.

[10] David J Field and Nuala Brady, "Visual sensitivity, blur and the sources of variability in the amplitude spectra of natural scenes," *Vision research*, vol. 37, no. 23, pp. 3367–3383, 1997.

[11] Zhou Wang and Xinli Shang, "Spatial pooling strategies for perceptual image quality assessment," in *Image Processing, 2006 IEEE International Conference on*. IEEE, 2006, pp. 2945–2948.

[12] Shiqi Wang, Jingjing Fu, Yan Lu, Shipeng Li, and Wen Gao, "Content-aware layered compound video compression," in *Circuits and Systems (ISCAS), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 145–148.

[13] X Yin, Kaizhu Huang, and H Hao, "Robust text detection in natural scene images," *Robust Text Detection in Natural Scene Images*, 2013.

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

[15] Anmin Liu, Weisi Lin, and Manish Narwaria, "Image quality assessment based on gradient similarity," *Image Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 1500–1512, 2012.

[16] Lin Zhang, Ying Shen, and Hongyu Li, "VSI: A visual saliency induced index for perceptual image quality assessment," *Image Processing, IEEE Transactions on*, vol. 23, no. 10, pp. 4270–4281, 2014.

[17] Video Quality Experts Group et al., "Final report from the video quality experts group on the validation of objective models of video quality assessment," *VQEG, Mar*, 2000.

[18] Shiqi Wang, Abdul Rehman, Zhou Wang, Siwei Ma, and Wen Gao, "Perceptual video coding based on SSIM-inspired divisive normalization," *Image Processing, IEEE Transactions on*, vol. 22, no. 4, pp. 1418–1429, 2013.

[19] Abdul Rehman and Zhou Wang, "SSIM-inspired perceptual video coding for HEVC," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*. IEEE, 2012, pp. 497–502.

[20] JCTVC, "HM15.0+RExt-8.0+SCM-2.0rc1," `https://hevc.hhi.fraunhofer.de/svn/svn\_HEVCSoft\\ware/tags/HM-15.0+RExt-8.0+SCM-2.0rc1`.