Objective Quality Assessment and Perceptual Compression of Screen Content Images

Shiqi Wang
City University of Hong Kong

Ke Gu
Beijing University of Technology

Kai Zeng
University of Waterloo

Zhou Wang
University of Waterloo

Weisi Lin
Nanyang Technological University

Screen content image (SCI) has recently emerged as an active topic due to the rapidly increasing demand in many graphically rich services such as wireless displays and virtual desktops. SCIs are often composed of pictorial regions and computer generated textual/graphical content, which exhibit different statistical properties that often lead to different viewer behaviors. Inspired by this, we propose an objective quality assessment approach for SCIs that incorporates both visual field adaptation and information content weighting into structural similarity based local quality assessment. Furthermore, we develop a perceptual screen content coding scheme based on the newly proposed quality assessment measure, targeting at further improving the SCI compression performance. Experimental results show that the proposed quality assessment method not only better predicts the perceptual quality of SCIs, but also demonstrates great potentials in the design of perceptually optimal SCI compression schemes.[1]

Recently, there has been an increasing demand to enable thin-clients to enjoy the computationally intensive and graphically rich services by instantly transmitting the complicated graphical interfaces to the clients. Such time variant interface can be rendered as a screen content image (SCI), which is a mixture of pictorial and computer generated textual/graphical regions. The

Published by the IEEE Computer Society
0272-1716/18/$33.00 ©2018 IEEE

quality of the SCIs directly determines the user experience of the screen remoting system. Therefore, an image quality assessment (IQA) model that can predict the perceptual quality of SCIs is desirable, which serves as a benchmark for monitoring, adjusting and optimizing the performance of the screen remoting systems.

In the past decades, there has been significant progress in the field of objective IQA.[1,2,3,4,5] However, most existing methods are designed and validated based on natural images, which do not always share the same properties of SCIs. Typically, the discontinuous-tone computer generated image is featured by repeated patterns, sharp edges and thin lines with few colors, while natural images usually have continuous-tone, smoother edges, thicker lines and more colors. Moreover, the acquisition of natural images may introduce noise due to the physical limitations of imaging sensors, while the screen content is usually noise free as they may be purely generated by computers. In view of these distinct properties of SCIs, in[6] a screen image quality assessment database (SIQAD) was created, which contains 20 reference and 980 distorted SCIs in total. The distorted images are generated by different distortion types including Gaussian noise, Gaussian blur, motion blur, contrast changing, JPEG, JPEG2000 and layer segmentation based coding. The reported low correlations between the scores of subjective and objective measures suggest that there is still large room to improve for SCI quality assessment.[6] In other words, IQA methods that suffice to provide useful quality evaluation of SCIs are largely lacking.

In this work, we study the characteristics of the SCIs and propose an IQA method that predicts SCI quality by incorporating viewing field adaption and local information content weighting. As widely hypothesized in computational vision science, the major task of the human visual system (HVS) when viewing a real scene is to act as an optimal information extractor, or an efficient coder.[7] This motivates us to evaluate the quality of SCIs with the strategy of local information content weighting. 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 than in textual regions.[8] A possible reason accounting for such observation is that the textual content is richer in salient stimuli. These observations further inspire us to introduce spatial adaptation in the local quality assessment approach.

In contrast to the numerous recent efforts in developing high efficiency SCI compression techniques, little has been dedicated to visual perception based SCI compression. This is due to the lack of trusted SCI IQA models that can provide essential guidance in optimizing advanced SCI coding schemes. Given our newly proposed SCI IQA method, we further incorporate it into a High Efficiency Video Coding (HEVC) screen content codec, targeting at improving the coding efficiency of SCIs. Specifically, we propose a novel perceptual SCI compression scheme inspired by the design philosophy of the divisive normalization transform,[9] which has been shown to be a useful framework that better accounts for the spatially varying distortion sensitivities of the HVS.

OBJECTIVE QUALITY ASSESSMENT OF SCIs

Characteristics of SCIs

We find that two statistical features are useful in differentiating the characteristics of pictorial and textual regions in an image, and also in the development of meaningful IQA and compression methods for SCIs.

Frequency Energy Falloff Statistics

It has long been discovered in the literature of natural scene statistics 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. By contrast, typical textual images generated by computers appear somewhat "unnatural." This inspires us to further examine such property on SCIs. Examples of natural and textual images are decomposed using

Fourier transform, as demonstrated in Figure 1. It is observed that the energy falloffs against spatial frequency 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 are peaks at mid and high frequencies. It is also interesting to observe that larger characters push the peak frequency towards lower frequencies, which further demonstrates the close relationship between the peak and the width and spacing of the strokes. Although such properties are not explicitly taken advantage of in the design of the proposed IQA method, these observations suggest that the statistical properties of textual images differ from natural images, motivating us to distinguish them in the design of quality assessment method.

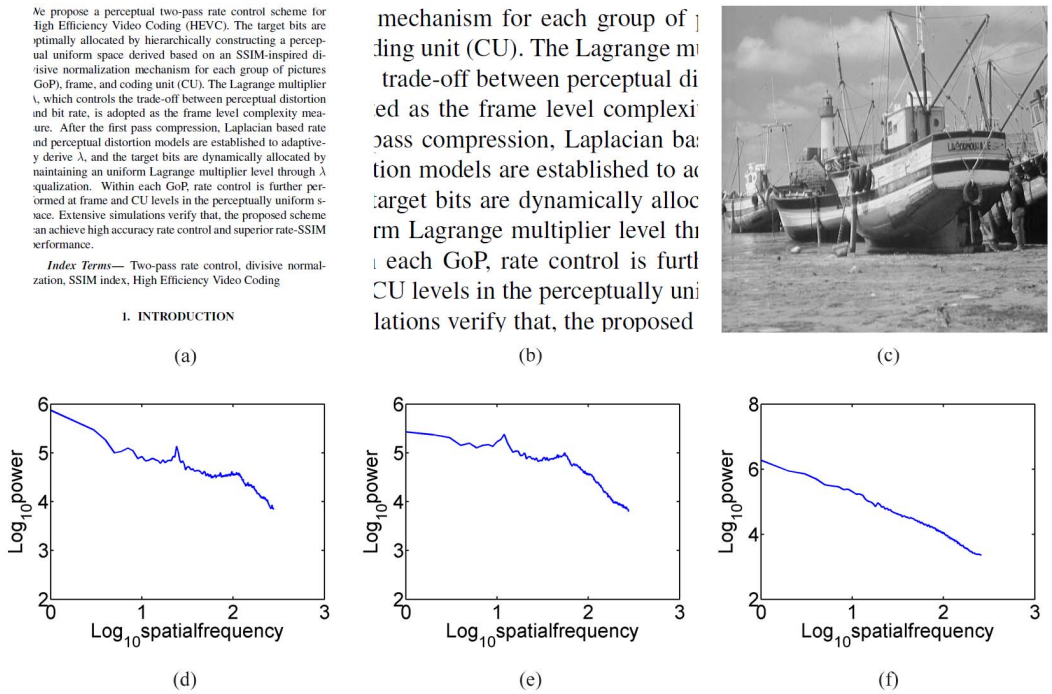

Figure 1. Examples of frequency energy falloffs of textual and natural images in log-log scale. (a) & (b) Textual images at different scales; (c) A natural image; (d) & (e) Frequency energy falloffs of textual images in (a) & (b); (f) Frequency energy falloff of the natural image (c).

Information Content of SCIs

An effective information content model[11] is obtained by locally modeling the input signal with a Gaussian source that is transmitted through a Gaussian noise channel to the receiver.[3] As such, the mutual information between the input and received signals is the amount of the perceived information content, which can be quantified by

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

where σ_p^2 is the variance within a local window x, and σ_n^2 is a constant parameter accounting for the noise level in the visual channel. An example of the local information maps computed using (1), together with the corresponding original images are shown in Figure 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 particular, since the local variances around high contrast edges are usually significant, higher information content can be observed from the information content map. As such, textual regions that contain abundant high contrast edges typically have higher local information content than pictorial regions. This is also consistent with

the recent findings regarding the saliency of webpages in,[12] in which it is shown that SCIs contain richer information in textual regions.

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

Figure 2. Examples of SCIs and the corresponding local information content maps (brighter indicates higher information content). (a)(c) SCIs; (b)(d) Corresponding information content maps.

Quality Assessment Model

The local quality prediction of SCIs is based on the structural similarity (SSIM) index,[1] which has been demonstrated to be an effective quality measure that achieves a good compromise between quality prediction accuracy and computational efficiency. Given two local image patches **x** and **y** extracted from the original and distorted images, respectively, the SSIM index between them is evaluated as

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (2)$$

where μ_x, σ_x, and σ_{xy} are the mean, standard deviation and cross correlation within a local window of 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, which are set to be

$$C_1 = (K_1 L)^2 \quad C_2 = (K_2 L)^2 \quad (3)$$

where L denotes the dynamic range of the pixel values. Parameters K_1 and K_2 are constants and are selected to be 0.01 and 0.03, respectively.

The distinct characteristics of SCIs described in Section 2.1 suggest that it is useful to differentiate the pictorial and textual content, such that their perceptual distortions can be evaluated in different ways. Another interesting property regarding the perception of the screen content is the extent of visual field when viewing SCIs. Monica Castelhano and Keith Rayner observed that the perceptual span in reading textual content is clearly smaller than that in natural scene perception or visual search.[8] This further motivates us to adapt the window size when accessing the local quality of textual and pictorial content.

In this work, instead of performing image segmentation that divides 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 Figure 2. Based on our analysis earlier, the textual regions are richer in saliency stimuli and typically have higher local information content. Therefore, we classify each 4 × 4 block by applying a threshold T_f on the sum of the information content in the block. Subsequently, the overall quality of the textual and pictorial regions Ω_T and Ω_P, denoted by S_T and S_P, respectively, are computed by applying spatially adaptive weighted pooling to access the relative weight of the local content within the textual or pictorial content,

$$S_T = \frac{\sum_{i\in\Omega_T} \mathrm{SSIM}_i \cdot \omega_i^\alpha}{\sum_{i\in\Omega_T} \omega_i^\alpha} \quad (4)$$

$$S_P = \frac{\sum_{j\in\Omega_P} \mathrm{SSIM}_j \cdot \omega_j^\alpha}{\sum_{j\in\Omega_P} \omega_j^\alpha}$$

where the parameter α is used to adjust the strength of weighting. The parameters T_f and α are selected empirically to be 30 and 0.3, respectively. Since textual content is perceived with smaller extend of visual field than pictorial regions, the local SSIM value is calculated by employing different sizes of Gaussian windows with different standard deviations (std), denoted by k_T and k_P, respectively. The local information ω_i and ω_j are calculated with their respective windows, within which the SSIM indices are computed. It is also worth mentioning that textual content is not the only difference between the natural images and SCIs, for example, SCIs often contain large flat areas. However, text is the most dominant characteristic in SCIs. It conveys meaningful information and meanwhile produces high perceptual contrast. Fortunately, this is captured by the information content measure that is used as a weighting factor in the proposed method.

The final SCI quality index (SQI) is given by a weighted average of S_T and S_P, which computes the relative weight between textual and pictorial region as a whole,

$$\mathrm{SQI} = \frac{S_T \cdot \mu_T + S_P \cdot \mu_P}{\mu_T + \mu_P} \tag{5}$$

where $\mu_T = \frac{1}{|\Omega_T|}\sum_{j\in\Omega_T}\omega_{u,j}^{\alpha}$ and $\mu_P = \frac{1}{|\Omega_P|}\sum_{j\in\Omega_P}\omega_{u,j}^{\alpha}$, respectively. These quantities measure the relative density of information content of the textual and pictorial regions, and for fair comparison, the Gaussian window size, denoted by k_U, used to compute ω_u needs to be uniform in both regions, and should be a compromise between k_T and k_P .

The window size parameters k_T , k_P and k_U are determined empirically. Based on our discussions earlier, it is a natural choice to let $k_T < k_P$, and k_U in-between. In our current implementation, we set $k_T = 0.5$, $k_P = 2.5$ and $k_U = 1.5$, respectively.

Validation

The SIQAD database was designed specifically for SCI quality assessment. It contains 980 images that are corrupted by seven distortion types. Full reference IQA algorithms including PSNR, SSIM,[1] IW-SSIM,[2] GSIM,[13] FSIM,[4] VSI,[5] and VIF[3] are used for comparison. In previous tests using subject-rated IQA databases, these state-of-the art IQA algorithms have been repeatedly proven to achieve high correlations with the mean opinion scores (MOS) of natural images.[14] Moreover, the specifically developed IQA measure SCI Perceptual Quality Assessment (SPQA)[6] is compared as well. Three evaluation metrics are employed to assess the performance of these IQA methods, including Pearson linear correlation coefficient (PLCC), Root mean-squared error (RMSE) and Spearman rank correlation coefficient (SRCC).

The PLCC is computed after a nonlinear mapping between the subjective and objective scores to evaluate the prediction accuracy. Given the raw objective scores, a logistic regression function is employed to generate the mapped scores, and PLCC is obtained by computing the correlation between the subjective and mapped objective scores. RMSE is subsequently calculated by comparing the subjective and objective scores after nonlinear mapping. SRCC is nonparametric rank order-based correlation metrics to assess prediction monotonicity. A better objective IQA measure should have higher PLCC and SRCC, but lower RMSE values.

As illustrated in Table 1, when all the test images are included in the evaluation, the proposed method clearly outperforms state-of-the-art quality assessment algorithms in terms of both prediction accuracy and monotonicity. Moreover, we examine the breakdown prediction performance for individual distortion types. The breakdown performance in terms of SRCC and PLCC are provided in Table 2, where in most cases, the proposed method is among the best. Scatter plots of human ratings versus raw objective predicted quality scores before nonlinear mapping for each FR methods are shown in Figure 3. It can be observed that the proposed method can accurately predict the MOS scores.

Table 1. Performance Comparison with State-of-the-Art FR Algorithms Based on the SIQAD Database.

IQA Methods	PSNR	SSIM	IWSSIM	GSIM	FSIM	VSI	VIF	SPQA	SQI
SRCC	0.5608	0.5836	0.6546	0.5483	0.5819	0.5381	0.8069	0.8416	**0.8548**
PLCC	0.5869	0.5912	0.6536	0.5686	0.5902	0.5568	0.8206	0.8584	**0.8644**
RMSE	11.5898	11.5450	10.8329	11.7750	11.5552	11.8904	8.1795	7.3421	**7.1982**

Table 2. Distortion Type Breakdown for IQA Performance Comparisons.

	Gaussian Noise		Gaussian Blur		Motion Blur		Contrast Change		JPEG		JPEG2000		Layer Coding	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
PSNR	**0.9052**	0.8790	0.8603	0.8577	0.7044	0.7132	0.7528	0.6828	0.7696	0.7569	0.7893	0.7746	0.7809	0.7928
SSIM	0.8657	0.8495	0.8530	0.8445	0.7464	0.7443	0.8106	0.7251	0.6921	0.6910	0.7037	0.6937	0.6601	0.6515
IWSSIM	0.8882	0.8743	0.9082	0.9060	0.8417	0.8421	**0.8411**	**0.7540**	0.7998	0.7981	0.8041	0.7988	0.8155	0.8214
GSIM	0.8511	0.8428	0.8830	0.8797	0.7750	0.7765	0.8169	0.7322	0.6765	0.6803	0.7244	0.7123	0.7231	0.7172
FSIM	0.8848	0.8705	0.8208	0.8221	0.7284	0.7239	0.8222	0.7146	0.6640	0.6637	0.7046	0.6851	0.7034	0.7055
VSI	0.8836	0.8655	0.8504	0.8495	0.7657	0.7658	0.7734	0.6459	0.7149	0.7196	0.7498	0.7299	0.7457	0.7419
VIF	0.9015	**0.8890**	0.9101	0.9062	0.8503	0.8504	0.7083	0.5269	0.8005	0.7932	0.8221	0.8151	**0.8414**	**0.8495**
SQI	0.8829	0.8602	**0.9202**	**0.9244**	**0.8789**	**0.8810**	0.7724	0.6677	**0.8218**	**0.8189**	**0.8271**	**0.8169**	0.8310	0.8432

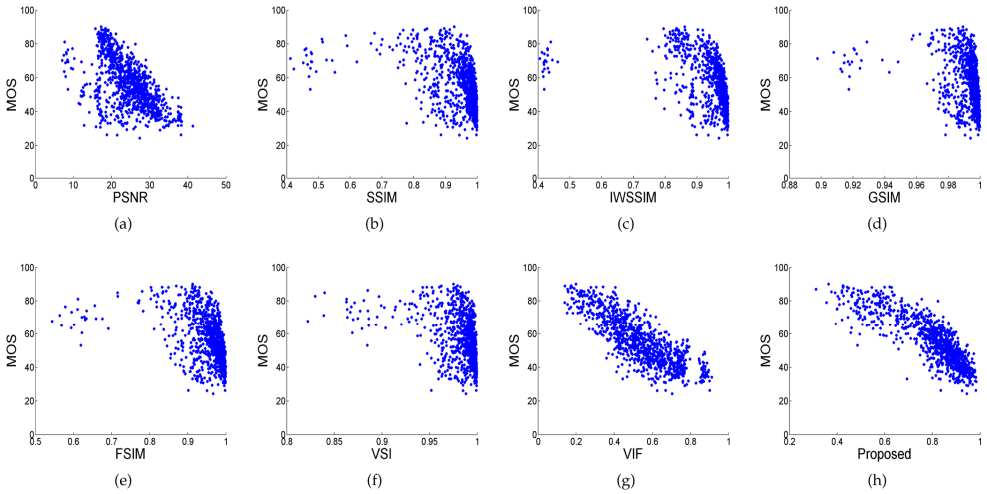

Figure 3. MOS versus model predictions.

Generally, a good IQA measure should be tolerant to small changes in parameter values. Therefore, the sensitivities of the parameters T_f and α on the IQA performance are examined. In particular, T_f varies from 20 to 40 with an interval of 5 and α varies from 0.1 to 0.5 with an interval of 0.1. The results are tabulated in Table 3 and Table 4, from which we can observe that the proposed method achieves considerably stable performance.

Table 3. Parameter sensitivity testing with the variation of T_f.

T_f	20	25	30	35	40
SRCC	0.8580	0.8577	0.8548	0.8489	0.8365
PLCC	0.8630	0.8645	0.8644	0.8620	0.8533
RMSE	7.2318	7.1946	7.1982	7.2558	7.4640

Table 4. Parameter sensitivity testing with the variation of α.

α	0.1	0.2	0.3	0.4	0.5
SRCC	0.8592	0.8586	0.8548	0.8498	0.8438
PLCC	0.8617	0.8648	0.8644	0.8620	0.8582
RMSE	7.2629	7.1866	7.1982	7.2569	7.3469

PERCEPTUAL SCI COMPRESSION

Divisive Normalization Based Video Coding

In the predictive video coding framework, previously coded frames are used to predict the current frame, and only the residuals after prediction are transformed and coded. In divisive normalization transform based video coding scheme,[15] the discrete cosine transform (DCT) coefficient of a residual block C_k is normalized by a positive perceptual normalization factor f to transform the DCT coefficients into a perceptually uniform domain:

$$C(k)' = C(k)/f \tag{6}$$

Subsequently, given the predefined quantization step Q_s, the quantization process of the normalized residuals is formulated as

$$Q(k) = \text{sign}\{C(k)\}\,\text{round}\left\{\frac{|C(k)'|}{Q_S} + p\right\} \tag{7}$$

$$= \text{sign}\{C(k)\}\,\text{round}\left\{\frac{|C(k)|}{Q_S \cdot f} + p\right\}$$

where p is the rounding offset in the quantization.

Correspondingly, at the decoder, the de-quantization and reconstruction of $C(k)$ is performed

$$R(k) = R(k)' \cdot f = Q(k) \cdot Q_S \cdot f$$

$$= \text{sign}\{C(k)\}\,\text{round}\left\{\frac{|C(k)|}{Q_S \cdot f} + p\right\} \cdot Q_S \cdot f \tag{8}$$

As such, the transform coefficients are converted into a perceptually uniform space by adaptively adjusting the quantization parameters for each coding unit (CU). The normalization factor f, which accounts for the perceptual importance, is derived from the SSIM index in DCT domain.[16] Specifically, given the reference and the reconstructed blocks, denoted by **x** and **y**, respectively, the DCT domain SSIM can be calculated as,

$$\text{SSIM(x,y)} = \left(1 - \frac{(X(0)-Y(0))^2}{X(0)^2 + Y(0)^2 + N \cdot C_1}\right) \times \left(1 - \frac{\frac{\sum_{k=1}^{N-1}(X(k)-Y(k))^2}{N-1}}{\frac{\sum_{k=1}^{N-1}X(k)^2 + Y(k)^2}{N-1} + C_2}\right) \tag{9}$$

where X and Y represent the DCT coefficient of x and y, respectively. Parameter N denotes the size of the block, and C_1, C_2 are constants according to the definition of the SSIM index.[1] Assuming each CU contains l DCT blocks, the normalization factors for AC coefficients are given by

$$f_{ac} = \frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{\frac{\sum_{k=1}^{N-1}(X_i(k)^2 + Y_i(k)^2)}{N-1}} + C_2}{\mathrm{E}\left(\sqrt{\frac{\sum_{k=1}^{N-1}(X(k)^2 + Y(k)^2)}{N-1}} + C_2\right)} \quad (10)$$

Practically, only the original block is used because the distorted one cannot be accessed before the actual encoding, and f_{ac} is applied to derive the quantization parameter (QP) offset for each CU.

Following the divisive normalization process, rate distortion optimization (RDO) is performed by minimizing the perceived distortion D with the bit rate R subject to a constraint R_c. This can be converted to an unconstrained optimization problem by

$$\min\{J\} \quad \text{where } J = D + \lambda \cdot R \quad (11)$$

where J denotes the rate-distortion (RD) cost and λ is known as the Lagrange multiplier which controls the tradeoff between R and D. Specifically, the distortion D is defined by computing the sum of squared difference (SSD) between the normalized original and distorted coefficients, which is given by

$$D = \sum_{i=1}^{l}\sum_{k=0}^{N-1}\left(C_i(k)' - R_i(k)'\right)^2 = \sum_{i=1}^{l}\sum_{k=0}^{N-1}\frac{\left(C_i(k) - R_i(k)\right)^2}{f_{ac}^2} \quad (12)$$

As the divisive normalization is performed to transform the DCT coefficients into a perceptually uniform space, the Lagrangian multiplier λ in rate distortion optimization is untouched in the encoder.

Perceptual SCI Compression

The main difference between SSIM and SQI may be well accounted for by the window adaption and information content weighting process. Specifically, in analogies to the SQI method, block type classification is firstly performed by evaluating the local information content in each block and then compare it with a predefined threshold. Subsequently, based on the design philosophy of SQI, the normalization factor for a textual block is given by

$$f_t = f_{ac} / g_t \quad (13)$$

where g_t denotes the relative importance of the local block in terms of the information content:

$$g_t = \sqrt{\frac{2\left(\frac{1}{l}\frac{1}{N}\sum_{i=1}^{l}\sum_{k=1}^{N}\omega_{i,k}^{\alpha}\right)\cdot\mu_T}{\left(\frac{1}{|\Omega_T|}\sum_{j\in\Omega_T}\omega_j^{\alpha}\right)\cdot(\mu_T + \mu_P)}} \quad (14)$$

where k is the spatial location index within a block of size N, and i is the block index within each CU.

Similarly, the normalization factor for a pictorial block is given by

$$f_p = f_{ac} / g_p \quad (15)$$

where

$$g_P = \sqrt{\frac{2\left(\frac{1}{l}\frac{1}{N}\sum_{i=1}^{l}\sum_{k=1}^{N}\omega_{i,k}^{\alpha}\right)\cdot\mu_P}{\left(\frac{1}{|\Omega_P|}\sum_{j\in\Omega_P}\omega_j^{\alpha}\right)\cdot(\mu_T + \mu_P)}} \quad (16)$$

It is worth noting that the local information here is computed with Gaussian window of larger size k_P .

The divisive normalization factors derived from SSIM and SQI for a typical SCI are given in Figure 4, where the divisive normalization factors are computed within each 4x4 block for better visualization. The results demonstrate that with the proposed method, we can assign smaller normalization factors to textual regions with high contrast edges, which are more sensitive to the HVS compared to the pictorial regions. Consequently, with the divisive normalization approach that is specifically designed for SCI compression, we are able to adapt the bit allocation process to improve the overall SCI quality.

Figure 4. Visualization of spatially adaptive divisive normalization factors for a typical SCI (darker pixels indicate higher normalization factors). (a) Original SCI. (b) Normalization factors derived from SSIM. (c) Normalization factors derived from SQI.

Experimental Results

We incorporate the proposed perceptual SCI coding approach into the newly developed HEVC screen content coding extension codec. The test images are in YUV4:4:4 format from both the SIQAD database and HEVC test sequences (the first frame of each sequence). They include common scenarios in screen image processing, such as web browsing, office software editing and video-conferencing. The R-D performance gain (BD-Rate) between the original HEVC encoder (anchor) and the proposed approach in terms of SQI is given in Table 5. 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 5. RD Performance for Different SCIs in terms of SQI.

Image	***BD-Rate***
Webpage	-4.6%
Digital magazine	-4.3%
PPT DOC XLS	-5.0%
Programming	-4.2%
Video Conferencing	-8.3%
Word Editing	-9.7%

We further carried a subjective test to verify the proposed perceptual SCI coding scheme. The subjective test is based on a two-alternative-forced-choice (2AFC) method, which has been

widely adopted in comparing the subjective quality of two video sequences.[15,17] Specifically, in each trial a subject is forced to choose the one he/she thinks to have better quality from a pair of compressed SCIs. We selected four pairs of SCIs, and each pair is repeated four times in random order. In total, 14 subjects were invited in the subjective test. The coding bits, SQI and the results of the subjective tests are reported in Table 6, where the percentage by which the subjects are in favor of the anchor against the proposed scheme are demonstrated. As can be observed, the SCIs are compressed at similar bit bits, and the subjects are inclined to select the proposed method to have better quality. These results provide useful evidence that the proposed method improves the coding performance in terms of subjective quality.

Table 6. Subjective Test Configurations and Results.

SCI	Anchor		Proposed		Percentage (In favor of Anchor)
	bpp	SQI	bpp	SQI	
PPT DOC XLS	0.260	0.8856	0.258	0.8937	23.21%
Program-ming	0.225	0.9286	0.230	0.9402	8.93%
Video Con-ferencing	0.378	0.9363	0.374	0.9454	14.29%
Word Edit-ing	0.265	0.8748	0.261	0.8983	5.36%

CONCLUSION

We propose an objective quality assessment method for SCIs and then employ it to optimize the encoding process of SCI compression. The quality assessment method differentiates textual and pictorial blocks, and applies different parameters in the computation of the structural similarity for local quality assessment. A local information content weighting scheme is further adopted to derive the optimal perceptual weights for spatial pooling. Experimental results show the superior performance of the proposed method in predicting the quality of SCIs, and also demonstrate its potential in improving the performance of SCI compression.

ACKNOWLEDGEMENT

Partial preliminary results of this work were presented at IEEE International Conference on Image Processing, Quebec, Canada, Sep. 2015. This work was also supported in part by National Natural Science Foundation of China under Grants 61703009.

REFERENCES

1. Z. Wang et al., “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004, pp. 600–612.
2. Z. Wang and Q. Li, “Information content weighting for perceptual image quality assessment,” *IEEE Transactions on Image Processing*, vol. 20, no. 5, 2011, pp. 1185–1198.

3. H.R. Sheikh and A.C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, 2006, pp. 430–444.
4. L. Zhang, L. Zhang, and X. Mou, "FSIM: a feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, 2011, pp. 2378–2386.
5. L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, 2014, pp. 4270–4281.
6. H. Yang, Y. Fang, and W. Lin, "Perceptual quality assessment of screen content images," *IEEE Transactions on Image Processing*, vol. 24, no. 11, 2015, pp. 4408–4421.
7. E.P. Simoncelli and B.A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, no. 1, 2001, pp. 1193–1216.
8. M.S. Castelhano and K. Rayner, "Eye movements during reading, visual search, and scene perception: An overview," *Cognitive and Cultural Influences on Eye Movements*, Psychology Press, 2009.
9. M.J. Wainwright and E.P. Simoncelli, "Scale mixtures of gaussians and the statistics of natural images," *Proceedings of the 1999 Conference on Advances in Neural Information Processing Systems* (NIPS 99), 1999, pp. 855–861.
10. D.J. Field and N. Brady, "Visual sensitivity, blur and the sources of variability in the amplitude spectra of natural scenes," *Vision research*, vol. 37, no. 23, 1997, pp. 3367–3383.
11. Z. Wang and X Shang, "Spatial pooling strategies for perceptual image quality assessment," *IEEE International Conference on Image Processing*, 2006, pp. 2945–2948.
12. C. Shen, X. Huang, and Q. Zhao, "Predicting eye fixations on webpage with an ensemble of early features and highlevel representations from deep network," *IEEE Transactions on Multimedia*, vol. 17, no. 11, 2015, pp. 2084–2093.
13. A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Transactions on Image Processing*, 2012, pp. 1500–1512.
14. W. Lin and C.-C. Jay Kou, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, 2011, pp. 297–312.
15. S. Wang et al., "Perceptual video coding based on SSIM-inspired divisive normalizatio," *IEEE Transactions on Image Processing*, vol. 22, no. 4, 2013, pp. 1418–1429.
16. S.S. Channappayya, A.C. Bovik, and R.W. Heath Jr., "Rate bounds on SSIM index of quantized images," *IEEE Transactions on Image Processing*, vol. 17, no. 9, 2008, pp. 1624–1639.
17. S. Wang et al., "SSIM-motivated rate-distortion optimization for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 4, 2012, pp. 516–529.

ABOUT THE AUTHORS

Shiqi Wang received a PhD in computer application technology from Peking University, Beijing, China, in 2014. He was a Postdoc Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. He is currently an Assistant Professor with the Department of Computer Science, City University of Hong Kong. His research interests include image/video quality assessment, compression and analysis. Contact him at sqwang1986@gmail.com.

Ke Gu received a PhD in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2015. He is currently with Beijing University of Technology. His research interests include quality assessment, contrast enhancement, and visual saliency. Contact him at guke@ntu.edu.sg.

Kai Zeng received a PhD in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2013, where he is currently a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering. His research interests include

image and video quality assessment and corresponding applications. Contact him at kzeng@uwaterloo.ca.

Zhou Wang received a PhD in electrical and computer engineering from The University of Texas at Austin, in 2001. He is currently a Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests include image processing, coding, and quality assessment; computational vision and pattern analysis; multimedia communications, and biomedical signal processing. Contact him at zhou.wang@uwaterloo.ca.

Weisi Lin received a PhD from King's College London, London, U.K, in 1993. He is currently a Professor with the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include image processing, visual quality evaluation, and perception-inspired signal modeling. Contact him at wslin@ntu.edu.sg.