PERCEPTUAL SCREEN CONTENT IMAGE QUALITY ASSESSMENT AND COMPRESSION

Shiqi Wang\textsuperscript{1}, Ke Gu\textsuperscript{1,2}, Kai Zeng\textsuperscript{1}, Zhou Wang\textsuperscript{1}, Weisi Lin\textsuperscript{3},

\textsuperscript{1}Dept. of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada
\textsuperscript{2}Institute of Image Comm. Info. Processing, Shanghai Jiao Tong University, Shanghai, China
\textsuperscript{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 $1/f^p$ 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 $1/f^p$ relationship. However, for textual images there is a peak at high frequency, which is some-
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

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

where \( \sigma_p \) is the local variance within a local window \( x \), and \( \sigma_n \) 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 \( x \) and \( y \) extracted from the original and distorted images, respectively, the SSIM index between them is evaluated as

\[ \text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}, \]

where \( \mu_x, \sigma_x \) and \( \sigma_{xy} \) 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 textual 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 \( 4 \times 4 \) block by applying a threshold on the mean of the information content in the block. The overall quality of the textual and pictorial regions \( \Omega_T \) and \( \Omega_P \), denoted by \( Q_T \) and \( Q_P \), respectively, are computed by information...
content-based local quality-weighted pooling:
\[ Q_T = \sum_{i \in \Omega_T} \text{SSIM}_i \cdot \omega_i^a \]
\[ Q_P = \sum_{j \in \Omega_P} \text{SSIM}_j \cdot \omega_j^a, \]
where the parameter \( \alpha \) is a constant used to adjust the strength of weighting and is set to 0.3. Since textual content is perceived with smaller extent of visual field than pictorial regions, the local SSIM map is calculated by employing different sizes of Gaussian windows and standard deviations (std) \( (l_t = 5, \text{std} = 0.5 \) for textual, and \( l_p = 17, \text{std} = 2.5 \) for pictorial). The local information \( \omega_i \) and \( \omega_j \) 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 \( Q_P \):
\[ Q_S = \frac{Q_T \cdot E(\omega_T^a) + Q_P \cdot E(\omega_P^a)}{E(\omega_T^a) + E(\omega_P^a)}, \]
where \( E(\omega_T^a) \) and \( E(\omega_P^a) \) 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 \( l_p \) (\( l = 11, \text{std} = 1.5 \)). 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 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 such, the quantization process of the normalized residuals for a given predefined \( Q_s \) can be formulated as
\[ Q(k) = \text{sign}(C(k)) \cdot \text{round}\{ \frac{|C(k)|}{Q_s} + p \} \]
\[ = \text{sign}(C(k)) \cdot \text{round}\{ \frac{|C(k)|}{Q_s \cdot f} + p \}, \]
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
\[ D = \sum_{k=0}^{N-1} (C(k)' - R(k)')^2 = \sum_{k=0}^{N-1} \frac{(C(k) - R(k))^2}{f^2}, \]
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 \( \lambda \) 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 \( i \)-th textual block is given by
\[ f = \frac{\sqrt{\sum_{k=1}^{N-1} (X(k)^2 + Y(k)^2) / N - 1} + C_2 \cdot g \cdot E(\omega_T^a)}{E(\sqrt{\sum_{k=1}^{N-1} (X(k)^2 + Y(k)^2) / N - 1} + C_2 \cdot (E(\omega_T^a) + E(\omega_P^a)))}, \]
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:
\[ g = \frac{\sum_{j=1}^{N} \omega_j^a}{N \cdot E(\omega_T^a)}. \]

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

Algorithm 1 Summary of SCI quality assessment method

Input: SCI

Output: \( Q_S \)

1: Compute the local information at each pixel location using \( 11 \times 11 \) Gaussian window with \( \text{std} = 1.5 \).
2: For each \( 4 \times 4 \) 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 \( 5 \times 5 \) Gaussian window with \( \text{std} = 0.5 \).
4: For pictorial blocks, compute \( Q_P \) using \( 17 \times 17 \) Gaussian window with \( \text{std} = 2.5 \).
5: Compute final \( Q_S \) using (4).
6: return \( Q_S \).
Table 1. Performance evaluation of the IQA models on SIQAD.

<table>
<thead>
<tr>
<th>Quality Measures</th>
<th>All</th>
<th>JPEG&amp;JPEG2000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PLCC</td>
<td>SRCC</td>
</tr>
<tr>
<td>PSNR</td>
<td>0.5788</td>
<td>0.5539</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.7445</td>
<td>0.7433</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.6047</td>
<td>0.5968</td>
</tr>
<tr>
<td>IW-SSIM</td>
<td>0.6408</td>
<td>0.6414</td>
</tr>
<tr>
<td>GSIM</td>
<td>0.5515</td>
<td>0.5311</td>
</tr>
<tr>
<td>FSIM</td>
<td>0.5741</td>
<td>0.5647</td>
</tr>
<tr>
<td>GMSD</td>
<td>0.7277</td>
<td>0.7168</td>
</tr>
<tr>
<td>VSI</td>
<td>0.5403</td>
<td>0.5199</td>
</tr>
<tr>
<td>VIF</td>
<td>0.8026</td>
<td>0.7857</td>
</tr>
<tr>
<td>Qs</td>
<td>0.8573</td>
<td>0.8456</td>
</tr>
</tbody>
</table>

Table 2. RD performance for different SCIs.

<table>
<thead>
<tr>
<th>Image</th>
<th>BD-Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ppt_doc.xls</td>
<td>-5.1%</td>
</tr>
<tr>
<td>sc_programming</td>
<td>-4.7%</td>
</tr>
<tr>
<td>video_conferencing</td>
<td>-7.2%</td>
</tr>
<tr>
<td>sc_web_browsing</td>
<td>-2.0%</td>
</tr>
<tr>
<td>wordEditing</td>
<td>-5.6%</td>
</tr>
<tr>
<td>twist_tunnel</td>
<td>-0.7%</td>
</tr>
<tr>
<td>news</td>
<td>-1.8%</td>
</tr>
<tr>
<td>sports</td>
<td>-4.9%</td>
</tr>
</tbody>
</table>

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.
5. REFERENCES


