Learning a Blind Quality Evaluation Engine of Screen Content Images

Ke Gu†, Guangtao Zhai†, Weisi Lin†, Xiaokang Yang‡, and Wenjun Zhang‡
†School of Computer Engineering, Nanyang Technological University, Singapore
‡Insti. of Image Commu. & Infor. Proce., Shanghai Jiao Tong University, China
†Shanghai Key Laboratory of Digital Media Processing and Transmissions
Email: gukesjtuee@gmail.com; guke.doctor@gmail.com
Telephone number: 086-13764402720

Abstract

We in this paper investigate how to blindly predict the visual quality of a screen content image (SCI). With the popularity of multi-client and remote-controlling systems, SCIs and the relevant applications have been a hot research topic. In general, SCIs contain texts or graphics in cartoons, ebooks or captures of computer screens. As for blind quality assessment (QA) of natural scene images (NSIs), it has been well established since NSIs possess certain statistical properties. SCIs however do not have reliable statistic models so far and thus the associated blind QA task is hard to be addressed. Aiming at solving this problem, we first extract 13 perceptual-inspired features with the free energy based brain theory and structural degradation model. In order to avoid the overfitting and guarantee the independence of training and testing samples, we then collect 100,000 images and use their objective quality scores computed via a high-accuracy full-reference QA method for SCIs as labels, before learn a new blind quality measure from aforementioned 13 features to the objective quality score. Experimental results performed on a large-scale screen image quality assessment database (SIQAD)
demonstrate that the proposed blind quality metric has a good correlation with human perception of quality, even superior to state-of-the-art full-, reduced- and no-reference QA algorithms.

Keywords: Screen content images (SCIs), image quality assessment (IQA), blind / no-reference (NR), statistical model, machine learning

1. Introduction

Screen content images (SCIs), a mixture of graphics, texts and natural scene images, typically include ebooks, emaps, cartoons and captures of computer screens. As the multi-client collaboration and communication systems, such as remote computing [1], cloud gaming [2] and cloud-guided enhancement [3], are increasingly widespread, SCIs have recently aroused broad interests from multimedia and image processing communities. During the process of acquisition, compression and transmission, SCIs are inevitably degraded by various distortions. For instance, contrast altered SCIs are produced for different settings of screen luminance and contrast. Compressed SCIs will generate blockiness and quantization. Noise often appears when SCIs are transmitted. Hence we in this paper propose an objective quality metric for the degradation measurement of SCIs.

Objective image quality assessment (IQA) can be divided into three types: full reference (FR), reduced reference (RR) and no reference (NR). With the known reference image, FR-IQA models compute the distance between the reference and distorted images. The variation in image structures is a good distance metric, and on this basis, some modern FR-IQA models have been devised in
[4, 5, 6, 7, 8, 9, 10, 11, 12]. Nonetheless, these methods do not perform nicely for the quality evaluation of SCIs. Therefore we have recently proposed an effective structure-induced quality metric (SIQM) [13], which combines the structural similarity measurement with the structural degradation model. Using the new large-scale screen image quality assessment database (SIQAD) [14, 15], the SIQM metric was shown to outperform state-of-the-art FR-IQA methods.

Using partial reference image or some extracted features, RR-IQA models work with various kinds of strategies. For example, using the advanced multi-resolution analysis tool — wavelet-based contourlet transform (WBCT) — for valid feature extraction, Gao et al. proposed the WBCT-based RR (RR-WBCT) metric [16]. The free energy based distortion metric (FEDM) [17] was devised to simulate the working of the brain’s perception to image quality for RR scenarios. Taking the low-level vision into account, Narwaria et al. developed the fourier transform based quality measure (FTQM) [18] based on the phase and magnitude of the Fourier transform. By largely reducing the amount of reference information used in SSIM, the structural degradation model (SDM) [19] succeeded in converting SSIM into the effective RR-IQA technique with only few features.

In most cases the (complete or partial) reference image is not available, and thus NR-IQA models become the only choice which has been intensively studied during the past few years. One type of classical blind IQA metrics is to predict the visual quality of specific distortion types, e.g. blockiness [20, 21], image sharpness / blurriness [22, 23, 24], and multiple distortions [25]. Another type of popular and mature strategy for the exploration of NR-IQA performs by extracting natural scene statistics (NSS) based features to be mapped to the visual
quality score with machine learning. In [26], Moorthy et al. have found that the subband coefficients after discrete wavelet transform (DWT) can be well-fitted using a generalized Gaussian distribution (GGD), and thus developed a two-stage system by distortion classification before quality prediction. In [27], Saad et al. observed that the coefficients after discrete cosine transform (DCT) have better regulations, and without identifying distortion types, they can directly predict the visual quality from local DCT coefficients. In [28], Mittal et al. discovered the NSS in the spatial domain, and thereby proposed an effective and efficient blind quality metric that is usually considered to be the benchmark in existing researches of NR-IQA methods.

As compared to the opinion-aware (OA) NR-IQA metrics stated above, there has been recently a trend towards constructing opinion-unaware (OU) blind IQA techniques without using human scored images. In [29], natural image quality evaluator (NIQE) was explored, without any prior knowledge of image contents or distortion categories, by quantifying the deviation between the model statistics of the distorted image and those extracted from a collection of pristine natural images. To make the NIQE more complete, Zhang et al. in [30] recently came up with the integrated-local NIQE (IL-NIQE) with three new statistical features (i.e. gradient statistics, statistics of Log-Gabor filter responses and statistics of colors) included.

Yet the NSS model is violated when the images do not come from a natural source, such as computer graphics and document images, or when natural images are corrupted [31], making the aforementioned NR-IQA models work not validly for SCIs. In our recent work, a new statistical model was built on natural scene im-
ages, which derives a high-accuracy blind quality metric [32]. It is surprising that
the statistical regulation approximately exists in lossless screen content images.
Hence we in this paper propose a novel blind quality measure for SCIs (BQMS),
by first extracting features based on a new screen content statistics model before
learning the module from the feature space to the quality score using a large set
of training data.

Compared with the previous works, the main contributions of this paper are
summarized as follows: 1) to the best of our knowledge, this work is the first
one blindly evaluating the quality of SCIs; 2) we construct the first screen content
statistics (SCS) model; 3) in contrast to previous works that only use a few hun-
dreds of training samples, a vast number of 100,000 training data are applied for
building our BQMS method, in order to avoid the overfitting in the learning; 4) the
proposed blind quality metric outperforms state-of-the-art FR-, RR- and NR-IQA
approaches using the new SIQAD database dedicated to SCIs [14, 15].

The remainder of this paper are arranged as follows: Section 2 presents the
statistic model of screen content images before constructs our blind IQA method.
In Section 3, thorough experiments are conducted using the large-scale SIQAD
database to verify the effectiveness of our BQMS with recently developed IQA
metrics. We conclude this paper in Section 4.

2. Methodology

2.1. Screen Content Statistics

We establish the SCS model using two sets of features of free energy measure
and structural degradation information. The free energy theory, from the perspec-
tive of human action, perception and learning, explains and unifies most existing brain principles [33]. Its core concept lies in that, in the brain, a so-called internal generative model controls the cognitive process, with a constructive way to predict meaningful parts and abandon uncertain residuals from an input visual signal. This residual “gap” always exists and it has been found to be highly related to the image quality [17].

To specify, the autoregressive (AR) model, due to its good descriptive ability [34, 35], is used to predict meaningful parts of an input image \( x \), defined as follows:

\[
x_n = \mathcal{X}^c(x_n) \mathbf{p} + q_n
\] (1)

where \( n \) is the pixel index. \( \mathcal{X}^c(x_n) \) is a vector of the \( c \) nearest pixel indices of \( x_n \). \( \mathbf{p} = (p_1, ..., p_c)^T \) is a vector of AR model coefficients. The superscript “\( T \)” indicates transpose. \( q_n \) is the error term. To solve \( \mathbf{p} \), we need to rewrite the linear system to be the matrix form:

\[
\hat{\mathbf{p}} = \arg \min_{\mathbf{p}} \| \mathbf{x} - \mathbf{X} \mathbf{p} \|_2
\] (2)

where \( \mathbf{x} = (x_1, ..., x_c)^T \); \( \mathbf{X}(t,: \) \( = \mathcal{X}^c(x_t) \). We use the least square method to find the solution to be \( \hat{\mathbf{p}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{x} \). In real application, an 8-th order AR model is trained in a local \( 7 \times 7 \) neighborhood before the pseudo-inverse of the \( 48 \times 8 \) matrix are solved via Gaussian eliminations. The predicted \( \hat{x} \) can be estimated as

\[
\hat{x}_n = \mathcal{X}^c(x_n) \hat{\mathbf{p}}.
\] (3)

Then, following the analysis in [17], the free energy of the image \( x \) can be ap-
proximated as the entropy of residual “gap”:

\[ F_e(x) = - \sum_i P_i(\bar{x}) \log P_i(\bar{x}) \]  

(4)

where \( \bar{x} = x - \hat{x} \) being the prediction error between the input image and its predicted one; \( P_i(\bar{x}) \) is the probability density of grayscale \( i \) in \( \bar{x} \).

Structural degradation model derives from the observation that, after the low-pass filtering, images corrupted by various distortion types and intensities will appear distinct degrees of spatial frequency reduction. For example, with the same low-pass filter, noisy images in general have higher degrees of frequency decrement than blurred images. As thus, we define the structural degradation information by

\[ S_\mu(x) = E\left( \frac{\sigma(\hat{\mu}_x \hat{\mu}_x) + \epsilon}{\sigma(\hat{\mu}_x)\sigma(\hat{\mu}_x) + \epsilon} \right) \]  

(5)

\[ S_\sigma(x) = E\left( \frac{\sigma(\hat{\sigma}_x \hat{\sigma}_x) + \epsilon}{\sigma(\hat{\sigma}_x)\sigma(\hat{\sigma}_x) + \epsilon} \right) \]  

(6)

where \( \hat{\mu}_x \) and \( \hat{\sigma}_x \) represent the local mean and variance of \( x \), computed by a normalized Gaussian kernel \( w = \{ w(u, v) | u = -U, ..., U, v = -V, ..., V \} \) with standard deviation of 1.5; \( \hat{\mu}_x \) and \( \hat{\sigma}_x \) are similarly defined using the impulse function instead of the Gaussian kernel; \( E(\cdot) \) is to compute the global average; \( \sigma_{(AB)} \) indicates the local covariance between \( A \) and \( B \); \( \epsilon \) is a small positive constant to avoid division-by-zero.

Further, we modify the structural degradation information with three considerations. First, we choose three couples of \((U, V)\) as \((1, 1), (3, 3)\) and \((5, 5)\) in the kernels, to introduce different amount of neighboring information. Second, \( S_\mu(x) \) and \( S_\sigma(x) \) are reversed when \( F_e(x) \) is bigger than \( T \), making the structural degradation information of images, which are corrupted by various distortion types such
Table 1: Definitions of the 12 structural degradation features for interior and exterior parts and three pairs of \((U, V)\).

<table>
<thead>
<tr>
<th></th>
<th>Interior parts</th>
<th>Exterior parts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(S_\mu)</td>
<td>(S_\sigma)</td>
</tr>
<tr>
<td>((U, V) = (1, 1))</td>
<td>(S_{i1}^\mu)</td>
<td>(S_{i2}^\mu)</td>
</tr>
<tr>
<td>((U, V) = (3, 3))</td>
<td>(S_{i2}^\mu)</td>
<td>(S_{i3}^\mu)</td>
</tr>
<tr>
<td>((U, V) = (5, 5))</td>
<td>(S_{i5}^\mu)</td>
<td>(S_{i5}^\sigma)</td>
</tr>
</tbody>
</table>

as blur and noise, have consistent changes with the visual quality. Third, \(S_\mu(x)\) and \(S_\sigma(x)\) are computed in the interior \(6 \times 6\) part and exterior block-edge part, to discriminate different frequency decrease in interior and exterior areas caused by JPEG compression. Table 1 presents all the 12 structural degradation features.

In order to validate the effectiveness of the screen content statistics, we have first downloaded thousands of “webpage” and “screen snap” images from the “Google Images” website, then used human eyes to observe the overall images one by one, and finally picked up 1,000 high-quality images. Those chosen images are of size from \(500 \times 221\) to \(4218 \times 1265\), part of which are exhibited in Fig. 1. It deserves to emphasize that there is not a bit overlapping between the chosen 1,000 high-quality images and the original images in the testing SIQAD database.

We then compare the structural degradation features \(S_\mu^\tau(x_0)\) and \(S_\sigma^\tau(x_0)\), where \(x_0\) means the original SCI and \(\tau = \{i_1, i_3, i_5, e_1, e_3, e_5\}\), with the free energy feature \(F_\epsilon(x_0)\) on the selected 1,000 original screen content images and draw the 12 scatter plots in Fig. 2. As shown, the linear dependence of the free energy feature and
Figure 1: Representative 37 screen content ("webpage" and "screen snap") images downloaded from the website of "Google Images".
Table 2: The estimates of model parameters $a_\tau$, $b_\tau$, $c_\tau$ and $d_\tau$ for $S^\tau_\mu$ and $S^\tau_\sigma$ ($\tau = \{i_1, i_3, i_5, e_1, e_3, e_5\}$) using the least square method.

<table>
<thead>
<tr>
<th></th>
<th>$a_\tau$</th>
<th>$b_\tau$</th>
<th>$c_\tau$</th>
<th>$d_\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S^i_{\mu}$</td>
<td>-15.690</td>
<td>16.281</td>
<td>-8.7722</td>
<td>9.3716</td>
</tr>
<tr>
<td>$S^i_{\mu}$</td>
<td>-6.3634</td>
<td>6.7923</td>
<td>-16.115</td>
<td>17.153</td>
</tr>
<tr>
<td>$S^i_{\mu}$</td>
<td>-15.358</td>
<td>15.872</td>
<td>-6.5959</td>
<td>7.0473</td>
</tr>
<tr>
<td>$S^{e_{1}}_{\mu}$</td>
<td>-15.630</td>
<td>16.229</td>
<td>-8.6480</td>
<td>9.2890</td>
</tr>
<tr>
<td>$S^{e_{3}}_{\mu}$</td>
<td>-6.3552</td>
<td>6.7896</td>
<td>-15.873</td>
<td>16.934</td>
</tr>
<tr>
<td>$S^{e_{5}}_{\mu}$</td>
<td>-15.353</td>
<td>15.869</td>
<td>-6.5785</td>
<td>7.0370</td>
</tr>
</tbody>
</table>

structural degradation features provides an opportunity to characterize distorted SCIs without original references. We fit the linear regression model:

$$F_e(x_0) = a_\tau \cdot S^\tau_\mu(x_0) + b_\tau$$  \hspace{1cm} (7)

$$F_e(x_0) = c_\tau \cdot S^\tau_\sigma(x_0) + d_\tau$$  \hspace{1cm} (8)

where $\tau = \{i_1, i_3, i_5, e_1, e_3, e_5\}$; the parameters $a_\tau$, $b_\tau$, $c_\tau$ and $d_\tau$ are obtained using the least square method, whose values are reported in Table 2.

2.2. Image Quality Evaluation

Based on the linear relationship of the free energy feature and the structural degradation features on original SCIs, we further define $N^\tau_\mu(x) = F_e(x) - (a_\tau \cdot S^\tau_\mu(x) + b_\tau)$ and $N^\tau_\sigma(x) = F_e(x) - (c_\tau \cdot S^\tau_\sigma(x) + d_\tau)$, since $N^\tau_\mu$ and $N^\tau_\sigma$ are close to zero for high-quality images (i.e. without distortions or with very few distortions), while they will be far from zero when distortions appear and become larger. In
Figure 2: Scatter plots of the structural degradation features $S_{\mu}^\tau(x)$ and $S_{\sigma}^\tau(x)$ ($\tau = \{i_1, i_3, i_5, e_1, e_3, e_5\}$) versus the free energy feature $F_e(x)$ on 1,000 original (high-quality) screen content images. The straight lines are fitted with the least square method.
Table 3: Summarization of the used thirteen features. $F_s$ indicates the features of structural degradation information including $S^\tau_{\mu}$ and $S^\tau_{\sigma}$.

<table>
<thead>
<tr>
<th>Feature ID</th>
<th>Feature Notation</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{01}$</td>
<td>$F_e$</td>
<td>Free energy entropy based on AR model</td>
</tr>
<tr>
<td>$f_{02} - f_{07}$</td>
<td>$N^\tau_{\mu} &amp; N^\tau_{\sigma}$, $\tau = {i_1, i_3, i_5}$</td>
<td>Difference of $F_s$ and $F_e$ in interior parts</td>
</tr>
<tr>
<td>$f_{08} - f_{13}$</td>
<td>$N^\tau_{\mu} &amp; N^\tau_{\sigma}$, $\tau = {e_1, e_3, e_5}$</td>
<td>Difference of $F_s$ and $F_e$ in exterior parts</td>
</tr>
</tbody>
</table>

addition, we also consider the free energy feature which has a good correlation with human ratings on noisy and blurred images [17]. Lastly, thirteen features used in our BQMS metric are summarized in Table 3.

The reason for selecting the adopted 13 features might be explained from the following two perspectives. The first one is that, for high-quality screen content images, there exist high correlations between structural degradation information and free energy feature, as given in Fig. 2, since the Gaussian convolution kernel used in the computation of structural degradation information is a uniform low-pass filter, and conversely, the AR model well preserves image textures but is not good at processing image edges. The second perspective is that, e.g. for white noise, the AR model is generally of the better denoising ability than the Gaussian filter, which gives rise to the deviations from the fitted red lines shown in Fig. 2. Based on these two considerations, we have a reason to believe the extracted features can be learned to infer the visual quality scores close to the real subjective opinions.

After the feature extraction, it needs to learn a regression module from fea-
tures to the image quality score. In order to avoid the problem of overfitting, this paper adopts a total of 100,000 images as training samples. To specify, we apply the aforesaid 1,000 high-quality SCIs as original references, which are exactly content-independent of the pristine images used in the testing SIQAD database. We then simulated the distorted SCIs via six typical distortion types, i.e. Gaussian noise (GN), Gaussian blur (GB), motion blur (MB), contrast change (CC), and JPEG and JPEG2000 (JP2K) compressions. Note that these six distortion types are also included in the SIQAD database. For each image, we create its distorted ones of each type on more than 15 distortion levels with Matlab embedded functions. We finally generate a training dataset consisting of 100,000 distorted screen content images.

But how to label these distorted SCIs? In [36], the authors, instead of training on human opinion scores, used synthetic scores yielded from FR-IQA approaches as training labels. Likewise, in this paper we apply the recently designed SIQM [13], which improves the SSIM weighted by the structural degradation measurement [19]. The SIQM method was proven to work validly for the screen content IQA, and thus the associated quality scores of distorted SCIs are used for approximating subjective ratings.

Having prepared the training set, we use the support vector regression (SVR) [37] to learn the module in the proposed BQMS. In reality, the SVR-based machine learning technology has penetrated into various kinds of research fields, such as used for image retrieval [38] and quality evaluation [39]. Given a training dataset \( D = \{(x_1, y_1), \ldots, (x_r, y_r)\} \), where \( x_i \) and \( y_i \) are respectively a feature vector of \( f_{01} - f_{13} \) in Table 3 and the target output of the SIQM score derived from
Figure 3: The primary framework of the blind BQMS method based on statistical model of screen content images.

For the \(i\)-th training image, let parameters \(\alpha > 0\) and \(\varepsilon > 0\), the standard form of SVR can be expressed as:

\[
\min_{w, \delta, v, v'} \frac{1}{2} w^T w + \alpha \left( \sum_{i=1}^{r} v_i + \sum_{i=1}^{r} v'_i \right)
\]

subject to

\[
w^T \phi(x_i) + \delta - y_i \leq \varepsilon + v_i,
\]

\[
y_i - w^T \phi(x_i) - \delta \leq \varepsilon + v'_i,
\]

\[
v_i, v'_i \geq 0, \quad i = 1, ..., r.
\]

where \(K(x_i, x_j) = \phi(x_i)^T \phi(x_j)\) is the kernel function, which is assigned as the Radial Basis Function (RBF) kernel having \(K(x_i, x_j) = \exp(-k ||x_i - x_j||^2)\).

The parameters \(\alpha, \varepsilon\) and \(k\) are to be determined using the training samples.

Finally, we illustrate the primary framework of the proposed blind BQMS technique in Fig. 3, in order to help readers readily understand how to apply the metric to predict the visual quality of a distorted screen content image.
3. Experimental Results

3.1. Evaluation Metrics, Databases and Protocols

Using the newly released SIQAD database [14, 15], the performance of our BQMS model is computed and compared with state-of-the-art IQA techniques, consisting of: 1) FR FSIM [6], GSI [7], IGM [9], VSI [10] and SIQM [13]; 2) RR FEDM [17], QFTB [18] and SDM [19]; 3) NR DIIVINE [26], BLIINDS-II [27], BRISQUE [28], NIQE [29], IL-NIQE [30] and NFERM [32]. The SIQAD database was constructed dedicated to the screen content IQA, which is made up of 980 distorted SCIs created by using seven distortion types at seven degradation levels to corrupt 20 pristine SCIs. Apart from the aforementioned six types of distortions (i.e. GN, GB, MB, CC, JPEG and JP2K), the SIQAD database also includes the Layer Segmentation based Coding (LSC).

As the video quality experts group (VQEG) suggested, we first remove the nonlinearity of objective quality predictions through a five-parameter nonlinear regression function [40]:

\[ v_c = \beta_1 \left( \frac{1}{2} - \frac{1}{1 + e^{\beta_2(v_o - \beta_3)}} \right) + \beta_4 v_o + \beta_5 \]  

(9)

where \( v_o, v_c \) and \( v_s \) respectively stand for the vectors of original IQA scores, converted IQA scores after the regression of Eq. (9) and subjective MOS values; model parameters \( \beta_1, \ldots, \beta_5 \) are decided during the curve fitting process. After the regression, we calculate four important performance indices, Spearman rank order correlation coefficient (SRC), Kendall’s rank-order correlation coefficient (KRC), Pearson linear correlation coefficient (PLC), and root mean-squared error.
(RMS) to testify and compare the proposed BQMS with other testing IQA metrics on the SIQAD database. The former two evaluations are used for prediction monotonicity, the third one for prediction accuracy, and the last one for prediction consistency. A value close to one for PLC, SRC and KRC while close to zero for RMS indicates the superior correlation performance in line with subjective human ratings. Interested readers can refer to [40] for a more detailed descriptions of these measures.

3.2. Validation of Screen Content Statistics Model

Additionally, we also check the robustness of the proposed screen content statistic model and justify the model parameters shown in Table 2. The overall 20 original screen content images contained in the SIQAD database were chosen for testing, since they are extremely different from the 1,000 high-quality “webpage” and “screen snap” images downloaded from the “Google Images” website. As illustrated in Fig. 4, we present the scatter plots and the fitted lines using the model parameters listed in Table 2. One can see that the used statistic model can be also applied to the new 20 screen content images with a small deviation.

3.3. Performance Measures and Comparisons

We adopt two manners in performance evaluations. The first evaluation is to compare our BQMS model and OA-NR-IQA algorithms (including DIIVINE, BLIINDS-II, BRISQUE and NFERM), coming from the typical training process. Based on the image scene, we randomly divide all the 980 SCIs in the SIQAD database into two groups, one of which covers 784 distorted SCIs associated to 16 original SCIs (i.e. 80% data for training) and the other of which covers 196
Figure 4: Scatter plots of the structural degradation features $S^\tau_{\mu}(x)$ and $S^\tau_{\sigma}(x)$ ($\tau = \{i_1, i_3, i_5, e_1, e_3, e_5\}$) versus the free energy feature $F_e(x)$ on 20 original screen content images in the SIQAD database. The red lines are based on model parameters in Table 2.
distorted SCIs associated to the rest 4 original SCIs (i.e. 20% data for testing). We repeat the aforesaid random 80% train - 20% test procedure 1,000 times and report the median result across the 1,000 iterations, to confirm that the testing method is robust across image scenes and not biased by some specific train-test splits. Table 4 lists the results of our BQMS approach and blind DIIVINE, BLIINDS-II, BRISQUE and NFERM metrics. As expected, the proposed BQMS model always obtains the highest scores across all.

Apart from the 80% train - 20% test procedure, other train-test procedures with 60%-40%, 40%-60%, 20%-80% are used to validate and compare BQMS with other four NR-IQA models, as reported in Table 4. Likewise, our BQMS algorithm has constantly achieved the optimal results. And moreover, one can easily see that, with the training data decreased and testing data increased, the performance of our model is slightly fluctuated whereas other four blind metrics show a large degree of performance reductions. The above phenomena might be explained by the fact that the NSS which two testing blind IQA metrics rely on does not come into existence in original screen content images, but conversely, our BQMS metric is established on a good screen content statistics model.

The second manner is to measure correlation performance of our BQMS and FR-, RR- and OU-NR-IQA models (including NIQE and IL-NIQE). As mentioned previously, rather than training on only hundreds of images, we utilize a large set of 100,000 distorted SCIs and the corresponding SIQM scores as training samples to learn a module for our (opinion-unaware) BQMS method, which is subsequently used for comparing with FR-, RR- and OU-NR-IQA techniques. From Table 5, we can easily find that the proposed BQMS algorithm, despite the
Table 4: Performance indices of our BQMS and two NR-IQA algorithms on the SIQAD database. We highlight the best performed metric with boldface.

<table>
<thead>
<tr>
<th>Type</th>
<th>80% - 20%</th>
<th>60% - 40%</th>
<th>40% - 60%</th>
<th>20% - 80%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PLC</td>
<td>SRC</td>
<td>KRC</td>
</tr>
<tr>
<td></td>
<td>0.6912</td>
<td>0.6587</td>
<td>0.4805</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.7255</td>
<td>0.6813</td>
<td>0.5020</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.7708</td>
<td>0.7237</td>
<td>0.5382</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.8007</td>
<td>0.7717</td>
<td>0.5815</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.8115</td>
<td>0.8005</td>
<td>0.6056</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5: Performance indices of our BQMS and FR- and RR-IQA models on the SIQAD database. We emphasize the top two quality metrics with boldface.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Type</th>
<th>PLC</th>
<th>SRC</th>
<th>KRC</th>
<th>RMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSIM [6]</td>
<td>FR</td>
<td>0.5746</td>
<td>0.5652</td>
<td>0.4092</td>
<td>11.612</td>
</tr>
<tr>
<td>GSI [7]</td>
<td>FR</td>
<td>0.5515</td>
<td>0.5311</td>
<td>0.3894</td>
<td>11.835</td>
</tr>
<tr>
<td>IGM [9]</td>
<td>FR</td>
<td>0.6287</td>
<td>0.6245</td>
<td>0.4594</td>
<td>11.033</td>
</tr>
<tr>
<td>VSI [10]</td>
<td>FR</td>
<td>0.5403</td>
<td>0.5199</td>
<td>0.3712</td>
<td>11.938</td>
</tr>
<tr>
<td>SIQM [13]</td>
<td>FR</td>
<td><strong>0.8522</strong></td>
<td><strong>0.8455</strong></td>
<td><strong>0.6529</strong></td>
<td><strong>7.4222</strong></td>
</tr>
<tr>
<td>FEDM [17]</td>
<td>RR</td>
<td>0.6670</td>
<td>0.5930</td>
<td>0.4097</td>
<td>10.571</td>
</tr>
<tr>
<td>QFTB [18]</td>
<td>RR</td>
<td>0.5798</td>
<td>0.5430</td>
<td>0.4165</td>
<td>11.559</td>
</tr>
<tr>
<td>SDM [19]</td>
<td>RR</td>
<td>0.5947</td>
<td>0.5891</td>
<td>0.4221</td>
<td>11.406</td>
</tr>
<tr>
<td>NIQE [29]</td>
<td>NR</td>
<td>0.3749</td>
<td>0.3568</td>
<td>0.2462</td>
<td>13.152</td>
</tr>
<tr>
<td>IL-NIQE [30]</td>
<td>NR</td>
<td>0.1497</td>
<td>0.0994</td>
<td>0.0706</td>
<td>14.027</td>
</tr>
<tr>
<td>BQMS (Proposed)</td>
<td>NR</td>
<td><strong>0.7549</strong></td>
<td><strong>0.7223</strong></td>
<td><strong>0.5299</strong></td>
<td><strong>9.3042</strong></td>
</tr>
</tbody>
</table>

The use of SIQM scores (instead of human opinions) as training labels, has obtained very encouraging performance results, noticeably better than OU-NR NIQE and IL-NIQE models. Due to the presence of the (complete or partial) original references for assistance in the FR- and RR-IQA techniques, they are considered hardly matchable with NR-IQA metrics. Despite this, our BQMS model is still superior to most state-of-the-art FR- and RR-IQA algorithms as well as just inferior to the SIQM method which was recently proposed and specific to the IQA of screen content images.

The statistical significance comparison is further conducted by comparing the
prediction errors between $v_c$ (converted scores) and $v_s$ (subjective scores) of each IQA model tested. Supposing that the prediction errors of each metric conform to the Gaussian distribution, we compute the f-test on the errors of our BQMS technique and other IQA methods. A value of “1” indicates that the metric in the row is significantly superior to the one in the column with a confidence greater than 95%, a value of “0” indicates that the metric in the row is significantly comparable to the one in the column, and a value of “−1” indicates that the metric in the row is statistically inferior to the one in the column. We list the results of statistical significance in Fig. 5. It can be viewed that the high performance has been obtained by our blind IQA metric, whose results are “1” in most conditions. Specifically, except the SIQM metric dedicated to the IQA of screen content, our
blind algorithm is completely statistically better than the FR-, RR- and NR-IQA metrics tested in this work.

At last, we straightforwardly compare the correlation performance in light of visualized scatter plots. As shown in Fig. 6, we display the scatter plots of subjective DMOS values and objective quality predictions of state-of-the-art FR VSI, FR SIQM, NR NIQE and our blind BQMS technique. Obviously, our method has acquired the impressive convergency and monotonicity, much better than recently
designed VSI and NIQE metrics while slightly inferior to the FR SIQM metric that is specific to the screen content IQA.

The quality assessment of screen content images is still in a preliminary stage, and there are some room for performance gain of the blind IQA model. The first suggestion about the future work is to improve the proposed screen content statistics model, e.g. to make the statistic model more robust across different contents and thus to reduce the deviations given in Fig. 4. And the second one is to devise higher-accuracy FR-IQA metrics for screen content, with whose scores as training labels to learn a more reliable blind quality method.

4. Conclusion

In this paper we have put forward a new blind quality metric for screen content images. Based on the new screen content statistics model, we have proposed 13 features and proven their effectiveness on the SIQAD database. In order to avoid the overfitting in training, we have used a large set of 100,000 SCIs and the associated SIQM scores to construct the BQMS technique. With respect to state-of-the-art RR- and FR-IQA methods, our BQMS algorithm has acquired better correlation performance in line with subjective quality scores. In the future, two plans will be carried out. One is to improve the screen content statistics model and the other is to employ more reliable training samples by advancing FR quality metrics of screen content images. Our source code will be available soon at http://www.ntu.edu.sg/home/wslin/Publications.htm and https://sites.google.com/site/guke198701/publications.
Acknowledgment

This work was supported in part by the National Science Foundation of China under Grant 61025005, Grant 61371146, Grant 61221001, and Grant 61390514, the Foundation for the Author of National Excellent Doctoral Dissertation of P.R. China under Grant 201339, and the Shanghai Municipal Commission of Economy and Informatization under Grant 140310.

5. Reference


