Evaluating Quality of Screen Content Images Via Structural Variation Analysis

Ke Gu, Junfei Qiao, *Member, IEEE*, Xiongkuo Min, Guanghui Yue, Weisi Lin, *Fellow, IEEE*, and Daniel Thalmann

Abstract—With the quick development and popularity of computers, computer-generated signals have drastically invaded into our daily lives. Screen content image is a typical example, since it also includes graphic and textual images as components as compared with natural scene images which have been deeply explored, and thus screen content image has posed novel challenges to current researches, such as compression, transmission, display, quality assessment, and more. In this paper, we focus our attention on evaluating the quality of screen content images based on the analysis of structural variation, which is caused by compression, transmission, and more. We classify structures into global and local structures, which correspond to basic and detailed perceptions of humans, respectively. The characteristics of graphic and textual images, e.g., limited color variations, and the human visual system are taken into consideration. Based on these concerns, we systematically combine the measurements of variations in the above-stated two types of structures to yield the final quality estimation of screen content images. Thorough experiments are conducted on three screen content image quality databases, in which the images are corrupted during capturing, compression, transmission, etc. Results demonstrate the superiority of our proposed quality model as compared with state-of-the-art relevant methods.

Index Terms—Computer-generated signals, screen content images, quality evaluation, structural variation, human visual system

1 Introduction

Natural scene images were once the dominant input visual signals that are reflected or emitted from natural objects or electronic devices (e.g., televisions) into our eyes. However, this situation was changed due to the emergence of computers. During the past decades, with the prompt development and popularity of computers and relevant Pads and cellulars, computer-created visual signals have been drastically invading into our daily lives. Such a typical example is screen content image, which can be attributed to a mixture of natural scene, graphic and textual images. Natural scene images have been widely and deeply studied, including transmission and compression [1], recognition [2], enhancement [3], saliency detection [4], smoothing [5], change blindness [6], etc. In contrast, the studies of screen content images are still under exploration, and have been gradually concerned by a rising number of researchers, e.g., quality evaluation [7], coding [8], segmentation [9], etc. In this paper we concentrate on quality evaluation of screen content images.

Most traditional Image Quality Assessment (IQA) metrics were mainly developed based upon the assumption that the Human Visual System (HVS) is highly adapted for extracting structural information from the scene. For illustration consider the subsequent IQA models. In [10], [11], [12], Wang *et al.* proposed Structural SIMilarity (SSIM) metric and its variants by comparing the deviation between a corrupted image and its associated reference image in terms of luminance, contrast and structural similarities. Subsequently, several works were devoted to systematically combining the structural measurement with some HVS characteristics, such as visual saliency and contrast sensitivity function, and therefore many advanced IQA methods have been proposed [13], [14], [15], [16], [17]. Besides, quite a few neuroscience-based quality metrics were also well established by exploiting near- and supra-threshold properties [18], [19] or approximating the working mechanism of free energy-based brain theory [20]. The above-mentioned IQA models were however developed aiming at natural scene images, and they were found to implement ineffectively in predicting the visual quality of screen content images [21], [22]. This motivates us to put forward a well-designed screen content IQA technique.

In general, screen content image is more complicated as compared with existing various kinds of images, since it may simultaneously contain natural scene areas, document areas and graphic areas. Despite high practical utility of screen content images, limited contributions were made to solve the

- *Manuscript received xxxx, 2017; revised xxxx, 2017; accepted xxxx, 2017. This work was supported in part by National Natural Science Foundation of China under Grants 61703009, 61533002, Nova Programme Interdisciplinary Cooperation Project under Grant Z161100004916041, Singapore MoE Tier 1 Project M4011379 and RG141/14.*

- *Ke Gu and Jun-Fei Qiao are with Beijing Key Laboratory of Computational Intelligence and Intelligent System, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China (e-mail: guke.doctor@gmail.com; junfeiq@bjut.edu.cn).*
- *Xiongkuo Min is with Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China (email: minxiongkuo@gmail.com).*
- *Guanghui Yue is with the School of Electrical and Information Engineering, Tianjing University, Tianjing 300072, China (e-mail: guanghuiyue1@gmail.com).*
- *Weisi Lin is with School of Computer Science and Engineering, Nanyang Technological University, Singapore, 639798 (e-mail: wslin@ntu.edu.sg).*
- *Daniel Thalmann is with EPFL, CH 1015 Lausanne, Switzerland, (e-mail: Daniel.Thalmann@epfl.ch).*

screen content IQA problem. In [21], the authors proposed Screen content Perceptual Quality Assessment (SPQA) model by first roughly dividing input images into pictorial and textual regions followed by comparing and combining the perceptual differences of the aforesaid two types of regions between the corrupted and uncorrupted screen content images to yield a single quality estimation. In [22], the authors adopt a simple idea by deploying adaptive window sizes of local filters to modify the classical SSIM metric [10]. To specify, a small-size kernel is used for textual areas while a large-size kernel is used for pictorial regions. Nonetheless, it was viewed that, in these two metrics, segmentation is required to distinguish textual and pictorial regions first. This on one hand noticeably raises the computational complexity, and on the other hand it may seriously degrade the performance of quality metrics due to mis-segmentation which mistakes pictorial regions for textual regions and vice versa. In [7], the authors designed a gradient direction-based screen content image quality metric, in which the gradient direction is extracted in accordance to local information followed by a deviation-inspired model for pooling towards deriving the final quality estimation. Via experiment, this algorithm was proven to operate simply and validly. However, it overlooks the influences of several types of distortions, e.g., luminance shift and contrast adjustment, which the gradient direction is not sensitive to.

Towards more accurately evaluating the quality of screen content images, this paper proposes a novel technique based on the analysis of structural variation. More specifically, we suppose that, given a visual signal, human beings first conduct basic perception of its global luminance, contrast, complexity, etc. Obviously, improper global luminance and contrast often lead to a large amount of quality degradation, while an image of high complexity itself usually has a very strong resistance to artifacts such as additive noise. We attribute this type of factors, which varies basic perception, to global structure. In the meantime, local information, e.g., edges and corners, plays a critical role in affecting human detailed perception. It is apparent that unsharp edges and unsuitably distributed corners generally decrease the favorable impression of an image. In our research, the factors related to detailed perception are attributed to local structure. By systematically incorporating the measurements of variations in global and local structures, we finally derive a single quality index of the input screen content image. For the readers' conveniences, we provide an illustration of the design theory behind the proposed quality metric in Fig. 1. Through extensive experiments conducted on Screen Image Quality Assessment Database (SIQAD) [21], Quality Assessment of Compressed Screen content images (QACS) [23], and Screen Content Transmission Loss (SCTL) [24], our metric performs better than current IQA measures, in performance and statistical comparisons.

The organization of this paper is arranged as follows. In Section 2, we present the motivation and implementation details of the proposed quality metric in turn. In Section 3, we conduct a comparison of our metric with recently proposed IQA methods on three screen content image databases [21], [23], [24]. In Section 4, we provide an overall conclusion of the whole paper.

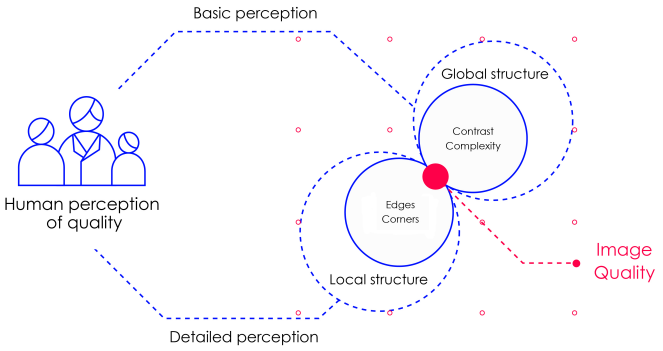


Fig. 1: Illustration of the basic design theory behind the proposed screen content IQA metric.

2 Screen Content IQA Model

2.1 Motivation

Recently the Friston's team provides a milestone brain theory, which is called free energy theory [25], [26]. On this basis, quite a few significant brain principles which refer to human perception, learning, thinking and action in biological and physical sciences are unified in this new free energy theory. Simply speaking, the free energy theory provides a primary assumption that the human brain implements the cognitive process on the basis of a mechanism called internal generative mechanism. Just resorting to this, the human brain leverages a constructive manner to predict the input visual signal. This process can be described by a probabilistic model composed of a prior component and a likelihood component. Towards deducing the posterior possibilities of the external visual input, human visual sensation inverts the likelihood component. Despite the fact that the brain is much more complicated beyond our current level of knowledge, we can reasonably hypothesize that there must exist a discrepancy between the input visual signal and the brain's internal generative model. The discrepancy gap was found to have a close relation with the quality of human perceptions [27].

Considering operational amenability, we suppose that the brain's internal generative mechanism is parametric, and it interprets the external input visual signal by controlling the vector of model parameters Θ. As for a visual signal $\mathcal{S}$, we define its "surprise" through integrating the joint distribution $\mathcal{P}_{(\mathcal{S},\Theta)}$ over the model parameters' space:

$$-\log \mathcal{P}_{(\mathcal{S})} = -\log \int \mathcal{P}_{(\mathcal{S},\Theta)}\, d\Theta. \tag{1}$$

This mathematical expression is not easy to understand. So we bring an assistant component $\mathcal{Q}(\Theta|\mathcal{S})$ and rewrite the equation as

$$-\log \mathcal{P}_{(\mathcal{S})} = -\log \int \mathcal{Q}_{(\Theta|\mathcal{S})} \frac{\mathcal{P}_{(\mathcal{S},\Theta)}}{\mathcal{Q}_{(\Theta|\mathcal{S})}}\, d\Theta. \tag{2}$$

Note that $\mathcal{Q}_{(\Theta|\mathcal{S})}$ is an assistant posterior distribution of the model parameters for the input signal $\mathcal{S}$. It can be regarded to be an approximate posterior to the true posterior of the model parameters $\mathcal{P}_{(\Theta|\mathcal{S})}$. The brain attempts to reduce the discrepancy gap between the approximate posterior $\mathcal{Q}_{(\Theta|\mathcal{S})}$ and the true posterior $\mathcal{P}_{(\Theta|\mathcal{S})}$ by altering the parameters Θ in

$\mathcal{Q}_{(\Theta|\mathcal{S})}$ for seeking the optimal explanation of the perceived visual signal $\mathcal{S}$.

Applying Jensen's inequality to Equation (2), we derive

$$-\log \mathcal{P}_{(\mathcal{S})} \leq -\int \mathcal{Q}_{(\Theta|\mathcal{S})} \log \frac{\mathcal{P}_{(\mathcal{S},\Theta)}}{\mathcal{Q}_{(\Theta|\mathcal{S})}} d\Theta. \tag{3}$$

According to the definition in statistical thermodynamics and physics [28], we define the right part of Equation (3) to be free energy:

$$\mathcal{F}_{(\Theta)} = -\int \mathcal{Q}_{(\Theta|\mathcal{S})} \log \frac{\mathcal{P}_{(\mathcal{S},\Theta)}}{\mathcal{Q}_{(\Theta|\mathcal{S})}} d\Theta. \tag{4}$$

Notice that, based on the Bayes' theorem, we have $\mathcal{P}_{(\mathcal{S},\Theta)} = \mathcal{P}_{(\Theta|\mathcal{S})}\mathcal{P}_{(\mathcal{S})}$, and thus rewrite Equation (4) as

$$\begin{aligned}
\mathcal{F}_{(\Theta)} &= \int \mathcal{Q}_{(\Theta|\mathcal{S})} \log \frac{\mathcal{Q}_{(\Theta|\mathcal{S})}}{\mathcal{P}_{(\mathcal{S})}\mathcal{P}_{(\Theta|\mathcal{S})}} d\Theta \\
&= \int \mathcal{Q}_{(\Theta|\mathcal{S})} \frac{1}{\log \mathcal{P}_{(\mathcal{S})}} d\Theta + \int \mathcal{Q}_{(\Theta|\mathcal{S})} \log \frac{\mathcal{Q}_{(\Theta|\mathcal{S})}}{\mathcal{P}_{(\Theta|\mathcal{S})}} d\Theta \\
&= -\log \mathcal{P}_{(\mathcal{S})} \int \mathcal{Q}_{(\Theta|\mathcal{S})} d\Theta + KL(\mathcal{Q}_{(\Theta|\mathcal{S})}||\mathcal{P}_{(\Theta|\mathcal{S})}) \\
&= -\log \mathcal{P}_{(\mathcal{S})} + KL(\mathcal{Q}_{(\Theta|\mathcal{S})}||\mathcal{P}_{(\Theta|\mathcal{S})}).
\end{aligned} \tag{5}$$

The above equation reveals that the free energy $\mathcal{F}_{(\Theta)}$ varies as $KL(\mathcal{Q}_{(\Theta|\mathcal{S})}||\mathcal{P}_{(\Theta|\mathcal{S})})$, which represents a Kullback-Leibler (KL) divergence of the approximate posterior against the true posterior. We note that $KL(\mathcal{Q}_{(\Theta|\mathcal{S})}||\mathcal{P}_{(\Theta|\mathcal{S})})$ is a non-negative component, so the free energy provides a strict upper bound. Only if the approximate posterior $\mathcal{Q}_{(\Theta|\mathcal{S})}$ equals to the true posterior $\mathcal{P}_{(\Theta|\mathcal{S})}$, $\mathcal{F}_{(\Theta)}$ achieves its minimal $-\log \mathcal{P}_{(\mathcal{S})}$. As for a fixed visual input $\mathcal{S}$, we deduce from Equation (5) that the free energy is suppressed by minimizing the divergence term; that is, the brain attempts to lower the KL divergence between the approximate density and its true posterior density when interpreting an external input visual signal.

Given a distorted image $\mathcal{D}$, it can be expressed by

$$\mathcal{D} = \mathcal{R} + \Delta, \tag{6}$$

where $\mathcal{R}$ is the reference image of $\mathcal{D}$ and Δ stands for the error difference. According to the free energy principle, the human brain actively restores the distorted image by lowering the error difference term, towards well visual perception or semantic understanding. Here we suppose that the true posterior corresponds to $\mathcal{R}$ and the inferred approximate posterior corresponds to $\mathcal{R}'$ that is restored by the human brain. The whole process of restoration is naturally accompanied with quality evaluation. In fact, due to some HVS characteristics, it does not require to restore all information of the image in the brain. For example, there is no necessity to recover the high-frequency noise existed in the reference image because the HVS is not sensitive to it [29]. The SSIM metric is such an example [10]. From the viewpoint of free energy principle, the design theory behind SSIM lies in that the brain restores the image in luminance, contrast and structure domains, and these three domains are selected by adjusting the vector of model parameters Θ. On this basis, we can also explain why the SSIM metric does not perform well on viewing distance-changed image database [30], which is mainly because there is no term (or feature) concerning viewing distance included in the SSIM metric. Facing different utilities, the human brain automatically adjusts the vector of model parameters Θ to choose proper features for visual perception.

2.2 Measurement of Structural Variation

When evaluating the quality of screen content images, the human brain adjusts the vector of model parameters Θ to understand the visual input from basic and detailed perceptions. The human brain is a long-term well trained organ, and thus it works in an extremely efficient manner. Based on this, we believe that the human brain first conducts a basic perception to the global structures of the given screen content image, since there is no need to perceive details for very low-contrast images (e.g., over-bright or over-dark images). After the basic perception, the human brain will selectively pay attention to some local structures to perceive variations in details. Systematically integrating the aforementioned two types of perceptions, the human brain will deliver the final measurement of image quality.

2.2.1 Variations in Global Structures

In the human basic perception of global structures, we first consider the influence of image global contrast on the visual quality. Apparently, very low contrast seriously degrades the image quality. In such condition, capturing details is almost impossible and understanding semantic information becomes a considerably hard task. We assume that the image $\mathcal{R}$ has the benchmark luminance and contrast, and the information entropy will change when luminance or contrast deviations occur. Hence we adopt the Variation Of Entropy (EOV), and we define this feature by

$$\mathcal{F}_1 = \frac{\mathcal{E}_{(\mathcal{D})} + \epsilon_1}{\mathcal{E}_{(\mathcal{R})} + \epsilon_1}, \tag{7}$$

where ϵ_1 is a very small constant for preventing its value too large, and $\mathcal{E}_{(\mathcal{D})}$ and $\mathcal{E}_{(\mathcal{R})}$ are the entropy values of reference and distorted images:

$$\mathcal{E} = -\int \mathcal{H}_{(\rho)} \log \mathcal{H}_{(\rho)} d\rho, \tag{8}$$

where $\mathcal{H}_{(\rho)}$ stands for the probability density of grayscale ρ. Towards observing the EGM's effect, we compare different types of distortions, including Contrast Change (CC), Motion Blur (MB), Gaussian Blur (GB), JPEG2000 compression (J2), JPEG compression (JP), Layer segmentation-backed Coding (LC) [31], White Noise (WN), HEVC Compression (HC) [32], SCC Compression (SC) [33], Transmission loss under HEVC compression (TH) [24], and Transmission loss under SCC compression (TS) [24]. The former seven distortions are from the SIQAD database [21], the eighth and ninth distortions are from the QACS database [23], and the last two distortions are from the SCTL database [24]. It can be found in Fig. 2[1] that, except contrast alteration which may reshape the histogram to

1. From the top to bottom, each box has five horizontal bars, which respectively correspond to the maximum, 75th percentile, median, 25th percentile and minimum values. Red crosses are outliers.

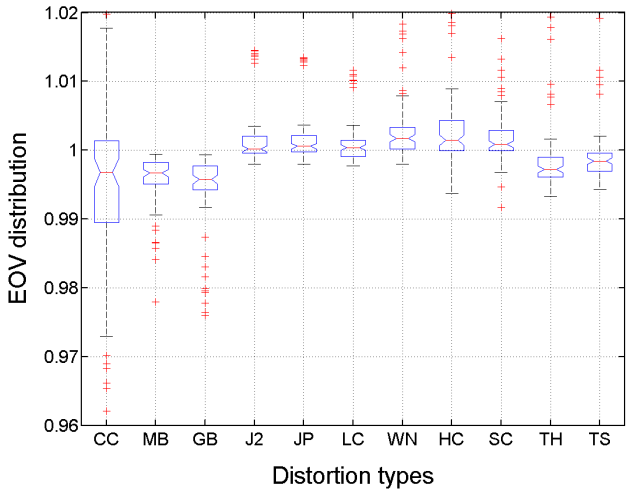


Fig. 2: Box plot of EOV distribution across Contrast Change (CC), Motion Blur (MB), Gaussian Blur (GB), JPEG2000 compression (J2), JPEG compression (JP), Layer segmentation-backed Coding (LC), White Noise (WN), HEVC Compression (HC), SCC Compression (SC), Transmission loss under HEVC compression (TH), and Transmission loss under SCC compression (TS) from three screen content image databases.

a large extent, other distortion types have little influences on the EOV's distribution. Besides information entropy applied in our study, we also consider several popular measures for histogram comparison, which involve Earth Mover's Distance [34], Kullback-Leibler (KL) divergence, and Jensen-Shannon (JS) divergence [35], but results illustrate that they do not introduce obvious performance gain while bring about much implementation cost.

The second feature we care about is the image complexity, which is an essential concept in human basic perception to visual stimulus but is substantially abstract and hard to be endowed with a definite definition. Generally speaking, high-complexity images are composed of more high-frequency information, such as edges and textures, which usually have strong noise masking effects. Clearly, an image having high self-description ability means it has low complexity. Compared with smooth regions, edges and textures are more difficult to self-described. We follow this idea to estimate the image complexity. The classical linear autoregressive (AR) model is first taken into account, since it is good at characterizing a broad scope of natural scenes [36] and also invariant to object transformations such as scaling, rotation and translation [37]. Considering an image, the AR model is constructed in each local patch:

$$r_i = \mathcal{V}_\phi(r_i) \cdot \mathbf{v} + d_i, \tag{9}$$

where r_i is the value of the pixel located at i in the given image; $\mathcal{V}_\phi(r_i)$ constitutes a vector of ϕ member neighborhood, $\mathbf{v}$ is a vector including ϕ AR parameters; d_i is the difference error term between the given pixel and its associated output prediction. In order to determine the optimal AR parameter vector $\mathbf{v}$, we construct a linear equation:

$$\mathbf{v}_{opt} = \arg\min_{\mathbf{v}} \left\| \mathbf{r} - \mathbf{R} \cdot \mathbf{v} \right\|_n, \tag{10}$$

where $\mathbf{r} = (r_1, r_2, ..., r_\delta)^T$ includes the surrounding δ pixels in a $\sqrt{\delta} \times \sqrt{\delta}$ block; $\mathbf{R}(i,:) = \mathcal{V}_\phi(r_i)$; n is the norm order. We assign $n = 2$ in this work, and thus we can leverage the least square method to find the solution of this linear equation to be $\mathbf{v}_{opt} = (\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T\mathbf{r}$. The AR model was found to perform well on textured regions whereas not good at edges, and thus we further introduce another classical bi-lateral (BL) filter to be combined with the AR model towards a tradeoff better filter. We can use Equation (9) to express the BL filter by replacing $\mathbf{v}$ and d_i with $\tilde{\mathbf{v}}$ and $\tilde{d}_i$, respectively. $\tilde{\mathbf{v}}$ indicates a group of BL filter's parameters, which are controlled by two components: one refers to the spatial distance of i and j (j is the index of a neighbour pixel of i); the other refers to the photometric distance between r_i and r_j. Based on this, we define the BL filter as

$$\tilde{g}_j = \exp\left\{ \frac{-\|i-j\|^2}{2\sigma_1^2} + \frac{-(r_i - r_j)^2}{2\sigma_2^2} \right\}, \tag{11}$$

where σ_1 and σ_2 are two constant variances used to balance the strength between the two distances in the equation. $\tilde{d}_i$ is also the difference term. For inheriting both the merits of AR and BL models, we introduce a linear fusion and derive the filtered image:

$$r'_i = \frac{1}{1+\omega_i}\left[\mathcal{V}_\phi(r_i)\hat{\mathbf{v}} + \omega_i \cdot \mathcal{V}_\phi(r_i)\tilde{\mathbf{v}} \right], \tag{12}$$

where ω_i is a space-variant non-negative weight that is used to control the relative contributions of the above two models. For simplicity we fix ω_i as 9 for emphasizing the significance of edges. As thus, we estimate the image complexity feature as follows:

$$\mathcal{F}_2 = -\int \mathcal{H}'_{(\rho)} \log \mathcal{H}'_{(\rho)} d\rho, \tag{13}$$

where $\mathcal{H}'_{(\rho)}$ denotes the probability density of grayscale ρ in the error map between the input image and its associated filtered version, i.e., $r_i^\epsilon = r_i - r'_i$.

2.2.2 *Variations in Local Structures*

When humans conduct detailed perception on local structures, variation in edges is our first concerned feature. There are many effective methods which can be deployed to measure the changes in edges, for instance, Canny operator [38], Sobel operator [39], SSIM metric [10], etc. In comparison, the Sobel operator has obvious advantages in efficacy and efficiency among the three. Furthermore, we compare the Sobel operator with its two modified versions, the Prewitt operator [39] and the Scharr operator [40]. It was observed in results that these three operators have very equivalent computational costs but the Scharr operator achieved a little higher performance than the other two. As thus, we apply the Scharr operator to the reference image $\mathcal{R}$ and derive

$$\mathcal{G}_{(\mathcal{R})} = \sqrt{\mathcal{G}^2_{(\mathcal{R},x)} + \mathcal{G}^2_{(\mathcal{R},y)}}, \tag{14}$$

where $\mathcal{G}_{(\mathcal{R},x)} = \mathcal{M} \otimes \mathcal{R}$ and $\mathcal{G}_{(\mathcal{R},y)} = \mathcal{M}^T \otimes \mathcal{R}$, with $\mathcal{M} = \frac{1}{16}[3,0,-3;10,0,-10;3,0,-3]$ and $\otimes$ indicates the convolution operation. Likewise, we can obtain $\mathcal{G}_{(\mathcal{D})}$ by convolving the distorted image $\mathcal{D}$ with the Scharr operator. Then

the variations in edges between the reference and distorted images can be measured by

$$\mathcal{A}_{(\mathcal{R},\mathcal{D})} = \frac{2\mathcal{G}_{(\mathcal{R})}\mathcal{G}_{(\mathcal{D})} + \epsilon_2}{\mathcal{G}^2_{(\mathcal{R})} + \mathcal{G}^2_{(\mathcal{D})} + \epsilon_2}, \tag{15}$$

where ϵ_2 is a small fixed positive number akin to ϵ_1. Notice that Equation (15) has three excellent characteristics of unique maximum, boundedness and symmetry. Actually, some recent studies have proven the effectiveness of the Scharr operator in capturing the local structural variations [13], [14], [15], [16]. Compared with those studies, the main difference is that we introduce two intrinsic attributes in human perceptions. One is that humans are more sensitive to abrupt local changes in a sequence of signals [41], [42]. We compare an image and its filtered version with a high-pass filter to find the local sharp areas. To specify, for the reference image $\mathcal{R}$, we generate its filtered image as $\mathcal{R}^* = \mathbf{h} \otimes \mathcal{R} = (1-\mathbf{g}) \otimes \mathcal{R} = \mathcal{R} - \mathbf{g} \otimes \mathcal{R} = \mathcal{R} - \mathcal{R}^+$, where $\mathbf{g}$ is a low-pass Gaussian function. As stated earlier, instead of the pixel domain, human beings pay more attention to changes in the structural domain. Therefore we compare $\mathcal{R}$ and $\mathcal{R}^*$:

$$\begin{aligned} \mathcal{B}'_{(\mathcal{R},\mathcal{R}^*)} &= \mathcal{A}_{(\mathcal{R},\mathcal{R})} - \mathcal{A}_{(\mathcal{R},\mathcal{R}^+)} \\ &= 1 - \frac{2\mathcal{G}_{(\mathcal{R})}\mathcal{G}_{(\mathcal{R}^+)} + \epsilon_2}{\mathcal{G}^2_{(\mathcal{R})} + \mathcal{G}^2_{(\mathcal{R}^+)} + \epsilon_2}. \end{aligned} \tag{16}$$

When human eyes stare at a fixation point on an image, all pixels in a proper-size of local window will be included for perception [43]. As for pictorial regions, the local window can be approximated as a circle, and thus the aforementioned rotationally symmetric Gaussian function is used. However, the situation becomes quite different when reading texts. The second attribute refers to the human eye movement in such condition. We assume that human eyes are adapted to saccade from left to right when reading texts. Hence the above local window should include the current fixation point and its right-side pixels. Due to the saccade problem, motion blur must be introduced anyway, and to cope with this problem, the HVS overly enhances the visual input signal beforehand with an associated high-pass filter. We similarly define its filtered image as $\mathcal{R}^\dagger = \mathbf{h}' \otimes \mathcal{R} = (1-\mathbf{g}') \otimes \mathcal{R} = \mathcal{R} - \mathbf{g}' \otimes \mathcal{R} = \mathcal{R} - \mathcal{R}^-$, where $\mathbf{g}'$ denotes a motion blur function, and we compare $\mathcal{R}$ and $\mathcal{R}^\dagger$:

$$\begin{aligned} \mathcal{B}''_{(\mathcal{R},\mathcal{R}^\dagger)} &= \mathcal{A}_{(\mathcal{R},\mathcal{R})} - \mathcal{A}_{(\mathcal{R},\mathcal{R}^-)} \\ &= 1 - \frac{2\mathcal{G}_{(\mathcal{R})}\mathcal{G}_{(\mathcal{R}^-)} + \epsilon_2}{\mathcal{G}^2_{(\mathcal{R})} + \mathcal{G}^2_{(\mathcal{R}^-)} + \epsilon_2}. \end{aligned} \tag{17}$$

We provide the illustration of the two kernels in Fig. 3. Integrating the two attributes in perceptions, we introduce a linear weighting function:

$$\mathcal{B}_{(\mathcal{R})} = \frac{1}{1+\alpha_i}\left[\mathcal{B}'_{(\mathcal{R},\mathcal{R}^*)} + \alpha_i \cdot \mathcal{B}''_{(\mathcal{R},\mathcal{R}^\dagger)}\right], \tag{18}$$

where α_i is a space-variant positive number for altering the two components' relative importance. We assign α_i as the unit based on a rough assumption that pictorial and textual parts

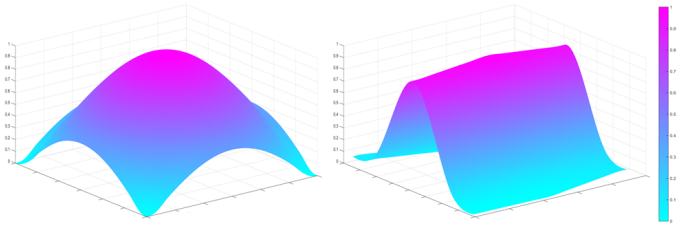


Fig. 3: Gaussian kernel (left) and motion blur kernel (right).

have almost equivalent sizes. We rewrite Equation (18) to be

$$\mathcal{B}_{(\mathcal{R})} = 1 - \frac{\mathcal{G}_{(\mathcal{R})}\mathcal{G}_{(\mathbf{g}\otimes\mathcal{R})} + \frac{1}{2}\epsilon_2}{\mathcal{G}^2_{(\mathcal{R})} + \mathcal{G}^2_{(\mathbf{g}\otimes\mathcal{R})} + \epsilon_2} - \frac{\mathcal{G}_{(\mathcal{R})}\mathcal{G}_{(\mathbf{g}'\otimes\mathcal{R})} + \frac{1}{2}\epsilon_2}{\mathcal{G}^2_{(\mathcal{R})} + \mathcal{G}^2_{(\mathbf{g}'\otimes\mathcal{R})} + \epsilon_2}. \tag{19}$$

Note that $\mathbf{g}$ and $\mathbf{g}'$ are fixed and Equation (19) is a function with only one variable of $\mathcal{R}$. We measure the variations in edges by modifying $\mathcal{A}_{(\mathcal{R},\mathcal{D})}$ with the weighting map $\mathcal{B}_{(\mathcal{R})}$:

$$\mathcal{F}_3 = \frac{\sum_i \mathcal{A}_{(\mathcal{R}_i,\mathcal{D}_i)} \cdot \mathcal{B}_{(\mathcal{R}_i)}}{\sum_i \mathcal{B}_{(\mathcal{R}_i)}}. \tag{20}$$

The fourth feature refers to the measurement of variations in corners. Corner reflects an intrinsic attribute of images, and it has been shown to perform validly in image quality evaluation [44]. Generally speaking, on one hand some distortion types (e.g., noise and blur) destroy the genuine corners which exist in the reference image but disappear after distortions introduced, and on the other hand some distortion types (e.g., block-based JPEG, H.264, and HEVC compressions) usually lead to the reduction of genuine corners while the generation of pseudo corners that occur at block boundaries. We therefore consider comparing the reference and distorted images in terms of the changes in corners to be an impacting factor of quality. More concretely, we rewrite the reference image in a matrix format $\mathcal{R} = [r_{ij}]$. We use the efficient Shi-Tomasi detector [45] to find corners and we thus denote the corner map $\mathcal{C}_{(\mathcal{R})}$:

$$c_{ij} = \begin{cases} 1 & \text{if } r_{ij} \in \mathcal{C}_{(\mathcal{R})} \\ 0 & \text{otherwise} \end{cases}, \tag{21}$$

where $r_{ij} \in \mathcal{C}_{(\mathcal{R})}$ means that a corner was detected at location (i,j). Likewise, we denote the corner map of the distorted image as $\mathcal{C}_{(\mathcal{D})}$. For illustration consider some commonly seen distortion types in Fig. 4, where we provide some examples to reflect the changes in corners of a reference image and its associated distorted images. Red dots denote the original genuine corners whereas blue dots denote newly generated pseudo corners. We can easily find the changes in corners before and after distortion introduction. As compared with the corner map of the reference image, several original genuine corners disappear and meanwhile some new pseudo corners occur due to the influence of distortions. Akin to Equation (15), we measure the variations occurred at corners between the reference and distorted images by

$$\mathcal{F}_4 = \sum_i \frac{2\mathcal{C}_{(\mathcal{R}_i)}\mathcal{C}_{(\mathcal{D}_i)} + \epsilon_3}{\mathcal{C}^2_{(\mathcal{R}_i)} + \mathcal{C}^2_{(\mathcal{D}_i)} + \epsilon_3}, \tag{22}$$

where ϵ_3 is a very small fixed number that is similar to ϵ_1. Notice that $\mathcal{C}_{(\mathcal{R}_i)}$ and $\mathcal{C}_{(\mathcal{D}_i)}$ are binary maps, so we have

Fig. 4: Comparison of genuine and pseudo corners on a reference image and its associated distorted images. Red dots denote the genuine corners while blue dots denote pseudo corners.

$\mathcal{C}_{(\mathcal{R}_i)}\mathcal{C}_{(\mathcal{D}_i)} \equiv \mathcal{C}_{(\mathcal{R}_i)} \bigcap \mathcal{C}_{(\mathcal{D}_i)}, \mathcal{C}^2_{(\mathcal{R}_i)} \equiv \mathcal{C}_{(\mathcal{R}_i)}$ and $\mathcal{C}^2_{(\mathcal{D}_i)} \equiv \mathcal{C}_{(\mathcal{D}_i)}$. We can simplify Equation (22) as

$$\mathcal{F}_4 = 2\sum_i \frac{\mathcal{C}_{(\mathcal{R}_i)} \bigcap \mathcal{C}_{(\mathcal{D}_i)} + \frac{1}{2}\epsilon_3}{\mathcal{C}_{(\mathcal{R}_i)} + \mathcal{C}_{(\mathcal{D}_i)} + \epsilon_3}. \tag{23}$$

2.3 Proposed Screen Content IQA Metric

From the viewpoints of basic and detailed perceptions, four features connected to global and local structures have been extracted. How to reliably combine them becomes an urgent and critical problem. To solve the problem, we first randomly select three reference screen content images from the SIQAD database, as shown in the leftmost column in Fig. 5. As for each reference image, we compute $\mathcal{F}_1$, $\mathcal{F}_3$ and $\mathcal{F}_4$ of its associated distorted images and display them in Fig. 5. From the left and right, the second, third and fourth columns correspond to $\mathcal{F}_1$, $\mathcal{F}_3$ and $\mathcal{F}_4$, respectively. Note that $\mathcal{F}_2$ only depends on the reference image, so we do not include it for comparison. Seeing the second column, the sample points are seemingly irregularly distributed. As stated before, this feature is sensitive to global image contrast change but disturbed very little by other types of distortions. We use two colors (red and blue) to separately represent contrast change and other distortion types. Points corresponding to contrast change have an approximating linear relationship while other points have almost equivalent values despite that the subjective evaluation ratings, namely differential mean opinion score (DMOS), are different. Next we observe the third and fourth columns. There exists an evident near-linear relationship in each plot. Those above linear relationships have the same ordering; that is each of them has a negative correlation with subjective DMOS values. As for the image complexity feature, it is used for normalization towards removing the disturbances of distinct image contents. Moreover, it is believed that the brain first conducts basic perception to the global structures of an image. When global contrast is extremely low or complexity is extremely high, which correspond to over-bright/dark or highly noised images, details included in the image cannot be perceived. In such condition, the brain will straightforwardly provide the image with a very low score; otherwise, detailed perception will be considered to estimate the variations in local structures and the quality will be predicted by systematically fusing the variations in both global and local structures. Via the analyses above, the final quality index is derived based on the subsequent function:

$$\mathcal{I}(\mathcal{R},\mathcal{D}) = \begin{cases} 0 & \text{if } \frac{\mathcal{F}_2}{\mathcal{F}_1} \geq \mathcal{T}_r \\ \frac{1}{\mathcal{F}_2^{\alpha}} \prod_{i=\{1,3,4\}} \mathcal{F}_i & \text{otherwise} \end{cases}, \tag{24}$$

where $\mathcal{T}_r$ is a constant threshold for judging whether the image has extremely low global contrast or extremely high complexity; α is a fixed adjusting operator to decrease the $\mathcal{F}_2$ value in order to make four features have comparable magnitudes.

3 Validations and Discussions

This section focuses on validating the performance of our proposed quality metric with 10 mainstream and state-of-the-art IQA models on three image quality assessment databases related to screen content images. For convenience we provide the proposed method with an abbreviated name, Structural Variation based Quality Index (SVQI).

3.1 Experimental Setup

Quality Models. Recent years have witnessed an increasing number of IQA models. The majority of them deliver well performance and meanwhile require few operating time. This paper selects 11 popular FR-IQA metrics for comparison.

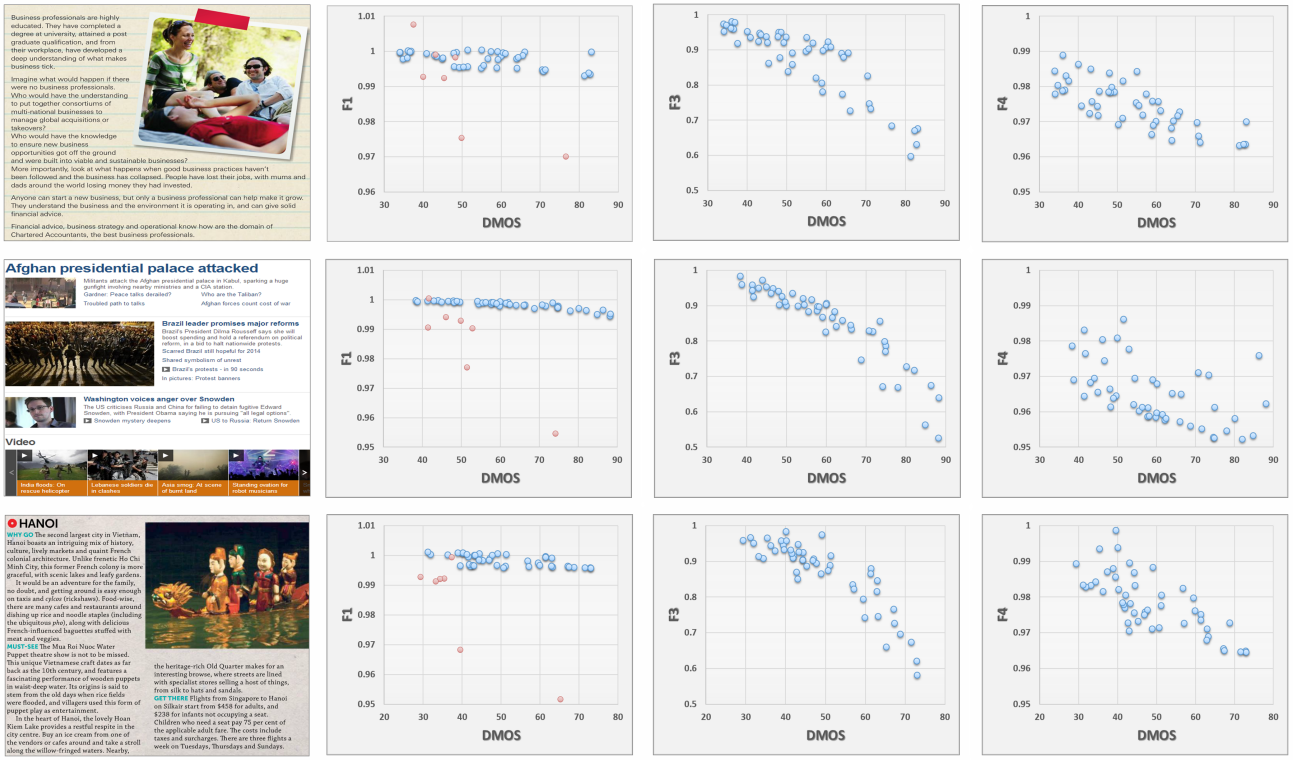


Fig. 5: Comparison of extracted global and local features on three reference screen content images chosen from the SIQAD database. From left to right, four columns are in turn associated to reference images, scatter plots of DMOS versus the features of $\mathcal{F}_1$, $\mathcal{F}_3$ and $\mathcal{F}_4$.

• Noise Quality Measure (NQM) [46], which incorporates Peli's contrast pyramid, variations in local luminance mean and contrast sensitivity, contrast masking effects and contrast interaction between spatial frequencies.

• SSIM [10], which compares the deviation between the reference and corrupted images in terms of luminance, contrast and structural similarities.

• Visual Information Fidelity in Pixel domain (VIFP) [47], which computes the ratio of the mutual information of a reference image and its associated distorted image to the reference image's information content.

• Visual Signal-to-Noise Ratio (VSNR) [18], which leverages near- and supra-threshold properties and visual masking and visual summation effects in the wavelet domain.

• Feature SIMilarity in Color domain (FSIMC) [13], which introduces phase congruency and gradient magnitude to characterize the local image quality since the HVS understands an image mainly relying on low-level features.

• Gradient Similarity Measurement (GSM) [48], which measures the changes of gradient similarity in contrast and structure of images considering the gradient about conveying visual information and favoring scene understanding.

• Gradient Magnitude Similarity Deviation (GMSD) [15], which assumes that the global variation of local quality can better reflect the overall image quality than the direct average pooling.

• Perceptual SIMilarity (PSIM) [14], which was developed based on the supposition that the human visual perception to image quality depends on measuring the variations of micro- and macro-structures.

• Visual Saliency induced Index (VSI) [16], which skillfully incorporates the variations in gradient magnitude and visual saliency caused by distortions to infer the image quality.

• SPQA, [21], which exploits segmentation to distinguish textual and pictorial regions and synthesizes the estimations of perceptual quality of each type of regions to derive an overall quality score.

• Gradient Direction-based Index (GDI) [7], which uses local information to extract the gradient direction followed by a deviation-inspired model for pooling.

Testing Datasets. To our best knowledge, there are three image quality databases related to screen content images. The first one is SIQAD [21], which was constructed by Nanyang Technological University (NTU) in the year of 2015. This database is composed of 20 reference screen content images and 980 corrupted images which were produced by applying 7 types of distortions (i.e., CC, MB, GB, J2, JP, LC and WN) at 7 intensities to those 20 reference images. More than 20 observers were invited to score the 980 images from 0 (the worst) to 10 (the best) with a unit step at a viewing distance about 2.25 time the screen height. The DMOS value of each image in this database is normalized to $[24.2, 90.1]$. The second and third databases were QACS [23] and SCTL

TABLE 1: Comparison of twelve IQA algorithms on screen content databases. We highlight the top three models in bold.

SIQAD [21]	NQM [46]	SSIM [10]	VIFP [47]	VSNR [18]	FSIMC [13]	GSM [48]	GMSD [15]	PSIM [14]	VSI [16]	SPQA [21]	GDI [7]	SVQI (Pro.)
SRC	0.6266	0.7583	**0.8451**	0.5704	0.5817	0.5483	0.7306	0.7056	0.5381	0.8416	**0.8436**	**0.8836**
KRC	0.4539	0.5682	**0.6516**	0.4381	0.4253	0.4054	0.5488	0.5393	0.3874	0.6803	**0.6486**	**0.6985**
PLC	0.6247	0.7615	0.8489	0.5878	0.5920	0.5686	0.7392	0.7144	0.5568	**0.8584**	**0.8515**	**0.8911**
MAE	8.7596	7.1854	**5.9342**	8.9277	8.9912	9.1663	7.3081	7.4771	9.2875	5.7890	**5.9744**	**5.2282**
RMS	11.177	9.2784	7.5650	11.580	11.537	11.775	9.6402	10.016	11.890	**7.3421**	**7.5055**	**6.4965**

QACS [23]	NQM [46]	SSIM [10]	VIFP [47]	VSNR [18]	FSIMC [13]	GSM [48]	GMSD [15]	PSIM [14]	VSI [16]	SPQA [21]	GDI [7]	SVQI (Pro.)
SRC	0.6603	0.8683	**0.9043**	0.7172	**0.9039**	0.8947	0.8769	0.8683	0.8719	0.8456	0.8632	**0.9194**
KRC	0.4749	0.6910	**0.7393**	0.5383	**0.7331**	0.7215	0.7010	0.6897	0.6941	0.6679	0.6812	**0.7623**
PLC	0.6837	0.8696	**0.9028**	0.7050	**0.9019**	0.8921	0.8746	0.8680	0.8715	0.8511	0.8669	**0.9158**
MAE	1.3350	0.8531	**0.7082**	1.1869	**0.7354**	0.7752	0.8156	0.8582	0.8337	0.9656	0.8742	**0.6608**
RMS	1.6189	1.0953	**0.9542**	1.5733	**0.9585**	1.0025	1.0755	1.1015	1.0879	1.1940	1.1059	**0.8909**

SCTL [24]	NQM [46]	SSIM [10]	VIFP [47]	VSNR [18]	FSIMC [13]	GSM [48]	GMSD [15]	PSIM [14]	VSI [16]	SPQA [21]	GDI [7]	SVQI (Pro.)
SRC	0.8337	0.8735	0.8595	0.8865	**0.9103**	0.8893	0.8870	0.8870	0.8729	0.8537	**0.8892**	**0.9134**
KRC	0.6354	0.6790	0.6411	0.7079	**0.7299**	0.7048	0.7085	0.7085	0.6858	0.6560	**0.7057**	**0.7357**
PLC	0.8448	0.8976	0.8937	0.9011	**0.9230**	0.9050	0.9038	0.9038	0.8843	0.8494	**0.9186**	**0.9345**
MAE	0.6830	0.5777	0.5898	0.5377	**0.4923**	0.5349	0.5220	0.5220	0.5886	0.6581	**0.5091**	**0.4566**
RMS	0.8677	0.7148	0.7273	0.7032	**0.6239**	0.6897	0.6938	0.6938	0.7570	0.8496	**0.6407**	**0.5771**

[24], which were established by Peking University in the year of 2016. The QACS database is composed of 24 reference screen content images of size 2560×1440, 1920×1080 and 1280×720, and 492 compressed images which were produced based on two advanced coding technologies (high-efficiency video coding and screen content compression). The total of 20 inexperienced subjects were asked for scoring the 492 images from the lowest 1 to the highest 10 at a viewing distance about 2.25 time the screen height. The MOS value of each image in this database ranges from $[1, 9.9]$. The SCTL database is composed of 20 reference screen content images and 160 contaminated images that were respectively generated by compressed by JPEG and H.264 followed by randomly discarding the coding blocks. The sizes of discarded blocks are separately 8×8 and 16×16 for JPEG and H.264. A test was conducted to invite 20 viewers to participate for scoring the quality of images and record in a 10-category discrete scale manner. The DMOS value of each image in this database is normalized to $[1.35, 7.15]$.

Performance Benchmarking. Five evaluation metrics for correlation performance comparison are routinely employed in IQA researches. The first and second indices are Spearman rank order correlation coefficient (SRC) or rank correlation coefficient and Kendall's rank-order correlation coefficient (KRC). SRC is a non-parametric test towards calculating the degree of association between two variables from the angle of prediction monotonicity, while, KRC evaluates the strength of dependence of two variables and it has stricter demands than SRC. SRC and KRC are defined as follows:

$$\text{SRC} = 1 - \frac{6\sum_{i=1}^{M} d_i^2}{M(M^2-1)}, \tag{25}$$

$$\text{KRC} = \frac{M_c - M_d}{\frac{1}{2}M(M-1)}, \tag{26}$$

where d_i is the difference between the i-th image's ranks in subjective and objective evaluations, M is the image numbers in the testing database, M_c and M_d indicate the numbers of concordant and discordant pairs in the data set, respectively. Before computing the latter three evaluation indices, it requires to eliminate the nonlinearity of objective quality predictions. One typical regression function is the five-parameter function

$$f(x) = \tau_1 \left(0.5 - \frac{1}{1+\exp^{\tau_2(x-\tau_3)}}\right) + \tau_4 x + \tau_5, \tag{27}$$

where x and $f(x)$ are the vectors of raw objective quality scores and converted scores after implementing Equation (27); we use the curve fitting process to compute the values of model parameters $\{\tau_1, ..., \tau_5\}$. The third one is Pearson Linear Correlation coefficient (PLC), which is devoted to estimating the prediction accuracy between the MOS/DMOS values and converted objective quality scores. The fourth and fifth indices are Mean Absolute Error (MAE) and Root Mean Square error (RMS), which measure the difference error between raw and converted objective quality estimations from the viewpoint of prediction consistency. Of the five performance evaluation indices stated above, a value approaching to one for PLC, SRC and KRC, and approaching to zero for MAE and RMS illustrates the superior correlation performance.

Performance Comparison. We illustrate the performance results of 11 IQA algorithms on three screen content image quality databases in Table 1. The top three IQA models are highlighted in bold in order for straightforward comparison. It is apparent that our proposed SVQI metric has constantly obtained the optimal correlation performance as compared with the testing IQA metrics. To specify, we only concentrate on the SRC index, and similar conclusions can be derived for other four indices. First, on the SIQAD database, our SVQI model has led to a performance gain larger than 16.5% than

TABLE 2: Mean Performance comparison of twelve IQA algorithms. We highlight the top three IQA models in boldface.

Average (I)	NQM [46]	SSIM [10]	VIFP [47]	VSNR [18]	FSIMC [13]	GSM [48]	GMSD [15]	PSIM [14]	VSI [16]	SPQA [21]	GDI [7]	SVQI (Pro.)
SRC	0.7069	0.8334	**0.8696**	0.7247	0.7986	0.7774	0.8315	0.8104	0.7610	0.8458	**0.8653**	**0.9055**
KRC	0.5214	0.6461	**0.6773**	0.5614	0.6294	0.6105	0.6528	0.6355	0.5891	0.6681	**0.6785**	**0.7322**
PLC	0.7177	0.8429	**0.8818**	0.7313	0.8056	0.7886	0.8392	0.8183	0.7709	0.8553	**0.8790**	**0.9138**
MAE	3.5925	2.8721	**2.4107**	3.5508	3.4063	3.4921	2.8819	2.9849	3.5699	2.4709	**2.4526**	**2.1152**
RMS	4.5546	3.6962	**3.0822**	4.6187	4.3730	4.4891	3.8031	3.9699	4.5784	3.1498	**3.0840**	**2.6548**

Average (II)	NQM [46]	SSIM [10]	VIFP [47]	VSNR [18]	FSIMC [13]	GSM [48]	GMSD [15]	PSIM [14]	VSI [16]	SPQA [21]	GDI [7]	SVQI (Pro.)
SRC	0.6571	0.8028	**0.8644**	0.6456	0.7110	0.6862	0.7900	0.7695	0.6715	0.8419	**0.8540**	**0.8973**
KRC	0.4780	0.6161	**0.6770**	0.4948	0.5480	0.5300	0.6103	0.5982	0.5091	0.6742	**0.6640**	**0.7214**
PLC	0.6641	0.8074	**0.8696**	0.6539	0.7178	0.6991	0.7962	0.7762	0.6838	0.8596	**0.8627**	**0.9028**
MAE	5.7295	4.6286	**3.8347**	5.7715	5.6691	5.7904	4.6855	4.8094	5.8861	3.8318	**3.9010**	**3.3834**
RMS	7.2850	5.9719	**4.9017**	7.4968	7.2778	7.4406	6.1811	6.4241	7.5422	4.8904	**4.9032**	**4.2262**

the benchmark SSIM method. Relative to the second- and third-ranking VIFP and GDI metrics, the performance gain of the proposed IQA model separately exceeds 4.5% and 4.7%. Second, let us see the performance evaluations on the QACS database. The SVQI model has introduced a performance gain beyond 5.8% as compared with the SSIM method. In contrast to the second and third performers, VIFP and FSIMC, the performance gain of our proposed SVQI metric is greater than 1.6% and 1.7%, respectively. Finally, on the SCTL database, the performance gain of our model is higher than 4.5% in comparison to the SSIM method, and larger than 0.34% and 2.7% in contrast to the second- and third-place FSIMC and GDI models, respectively.

Towards a comprehensive comparison, we further report in Table 2 two commonly used mean performance evaluations. We define the mean performance evaluation as

$$\bar{\xi} = \frac{\sum_{i=1}^{3} \xi_i \cdot \pi_i}{\sum_{i=1}^{3} \pi_i}, \tag{28}$$

where π_i are weights, and ξ_1, ξ_2 and ξ_3 are the performance indices for SIQAD, QACS and SCTL databases, respectively. The first average, called *Average (I)*, is computed by assigning three weights π_i as the unit, and the second average, called *Average (II)*, is computed by assigning three weights as the number of images in each database, namely 980 for SIQAD, 492 for QACS, and 160 for SCTL. It can be found in Table 2, we are able to derived similar observations. We also bold the best performing three models for easy comparison. Our proposed SVQI metric has attained noticeably high performance, much superior to other competing IQA models. We also just pay attention to the SRC index. In contrast to the benchmark SSIM method, our SVQI has resulted in 8.6% performance increase on *Average (I)* and 11.7% performance increase on *Average (II)*. VIFP and GDI are the second and third best performing metrics. Relative to these two, the performance gain of the proposed IQA metric is respectively greater than 4.1% and 4.6% on *Average (I)*, as well as 3.8% and 5.0% on *Average (II)*.

Statistical Significance. We further implement the f-test to compare the statistical significance of our proposed SVQI metric with other IQA approaches on average. Each metric's prediction residual is compared in the f-test. We denote by f the ratio between two residual variances and denote by f_{ct} the critical threshold. If f is greater than f_{ct}, we consider that there is a significant performance difference between two testing IQA algorithms. We assign the confidence level to be 95% as usual. In Table 3, we provide the results of statistical significance comparison. Red ('+1') means the metric in the row is significantly better than the metric in the column; blue ('-1') means the metric in the row is significantly worse than the metric in the column; gray ('0') means the metric in the row is significantly equivalent to the metric in the column. It can be observed that our model is completely statistical better the overall IQA models tested using the f-test.

TABLE 3: Statistical significance comparison of the proposed SVQI metric and competing IQA models with the f-test. Red ('+1') means the metric in the row is significantly better than the metric in the column; blue ('-1') means the metric in the row is significantly worse than the metric in the column; gray ('0') means the metric in the row is significantly equivalent to the metric in the column.

Average	NQM	SSIM	VIFP	VSNR	FSIMC	GSM	GMSD	PSIM	VSI	SPQA	GDI	SVQI
NQM	—	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
SSIM	+1	—	-1	+1	+1	+1	+1	+1	+1	-1	-1	-1
VIFP	+1	+1	—	+1	+1	+1	+1	+1	+1	+1	0	-1
VSNR	+1	-1	-1	—	0	0	-1	-1	0	-1	-1	-1
FSIMC	+1	-1	-1	0	—	0	-1	0	0	-1	-1	-1
GSM	+1	-1	-1	0	0	—	-1	-1	0	-1	-1	-1
GMSD	+1	-1	-1	+1	+1	+1	—	+1	+1	-1	-1	-1
PSIM	+1	-1	-1	+1	0	+1	-1	—	+1	-1	-1	-1
VSI	+1	-1	-1	0	0	0	-1	-1	—	-1	-1	-1
SPQA	+1	+1	-1	+1	+1	+1	+1	+1	+1	—	-1	-1
GDI	+1	+1	0	+1	+1	+1	+1	+1	+1	+1	—	-1
SVQI	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	—

Scatter Plots. Scatter plot is a popular way for visualized comparison. In this paper we provide in Fig. 6 the objective quality predictions of 9 testing quality metric on the SIQAD database. The nine IQA models included refer to SSIM, VSNR, FSIMC, GSI, GMSD, PSIM, VSI, GDI and the proposed

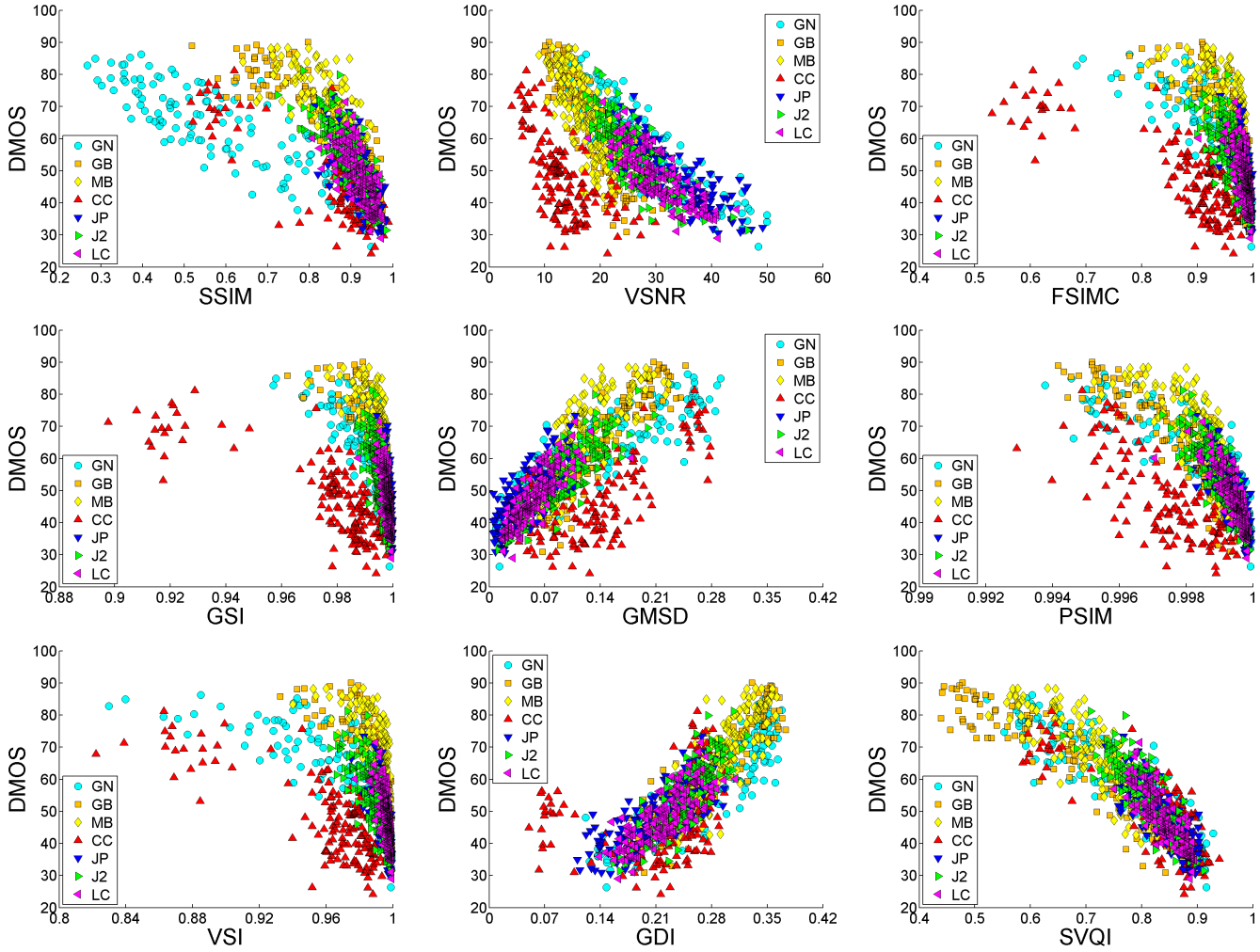


Fig. 6: Scatter plots of DMOS versus SSIM, VSNR, FSIMC, GSI, GMSD, PSIM, VSI, GDI and our proposed SVQI models on the large-scale SIQAD database. CC: red; MB: yellow; GB: orange; J2: green; JP: blue; LC: magenta; WN: cyan.

SVQI algorithms. In each scatter plot, we deploy different colors to label the sample points that correspond to different types of distortions: red for CC, yellow for MB, orange for GB, green for J2, blue for JP, magenta for LC, cyan for WN. Of course, an excellent IQA metric should predict the image quality consistently across different distortion categories. One can see from Fig. 6 that the sample points in the scatter plot of our proposed SVQI model are more robust across different types of distortions, which demonstrates its better consistency in prediction performance. Particularly, as for the data set associated to contrast change, the sample points of the proposed IQA model are quite close to the other six types of distortions, whereas those of the majority of other testing IQA methods are rather far from the other six types of distortions. According to this, we believe that our proposed SVQI metric yields a higher correlation performance.

Implementation. This paper also introduces more image results distorted by JPEG, H.264 and HEVC standards, which would be beneficial for interested readers to better understand how our quality metric works. JPEG is controlled by the parameter Q while H.264 and HEVC are controlled by the parameter QP. As exhibited in Fig. 7, the six images in the top row are corrupted by the JPEG compression. From the left to right, these images separately correspond to six Q values, from 25 to 0 with an interval of 5. We can find that, as the distortion intensity increases, or equivalently, the Q decreases, the associated quality score of SVQI declines. This reflects the effectiveness of our IQA model for assessing the JPEG compression. In Fig. 7, the six images in the middle bottom row are distorted by the H.264 compression. From the left to right, the images are respectively associated to six QP values, from 40 to 50 with an interval of 2. The images show a reducing trend in terms of visual quality, and likewise, the SVQI metric delivers faithful quality prediction values, which proves the SVQI's reliability. Based on the HEVC standard, the six compressed images in the bottom row show similar results with the middle bottom. Further, we compare the two JPEG and H.264 compressed images, (b) and (i), which have similar predicted quality scores, and it can be found that the pictorial parts in (i) are better preserved than those in (b), while, the results for textual parts are contrary. Besides, we also compare two images (a) and (o), which are distorted by JPEG and HEVC compressions, and similar findings and results can be found.

JPEG compression :

(a) Q:25, SVQI:0.8321 (b) Q:20, SVQI:0.8197 (c) Q:15, SVQI:0.8001 (d) Q:10,SVQI:0.7719 (e) Q:5, SVQI:0.7346 (f) Q:0, SVQI:0.6786

H.264 compression :

(g) Q:40, SVQI:0.8631 (h) QP:42, SVQI:0.8403 (i) QP:44, SVQI:0.8136 (j) QP:46, SVQI:0.7878 (k) QP:48, SVQI:0.7469 (l) QP:50, SVQI:0.6976

HEVC compression :

(m) Q:40, SVQI:0.8841 (n) QP:42, SVQI:0.8638 (o) QP:44, SVQI:0.8332 (p) QP:46, SVQI:0.7965 (q) QP:48, SVQI:0.7560 (r) QP:50, SVQI:0.6988

Fig. 7: Exemplified screen content images corrupted by JPEG, H.264 and H.265 compressions and the associated quality scores.

Application. In addition to the crucial values in theory, a good research should have important applications. According to the current researches, the full reference quality measures have two main application scenarios. One is used to guide the image/video compression [49], HDR tone mapping [50], and image fusion [51], since in this case the reference image is fully known and available. With this concern, the proposed SVQI metric can point out which one is the best among many compression technologies and be used to guide the parameter optimization in image/video coding. The other one is the recently designed weak supervision framework dedicated for learning the robust no reference IQA models [52]. To address the expensive, time-consuming and labor-intensive problems when conducting subjective assessment, the quality estimations of a high-accuracy full reference quality method can be treated as weak labels of a large quantity of training images and used for learning the features and developing the no reference models. Finally, it is worthy to stress that, in real applications, we suggest to recognize the flow direction of the text before applying the model of proper direction.

Future work. Our SVQI metric is proposed to merge four features with a manual and empirical function, i.e., Equation (24). Clearly, a better way is to intelligently find the strategy of feature combination, such as distance metric learning. There exist many excellent metric learning methods [53], [54], [55], but they were proposed to address the classification problem. We have attempted to normalize the subjective quality scores to be 101 classes from 0 to 100 with an interval of 1 and thus transfer the regression problem to be the classification issue. However, the experimental results were poor. The support vector regressor (SVR) is essentially treated as the metric learning and used for regression [56]. We also checked the results by using SVR to combine the four features used, but the performance results are respectively 0.8459, 0.8337 and 0.8687 in terms of SROCC. By contrast with Table 1, the SVR-based combination scheme is far less than the proposed SVQI metric. In the future, we will focus on how to better integrate features with the metric learning in IQA tasks.

4 Conclusion

We in this paper have devised a novel method for evaluating the quality of screen content images. The past few decades have witnessed the dramatic development and popularity of computer-generated signals, which have been invading into our daily lives at a quick pace. Screen content image including natural scene, graphic and textual images is such a typical example. Due to the remarkable difference as compared with traditional natural scene images, screen content images have posed new challenges and the associated quality evaluation problem deserves broad attention. Towards this, this paper has proposed an IQA metric called Structural Variation based Quality Index (SVQI) from the perspectives of basic and detailed perceptions of humans. The proposed SVQI metric systematically combines the measurements of variations in global structures and local structures (including global luminance, contrast, complexity, edges and corners) to predict the final quality score of a screen content image. We carry out sufficient experiments using three databases related to screen content images. Our proposed SVQI model has been demonstrated of the better performance than mainstream and state-of-the-art IQA models. For promoting relevant researches, our implementation code will be released to the public soon at https://sites.google.com/site/guke198701/publications.

REFERENCES

[1] H. R. Wu, A. Reibman, W. Lin, F. Pereira, and S. S. Hemami, "Perceptual visual signal compression and transmission," *Proceedings of the IEEE* (invited paper), vol. 101, no. 9, pp. 2025-2043, Sep. 2013.

[2] D. Tao, X. Li, X. Wu, and S. J. Maybank "General tensor discriminant analysis and gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700-1715, Aug. 2007.

[3] T. Arici, S. Dikbas, and Y. Altunbasak, "A histogram modification framework and its application for image contrast enhancement," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 1921-1935, Sep. 2009.

[4] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569-582, Mar. 2015.

[5] Y. Zang, H. Huang, L. Zhang, "Guided adaptive image smoothing via directional anisotropic structure measurement," *IEEE Trans. Visual. Comput. Graphics*, vol. 21, no. 9, pp. 1015-1027, Sept. 2015.

[6] L.-Q. Ma, K. Xu, T.-T. Wong, B.-Y. Jiang, and S.-M. Hu, "Change blindness images," *IEEE Trans. Visual. Comput. Graphics*, vol. 19, no. 11, pp. 1808-1819, Nov. 2013.

[7] Z. Ni, L. Ma, H. Zeng, C. Cai, and K.-K. Ma, "Gradient direction for screen content image quality assessment," *IEEE Sig. Process. Lett.*, vol. 23, no. 10, pp. 1394-1398, Oct. 2016.

[8] T. Lin, P. Zhang, S. Wang, K. Zhou, and X. Chen, "Mixed chroma sampling-rate high efficiency video coding for full-chroma screen content," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 1, pp. 173-185, Jan. 2013.

[9] S. Minaee and Y. Wang, "Screen content image segmentation using robust regression and sparse decomposition," *IEEE J. Emerg. Sel. T. Circuits Syst.*, vol. 6, no. 4, pp. 573-584, Dec. 2016.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600-612, Apr. 2004.

[11] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," *in Proc. IEEE Asilomar Conf. Signals, Syst., Comput.*, pp. 1398-1402, Nov. 2003.

[12] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Sig. Process.: Image Commun.*, vol. 19, no. 2, pp. 121-132, Feb. 2004.

[13] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378-2386, Aug. 2011.

[14] K. Gu, L. Li, H. Lu, X. Min, and W. Lin, "A fast reliable image quality predictor by fusing micro- and macro-structures," *IEEE Trans. Ind. Electron.*, vol. 64, no. 5, pp. 3903-3912, May 2017.

[15] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684-695, Feb. 2014.

[16] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270-4281, Oct. 2014.

[17] K. Gu, S. Wang, G. Zhai, W. Lin, X. Yang, and W. Zhang, "Analysis of distortion distribution for pooling in image quality prediction," *IEEE Trans. Broadcasting*, vol. 62, no. 2, pp. 446-456, Jun. 2016.

[18] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284-2298, Sept. 2007.

[19] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, Mar. 2010. [Online], Available: http://vision.okstate.edu/csiq

[20] J. Wu, W. Lin, G. Shi, and A. Liu, "Perceptual quality metric with internal generative mechanism," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 43-54, Jan. 2013.

[21] H. Yang, Y. Fang, and W. Lin, "Perceptual quality assessment of screen content images," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4408-4421, Nov. 2015.

[22] S. Wang, K. Gu, K. Zeng, Z. Wang, and W. Lin, "Objective quality assessment and perceptual compression of screen content images," *IEEE Computer Graphics and Applications*, DOI: 10.1109/MCG.2016.46, 2017, to appear.

[23] S. Wang, K. Gu, X. Zhang, W. Lin, L. Zhang, S. Ma, and W. Gao, "Subjective and objective quality assessment of compressed screen content images," *IEEE J. Emerg. Sel. T. Circuits Syst.*, vol. 6, no. 4, pp. 532-543, Dec. 2016.

[24] S. Wang, K. Gu, X. Zhang, W. Lin, S. Ma, and W. Gao, "Reduced-reference quality assessment of screen content images," *IEEE Trans. Circuits Syst. Video Technol.*, DOI: 10.1109/TCSVT.2016.2602764, 2017, to appear.

[25] K. Friston, J. Kilner, and L. Harrison, "A free energy principle for the brain," *Journal of Physiology Paris*, vol. 100, pp. 70-87, 2006.

[26] K. Friston, "The free-energy principle: A unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, pp. 127-138, 2010.

[27] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 50-63, Jan. 2015.

[28] R. P. Feynman, *Statistical Mechanics: A Set of Lectures*, 2nd ed. Boulder, CO: Westview, 1998.

[29] C. F. Hall and E. L. Hall, "A nonlinear model for the spatial characteristics of the human visual system," *IEEE Trans. Syst., Man, Cybern.*, vol. 7, no. 3, pp. 161-170, Mar. 1977.

[30] K. Gu, M. Liu, G. Zhai, X. Yang, and W. Zhang, "Quality assessment considering viewing distance and image resolution," *IEEE Trans. Broadcasting*, vol. 61, no. 3, pp. 520-531, Sep. 2015.

[31] Z. Pan, H. Shen, S. Li, and N. Yu, "A low-complexity screen compression scheme for interactive screen sharing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 6, pp. 949-960, Jun. 2013.

[32] G. J. Sullivan, J. R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649-1668, Dec. 2012.

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

[34] O. Pele and M. Werman, "Fast and robust earth mover's distances," *in Proc. IEEE Int. Conf. Computer Vision*, pp. 460-467, Sept. 2009.

[35] D. H. Johnson and S. Sinanović, "Symmetrizing the Kullback-Leibler distance," *IEEE Trans. Information Theory*, 2001.

[36] K. Gu, G. Zhai, W. Lin, X. Yang, and W. Zhang, "No-reference image sharpness assessment in autoregressive parameter space," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3218-3231, Oct. 2015.

[37] I. Sekita, T. Kurita, and N. Otsu, "Complex autoregressive model for shape recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 4, pp. 489-496, Apr. 1992.

[38] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679-698, Nov. 1986.

[39] R. Jain, R. Kasturi, and B. G. Schunck, *Machine Vision.* New York: McGraw-Hill, 1995.

[40] E. Bruce Goldstein, Sensation and Perception. 1979.

[41] J. Miller, "The control of attention by abrupt visual onsets and offsets, *Perception & Psychophysics*, vol. 45, no. 6, pp. 567-571, Nov. 1989.

[42] S. Harnad, "Categorical perception, *Encyclopedia of cognitive science*, 2003.

[43] D. Pelli and K. Tillman, "The uncrowded window of object recognition," *Nature Neuroscience*, vol. 11, no. 10, pp. 1129-1135, Nov. 2008.

[44] L. Li, W. Lin, and H. Zhu, "Learning structural regularity for evaluating blocking artifacts in JPEG images," *IEEE Sig. Process. Lett.*, vol. 21, no. 8, pp. 918-922, Aug. 2014.

[45] J. Shi and C. Tomasi, "Good features to track," *in Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 593-600, 1994.

[46] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 636-650, Apr. 2000.

[47] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430-444, Feb. 2006.

[48] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500-1512, Apr. 2012.

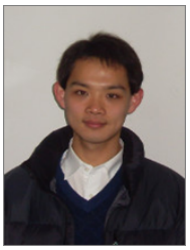
[49] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "Perceptual video coding based on SSIM-inspired divisive normalization," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1418-1429, Apr. 2013.

[50] L. Xie, X. Zhang, S. Wang, S. Wang, X. Zhang, and S. Ma, "Perceptually optimized sparse coding for HDR images via divisive normalization," *in Proc. IEEE Vis. Commun. Image Process.*, pp. 1-5, Nov. 2016.

[51] K. Ma, H. Li, H. Yong, Z. Wang, D. Meng, and L. Zhang, "Robust multi-exposure image fusion: A Structural patch decomposition approach," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2519-2532, May 2017.

[52] K. Gu, D. Tao, J.-F. Qiao, and W. Lin, "Learning a no-reference quality assessment model of enhanced images with big data," *IEEE Trans. Neural Netw. Learning Syst.*, DOI: 10.1109/TNNLS.2017.2649101, 2017, to appear.

[53] E. P. Xing, A. Y. Ng, M. I. Jordan and S. Russell, "Distance metric learning with application to clustering with side-information," *in Proc. Adv. Neural Inf. Process. Syst.*, pp. 521-528, Dec. 2003.
[54] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," *Int. Conf. Mach. Learn.*, pp. 209-216. Jul. 2007.
[55] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10. pp. 207-244, Feb. 2009.
[56] C-C. Chang and C-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intelligent Systems and Technology*, vol. 2, no. 3, Apr. 2011.



Ke Gu received the B.S. and PhD degrees in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2009 and 2015. He is currently the associated editor for IEEE Access, and is the reviewer for 20 top SCI journals. His research interests include quality perception, image processing, and machine learning. Dr. Gu received the Best Paper Award at the IEEE International Conference on Multimedia and Expo (ICME) in 2016, and received the excellent Ph.D. thesis award from the Chinese Institute of Electronics (CIE) in 2016. He is the leading special session organizer in VCIP2016 and ICIP2017.

Jun-Fei Qiao (M'11) received the B.E. and M.E. degrees in control engineer from Liaoning Technical University, Fuxin, China, in 1992 and 1995, respectively, and the Ph.D. degree from Northeast University, Shenyang, China, in 1998. He was a Post-Doctoral Fellow with the School of Automatics, Tianjin University, Tianjin, China, from 1998 to 2000. He joined the Beijing University of Technology, Beijing, China, where he is currently a Professor. He is the Director of the Intelligence Systems Laboratory. His current research interests include neural networks, intelligent systems, self-adaptive/learning systems, and process control systems. Prof. Qiao is a member of the IEEE Computational Intelligence Society. He is a Reviewer for more than 20 international journals, such as the IEEE Transactions on Fuzzy Systems and the IEEE Transactions on Neural Networks and Learning Systems.

Xiongkuo Min received the B.E. degree in electronic engineering from Wuhan University, Wuhan, China, in 2013. He is currently pursuing the Ph.D. degree at the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China. From 2016 to current, he was a visiting student at the Department of Electrical and Computer Engineering, University of Waterloo, Canada. He received the best student paper award of ICME 2016. His research interests include image and video quality assessment, visual attention modeling and perceptual signal processing.

Guanghui Yue received the B.Eng. and M.S. degrees from Tianjin University, Tianjin, China, in 2014 and 2017, repectively, where he is currently pursuing the Ph.D. degree. From Dec. 2017 to current, he was visiting at the School of Computer Engineering, Nanyang Technological University, Singapore. His current research interests include image quality evaluation, signal and image processing, computer vision and machine learning.

Weisi Lin (F'16) received the Ph.D. degree from Kings College London. He is currently an Associate 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, with more than 340 refereed papers published in international journals and conferences. He has been on the Editorial Board of the IEEE T-IP, T-MM (2011-2013), SPL, and JVCI. He has been elected as an APSIPA Distinguished Lecturer (2012/13). He served as a Technical-Program Chair for Pacific-Rim Conference on Multimedia 2012, the IEEE International Conference on Multimedia and Expo 2013, and the International Workshop on Quality of Multimedia Experience 2014. He is a fellow of Institution of Engineering Technology, an Honorary Fellow of the Singapore Institute of Engineering Technologists, and a Fellow of IEEE.

Daniel Thalmann is a Swiss and Canadian Computer Scientist. He is one of the most highly cited scientists in Computer Graphics. He is currently Honorary Professor at EPFL, Switzerland, and Director of Research Development at MIRALab Sarl. Pioneer in research on Virtual Humans, his current research interests also include social robots, crowd simulation and Virtual Reality. Daniel Thalmann has been the Founder of The Virtual Reality Lab (VRlab) at EPFL, Switzerland, Professor at The University of Montreal and Visiting Professor/Researcher at CERN, University of Nebraska, University of Tokyo, and National University of Singapore. From 2009 to 2017, he was Visiting Professor at the Institute for Media Innovation, Nanyang Technological University, Singapore. He is coeditor-in-chief of the Journal of Computer Animation and Virtual Worlds, and member of the editorial board of 12 other journals. Daniel Thalmann was Program Chair and CoChair of several conferences including IEEE-VR, ACM-VRST, and ACM-VRCAI. Daniel Thalmann has published more than 600 papers in Graphics, Animation, and Virtual Reality. He is coeditor of 30 books, and coauthor of several books including 'Crowd Simulation' (second edition 2012) and 'Stepping Into Virtual Reality' (2007), published by Springer. He received his PhD in Computer Science in 1977 from the University of Geneva and an Honorary Doctorate from University Paul-Sabatier in Toulouse, France, in 2003. He also received the Eurographics Distinguished Career Award in 2010, the 2012 Canadian Human Computer Communications Society Achievement Award, and the CGI 2015 Career Achievement. More details on http://en.wikipedia.org/wiki/Daniel_Thalmann