Quality Assessment of Screen Content Images via Convolutional-Neural-Network-Based Synthetic/Natural Segmentation

Yi Zhang, Damon M. Chandler, *Senior Member, IEEE*, and Xuanqin Mou, *Member, IEEE*

***Abstract*— The recent popularity of remote desktop software and live streaming of composited video has given rise to a growing number of applications that make use of the so-called *screen content images* that contain a mixture of text, graphics, and photographic imagery. Automatic quality assessment (QA) of screen-content images is necessary to enable tasks, such as quality monitoring, parameter adaptation, and other optimizations. Although QA of natural images has been heavily researched over the last several decades, the QA of screen content images is a relatively new topic. In this paper, we present a QA algorithm called convolutional neural network-based screen content image quality estimator (CNN-SQE), which operates via a fuzzy classification of screen content images into plain-text, computer-graphics/cartoons, and natural-image regions. The first two classes are considered to contain synthetic content (text/graphics), and the latter two classes are considered to contain naturalistic content (graphics/photographs), where the overlap of the classes allows the computer graphics/cartoons segments to be analyzed by both text-based and natural-image-based features. We present a CNN-based approach for the classification, an edge-structure-based quality degradation model, and a region-size-adaptive quality-fusion strategy. As we will demonstrate, the proposed CNN-SQE algorithm can achieve better/competitive performance as compared with the other state-of-the-art QA algorithms.**

***Index Terms*— Screen content image, full reference quality assessment, convolutional neural network, local entropy.**

I. Introduction

THE prevalence of modern multimedia communication devices connected to the Internet (e.g., mobile phone, tablet, laptop, smart-home displays, etc.) have enabled users to accomplish many complicated communication tasks, such as visual screen sharing, cloud computing and gaming, remote conferencing and education, online product advertising, etc. Consequently, the visual content presented to consumers is not limited to natural images, but can contain a mixture of sources which include natural images, computer-generated graphics, text, charts, maps, users' hand-writings and drawings, and even some special symbols or patterns (e.g., logos, bar codes, QR codes, etc.) [1]. Such visual content is commonly referred to as *screen content images* (SCIs).

For some types of SCIs, the compositing is performed at the receiver side, in which case many computer-generated aspects can be rendered on the receiving device, and thus only the natural imagery is subjected to distortion via transmission/compression. However, for many other types of SCIs, the compositing is performed at the sender side, and thus the entire SCI is subjected to distortion due to processing, transmission, and compression. For example, normal web pages are rendered at the receiver side, and thus it is uncommon to see e.g., compression artifacts in the text regions. However, when watching an online news broadcast, all parts of SCI typically contain artifacts. These distortions, along with other processing artifacts and possibly improper settings of the local devices, inevitably introduce quality degradation of the received SCIs, which ultimately affects the user experience. Therefore, accurate predictions of the visual qualities of SCIs at the receiver are crucial.

Although quality assessment (QA) of natural images (NIs) has been intensively studied over the last several decades (see [2] and [3] for an overview), QA of SCIs still remains extremely challenging. The main challenge arises from the fact that a screen content image may contain not only natural imagery, but also computer-generated content such as text, charts, graphics, and icons, which have significantly different properties compared to natural content. Specifically, natural images typically contain relatively smooth edges, complicated textures, and thick lines with virtually unlimited colors, whereas computer-generated content typically contains very sharp edges, relatively uncomplicated shapes, thin lines with few colors, and even single-pixel-wide single-color lines [4]. As a result, some statistical properties of SCIs also differ significantly from those of NIs. For example, as noted by Guo *et al.* [5], the local image activity measure (LIAM) [6] of NIs tends to have smaller values than of SCIs, and thus the LIAM histograms of NIs and SCIs are different. Because of these obvious distinctions, most existing IQA algorithms designed for natural images are not quite effective for QA of SCIs.

To investigate QA of SCIs, three screen content image databases have been developed (i.e., SIQAD [7], [8], SCD [9], SCID [1]), and several screen content image quality assessment (SC-IQA) algorithms have been proposed.

Manuscript received December 26, 2017; revised May 10, 2018 and June 19, 2018; accepted June 20, 2018. Date of publication June 28, 2018; date of current version July 17, 2018. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFA0202003 and in part by the National Natural Science Foundation of China under Grant 61571359. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lei Zhang. *(Corresponding author: Yi Zhang.)*

Y. Zhang and X. Mou are with the School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: yi.zhang.osu@xjtu.edu.cn).

D. M. Chandler is with the Department of Electrical and Electronic Engineering, Shizuoka University Hamamatsu, Hamamatsu 432-8561, Japan.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2851390

Among these approaches, one common idea is to measure the structure similarity (SSIM) [10] between quality-related features computed for both the reference and the distorted SCIs, and then the final quality score is taken as a weighted sum of the estimated feature-similarity map (e.g., [1] and [11]–[14]). Frequently used features for this type of approach include luminance, contrast, gradients, and some specific edge characteristics (e.g., edge contrast, edge width, and edge direction). Despite the various features employed, these methods do not perform as well as expected because they do not take into account how the human visual system (HVS) perceives different regions of SCIs. For example, it is commonly known that the visual quality of a text region is judged based mainly on its readability, regardless of the chromaticity/saturation change, while the quality of a picture region is judged based mainly on its appearance. As a result, it was reported in [8] that the visual qualities of text and picture regions are distinctively affected by different distortions. Also, it was reported in [8] that the perceptual quality of text portion has a higher correlation with the entire SCI quality as compared to the picture portion, which partly demonstrates that text is more important than picture in human subjective quality evaluation of SCIs when both portions are presented. All of these findings seem to indicate that it is crucial to perform separate analyses of different regions of SCIs, which ultimately leads to the second type of SC-IQA approach.

The second type of SC-IQA approach (e.g., [7], [8], and [15]–[17]) attempts to separately measure the quality degradations of various kinds of regions (e.g., text and picture regions) based on an algorithm-predicted segmentation map, and then systematically integrates the quality estimate of each region to infer the overall quality score. Compared with the first type of approach, these segmentation-based methods represent significant progress in mimicking the way the HVS judges SCI quality. However, major limitations exist for current segmentation models, which categorize local SCI regions as either text or non-text (i.e., picture) parts, and without any exception. One potential limitation is that such a strict two-class segmentation scheme will ignore any SCI regions that contain computer graphics, charts, maps, and cartoons, all of which share similarities with text regions in terms of sharp edges, thin strokes, and limited amount of colors. For example, as shown in Figure 1, the reference SCI contains decorative patterns on its two sides, which are apparently not text, but display text-related features. Although these decorative patterns, along with many other computer graphics/cartoons, are always labeled as non-text regions in existing SC-IQA algorithms, we believe that these regions should also be analyzed in a similar fashion as the text regions in order to better capture the perception of text-specific distortion. Another potential limitation is that the current segmentation scheme does not filter out flat/solid-color regions, while common sense suggests that these regions play little or less important roles in guiding the HVS in judging SCI quality.

To overcome these potential limitations, we present in this paper an improved segmentation-based full-reference (FR) SC-IQA algorithm, which employs a convolutional neural network (CNN) for SCI segmentation and an edge-structure similarity index for SCI quality estimation. Compared with existing methods of the second type of SC-IQA approach, our method has the following advantages/improvements:

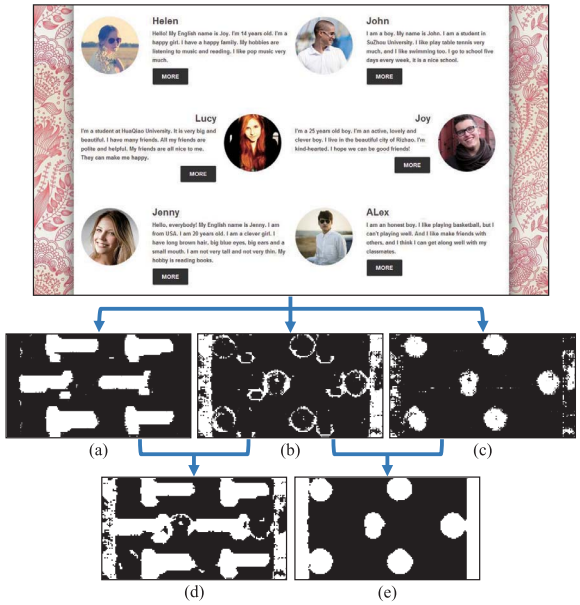


Fig. 1. Illustration of a reference SCI (*SCI03.bmp*) in the SCID database [1] being segmented into three types of regions: (a) plain text, (b) computer graphics/cartoons, and (c) natural images. The plain text and computer graphics/cartoons regions are grouped together to form the SCI synthetic region (d); the natural images and computer graphics/cartoons regions are grouped together to form the SCI natural region (e) (i.e., both the synthetic and the natural regions contain the computer graphics/cartoons area in common). Note that pixels in (a)-(c) represent the raw outputs of the CNN model, and (d)-(e) are the binary synthetic and natural segmentation maps computed by Eqs. (10) and (11).

First, to avoid the strict two-class segmentation commonly adopted in previous works, we propose to classify an SCI region into three categories: (1) plain text, (2) computer graphics/cartoons, and (3) natural images, based on which a non-strict, two-category (i.e., "synthetic" and "natural") segmentation scheme is constructed. To facilitate this synthetic-natural segmentation, the proposed CNN-based segmentation model is more accurate and robust in segmenting different SCI regions by taking advantage of the power of machine learning, while most previous models only rely on the empirically-selected thresholds functioning for some specifically-designed feature maps (e.g., local information content [16] and LIAM [6]).

As shown in Figure 1, the plain text and decorative patterns are grouped together to form the SCI synthetic region; the natural images and decorative patterns are grouped together to form the SCI natural region. Accordingly, in our proposed fuzzy segmentation scheme, both the synthetic and natural regions will contain the computer graphics/cartoons area (i.e., decorative patterns) in common. Such an overlap between the two categories is useful, because an SCI may contain no (or minimal) plain text and/or natural images, in which case the quality estimates of the plain-text and/or natural-image region are not representative of the overall SCI quality. The proposed segmentation scheme also has the advantage of being able to facilitate the training data collection, because in some cases it is difficult to distinguish between plain text and cartoon text.

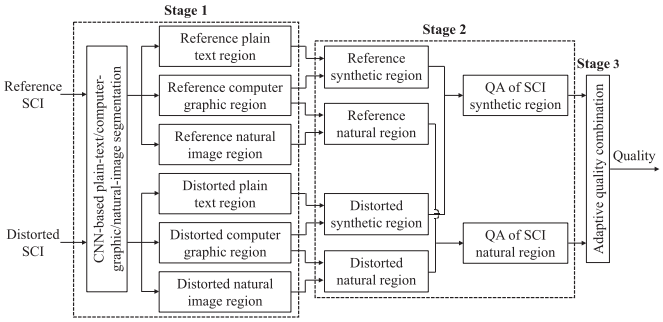


Fig. 2. Block diagram of the CNN-SQE algorithm.

Second, to prevent the large redundant flat/solid-color area in SCIs from being considered for QA, we incorporate a preprocessing stage, which selects candidate SCI regions for CNN classification. This process also accelerates the algorithm to some extent, as parts of the image regions are filtered out at the very beginning. Towards this end, we employed the local standard deviation (LSD) and local entropy features for candidate patch selection. As we will demonstrate in Section III-A, both features are effective in differentiating flat/solid-color regions in SCIs. This preprocessing stage was not seen in previous SC-IQA works.

Finally, to improve the quality predicting performance on SCIs, we propose more effective quality-related features and a quality-fusion strategy. Specifically, we propose a measure of edge-structure degradation which operates based on LoG filters and contrast maps. In the quality combination stage, we propose an adaptive quality-fusion method, which assumes that the influence of each type of region on the overall SCI quality adapts based on that region's area, as well as its importance towards the QA process. As we will demonstrate in Section IV, these improvements can yield better/competitive quality predicting performance as compared to existing NI and SCI QA algorithms.

Our method, called CNN-based Screen content image Quality Estimator (CNN-SQE), employs three main stages to assess SCI quality, as shown in Figure 2. The first stage performs image segmentation. In this stage, an eight-layer CNN model is used to soft-classify the candidate image patch into one of the three categories: (1) plain text, (2) computer graphics and cartoons, and (3) natural images. Then, the patches belonging to the first and second categories will be grouped together to form the *synthetic* region, and patches belonging to the second and third categories will be grouped together to form the *natural* region. The second stage performs QA of each segmented region by using our measure of edge-structure degradation. Finally, the quality estimates obtained from the two regions are adaptively combined to yield the overall SCI quality estimate.

This paper is organized as follows. Section II reviews existing approaches for FR SC-IQA. Section III describes details of the proposed CNN-SQE algorithm. In Section IV, we analyze and discuss the performance of CNN-SQE on various screen content image quality databases. General conclusions are presented in Section V.

II. Previous Work

In this section, we provide a brief review of current FR SC-IQA algorithms. As mentioned in Section I, these previous techniques can be roughly classified into two categories based on whether or not different regions of SCIs are taken into account for quality estimation: (1) feature-based approaches, and (2) segmentation-based approaches.

A. Feature-Based SC-IQA Approach

As stated in Section I, feature-based SC-IQA approaches operate by extracting various quality-related features from both the reference and distorted SCIs, and then outputting the final quality score by collapsing the estimated feature-similarity map.

For instance, Gu *et al.* [11] proposed an SC-IQA method called SIQM, which combines SSIM with a measure of structural degradation. Specifically, it is argued in [11] that SSIM can be effective for SCIs if the normal spatial pooling is adjusted to take into account the different types of SCI regions. To this end, SIQM modifies the SSIM pooling stage to use spatial weights derived from a measure of structural degradation (defined as the SSIM between the reference image and a blurred version of the reference image). Gu *et al.* [12] proposed another SC-IQA method called SQMS, which employs efficient convolution operations to produce the gradient magnitude structural similarity map, and then collapses the map into a quality score by using the visual saliency estimates.

Ni *et al.* [13] proposed an SC-IQA method called GSS, which operates by using the gradient direction similarity and gradient magnitude similarity between the reference and distorted SCIs. Then, a deviation-based pooling strategy is utilized to yield the overall quality estimate by combing the two similarity maps. Ni *et al.* [14] proposed another SC-IQA method called EMSQA, which operates based on changes in edge contrast and edge width. The method computes two types of maps from the reference and distorted images: (1) an edge-contrast map, and (2) an edge-width map. Changes in these maps for the reference vs. distorted SCIs are then quantified via edge-contrast similarity and edge-width similarity maps. These latter two maps are combined into a final map via a weighted product, and then the final map is pooled over space in a weighted fashion by using weights based on the edge-width map. Ni *et al.* [1] combined the approaches in [13] and [14] to build a more efficient SC-IQA model called ESIM, which operates based on edge contrast, edge width, and edge direction information. Similar edge-contrast, edge-width, and edge direction similarity maps are computed and pooled (following [14]) to yield the final quality estimates of SCIs.

B. Segmentation-Based SC-IQA Approach

Segmentation-based SC-IQA approaches separately measure the quality degradations corresponding to different SCI regions by using different quality-related features and/or feature similarity maps. The overall quality is a combination of the various quality scores estimated from the different regions.

For instance, Yang *et al.* [8] proposed the first SC-IQA method called SPQA, which generates a segmentation map of the textual and pictorial regions from the reference SCI based on the algorithm proposed in [6]. Next, the quality map for the pictorial regions is generated based on a measure of sharpness similarity, and the quality map for the textual regions is generated based on measures of sharpness similarity and luminance similarity. Finally, the overall quality score is computed by combining and pooling the quality maps over space via an activity-map-based pooling scheme.

Wang *et al.* [16] proposed an SC-IQA method called SQI, which employs an information content model [18] to classify local 4×4 image blocks into texture or pictorial regions via an information content threshold. Next, SQI operates based on two key features: (1) the fact that the power spectrum of textual images differs from the characteristic $1/f^2$ power spectrum of natural scenes; and (2) the fact that different regions of SCIs convey different amounts of information, and thus a measure of such information should be used to guide the QA process. Consequently, SQI employs two separate SSIM-based quality estimates to capture the aforementioned properties, and then the two measures are combined into the the overall quality index.

Fang *et al.* [17] proposed a full-reference SC-IQA method called SFUW, which operates based on structural features and uncertainty weighting. In this method, the reference and distorted images are divided into textual and pictorial regions based on the algorithm proposed in [19]. For the textual regions, a gradient-based structural similarity measure (for reference vs. distorted) is used to estimate the quality. For the pictorial regions, a separate similarity measure that uses structural and luminance features is used to estimate the quality. Finally, the overall quality is estimated via an uncertainty weighting fusion method.

In the following section, we describe the proposed CNN-SQE algorithm which is also considered as a segmentation-based approach. As we will describe, the key contribution of CNN-SQE over prior works is that a more powerful machine-learning-based approach is developed, which allows a finer classification of the SCI into plain-text, computer graphics/cartoons, and natural-image regions. Based on these three classes, we are able to segment an SCI into synthetic and natural regions, in which computer graphics/cartoons are considered to fall into both of these categories. In the following section, we present our CNN-based approach for the classification, an edge-structure-based quality degradation model, and a region-size-adaptive quality-fusion strategy.

III. Algorithm

The proposed CNN-SQE algorithm operates by first dividing an input SCI into plain-text, computer-graphic/cartoon, and natural-image regions, from which the synthetic/natural segmentation maps are constructed. Then, the visual quality of both synthetic and natural regions are estimated by gradient-weighted edge-structure similarity maps computed from the reference and distorted SCIs. Finally, the two quality estimates obtained from the two regions are adaptively combined to yield the overall SCI quality score. In the following subsections, we first describe our CNN-based SCI segmentation model, and then provide details for the SCI quality estimation and combination stages.

A. CNN-Based SCI Segmentation

The proposed CNN segmentation model takes as input a locally normalized 48×48 image patch, and outputs three probabilities, which correspond to the three categories. This is achieved by learning the CNN model from training data. Then, the synthetic and natural segmentation maps are constructed by adding and thresholding the corresponding probability values. As we mentioned in Section I, candidate patches are first selected before CNN classification such that the flat/solid-color regions are removed.

1) Image Preprocessing: The image preprocessing stage includes local normalization and candidate patch selection. Specifically, given a gray scale image $I(i, j)$, we first compute the locally normalized pixel values via local mean subtraction and divisive normalization [20] defined as:

$$\hat{I}(i, j) = \frac{I(i, j) - \mu(i, j)}{\sigma(i, j) + C}, \tag{1}$$

where

$$\mu(i, j) = \sum_{k=-K}^{K} \sum_{l=-L}^{L} \omega_{k,l} I_{k,l}(i, j), \tag{2}$$

and

$$\sigma(i, j) = \sqrt{\sum_{k=-K}^{K} \sum_{l=-L}^{L} \omega_{k,l} (I_{k,l}(i, j) - \mu(i, j))^2}. \tag{3}$$

Here, $\omega = \{\omega_{k,l} | k = -K, \cdots, K, l = -L, \cdots, L\}$ is a 2D circularly-symmetric Gaussian weighting function sampled out to three standard deviations and rescaled to unit volume; $I_{k,l}(i, j) = I(i+k, j+l)$ denotes the local image pixel value. K and L represent the normalization window sizes; and $C = 1$ is a constant that prevents division by zero. As in [21], we also define $K = L = 3$.

Next, the candidate locally normalized synthetic and natural patches are selected based on the local standard deviation (LSD) and local entropy, respectively. In our work, the LSD is computed by using Eq. (3), and the local entropy is computed for each 16×16 image patch (with eight pixels overlap) defined as:

$$E = -\sum_{i=0}^{N_p} p_i \cdot \log_2 p_i, \tag{4}$$

where N_p is the maximum pixel value in the luminance patch; and p_i denotes the probability that the pixel value equals i in the patch. As shown in Figure 3, the LSD feature is more effective at highlighting plain text, whereas the local entropy feature is more effective at highlighting natural images. We select only those patches that have suprathreshold averaged LSD or local entropy values as the to-be-classified input patches to the CNN. Here, we empirically select the threshold as

$$T = \theta_1 \cdot \delta_{max}, \tag{5}$$

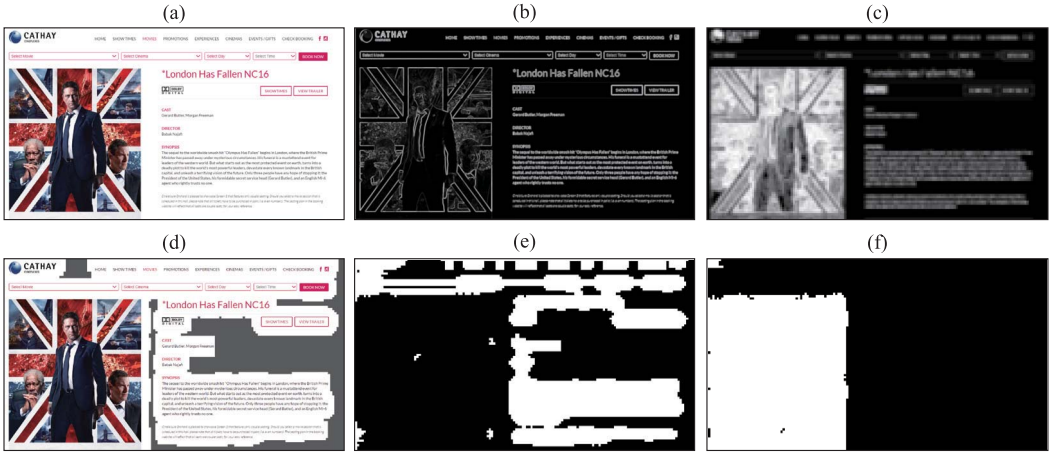


Fig. 3. A reference SCI from the SIQAD database [7], [8], and its computed LSD map, local entropy map, selected region (for CNN classification), and CNN-based synthetic/natural segmentation maps (denoted by $\mathbb{L}_{syn}$ and $\mathbb{L}_{nat}$, respectively). Observe that the LSD feature is more effective at highlighting plain text, whereas the local entropy feature is more effective at highlighting natural images. As a result, large redundant flat/solid-color areas in the reference image (marked by gray color in (d)) are filtered out before CNN classification. Note that in the $\mathbb{L}_{syn}$ map, a value of 1 indicates a synthetic region, and in $\mathbb{L}_{nat}$ map, a value of 1 indicates a natural region. (a) Reference SCI. (b) LSD. (c) Local entropy. (d) Region for classification. (e) $\mathbb{L}_{syn}$. (f) $\mathbb{L}_{nat}$.

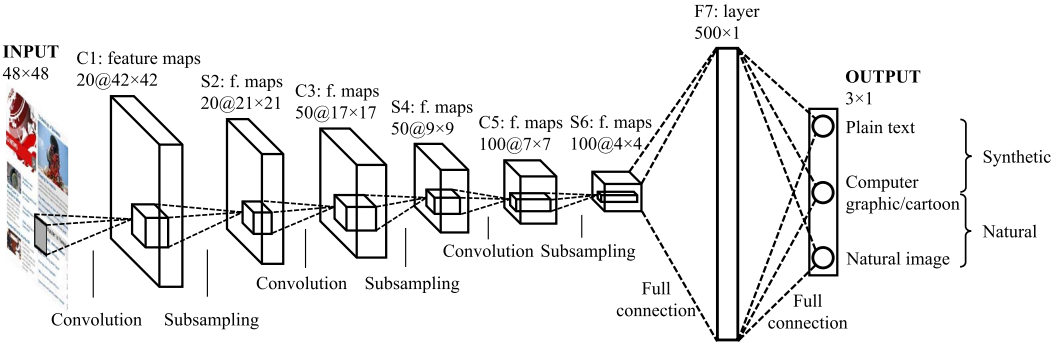


Fig. 4. An illustration of the architecture of our CNN model.

where $\theta_1 = 0.25$; δ_{max} denotes the maximum LSD/local entropy values. This scheme ensures that flat/solid-color regions are not used.

2) Network Architecture: The architecture of our proposed CNN model for SCI segmentation is shown in Figure 4, from which it can be observed that the network consists of eight layers, not counting the input: three convolutional layers (denoted by Cx), three sub-sampling layers (denoted by Sx), one fully-connected layer (denoted by Fx), and one softmax classification layer (here, "x" denotes the layer index). The dimensions of the proposed CNN model are as follows: $48 \times 48 \Rightarrow 42 \times 42 \times 20 \Rightarrow 21 \times 21 \times 20 \Rightarrow 17 \times 17 \times 50 \Rightarrow 9 \times 9 \times 50 \Rightarrow 7 \times 7 \times 100 \Rightarrow 4 \times 4 \times 100 \Rightarrow 500 \Rightarrow 3$. Note that all layers except the sub-sampling layers contain trainable parameters (weights and bias).

The three convolutional layers can be expressed as

$$F_i(X) = W_i * X + B_i, i \in \{1, 2, 3\}, \tag{6}$$

where W_i and B_i represent the filter weights and biases of the i^{th} layer, respectively; F_i is the i^{th} output feature maps; and "$*$" denotes the convolution operation. W_i contains n_i filters of support $f_i \times f_i \times n_{i-1}$, where f_i is the spatial support of a filter and n_i is the number of filters. In our work, we use the following parameters: $f_1 = 7$, $f_2 = 5$, $f_3 = 3$, $n_1 = 20$, $n_2 = 50$, and $n_3 = 100$. Thus, the number of trainable parameters for the three convolutional layers are 1000, 25050, and 45100 respectively.

In our proposed CNN architecture, each convolutional layer is followed by a sub-sampling layer, wherein the pooling operation is applied on each feature map to reduce the filter responses to a lower dimension, and then the results are passed through a nonlinear activation function. Specifically, each non-overlapping 2×2 receptive fields in the feature map is pooled into one max value, and the Rectified Linear Units [22] is employed for nonlinear activation due to its effectiveness and efficiency [23]. Note that this pooling strategy is particularly helpful for object classification/recognition, as objects can typically be modeled as multiple parts organized in a certain spatial order. Thus, the feature maps in the sub-sampling layers have approximately half the number of rows and columns as the corresponding convolutional layers.

The fully-connected layer F7 contains 500 units, each of which is connected to a 4×4 neighborhood on all 100 of S6's feature maps. Hence, layer F7 has 800500 (4×4×100 feature maps × 500 units + 500 bias) trainable parameters. As in classical neural networks, units in layers up to F7 compute a dot product between their input vector and their weight vector to which a bias is added. In the softmax classification layer, the weighted sum, which is computed from the 500 units in layer F7 and denoted by z_k for unit k, is then passed through a

softmax function to produce the state of unit k, denoted by σ_k:

$$\sigma_k = \frac{\exp(z_k)}{\sum_{j=1}^{m} \exp(z_k)}, k = 1, 2, \cdots, m \tag{7}$$

where $m = 3$ denotes the three categories. Consequently, the network has three outputs, each of which represents the corresponding classification probability. The number of trainable parameters in softmax classification layer is 1503 (500 units $\times$ 3 weights/unit + 3 bias).

3) Network Training and SCI Segmentation: The training data consists of locally normalized image patches of size 48×48 pixels taken from 600 SCIs (with 16 pixels overlap). In order to simplify the ground-truth labeling of each patch, these 600 SCIs were carefully selected such that each SCI contains only one of the three categories. Specifically, among these 600 SCIs, 200 images contain the plain text only; 200 images contain the computer graphics and cartoons only; and the remaining 200 images contain the natural scenes only. As mentioned previously, test SCIs are first preprocessed to remove flat/solid-color regions before being fed into CNN for classification. Thus, when building the training data, only those patches containing sharp regions and/or apparent textures/edges were selected as the training data. Consequently, our training data consisted of 394,480 image patches in total.

We trained our model on Caffe [24] using stochastic gradient descent with a batch size of 128 examples, a momentum of 0.9, and a weight decay of 0.0005. The update rule for weight w was

$$v_{i+1} = 0.9 \cdot v_i - 0.0005 \cdot \epsilon \cdot w_i - \epsilon \cdot \left\langle \frac{\partial L}{\partial w}|_{w_i} \right\rangle_{D_i} \tag{8}$$

$$w_{i+1} = w_i + v_{i+1}, \tag{9}$$

where i is the iteration index; v is the momentum variable; ϵ is the learning rate; and $\left\langle \frac{\partial L}{\partial w}|_{w_i} \right\rangle_{D_i}$ is the average over the i^{th} batch D_i of the derivative of the objective with respect to w, evaluated at w_i. We found that a small amount of weight decay can benefit the learning process by reducing the model's training error.

We initialized the weights in each layer from a zero-mean Gaussian distribution with a standard deviation of 0.001 for layer C1, 0.01 for layer C3 and C5, and 0.1 for layer F7. We initialized the neuron biases in all layers to zero, and set the learning rate as a small constant value of 10^{-4} in the training phase. The network was trained one million times on an NVIDIA GeForce GTX 960 GPU which took approximately six hours. Figure 5 shows the 20 convolutional kernels of size 7×7 learned by the first convolutional layer on the 48×48 input patches, and we observe that most of these kernels present obvious structures related to the frequency and orientation selection. Figure 6 shows the feature maps output by the three convolutional layers after the trained model was applied on a sample image patch. As observed, lower convolutional layers extract simple low-level features such as edges, and higher convolutional layers extract more complex and abstract high-level features.

With the trained CNN model, input 48×48 patches (40 pixels overlap when taken from the reference SCI) are soft-classified into one of the three categories: plain text, computer graphics/cartoons, and natural images. Let p_i ($i = 1, 2, 3$) denotes the probability that a candidate patch falls into the i^{th} category. Then, the synthetic and natural segmented labels of that patch (denoted by L_{syn} and L_{nat}, respectively) are given by

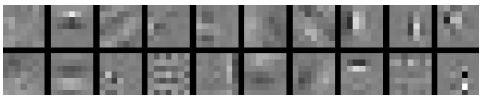


Fig. 5. Visualization of the 20 convolutional kernels of size 7×7 learned by the first convolutional layer (C1 in Figure 4) on the 48×48 input patches.

$$L_{syn} = \begin{cases} 1 & (p_1 + p_2 > \theta_2) \& (\delta_{LSD} > T_{LSD}) \\ 0 & otherwise \end{cases} \tag{10}$$

$$L_{nat} = \begin{cases} 1 & (p_2 + p_3 > \theta_2) \& (\delta_{ENT} > T_{ENT}) \\ 0 & otherwise \end{cases} \tag{11}$$

where $\theta_2 = 0.75$; δ_{LSD} and δ_{ENT} denote, respectively, the average LSD and local entropy values of the patch; T_{LSD} and T_{ENT} denote the corresponding thresholds computed through Eq. (5). For those patches that are not input into the CNN, their segmented labels are set to be zero. Finally, we form the synthetic and natural segmentation maps (denoted by $\mathbb{L}_{syn}$ and $\mathbb{L}_{nat}$, respectively), both of which will be used for the subsequent QA stage. Sample synthetic and natural segmentation maps are shown in Figure 3. More discussions about the thresholds (θ_1 and θ_2) and the segmentation results are provided at http://vision.eng.shizuoka.ac.jp/CNNSQE/.

Compared to using the three regions (plain text, computer graphics/cartoons, and natural images) directly for QA, the proposed non-strict, two-category (synthetic and natural) segmentation scheme has the following advantages. First, the two-category segmentation scheme is more tolerant to classification errors, meaning that the QA performance can be decent even if the preliminary three-region segmentation is not perfect. Such a property is important, as classification errors can occur frequently for SCIs. For example, as shown in Figure 7, the computer graphics/cartoons region on the top-right corner of image "cim5.bmp" is misclassified as a natural image due to its relatively heavy textured painting. However, the two-category segmentation maps are still reasonable and not affected much by the error. Second, the two-category segmentation scheme is more tolerant to different SCI contents. Such a property is also important, because an SCI may contain no (or minimal) plain text and/or natural images, in which case the quality estimates of only the plain-text and/or natural-image region are not representative of the overall SCI quality. For example, as shown in Figure 7, there is no natural-image region in image "cim10.bmp". However, the *natural* part of "cim10.bmp" still contains a large area (and thus sufficient distortion information) for a potentially good quality estimate.

B. Quality Assessment of SCIs

Traditionally, when designing an IQA algorithm, a single, unified approach is often sought. However, as mentioned in Section I, the HVS tends to focus on different aspects when

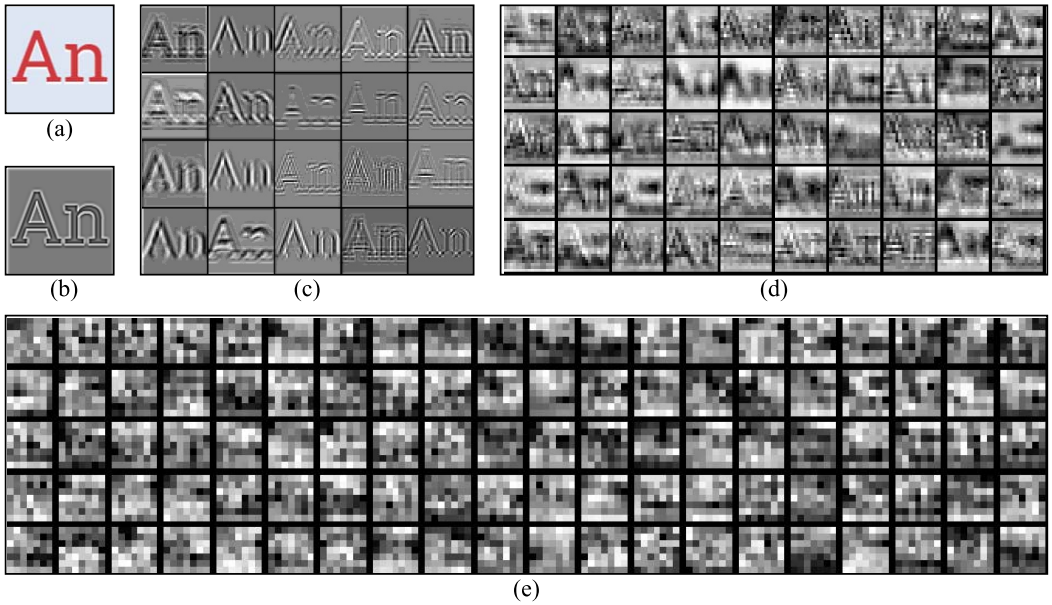


Fig. 6. Visualization of feature maps output by the three convolutional layers (C1, C3, and C5 in Figure 4) after the trained model was applied on a sample image patch. (a) original image patch; (b) image patch after local normalization using Eq. (1); (c)-(e) feature maps output by the three convolutional layers, respectively. Note that lower convolutional layers extract simple low-level features such as edges, and higher convolutional layers extract more complex and abstract high-level features.

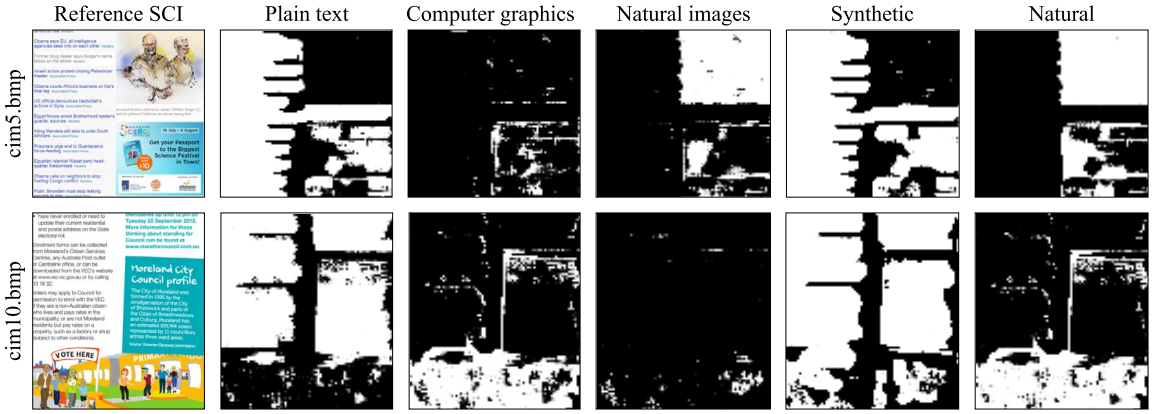


Fig. 7. Two images from the SIQAD database [7], [8], and their corresponding CNN-based segmentation maps. Note that although the computer graphics region on the top-right corner of image "cim5.bmp" is misclassified as natural images, and image "cim10.bmp" does not contain natural-image regions, the two-category segmentation maps for the two SCIs are still effective for potentially good quality estimates.

judging the qualities of different types of SCI regions. In our algorithm, we use a common set of features which are effective for QA, but which have been customized and combined in SCI-region-specific fashions in an attempt to capture, in part, the different HVS analyses. Figure 8 shows a block diagram of this stage, and we provide the algorithm details in the following subsections.

1) QA of SCI Synthetic Region: The synthetic regions of an SCI can be fairly well represented as a collection of relatively sharp edges. Thus, we first use LoG filter applied on contrast maps to extract the edge information of the reference and distorted SCIs. Then, we quantify the edge-structure degradation by a similarity measure computed on the local luminance distance (LLD) of the edge map. To allow the text region quality degradation to better represent the overall SCI quality degradation, an averaged gradient similarity measure computed on the whole SCI is incorporated as a factor to augment the edge-structure similarity map. Finally, the quality estimate of the SCI synthetic region is computed via a weighted sum of the augmented edge-structure similarity map.

Specifically, given an image $\mathbf{I}$, the gradient magnitude is computed by

$$\mathbf{G} = \sqrt{\mathbf{I}_x^2 + \mathbf{I}_y^2}, \tag{12}$$

where

$$\mathbf{I}_x = \mathbf{h}_x * \mathbf{I} = \frac{1}{16}\begin{bmatrix} +3 & 0 & -3 \\ +10 & 0 & -10 \\ +3 & 0 & -3 \end{bmatrix} * \mathbf{I} \tag{13}$$

$$\mathbf{I}_y = \mathbf{h}_y * \mathbf{I} = \frac{1}{16}\begin{bmatrix} +3 & +10 & +3 \\ 0 & 0 & 0 \\ -3 & -10 & -3 \end{bmatrix} * \mathbf{I}. \tag{14}$$

Here, $\mathbf{I}_x$ and $\mathbf{I}_y$ denote the gradient features along the x- and y-axes; "$*$" denotes the convolution operation. Then, the gradient similarity map of the reference and distorted SCI can be

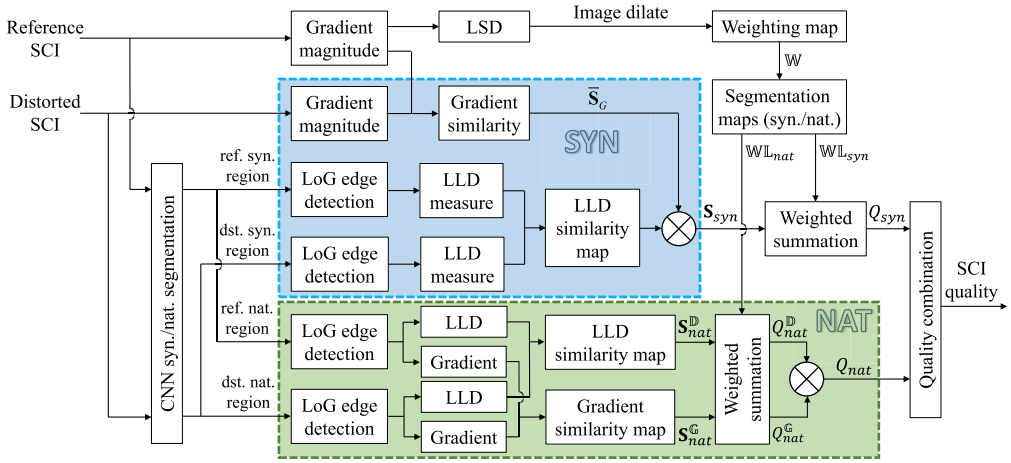


Fig. 8. Block diagram of the quality assessment stage in the proposed CNN-SQE model.

measured by

$$\mathbf{S}_G = \frac{2\mathbf{G}_r \cdot \mathbf{G}_d + C_1}{\mathbf{G}_r^2 + \mathbf{G}_d^2 + C_1}, \tag{15}$$

where $\mathbf{G}_r$ and $\mathbf{G}_d$ denote the gradient magnitude maps of the reference and distorted SCIs, respectively; and $C_1 = 250$ is a constant value. Note that the mean value of $\mathbf{S}_G$ will be later used to augment the edge-structure similarity map, and $\mathbf{G}_r$ will be later used to generate the weighting map.

Our next step is to estimate the quality degradation on the synthetic regions. Towards this end, we first apply a LoG filter on the contrast map to extract the edge features from the SCI synthetic regions. Let $\mathbb{R}_r^{syn}$ and $\mathbb{R}_d^{syn}$ denote the synthetic regions of the reference and distorted SCIs, and their corresponding contrast maps (denoted by $\mathbb{C}_r^{syn}$ and $\mathbb{C}_d^{syn}$, respectively) are given by

$$\mathbb{C}_r^{syn} = \mathbb{R}_r^{syn} - \bar{\mathbb{R}}_r^{syn} \tag{16}$$

$$\mathbb{C}_d^{syn} = \mathbb{R}_d^{syn} - \bar{\mathbb{R}}_r^{syn}, \tag{17}$$

where $\bar{\mathbb{R}}_r^{syn} = \mathbb{R}_r^{syn} * \mathbf{h}_l$, and $\mathbf{h}_l$ denotes a circular averaging filter within the square matrix of size $(2l+1)^2$ pixels. The rotationally symmetric LoG filter is defined as

$$\begin{aligned} LoG &= \frac{\partial^2}{\partial x^2} G_\sigma(x,y) + \frac{\partial^2}{\partial y^2} G_\sigma(x,y) \\ &= \frac{x^2+y^2-2\sigma^2}{\sigma^4} e^{-(x^2+y^2)/2\sigma^2}, \end{aligned} \tag{18}$$

where $G_\sigma(x,y) = e^{-(x^2+y^2)/2\sigma^2}$ is a normalized Gaussian function with standard deviation $\sigma = 1.35$. Then, we compute the edge features of the reference and distorted SCIs' synthetic regions (denoted by $\mathbb{E}_r^{syn}$ and $\mathbb{E}_d^{syn}$, respectively) via

$$\mathbb{E}_r^{syn} = \mathbb{C}_r^{syn} * LoG \tag{19}$$

$$\mathbb{E}_d^{syn} = \mathbb{C}_d^{syn} * LoG. \tag{20}$$

To characterize the edge-structure change caused by distortion, we further extract the LLD feature from the obtained edge feature maps via

$$\mathbb{D}_r^{syn} = \left| \mathbb{E}_r^{syn} - \bar{\mathbb{E}}_r^{syn} \right| \tag{21}$$

$$\mathbb{D}_d^{syn} = \left| \mathbb{E}_d^{syn} - \bar{\mathbb{E}}_d^{syn} \right|, \tag{22}$$

where $\bar{\mathbb{E}}_r^{syn}$ and $\bar{\mathbb{E}}_d^{syn}$ denote the local averaged value of the reference and distorted edge feature maps computed by

$$\bar{\mathbb{E}}_r^{syn} = \mathbb{E}_r^{syn} * \mathbf{h}_l \tag{23}$$

$$\bar{\mathbb{E}}_d^{syn} = \mathbb{E}_d^{syn} * \mathbf{h}_l. \tag{24}$$

Then, the edge-structure similarity map of the synthetic region (denoted by $\mathbf{S}_{syn}$) is computed by

$$\mathbf{S}_{syn} = \bar{\mathbf{S}}_G \cdot \frac{2\sigma_{\mathbb{D}_r^{syn}\mathbb{D}_d^{syn}} + C_2}{\sigma^2_{\mathbb{D}_r^{syn}} + \sigma^2_{\mathbb{D}_d^{syn}} + C_2}, \tag{25}$$

where $\bar{\mathbf{S}}_G$ is the mean value of $\mathbf{S}_G$; $\sigma_{\mathbb{D}_r^{syn}}$ and $\sigma_{\mathbb{D}_d^{syn}}$ denote the local standard deviation of the LLD feature maps $\mathbb{D}_r^{syn}$ and $\mathbb{D}_d^{syn}$; $\sigma_{\mathbb{D}_r^{syn}\mathbb{D}_d^{syn}}$ is the covariance of $\mathbb{D}_r^{syn}$ and $\mathbb{D}_d^{syn}$; and C_2 is a small constant value. Sample $\mathbf{S}_{syn}$ maps for the seven distorted versions of one reference image (*cim19.bmp*) in the SIQAD database [7], [8] are shown in Figure 9.

Note that $\sigma_{\mathbb{D}_r^{syn}}$, $\sigma_{\mathbb{D}_d^{syn}}$, and $\sigma_{\mathbb{D}_r^{syn}\mathbb{D}_d^{syn}}$ are computed in the same way as in [10]; the only differences are the window size (7×7) and standard deviation (0.5 samples) of the Gaussian weighting function. Also note that we use large-size filters when characterizing edges within local image areas, but use small-size filters to model the edge structure due to the thin lines featured by the text. In our implementation, the LoG and circular averaging filters employed for edge detection are 11×11 pixels in size ($l = 5$), and the circular averaging filters employed for estimating $\bar{\mathbb{E}}_r^{syn}$ and $\bar{\mathbb{E}}_d^{syn}$ are 7×7 pixels in size ($l = 3$).

The final step is to weight and sum the edge-structure similarity map $\mathbf{S}_{syn}$ to yield the synthetic-region quality estimate. Towards this end, we first compute the weighting map $\mathbb{W}$ from the LSD of $\mathbf{G}_r$, which is computed the same way as in Eq.(3). Motivated by [12] that humans identify information through the fixation and saccade behaviors of eyes, the obtained LSD map is subjected to morphological dilations with both the "disk" and "plus-sign" shaped structural elements, and the weighting map is the average of the two dilated LSD of $\mathbf{G}_r$ maps. Then, the quality degradation of synthetic region

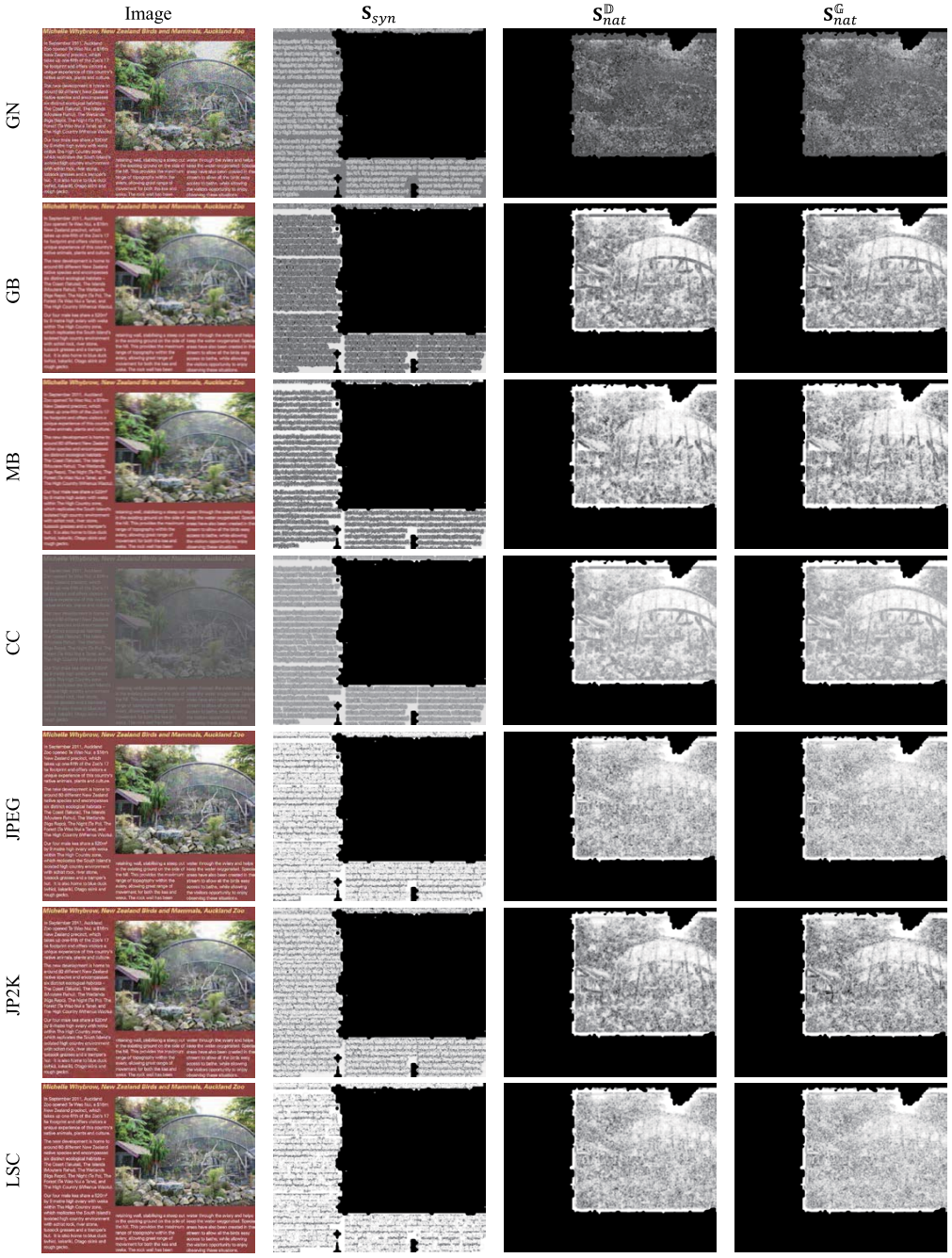


Fig. 9. Sample $\mathbf{S}_{syn}$, $\mathbf{S}_{nat}^{\mathbb{D}}$, and $\mathbf{S}_{nat}^{\mathbb{G}}$ maps for the seven distorted versions of one reference image (*cim19.bmp*) in the SIQAD database [7], [8]. The first column shows the seven distorted versions of *cim19.bmp*: Gaussian noise (GN), Gaussian blurring (GB), motion blurring (MB), contrast change (CC), JPEG compression, JPEG2000 compression (JP2K), and Layer-segmentation based compression (LSC). The second to fourth columns show respectively the corresponding $\mathbf{S}_{syn}$, $\mathbf{S}_{nat}^{\mathbb{D}}$, and $\mathbf{S}_{nat}^{\mathbb{G}}$ maps, where brighter pixels indicate higher similarity with the reference image.

(denoted by Q_{syn}) is given by

$$Q_{syn} = \frac{\sqrt{\sum_k \left[1 - \mathbf{S}_{syn}(k)\right]^2 \mathbb{W}(k)\mathbb{L}_{syn}(k)}}{\sqrt{\sum_k \mathbb{W}(k)\mathbb{L}_{syn}(k)}}, \tag{26}$$

where k is a spatial index.

2) QA of SCI Natural Region: As mentioned previously, the natural region of an SCI differs significantly from its synthetic region in color, texture, and line properties; thus, we model the edge-structure degradation in natural regions by using slightly different approaches. Specifically, given the natural regions of the reference and distorted SCIs (denoted by $\mathbb{R}_r^{nat}$ and $\mathbb{R}_d^{nat}$, respectively), the first step is to compute the contrast maps given by

$$\mathbb{C}_r^{nat} = \left(\mathbb{R}_r^{nat} + C_3\right) / \left(\bar{\mathbb{R}}_r^{nat} + C_3\right) \tag{27}$$

$$\mathbb{C}_d^{nat} = \left(\mathbb{R}_d^{nat} + C_3\right) / \left(\bar{\mathbb{R}}_r^{nat} + C_3\right), \tag{28}$$

where $\bar{\mathbb{R}}_r^{nat} = \mathbb{R}_r^{nat} * \mathbf{h}_l$ represents the local averaged value of $\mathbb{R}_r^{nat}$; and $C_3 = 80$ is a constant value. The edge features of the reference and distorted SCIs' natural regions (denoted by $\mathbb{E}_r^{nat}$ and $\mathbb{E}_d^{nat}$, respectively) are then computed by

$$\mathbb{E}_r^{nat} = \mathbb{C}_r^{nat} * LoG \tag{29}$$

$$\mathbb{E}_d^{nat} = \mathbb{C}_d^{nat} * LoG, \tag{30}$$

where the LoG filter was previously defined in Eq. (18) with the same kernel size (11×11; $l = 5$) but a different standard deviation ($\sigma = 0.9$).

Compared with the edge feature extraction in synthetic regions [Eqs. (16)-(20)], one may notice that different equations [Eqs. (27)-(28)] are applied to compute the contrast maps and a different standard deviation ($\sigma = 0.9$) of the LoG filter is employed to compute the edge feature in natural regions. This is due to the fact that in natural regions (especially natural images), the central pixel often has a high correlation with its surrounding pixels, while in synthetic regions (especially plain text), such high correlations do not exist. Specifically, in synthetic regions, the central pixel value often differs considerably from its local background (due to the abrupt changes on sharp edges), in which case a *subtraction* operation is more suitable for computing the contrast, because a *division* operation will produce very large or very small contrast values. In comparison, the difference between neighboring pixel values in natural regions is often very minor, and thus a *division* operation is more suitable than a *subtraction* operation in strengthening such differences. Similarly, for the two different LoG filters applied on the contrast maps of the two different regions, we found that a smaller σ value has to be used for computing the edge feature in natural regions so that the tiny difference between neighboring pixel values is enlarged. Additional results tested with different σ values for the LoG filters are provided at http://vision.eng.shizuoka.ac.jp/CNNSQE/.

Next, similar to Section III-B1, we characterize the edge-structure change in natural regions by utilizing the LLD and gradient features computed from the edge feature map. Specifically, the LLD features (denoted by $\mathbb{D}_r^{nat}$ and $\mathbb{D}_d^{nat}$, respectively) are computed the same way as in Section III-B1, and the only difference is the circular averaging filter size (15×15; $l = 7$). The gradient features (denoted by $\mathbb{G}_r^{nat}$ and $\mathbb{G}_d^{nat}$ for the reference and distorted SCIs, respectively) are computed by

$$\mathbb{G}_r^{nat} = \left|\mathbf{E}_{rx}^{nat}\right| + \left|\mathbf{E}_{ry}^{nat}\right| \tag{31}$$

$$\mathbb{G}_d^{nat} = \left|\mathbf{E}_{dx}^{nat}\right| + \left|\mathbf{E}_{dy}^{nat}\right|, \tag{32}$$

where

$$\mathbf{E}_{rx}^{nat} = \mathbf{h}_x * \mathbb{E}_r^{nat} = \begin{bmatrix}-1/2 & 0 & 1/2\end{bmatrix} * \mathbb{E}_r^{nat} \tag{33}$$

$$\mathbf{E}_{ry}^{nat} = \mathbf{h}_y * \mathbb{E}_r^{nat} = \begin{bmatrix}-1/2 & 0 & 1/2\end{bmatrix}' * \mathbb{E}_r^{nat} \tag{34}$$

$$\mathbf{E}_{dx}^{nat} = \mathbf{h}_x * \mathbb{E}_d^{nat} = \begin{bmatrix}-1/2 & 0 & 1/2\end{bmatrix} * \mathbb{E}_d^{nat} \tag{35}$$

$$\mathbf{E}_{dy}^{nat} = \mathbf{h}_y * \mathbb{E}_d^{nat} = \begin{bmatrix}-1/2 & 0 & 1/2\end{bmatrix}' * \mathbb{E}_d^{nat}. \tag{36}$$

Next, the edge-structure similarity maps corresponding to the two features in the natural region are computed by

$$\mathbf{S}_{nat}^{\mathbb{D}} = \frac{2\sigma_{\mathbb{D}_r^{nat}\mathbb{D}_d^{nat}} + C_4}{\sigma_{\mathbb{D}_r^{nat}}^2 + \sigma_{\mathbb{D}_d^{nat}}^2 + C_4} \tag{37}$$

$$\mathbf{S}_{nat}^{\mathbb{G}} = \frac{2\sigma_{\mathbb{G}_r^{nat}\mathbb{G}_d^{nat}} + C_4}{\sigma_{\mathbb{G}_r^{nat}}^2 + \sigma_{\mathbb{G}_d^{nat}}^2 + C_4}, \tag{38}$$

where $\sigma_{\mathbb{D}_r^{nat}}$, $\sigma_{\mathbb{D}_d^{nat}}$, $\sigma_{\mathbb{G}_r^{nat}}$, and $\sigma_{\mathbb{G}_d^{nat}}$ denote the local standard deviation of the LLD and gradient feature maps; $\sigma_{\mathbb{D}_r^{nat}\mathbb{D}_d^{nat}}$ and $\sigma_{\mathbb{G}_r^{nat}\mathbb{G}_d^{nat}}$ are the covariance of corresponding feature maps computed between reference and distorted SCIs; and C_4 is a small constant value. Sample $\mathbf{S}_{nat}^{\mathbb{D}}$ and $\mathbf{S}_{nat}^{\mathbb{G}}$ maps for the seven distorted versions of one reference image (*cim19.bmp*) in the SIQAD database [7], [8] are shown in Figure 9.

The final step is to weight and sum the two edge-structure similarity maps $\mathbf{S}_{nat}^{\mathbb{D}}$ and $\mathbf{S}_{nat}^{\mathbb{G}}$ to two scalars, which are then combined to yield the natural-region quality estimate. Here, we use the same weighting map $\mathbb{W}$ computed in Section III-B1. Therefore, the quality degradation of natural region (denoted by Q_{nat}) is given by

$$Q_{nat} = \sqrt{Q_{nat}^{\mathbb{D}} \times Q_{nat}^{\mathbb{G}}}, \tag{39}$$

where

$$Q_{nat}^{\mathbb{D}} = \frac{\sqrt{\sum_k \left[1 - \mathbf{S}_{nat}^{\mathbb{D}}(k)\right]^2 \mathbb{W}(k)\mathbb{L}_{nat}(k)}}{\sqrt{\sum_k \mathbb{W}(k)\mathbb{L}_{nat}(k)}}, \tag{40}$$

$$Q_{nat}^{\mathbb{G}} = \frac{\sqrt{\sum_k \left[1 - \mathbf{S}_{nat}^{\mathbb{G}}(k)\right]^2 \mathbb{W}(k)\mathbb{L}_{nat}(k)}}{\sqrt{\sum_k \mathbb{W}(k)\mathbb{L}_{nat}(k)}}, \tag{41}$$

and k is a spatial index.

Note that in the analysis of SCI natural regions, we use filters with larger sizes in order to characterize the more complicated textures and thicker lines. Specifically, the two circular averaging filters are both 15×15 pixel size. The LoG filter has the same kernel size as in Section III-B1, but with a different standard deviation ($\sigma = 0.9$). The Gaussian weighting function used for computing $\mathbf{S}_{nat}^{\mathbb{D}}$ and $\mathbf{S}_{nat}^{\mathbb{G}}$ has the same window size (11×11) and standard deviation (1.5 samples) as in [10].

3) Quality Combination: Given the quality estimates of both the synthetic and natural regions, the final stage is to combine these scores into an overall perceived distortion estimate (denoted by CNN-SQE), which is computed as the weighted product of Q_{syn} and Q_{nat}:

$$\text{CNN-SQE} = \left(Q_{syn}\right)^{\alpha} \times \left(Q_{nat}\right)^{1-\alpha}. \tag{42}$$

Here, α is a weighting parameter determined by the areas of the two regions, and is computed via a sigmoid transducer function:

$$\alpha = \frac{A}{1 + e^{t_1(\omega - t_2)}} + (1 - A), \tag{43}$$

where $\omega = \sum \mathbb{L}_{syn} / \left(\sum \mathbb{L}_{syn} + \sum \mathbb{L}_{nat}\right)$ denotes the area percentage of the synthetic region; $t_1 = -5$ and $t_2 = 0.5$ are the two parameters that control the shape of the sigmoid transducer function curve; and $A = 0.7$ is a parameter that controls the position and dynamic range of the curve. Smaller values of CNN-SQE denote predictions of higher SCI quality.

As shown in Figure 10, when mapping ω into α through the blue curve ($A = 1$), the importance of synthetic and natural regions are equally considered, because the two regions are assigned with the same weight ($\alpha = 0.5$) when their sizes are the same ($\omega = 0.5$). In view of the importance of plain text, we use the red curve ($A = 0.7$) such that α is greater

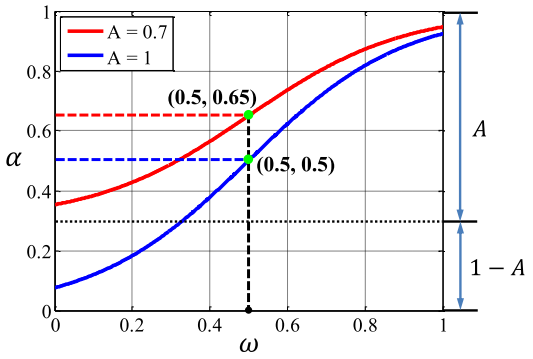


Fig. 10. Plots of two sigmoid transducer function curves that map ω to α via Eq. (43) with different "A" values. Note that the importance of synthetic and natural regions are equally treated when using the blue curve for mapping, while the red curve gives more consideration to the importance of text.

than 0.5 when ω equals to 0.5. In fact, the item "$1 - A$" in Eq. (43) aims to add more weight to the synthetic region. A smaller value of A results in a larger α value for a certain ω value, and thus the more important role of the plain text is taken into account.

It is important to note that by using Eqs. (42) and (43), we are assuming that the overall quality of an SCI is jointly affected by three factors: the perceived quality of the synthetic and natural regions, the areas of the two regions, and the importance of plain text. Admittedly, there are many other factors which the HVS uses to judge the quality of SCIs, including the positions of texts, the fonts and sizes of characters, and the contents of pictures, etc. However, as a preliminary attempt to solve this problem, we focus on seeking simple and general principles/rules that are effective for QA of most SCIs, rather than any specific factor. We have found that combining quality scores of different regions based on their sizes is a useful idea that has been adopted in many previous SC-IQA works (e.g., [8], [16], and [17],, etc.) though other features may also be used. The results that we will show in Section IV seem to suggest that our assumption is reasonable. A performance analysis of CNN-SQE with different t_1, t_2, and A values is provided at http://vision.eng.shizuoka.ac.jp/CNNSQE/.

IV. Results and Analysis

In this section, we analyze CNN-SQE's ability to predict SCI quality by using various SCI databases. For this task, we test CNN-SQE on three publicly available SCI databases: (1) the SIQAD database [7], [8]; (2) the SCD [9] database; and (3) the SCID database [1]. We also compare the performance of CNN-SQE to other FR IQA algorithms.

A. Screen Content Database

The SIQAD database contains 20 original SCIs and 980 distorted versions generated with seven distortion types at seven different levels. The seven distortion types are: Gaussian noise (GN), Gaussian blurring (GB), motion blurring (MB), contrast change (CC), JPEG compression, JPEG2000 compression (JP2K), and Layer-segmentation based compression (LSC). The database contains SCIs with different resolutions, and the subjective ratings (in terms of DMOS values) were collected using a single-stimulus scaling paradigm.

The SCD database contains 24 original SCIs and 492 compressed versions generated by two coding technologies: (1) high-efficiency video coding (HEVC), and (2) HEVC screen content compression (HEVC-SCC), which is an extension of HEVC to support 4:4:4 format compression. For each distortion type, the quantization parameter ranges from 30 to 50 to create 11 different levels. The database contains SCIs with three different resolutions: 2560×1440, 1920×1080, and 1280×720. The subjective ratings (in terms of MOS values) were collected using a single-stimulus scaling paradigm.

The SCID database contains 40 original SCIs and 1800 distorted versions generated with nine distortion types at five different levels. The nine distortion types are: GN, GB, MB, CC, color saturation change (CSC), color quantization with dithering (CQD), JPEG compression, JPEG2000 compression, and HEVC-SCC. All SCIs in the database are of resolution 1920×1080. The subjective ratings (in terms of MOS values) were collected using a double-stimulus scaling paradigm.

Note that the original SCIs in all three databases were gathered through screen snapshots from web-pages, powerpoint slides, PDF files, digital magazines, etc., with various combinations of text, graphics and pictures. Also note that the color saturation change in the SCID database does not introduce any visible distortions when images are viewed in grayscale. Thus, we do not test CSC distortion in our work.

B. Algorithms and Performance Measures

We compared CNN-SQE with various FR and NR IQA algorithms. Among the 12 FR IQA algorithms, six are classical quality methods traditionally designed for QA of NIs, and the other six are specifically designed for QA of SCIs. The six classical quality metrics were SSIM [10], multi-scale structure similarity (MS-SSIM) [25], feature similarity index (FSIM) [26], gradient similarity (GSIM) [27], visual saliency-induced index (VSI) [28], and gradient magnitude similarity deviation (GMSD) [29]. The six SC-IQA algorithms were SCI quality index (SQI) [16], gradient structure similarity (GSS) [13], structure-induced quality metric (SIQM) [11], saliency-guided quality measure (SQMS) [12], structure features and uncertainty weighting metric (SFUW) [17], and edge similarity (ESIM) [1]. The two NR IQA algorithms were visual importance and distortion guided deep image quality assessment framework (VIDGIQA) [30] and integrated local natural image quality evaluator (IL-NIQE) [31].

A QA algorithm might yield quality predictions that are nonlinearly related to the actual MOS/DMOS values, and such nonlinearity can lead to a seemingly poor prediction accuracy/consistency. To solve this problem, we applied a logistic transform to each algorithms' raw predicted scores before evaluating the performance of a particular QA method on a particular database. As recommended in [32], the logistic transform was a four-parameter sigmoid given by

$$f(x) = \frac{\tau_1 - \tau_2}{1 + \exp(-\frac{x-\tau_3}{|\tau_4|})} + \tau_2, \tag{44}$$

TABLE I

PERFORMANCE COMPARISON OF SEVEN ABRIDGED VERSIONS OF CNN-SQE, AMONG WHICH SIX VERSIONS ESTIMATE SCI QUALITY BY ANALYZING ONE/TWO OF THE THREE CNN-BASED SEGMENTED REGIONS, AND THE LAST ONE (ICM-SQE) ESTIMATES SCI QUALITY BY USING THE INFORMATION CONTENT MODEL FOR SCI REGION SEGMENTATION. FOR REFERENCE, THE RESULTS OF CNN-SQE ARE ALSO INCLUDED

		PT-SQE	NI-SQE	CGC-SQE-1	CGC-SQE-2	Syn-SQE	Nat-SQE	ICM-SQE	CNN-SQE
SROCC	SIQAD	0.8792	0.8572	0.8403	0.8423	0.8795	0.8632	0.8646	0.8943
	SCD	0.9154	0.8192	0.9016	0.8961	0.9189	0.8747	0.9057	0.9310
	SCID	0.8973	0.8749	0.8797	0.8734	0.8990	0.8817	0.8937	0.9139
PLCC	SIQAD	0.8906	0.8738	0.8578	0.8534	0.8911	0.8784	0.8798	0.9042
	SCD	0.9124	0.8203	0.8994	0.8911	0.9153	0.8750	0.9040	0.9291
	SCID	0.8973	0.8750	0.8808	0.8744	0.8992	0.8819	0.8933	0.9147
KROCC	SIQAD	0.6928	0.6684	0.6473	0.6508	0.6931	0.6762	0.6733	0.7152
	SCD	0.7514	0.6337	0.7300	0.7207	0.7599	0.6987	0.7363	0.7747
	SCID	0.7131	0.6833	0.6888	0.6804	0.7158	0.6907	0.7085	0.7352
RMSE	SIQAD	6.5094	6.9609	7.3572	7.4613	6.4964	6.8406	6.8050	6.1147
	SCD	0.9082	1.2687	0.9697	1.0068	0.8934	1.0742	0.9484	0.8207
	SCID	6.2915	6.9005	6.7487	6.9171	6.2360	6.7200	6.4074	5.7609

where x denotes the raw predicted score, and where τ_1, τ_2, τ_3, and τ_4 are free parameters selected to provide the best fit of the predicted scores to the MOS/DMOS values.

After non-linear regression, four criteria were used to measure the prediction monotonicity and prediction accuracy of each algorithm: (1) the Spearman Rank-Order Correlation Coefficient (SROCC), (2) the Pearson Linear Correlation Coefficient (PLCC), (3) the Kendall Rank-Order Correlation Coefficient (KROCC), and (4) the Root Mean Square Error (RMSE). Note that the logistic transform in Eq. (44) will affect only PLCC and RMSE, not SROCC and KROCC. A value close to 1 for SROCC, PLCC, KROCC, and close to 0 for RMSE indicate superior quality prediction.

C. Contributions of CNN-Based Segmented Regions

To analyze the contribution of each SCI region (plain text, compute graphics/cartoons, and natural images) toward the overall performance, we created six versions of CNN-SQE in which each version estimates SCI quality by analyzing only one/two of the three segmented regions. We also created another version of CNN-SQE, which employs the same QA settings and procedures as in Section III-B, but utilizes a different classifier (information content model [16]) to perform the SCI region segmentation. In total, the following seven abridged versions were created:

- Plain text only (PT-SQE);
- Natural images only (NI-SQE);
- Computer graphics/cartoons using QA method in Section III-B1 (CGC-SQE-1);
- Computer graphics/cartoons using QA method in Section III-B2 (CGC-SQE-2);
- Synthetic (plain text + computer graphics/cartoons) (Syn-SQE);
- Natural (natural images + computer graphics/cartoons) (Nat-SQE);
- Information content model (ICM-SQE);

The first six versions operate by using the same CNN model for region segmentation and the same parameter settings for quality-related feature extraction. The last version (ICM-SQE) operates by first classifying each 4×4 image block into one of the three categories (plain text, compute graphics/cartoons, and natural images) based on thresholding the information content map, and then combines different blocks to build synthetic and natural regions. The QA methods in Section III-B1 and Section III-B2 with the same parameter settings were then applied on the two regions, from which the final quality score was computed by Eq. (42). The testing was performed on the SIQAD, SCD, and SCID databases, and the results of this analysis are shown in Table I in terms of SROCC, PLCC, KROCC, and RMSE. For reference, also shown in Table I are the testing results of the original CNN-SQE algorithm.

As shown in Table I, observe that the additional analysis of computer graphics/cartoons region can improve the predictive performance of using only the plain text or natural images, which demonstrates the effectiveness of our proposed non-strict segmentation scheme. Specifically, for the SIQAD database, we observe little performance improvement, which is as expected, because SCIs in this database contain only a small amount of computer graphics/cartoons. As a result, CGC-SQE-1 and CGC-SQE-2 perform worse than PT-SQE and NI-SQE. In comparison, for SCD and SCID which contain images with larger areas of computer graphics/cartoons, the performance is greatly improved after region combination. We also observe that for SCD and SCID, CGC-SQE-1 and/or CGC-SQE-2 perform worse than PT-SQE, but better than NI-SQE, which suggests that computer graphics/cartoons are more important than natural images but less crucial than plain text in determining the qualities of SCIs in these two databases.

Comparing testing results of Syn-SQE, Nat-SQE, and ICM-SQE with CNN-SQE, we further observe that combining the quality analyses of the synthetic and natural regions together can largely improve the performance, but an inappropriate synthetic/natural segmentation map such as that obtained by using the ICM can decrease the performance. This fact demonstrates the effectiveness of the proposed CNN-based segmentation model and the adaptive quality-fusion strategy. Moreover, as shown in Table I, on all three databases, Syn-SQE performs better than Nat-SQE, and PT-SQE performs better than NI-SQE, which partially suggests that humans judge SCI quality based more on synthetic (or plain text) regions than natural (or natural image) regions when both are presented. In summary, the proposed segmentation scheme along with the adaptive quality-fusion strategy both play important roles in estimating SCI quality.

D. Overall Performance

The overall testing results on the SIQAD, SCD, and SCID databases are shown in Table II in terms of SROCC, PLCC, KROCC, and RMSE. Italicized entries denote the FR IQA

TABLE II

OVERALL PERFORMANCES OF CNN-SQE AND OTHER FR AND NR IQA ALGORITHMS ON THE SIQAD, SCD, AND SCID DATABASES. ITALICIZED ENTRIES DENOTE QA ALGORITHMS DESIGNED FOR NIS. ENTRIES MARKED BY "∗" DENOTE THE NR ALGORITHMS. RESULTS OF THE BEST-PERFORMING IQA ALGORITHM ARE BOLDED

		SSIM	*MS-SSIM*	*FSIM*	*GMSD*	*GSIM*	*VSI*	IL-NIQE*	VIDGIQA*	SQI	GSS	SIQM	SQMS	SFUW	ESIM	CNN-SQE
SROCC	SIQAD	*0.758*	*0.618*	*0.582*	*0.731*	*0.548*	*0.538*	0.321	0.524	0.855	0.836	0.858	0.880	0.869	0.863	**0.894**
	SCD	*0.870*	*0.895*	*0.911*	*0.877*	*0.895*	*0.872*	0.256	0.024	0.908	0.864	0.895	0.910	0.895	0.921	**0.931**
	SCID	*0.746*	*0.775*	*0.797*	*0.872*	*0.716*	*0.753*	0.125	0.409	0.881	0.775	0.866	0.897	0.856	0.894	**0.914**
	Average	*0.770*	*0.744*	*0.746*	*0.827*	*0.691*	*0.703*	0.209	0.384	0.877	0.809	0.868	0.894	0.866	0.889	**0.910**
PLCC	SIQAD	*0.758*	*0.617*	*0.584*	*0.738*	*0.565*	*0.535*	0.373	0.602	0.864	0.846	0.862	0.887	0.882	0.879	**0.904**
	SCD	*0.859*	*0.725*	*0.772*	*0.874*	*0.787*	*0.772*	0.272	0.027	0.903	0.865	0.886	0.903	0.895	0.921	**0.929**
	SCID	*0.759*	*0.787*	*0.807*	*0.872*	*0.729*	*0.755*	0.229	0.414	0.884	0.774	0.870	0.897	0.859	0.893	**0.915**
	Average	*0.775*	*0.723*	*0.730*	*0.830*	*0.686*	*0.687*	0.282	0.412	0.881	0.812	0.870	0.895	0.872	0.893	**0.914**
KROCC	SIQAD	*0.560*	*0.459*	*0.425*	*0.549*	*0.405*	*0.387*	0.228	0.370	0.667	0.640	0.668	0.694	0.681	0.674	**0.715**
	SCD	*0.693*	*0.723*	*0.743*	*0.701*	*0.721*	*0.694*	0.175	0.015	0.740	0.682	0.724	0.747	0.729	0.762	**0.775**
	SCID	*0.552*	*0.574*	*0.598*	*0.679*	*0.515*	*0.554*	0.089	0.284	0.697	0.569	0.678	0.712	0.656	0.706	**0.735**
	Average	*0.577*	*0.561*	*0.566*	*0.641*	*0.513*	*0.524*	0.147	0.268	0.694	0.610	0.682	0.712	0.676	0.705	**0.735**
RMSE	SIQAD	*9.332*	*11.263*	*11.618*	*9.658*	*11.814*	*12.093*	13.278	11.429	7.200	7.629	7.248	6.609	6.735	6.828	**6.115**
	SCD	*1.134*	*2.218*	*2.218*	*1.079*	*2.218*	*2.218*	2.135	2.218	0.955	1.112	1.027	0.953	0.989	0.864	**0.821**
	SCID	*9.282*	*8.794*	*8.425*	*6.966*	*9.755*	*9.347*	13.876	12.974	6.666	9.020	7.035	6.288	7.310	6.403	**5.761**
	Average	*7.993*	*8.529*	*8.449*	*6.882*	*9.205*	*9.082*	11.805	10.758	5.922	7.309	6.141	5.536	6.114	5.651	**5.083**

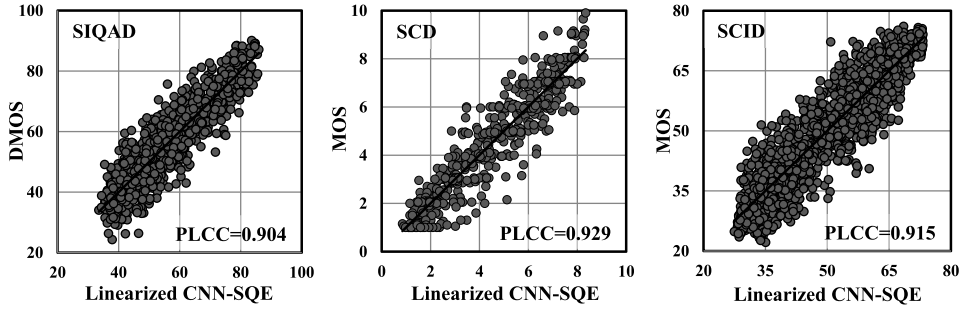


Fig. 11. Scatter plots of objective scores predicted by CNN-SQE after a logistic transform versus subjective scores on different SCI databases. Note that the x-axis across all three figures represents the predicted value transformed via Eq.(44); the y-axis represents the true DMOS value for the SIQAD database, true MOS value for the SCD and SCID databases.

algorithms that are designed for QA of NIs. Entries marked by "∗" denote the NR IQA algorithms. Results of the best-performing IQA algorithm are bolded. As mentioned previously, we did not test CSC distortion in the SCID database, and thus the total number of distorted images being tested in SCID was 1600.

As shown in Table II, it is clear that CNN-SQE performs the best in predicting the quality of SCIs as compared with other methods. It improves upon GSS and SIQM, and is superior to the six NI-based QA methods considered. Specifically, CNN-SQE outperforms all the other FR/NR IQA algorithms in terms of all four performance criteria on all the three databases. As reported in [8] and also suggested by Table I, the perceptual quality of text portion has a higher correlation with the entire SCI quality as compared with the image portion. Accordingly, the six NI-based QA methods perform less effectively on the three SCI databases, and especially weak on SIQAD, which contains mostly the text-dominant SCIs. In comparison, the SC-IQA algorithms (such as SQMS and ESIM) extract text-aware quality-related features and thus these algorithms can achieve relatively better performance than the NI-based methods. The last rows of the SROCC, PLCC, KROCC, and RMSE results in Table II show the average SROCC, PLCC, KROCC, and RMSE, where the averages are weighted by the number of distorted images tested in each database. On an average, CNN-SQE also demonstrates the best FR IQA performance.

To visualize the performance yielded by CNN-SQE, Figure 11 shows scatter-plots of logistic-transformed CNN-SQE quality predictions vs. subjective ratings (MOS or DMOS) on the three test databases. Although for each database, there are some images whose quality scores are predicted far from their true MOS/DMOS values, overall, the proposed CNN-SQE algorithm can predict quality well across the range of MOS/DMOS.

E. Statistical Significance

To quantify whether or not the numerical difference between IQA algorithms' performances were statistically significant, we performed two types of the statistical significance test. One is a t-test [33] and the other one is an F-test.

A t-test was employed to compare the mean correlations between different algorithms. To this end, for each SCI database, we randomly selected 75% of the distorted images for testing and repeated the procedure 1000 times, which give rise to 1000 SROCC values for each algorithm. Please note that for each distorted image in a database, the quality score given by a specific IQA algorithm is fixed, and thus the SROCC value is affected only by which images were selected. The null hypothesis is that the mean SROCC values of two algorithms are equal with a confident of 95%. The alternate hypothesis is that the mean SROCC values of two algorithms are different. The test statistic is calculated as

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{45}$$

TABLE III

COMPARISON OF THE STATISTICAL SIGNIFICANCE (t-TEST) OF CNN-SQE AND OTHER IQA ALGORITHMS ON THE SIQAD, SCD, SCID DATABASES

		SSIM	MS-SSIM	FSIM	GMSD	GSIM	VSI	IL-NIQE	VIDGIQA	SQI	GSS	SIQM	SQMS	SFUW	ESIM	CNN-SQE
SIQAD	CNN-SQE	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	–
	Mean	0.759	0.618	0.583	0.730	0.549	0.538	0.324	0.527	0.855	0.837	0.859	0.881	0.870	0.864	0.894
	Median	0.759	0.618	0.582	0.730	0.548	0.538	0.323	0.527	0.855	0.837	0.858	0.880	0.868	0.864	0.894
SCD	CNN-SQE	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	–
	Mean	0.872	0.897	0.911	0.878	0.896	0.873	0.257	0.045	0.908	0.866	0.897	0.911	0.897	0.922	0.931
	Median	0.872	0.897	0.910	0.875	0.895	0.871	0.260	0.038	0.908	0.865	0.896	0.910	0.896	0.919	0.931
SCID	CNN-SQE	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	–
	Mean	0.746	0.775	0.796	0.872	0.716	0.753	0.124	0.409	0.881	0.775	0.866	0.897	0.856	0.894	0.914
	Median	0.746	0.775	0.797	0.872	0.716	0.753	0.122	0.406	0.880	0.774	0.866	0.897	0.856	0.894	0.914

TABLE IV

COMPARISON OF THE STATISTICAL SIGNIFICANCE (F-TEST) OF CNN-SQE AND OTHER IQA ALGORITHMS ON THE SIQAD, SCD, SCID DATABASES

		SSIM	MS-SSIM	FSIM	GMSD	GSIM	VSI	IL-NIQE	VIDGIQA	SQI	GSS	SIQM	SQMS	SFUW	ESIM	CNN-SQE
SIQAD	CNN-SQE	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	–
	JBSTAT	37.5	64.2	28.0	99.6	23.1	6.0	13.6	*5.0*	59.0	*3.7*	*52.1*	*2.5*	*1.2*	*1.7*	*1.4*
SCD	CNN-SQE	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	0	–
	JBSTAT	32.1	21.5	21.5	94.4	21.5	21.5	21.1	20.9	*8.0*	*3.0*	20.5	29.8	35.3	58.4	15.7
SCID	CNN-SQE	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	–
	JBSTAT	69.0	48.0	21.4	6.3	30.5	*2.7*	78.4	*35.9*	127.5	28.4	145.7	*9.4*	*1.3*	*0.1*	*0.2*

where $\bar{X}_i$ and s_i^2 $(i = 1, 2)$ denote, respectively, the mean and the unbiased estimate of the variance of the 1000 SROCC values; and $n_1 = n_2 = 1000$.

An F-test was employed to compare the prediction residuals of two algorithms, assuming that the residuals are Gaussian-distributed. The test statistic is the ratio of two algorithms' residual variances (errors in predictions), denoted by $F = \sigma_A^2/\sigma_B^2$. A smaller residual variance indicates a better prediction. Values of $F > F_{critical}$ (or $F < 1/F_{critical}$) indicate that at a given confidence level, method A has significantly larger (or smaller) residuals than method B, where $F_{critical}$ is computed based on the number of residuals and the confidence level. Note that the significance test is often inconclusive if residuals are not Gaussian. In this paper, we used the Jarque-Bera (JB) statistic [34] to measure the Gaussianity of the residuals. A smaller value of the JB statistic denotes less deviation from Gaussianity, and vice versa.

The overall statistical performance of CNN-SQE as compared with other IQA algorithms on the SIQAD, SCD, and SCID databases is shown in Tables III (t-test results) and IV (F-test results), in which "+1", "0", "−1" indicates that CNN-SQE is statistically superior, equivalent, or inferior to the test algorithm with confidence greater than 95%. Note that for t-test, we also provided in Table III the mean and median of the 1000 SROCC values for each algorithm. For F-test, we included in Table IV the JB statistic measures of Gaussianity. Italicized JB entries denote that the residuals can be deemed Gaussian with 95% confidence.

As shown in Tables III and IV, CNN-SQE is statistically the best-performing algorithm on the SIQAD and SCID databases. On the SCD database, the t-test results indicate that CNN-SQE performs statistically the best, while the F-test results indicate that CNN-SQE has statistically the same performance as ESIM (which might due to the fact SCD contains relatively a small number of the compressed SCIs only). Also, note that on the SIQAD and SCID databases, both ESIM and CNN-SQE, and along with some other algorithms, have Gaussian residuals as denoted by their JB statistics. However, on the SCD database, only GSS has Gaussian residuals.

F. Performance on Individual Distortion Types

We also report the performance of CNN-SQE and other FR and NR IQA algorithms on subsets of the SIQAD, SCD, and SCID databases corresponding to each individual distortion type. Test results of seven distortion types in SIQAD, two in SCD, and eight in SCID are presented in Table V in terms of SROCC values (the PLCC, KROCC, and RMSE values follow similar trends). Bold entries denote the best-performing IQA algorithm for each distortion type on each database.

As shown in Table V, CNN-SQE provides better or competitive predictions in comparison to the six SC-IQA algorithms on most distortion types. Compared with the six NI-based QA algorithms (especially FSIM), we observe that CNN-SQE does not always perform better, and even performs slightly worse on certain distortion types (e.g., JPEG2000 compression and HEVC screen content compression in the SCID database). This interesting observation, together with results in Table II, indicate that these NI-based QA methods can sometimes work effectively at evaluating degradation levels of certain distortion types. However, they have difficulties in bringing all quality measures on the same scale when viewing the database as a whole. One possible reason is that different distortion types can have varying degrees of impact on the perceived qualities of different regions in SCIs. Thus, the quality estimate given by the NI-based QA method may be consistent with the ground-truth quality of the natural image region (when the region is viewed separately alone), but not the whole SCI (which also includes plain text and computer graphics/cartoons regions), when different distortion types are considered. Also observe from Table V that most IQA algorithms, including CNN-SQE, perform less effectively on CC images. This is due to the fact that CC does not affect the image's edge-structure. In summary, when looking at the performance on individual distortion types, CNN-SQE still demonstrates competitive FR IQA performance.

G. Computational Complexity

In this section, we analyze the computational complexity of the proposed CNN-SQE algorithm. To this end, we first mea-

TABLE V

SROCC VALUES OF CNN-SQE AND OTHER FR AND NR IQA ALGORITHMS ON DIFFERENT TYPES OF DISTORTION ON THE SIQAD, SCD, AND SCID DATABASES. ITALICIZED ENTRIES DENOTE QA ALGORITHMS DESIGNED FOR NIS. ENTRIES MARKED BY "∗" DENOTE THE NR ALGORITHMS. RESULTS OF THE BEST-PERFORMING IQA ALGORITHM ARE BOLDED

		SSIM	*MS-SSIM*	*FSIM*	*GMSD*	*GSIM*	*VSI*	IL-NIQE*	VIDGIQA*	SQI	GSS	SIQM	SQMS	SFUW	ESIM	CNN-SQE
SIQAD	GN	*0.868*	*0.868*	*0.871*	*0.886*	*0.843*	*0.866*	0.816	0.705	0.860	0.852	0.871	0.886	0.880	0.876	**0.893**
	GB	*0.894*	*0.889*	*0.822*	*0.912*	*0.880*	*0.850*	0.456	0.746	0.924	0.908	0.923	0.915	0.899	0.924	**0.924**
	MB	*0.806*	*0.822*	*0.724*	*0.844*	*0.776*	*0.766*	0.446	0.393	0.881	0.838	0.858	0.869	0.841	0.894	**0.904**
	CC	*0.641*	***0.749***	*0.715*	*0.544*	*0.733*	*0.646*	0.044	0.029	0.668	0.454	0.693	0.695	0.687	0.611	0.665
	JPEG	*0.757*	*0.777*	*0.664*	*0.771*	*0.680*	*0.720*	0.287	0.334	0.819	0.797	0.811	0.789	0.729	0.796	**0.847**
	JP2K	*0.757*	*0.786*	*0.686*	*0.844*	*0.712*	*0.730*	0.381	0.277	0.817	0.810	0.811	0.819	0.819	0.782	**0.862**
	LSC	*0.735*	*0.775*	*0.706*	*0.859*	*0.717*	*0.742*	0.167	0.313	0.843	0.816	0.806	0.829	0.746	0.796	**0.887**
SCD	HEVC	*0.865*	*0.897*	*0.913*	*0.888*	*0.903*	*0.854*	0.116	0.119	0.910	0.888	0.904	0.917	0.896	0.923	**0.933**
	HEVC-SCC	*0.869*	*0.888*	*0.905*	*0.860*	*0.884*	*0.903*	0.367	0.004	0.905	0.840	0.883	0.900	0.890	0.920	**0.924**
SCID	GN	*0.917*	*0.932*	*0.938*	*0.934*	*0.911*	*0.946*	0.701	0.788	0.923	0.697	0.913	0.915	0.929	0.943	**0.949**
	GB	*0.869*	*0.899*	*0.847*	*0.793*	*0.842*	*0.822*	0.200	0.584	0.895	0.886	**0.923**	0.910	0.897	0.869	0.907
	MB	*0.859*	*0.892*	*0.837*	*0.815*	*0.819*	*0.801*	0.240	0.407	0.842	0.873	**0.901**	0.881	0.812	0.860	0.878
	CC	*0.659*	*0.837*	***0.847***	*0.578*	*0.830*	*0.816*	0.021	0.055	0.741	0.364	0.743	0.803	0.744	0.618	0.745
	JPEG	*0.848*	*0.918*	*0.940*	*0.934*	*0.937*	*0.914*	0.274	0.678	0.911	0.889	0.916	0.924	0.900	0.934	**0.947**
	JP2K	*0.843*	*0.907*	***0.948***	*0.928*	*0.935*	*0.931*	0.097	0.400	0.896	0.837	0.893	0.932	0.910	0.931	0.939
	HEVC-SCC	*0.826*	*0.866*	***0.909***	*0.896*	*0.873*	*0.893*	0.188	0.497	0.858	0.797	0.852	0.867	0.808	0.904	0.898
	CQD	*0.778*	*0.865*	***0.906***	*0.905*	*0.871*	*0.882*	0.008	0.250	0.873	0.741	0.830	0.891	0.813	0.887	0.904

TABLE VI

INFORMAL COMPLEXITY ANALYSIS OF CNN-SQE. TABULATED VALUES REFLECT THE PERCENTAGE OF TIME DEVOTED TO EACH STAGE IN CNN-SQE

CNN-SQE	Percentage of time
SCI segmentation	95.4
QA of synthetic/natural region	4.6
Quality combination	0.0

TABLE VII

A COMPARISON OF THE RUNTIME REQUIREMENTS (SECONDS/IMAGE) FOR SIX FR SC-IQA ALGORITHMS ON DIFFERENT IMAGE SIZES

	180×320	360×640	540×960	720×1280	900×1600
SQI [16]	0.05	0.16	0.36	0.63	0.96
GSS [13]	0.02	0.07	0.12	0.32	0.32
SIQM [11]	0.01	0.05	0.11	0.19	0.29
SQMS [12]	0.01	0.03	0.06	0.12	0.19
SFUW [17]	7.69	31.40	70.00	123.07	189.68
ESIM [1]	0.16	0.71	1.49	3.38	4.90
CNN-SQE	1.65	5.92	13.08	23.15	37.04

sured the relative percentage of time required by each stage of CNN-SQE on a 720×1280 image. We also compared the overall computational complexity of CNN-SQE with other FR SC-IQA algorithms on images with different sizes (180×320, 360×640, 540×960, 720×1280, 900×1600). All of these tests were performed by running unoptimized MATLAB code on a modern desktop computer [Intel(R) Core(TM) i7-4790K CPU at 4.00 GHz, 16.0 GB RAM, Windows 10 Pro 64-bit]. The results are shown in Table VI and Table VII, respectively.

As shown in Table VI, most of the time required by CNN-SQE is spent on SCI segmentation, which results in a relatively longer operating time as compared with other FR SC-IQA algorithms (as shown in Table VII). However, CNN-SQE still runs faster than SFUW, and can be easily accelerated by decreasing the number of overlapping pixels between neighboring blocks in the SCI segmentation stage. To demonstrate, Table VIII shows the runtime requirements (seconds/image) for CNN-SQE with different block overlaps on different sizes of images, as well as their corresponding SROCC values on the SIQAD, SCD, and SCID databases (note that during this test, the local entropy is computed for non-overlapping 16×16 image patches). We observe that when the amount of block overlap changes from 36 to 8 pixels, the runtime required by CNN-SQE decreases significantly while the QA performances on the three databases are not seriously affected. Thus, although we set large block overlaps (40 pixels for SCI region segmentation and eight pixels for local entropy computation) throughout the paper in order for clear visualization and illustration of the SCI segmentation stage, in practice these overlaps can be adjusted without hurting the QA performance and the algorithm can run much faster.

TABLE VIII

RUNTIME REQUIREMENTS (SECONDS/IMAGE) FOR CNN-SQE WITH DIFFERENT BLOCK OVERLAPS IN SCI SEGMENTATION AND THEIR CORRESPONDING PERFORMANCES ON THE SIQAD, SCD, AND SCID DATABASES

# of overlap pixels		36	32	24	16	8	0
SROCC	SIQAD	0.894	0.894	0.894	0.894	0.893	0.892
	SCD	0.931	0.932	0.933	0.932	0.931	0.934
	SCID	0.913	0.913	0.913	0.914	0.914	0.913
Runtime (seconds/ image)	180×320	0.59	0.42	0.24	0.19	0.15	0.14
	360×640	2.74	1.66	0.93	0.62	0.49	0.44
	540×960	5.92	3.62	1.95	1.36	1.09	0.94
	720×1280	10.51	6.37	3.43	2.42	1.86	1.63
	900×1600	16.45	10.10	5.32	3.74	2.96	2.51

V. CONCLUSION

This paper presented an FR algorithm for quality assessment of screen content images based on separate measures of the edge-structure degradations for different segmented regions. Our method, called CNN-SQE, operates under the principle that a screen content image often contains plain text and computer graphics/cartoons in addition to natural images, and thus quality estimation should be performed separately for different regions in order to achieve improved quality predictive performance. To this end, CNN-SQE operates via three main stages: (1) image segmentation, (2) quality assessment of each segmented region, and (3) quality combination. The first stage employs local standard deviation, local entropy, and a convolutional neural network to generate the synthetic/natural segmentation maps, which are then incorporated in the second stage to guide the quality assessment task. In the second stage, the Laplacian of Gaussian filters are first applied on the contrast maps to extract the edge information, and then the edge-structure degradation for each region is modeled and measured to produce the region-based quality degradation scores. The final stage adaptively combines the various quality estimates obtained from each region to yield an overall quality degradation estimate based on the region area. Experimental results tested on various screen content image quality databases demonstrate the effectiveness of our proposed CNN-SQE model.

REFERENCES

[1] Z. Ni, L. Ma, H. Zeng, J. Chen, C. Cai, and K.-K. Ma, "ESIM: Edge similarity for screen content image quality assessment," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4818–4831, Oct. 2017.

[2] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *J. Vis. Commun. Image Represent.*, vol. 22, no. 4, pp. 297–312, 2011.

[3] D. M. Chandler, "Seven challenges in image quality assessment: Past, present, and future research," *ISRN Signal Process.*, vol. 2013, Nov. 2013, Art. no. 905685.

[4] 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.

[5] X. Guo, L. Huang, K. Gu, L. Li, Z. Zhou, and L. Tang, "Naturalization of screen content images for enhanced quality evaluation," *IEICE Trans. Inf. Syst.*, vol. 100, no. 3, pp. 574–577, 2017.

[6] H. Yang, W. Lin, and C. Deng, "Image activity measure (IAM) for screen image segmentation," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 1569–1572.

[7] H. Yang, Y. Fang, W. Lin, and Z. Wang, "Subjective quality assessment of screen content images," in *Proc. IEEE 6th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Sep. 2014, pp. 257–262.

[8] 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, Aug. 2015.

[9] S. Shi, X. Zhang, S. Wang, R. Xiong, and S. Ma, "Study on subjective quality assessment of screen content images," in *Proc. Picture Coding Symp. (PCS)*, May/Jun. 2015, pp. 75–79.

[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] K. Gu, S. Wang, G. Zhai, S. Ma, and W. Lin, "Screen image quality assessment incorporating structural degradation measurement," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2015, pp. 125–128.

[12] K. Gu *et al.*, "Saliency-guided quality assessment of screen content images," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1098–1110, Jun. 2016.

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

[14] Z. Ni, L. Ma, H. Zeng, C. Cai, and K.-K. Ma, "Screen content image quality assessment using edge model," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 81–85.

[15] S. Wang, K. Gu, K. Zeng, Z. Wang, and W. Lin, "Perceptual screen content image quality assessment and compression," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 1434–1438.

[16] S. Wang, K. Gu, K. Zeng, Z. Wang, and W. Lin, "Objective quality assessment and perceptual compression of screen content images," *IEEE Comput. Graph. Appl.*, vol. 38, no. 1, pp. 47–58, Jan./Feb. 2018.

[17] Y. Fang, J. Yan, J. Liu, S. Wang, Q. Li, and Z. Guo, "Objective quality assessment of screen content images by uncertainty weighting," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 2016–2027, Apr. 2017.

[18] Z. Wang and X. Shang, "Spatial pooling strategies for perceptual image quality assessment," in *Proc. Int. Conf. Image Process.*, Oct. 2016, pp. 2945–2948.

[19] H. Yang, S. Wu, C. Deng, and W. Lin, "Scale and orientation invariant text segmentation for born-digital compound images," *IEEE Trans. Cybern.*, vol. 45, no. 3, pp. 519–533, Mar. 2015.

[20] D. L. Ruderman, "The statistics of natural images," *Netw., Comput. Neural Syst.*, vol. 5, no. 4, pp. 517–548, 1994.

[21] A. Mittal, G. S. Muralidhar, J. Ghosh, and A. C. Bovik, "Blind image quality assessment without human training using latent quality factors," *IEEE Signal Process. Lett.*, vol. 19, no. 2, pp. 75–78, Feb. 2012.

[22] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[24] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[25] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. IEEE 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Nov. 2003, pp. 1398–1402.

[26] 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.

[27] 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.

[28] 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, Aug. 2014.

[29] 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.

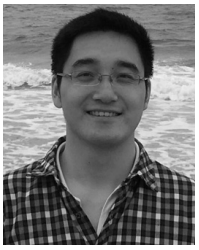
[30] J. Guan, S. Yi, X. Zeng, W.-K. Cham, and X. Wang, "Visual importance and distortion guided deep image quality assessment framework," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2505–2520, Nov. 2017.

[31] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[32] VQEG. (Aug. 2003). *Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II*. [Online]. Available: http://www.vqeg.org

[33] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, FL, USA: CRC Press, 2003.

[34] C. M. Jarque and A. K. Bera, "Efficient tests for normality, homoscedasticity and serial independence of regression residuals," *Econ. Lett.*, vol. 6, no. 3, pp. 255–259, 1980.



Yi Zhang received the B.S. and M.S. degrees in electrical engineering from Northwestern Polytechnical University, Xi'an, China, in 2008 and 2011, respectively, and the Ph.D. degree in electrical engineering from Oklahoma State University, Stillwater, OK, USA, in 2015. From 2016 to 2018, he was a Post-Doctoral Research Associate with the Department of Electrical and Electronic Engineering, Shizuoka University, Japan. He is currently a Faculty Member with the School of Electronic and Information Engineering, Xi'an Jiaotong University, China. His research interests include 2D/3D image processing, machine learning, pattern recognition, and computer vision.

Damon M. Chandler received the B.S. degree in biomedical engineering from Johns Hopkins University, Baltimore, MD, USA, in 1998, and the M.Eng., M.S., and Ph.D. degrees in electrical engineering from Cornell University, Ithaca, NY, USA, in 2000, 2003, and 2005, respectively. From 2005 to 2006, he was a Post-Doctoral Research Associate with the Department of Psychology, Cornell University. From 2006 to 2015, he was with the Faculty of the School of Electrical and Computer Engineering, Oklahoma State University, USA. He is currently an Associate Professor with the Department of Electrical and Electronic Engineering, Shizuoka University, Japan. His research interests include image processing, data compression, computational vision, natural scene statistics, and visual perception.

Xuanqin Mou (M'08) has been with the School of Electronic and Information Engineering, Institute of Image Processing and Pattern Recognition (IPPR), Xi'an Jiaotong University, since 1987. He has been an Associate Professor since 1997 and a Professor since 2002. He is currently the Director of IPPR and the Director of the National Data Broadcasting Engineering and Technology Research Center. He has authored or co-authored over 200 peer-reviewed journals or conference papers. He served as a member of the 12th Expert Evaluation Committee for the National Natural Science Foundation of China, the Executive Committee of the China Society of Image and Graphics, and the Executive Committee of the Chinese Society for Stereology, and the Deputy Director of the CT committee of the Chinese Society for Stereology. He received the Yung Wing Award for excellence in education, the KC Wong Education Award, the Technology Academy Award for invention by the Ministry of Education of China, and the Technology Academy Awards from the Government of Shaanxi Province, China.