Deep Convolutional Neural Models for Picture Quality Prediction

Jongyoo Kim, Member, IEEE, Hui Zeng, Deepti Ghadiyaram, Student Member, IEEE,
Sanghoon Lee, Senior Member, IEEE, Lei Zhang, Senior Member, IEEE,
and Alan C. Bovik, Fellow, IEEE,

Abstract

Convolutional neural networks (CNNs) have been shown to deliver standout performance in a wide variety of visual information processing applications. However, this rapidly developing technology has not yet been applied with systematic energy to the problem of picture quality prediction, primarily because of limitations imposed by a lack of adequate ground truth human subjective data. This situation has begun to change with the development of promising data-gathering methods that are driving new approaches to deep learning-based perceptual picture quality prediction. Here we assay progress in this rapidly evolving field, focussing in particular on new ways to collect large quantities of ground truth data, and on recent CNN-based picture quality prediction models that deliver excellent results on a large real-world picture quality database.

Index Terms

Convolutional neural network, deep learning, perceptual picture quality, image quality assessment, no-reference image quality assessment.

I. INTRODUCTION

Recent years have seen significant efforts applied to the development of successful models and algorithms that can automatically and accurately predict the perceptual quality of 2D and 3D digital images...
and videos as reported by human viewers [1]. Concurrently, there has been a tremendous surge of work on exploiting large datasets of annotated image data as inputs to deep neural networks, towards solving such challenging problems as image classification and recognition [2]. These efforts have often produced dramatic improvement relative to the state-of-the-art. While in some regards it is not unexpected that very deep models having universal representation capability, trained on very massive image datasets to optimize hundreds of thousands or millions of parameters on exceptionally fast graphical computing architectures, should produce excellent results. Yet the generalization capability of these models is remarkable.

Yet, until recently, there has been limited effort directed towards marrying picture quality prediction with deep networks, although in principal, this could lead to greatly improved performance. The practical significance of the problem, and the relative ease of implementing algorithms learned on deep architectures make this a compelling issue. Indeed, the explosive consumption of visual media in recent years, owing to advances in digital camera technology, digital television, streaming video services, and social media applications, is driving a critical need for improved picture quality monitoring. For example, around 6 billion digital photos have been uploaded to Flickr, 300 hours of videos are uploaded to YouTube every minute, and streaming Netflix videos alone account for more than a third of North American bandwidth consumption. The pipelines from picture content generation to consumption are fraught with numerous sources of distortions, including blur, noise, and artifacts arising from such processes as compression, scaling, format conversion, color modification, and so on. Multiple interacting distortions are often present, which greatly complicates the problem. Picture quality models that can accurately predict human quality judgments can be used to greatly improve consumer satisfaction, via automatic monitoring of the qualities of massively distributed pictures and videos, and to perceptually benchmark picture processing algorithms such as compression engines, denoising algorithms, and super-resolution systems that substantially affect viewed picture quality. While many successful picture quality models have been devised, the problem is hardly solved, and there remains significant scope for improvement [3]. Deep learning engines offer a potentially powerful framework for achieving sought-after gains in performance; however, as we shall explain, progress has been limited by a lack of adequate amounts of distorted picture data and ground truth subjective quality scores, which are much harder to acquire than other kinds of labeled image data. Further, typical data-augmentation strategies such as those used for machine vision are of little use on this problem.

---

1https://www.flickr.com/photos/franckmichel/6855169886
2http://www.statisticbrain.com/youtube-statistics/
A. Perceptual Picture Quality Prediction

Picture quality models are generally classified according to whether an ostensibly pristine reference image is available for comparison. Full-reference and reduced-reference models assume that a reference is available; otherwise, the model is no-reference, or blind. Reference models are generally deployed when a process is applied to an original image, such as compression or enhancement. No-reference models are applied when the quality of an original image is suspect, as in a source inspection process, or when analyzing the output of a digital camera. Generally, no-reference prediction is a more difficult problem.

Both reference and no-reference picture quality models rely heavily on principles of computational visual neuroscience, and/or on highly regular models of natural picture statistics [1]. Given the unavailability of the useful information afforded by reference comparisons, successful no-reference models have instead used powerful, but shallow regression engines to achieve results that approach the prediction accuracy of reference quality predictors.

B. Deep Learning and Convolutional Neural Networks

The use of deep learning has had a transformative impact on such difficult problems as speech recognition and image classification. One of the principal advantages of deep learning models are the remarkable generalization capabilities that they can acquire when they are trained on large scale labeled datasets. Whereas models learned using conventional machine learning methods are heavily dependent on the determination of, and discrimination capability of sophisticated training features, by contrast, deep learning models employ multiple levels of linear and nonlinear transformations to generate highly general data representations, greatly decreasing dependence on the selection of features, which are often reduced simply to raw pixel values [2], [4]. In particular, deep convolutional neural networks (CNNs) optimized for image recognition and classification have greatly outperformed conventional methods. Open source frameworks such as TensorFlow [5] have greatly increased the accessibility of deep learning models, and their application to diverse image processing and analysis problems has greatly expanded.

C. Convolutional Neural Networks

(Instead of merging.)

Unlike traditional neural networks, CNNs can be adapted to effectively process high-dimensional, raw image data such as RGB pixel values. Two key ideas underlie a convolutional layer: local connectivity and shared weights. Each output neuron of a convolutional layer is computed only on a locally connected subset of the input, called a local receptive field (drawing from vision science terminology). However, by stacking multiple convolutional layers, the effective receptive fields may enlarge to capture global picture
characteristics. Usually, the parameters in a layer (i.e. filter weights) are shared across the entire visual field to limit their number. A common conception is that CNNs resemble precessing by neurons in visual cortex. This idea largely arises from the observation that, in deep convolutional networks deploying many layers of adaptation on images, early layers of processing often resemble the profiles of low-level cortical neurons in area V1, viz., directionally tuned Gabor filters [6], or neurons in visual area V2 implicated in assembling low-level representations of image structure [7]. At early layers of network abstraction, these “perceptual” attributes make them appealing tools for adaption to the picture quality prediction problem.

An example of a CNN structure similar to those studied here is shown in Fig. 1, where the learned kernels and obtained feature maps when the model is trained for the picture quality prediction task, are also shown. Generally, a CNN model consists of several convolutional layers followed by fully-connected layers. Some convolutional layers may be followed by pooling payers which reduce the sizes of the feature maps. The fully-connected layers are essentially traditional neural networks, where all of the neurons in a previous layer are connected to every neuron in a current layer.
II. OVERVIEW OF THE PROBLEM

Machine learning has played an important role in the development of modern picture quality models. Although these models have been largely driven by features drawn from meaningful quantitative perceptual models, mapping them against the wide variety of generally non-linear, often commingled, and poorly understood distortions that occur in practice is a formidable problem. Sophisticated, yet shallow mapping engines such as support vector regressors (SVR), have produced good prediction results (against human quality opinions), yet there remains substantial room for improvement [3], which greatly motivates the study of deep learners for this problem. Figure 2 shows conceptual flow diagrams of reference and no-reference learned picture quality predictors. A major difference of deep CNN models is the lack of a feature extraction stage, although preprocessing steps may still be put to effective use. In a deep CNN, features conducive to effective picture quality prediction are ostensibly learned by the network during the training process. The pre-processing stages may include, for example, color conversion, debiasing, (divisive) normalization, or a domain transformation.

Most popular learned picture quality prediction models operate by regressing an extracted perceptual feature vector onto recorded subjective scores. Typically shallow regressors, such as SVRs, general regression neural networks, or random forests have been used [8]–[10]. A deep CNN model can instead alternate feature extraction and regression stages. High-dimensional (raw or pre-processed pixel values) input data can be fed into the CNN, and over many iterations or epochs of training on a large dataset, useful image representations are learned automatically. In the early layers of a deep CNN, low-level encoding or sparsifying features are learned, possibly followed by intermediate descriptors of feature correlations [7]. In the deeper layers, the learned features contain more abstract information that can capture relationships between image distortions and human perceptions of them. In a CNN, differentiable feature aggregation or pooling stages are interspersed with feature extraction and regression stages, enabling effective end-to-end optimization. However, despite significant successes on a wide array of other image analysis problems, the application of deep learning networks to the picture quality prediction problem has been complicated by a number of significant obstacles which we summarize in the following.

A. Challenges Confronting CNN-Based Picture Quality Prediction

1) Lack of training data: The performance of deep learning models generally depends heavily on the size of the available training dataset(s). Currently available legacy public-domain subjective picture quality databases such as LIVE IQA [11] and TID2013 [12] are far too small to effectively train deep learning models. For example, the LIVE IQA and TID2013 databases each contain fewer than a microscopic 30 unique image contents, and no more than 24 different types of distortions per image, all of which are
Fig. 2. Flow chart comparisons of traditional learning-based and CNN-based reference and no-reference picture quality models. Blue boxes indicate learned processes.

Synthetic. Even the recent LIVE “In the Wild” Challenge Database (hereafter “LIVE Challenge”) [3], the largest available resource in these dimensions (with nearly 1,200 unique picture contents, each afflicted by a unique, unknown combination of authentic distortions, judged by more than 350,000 unique human subjects) is of insufficient size, although it provides an excellent challenge for any no-reference model. By comparison, recognition datasets such as ImageNet [13] contain tens of millions of labeled images. Creating larger subjective quality datasets is a formidable problem. Controlled laboratory studies like [11], [12] are out of the question, and even the study in [3] exhausted the pool of high-quality human subjects available on Amazon Mechanical Turk.

Common strategies for attacking this problem are data augmentation techniques, which seek to multiply the image data via rotations, cropping, reflections, and so on. Unfortunately, with the likely exception of horizontal reflections, which we use later, applying these kinds of transformations to an image will generally significantly change its perceived quality. While generating large numbers of picture content is simple, ensuring adequate distortion diversity and realism is much harder. Obtaining adequate human subjective labels is beyond current reach.

In another common strategy, the images being trained on are divided into many small patches. However, this approach produces another problem, that distinct ground truth subjective labels are not available for the patches. In every experimental scenario to date, human subjects supply a single scalar subjective score
on each global image. Since images, distortions of images, and human perceptions of both are all highly non-stationary, the scores that subjects would apply to an image patch will generally differ greatly from those applied to the entire image. Obtaining human judgments of image patch quality is not practical, nor would it reduce the overhead of acquiring human scores.

2) Different purposes of feature learning: One way to try to overcome the lack of an adequate training dataset is to utilize unsupervised learning, such as a restricted Boltzmann machine or an autoencoder [4]. Unsupervised methods attempt to learn meaningful features from unlabeled data. For example, on the image recognition task, the learning process may promote shape-related features while suppressing small variations. For example, a denoising autoencoder can be trained to reconstruct an original image from a noisy one by enforcing robustness against small corruptions of the input data, or by adding a regularization term to the objective function. By contrast, the representations learned by a picture quality predictor must be particularly sensitive to local and global degrees of distortion, as well as to perceived interactions between content and distortion. Successful, generalizable, deep unsupervised picture quality prediction models have not yet been reported.

3) Difficulty of modeling perception: The perception of picture distortion engages multiple complex processes along the visual pathway, including bandpass, multiscale, and directional decompositions [6], local nonlinearities, and normalization mechanisms. For example, contrast masking [14], whereby the spatially localized energy of image content can reduce or eliminate the visibility of distortions, is well-explained by a local cortical divisive normalization model [15]. Successful reference and no-reference picture quality models [8], [9], [14], [16], generally approximate these perceptual mechanisms by various models; however, errors in these approximations, along with a lack of information describing other relevant, perhaps higher-level processes, still limit their prediction efficacy [3].

B. Overview of Conventional Learning-Based Picture Quality Predictors

The most successful reference picture quality predictors, such as those deployed by the television industry, are not learned models, but instead compute similarity or error measures modulated by perceptual criteria in some manner. Performance is high since a reference error, whether implicit or explicit, is available to be analyzed using perceptual models. No-reference models operate without the benefit of an implied error signal, so their design has relied heavily on machine learning. Broadly, these models deploy perceptually relevant, low-level feature extraction mechanisms based on simple, yet highly regular,

---

4Such as the Emmy-winning SSIM model [14] and the VIF index [17], which is at the core of the VMAF processing system that quality-controls all Netflix content encodes.
parametric models of good-quality pictures. These “natural scene statistics” (NSS) models are predictably altered by the presence of distortions [17]. Simply stated, high-quality images subjected to bandpass filtering, followed by local energy normalization, become substantially decorrelated and gaussianized, while distorted images tend not to obey this model. Picture quality prediction models of this type have been developed in the wavelet [17], discrete cosine transform, and spatial domains [8]. Other approaches have used unsupervised learning to construct visual codebooks by clustering on normalized image patches. Quality prediction is then accomplished by soft-assigning coded image patches, followed by max pooling, then regressing onto subjective quality scores [10]. The FRIQUEE model [18] achieved state-of-the-art performance on the LIVE Challenge database [3] by regressing on a “bag” of NSS features drawn from diverse color spaces and perceptually-motivated transform domains.

There have also been attempts to apply various deep learning models to the no-reference picture quality prediction problem. For example, Hou et al. trained a deep belief network (DBN) on wavelet domain NSS features to classify distorted images into five discrete score categories [16], while Li et al. regressed shearlet NSS features onto subjective scores using a stacked auto-encoder [19]. While competitive performance was obtained by many of these models on legacy databases, none has been trained on very large datasets or tested on LIVE Challenge [3], and none have deployed deep CNNs.

III. CNN-BASED PICTURE QUALITY PREDICTION

A. CNN-Based No-Reference Picture Quality Models

As mentioned, several CNN-based picture quality prediction models have attempted to use patch-based labeling to increase the set of informative training samples. Generally, two types of training approaches have been used: patch-wise and image-wise, as depicted in Fig. 3. In the former, each image patch is independently regressed onto its target. In the latter, the patch features or predicted scores are aggregated or pooled, then regressed onto a single ground truth subjective score.

The first application of a spatial CNN model to the picture quality prediction problem was reported in [20], wherein a high-dimensional input image was directly fed into a shallow CNN model without finding handcrafted features. To obtain more data, each input image was subdivided into small patches as a method of data augmentation, each being assigned the same subjective quality score during training. The CNN model included local divisive normalization stage as well as both max and min pooling stage. Patch-wise training was used, and during application, the predicted patch scores were averaged to obtain a single picture quality score.

Li et al. utilized a deep CNN model that was pre-trained on the ImageNet dataset [21]. A network-in-network (NiN) structure was used to enhance the abstraction ability of the model. The final layer of the
pre-trained model was replaced by regression layers which mapped the learned features onto subjective scores. As in [20], image patches were regressed onto identical subjective quality scores during training.

The labeling of local patches with global subjective quality scores during training is quite problematic. While the reported prediction accuracy of these model was competitive with that of handcrafted feature-based quality prediction models, it is not reasonable to expect local image quality to closely agree with global subjective scores, even when synthetic distortions are applied homogeneously. Picture quality is inevitably space-varying, because of the high degree of non-stationarity of picture contents and the complex perceptual interactions between content and distortions (such as masking). Towards solving this problem, image-wise training strategies have been studied.

Bosse et al. deployed a deeper, 12-layer CNN model fed only by raw RGB image patches to learn a no-reference picture quality model [22]. They proposed two training strategies: ‘patch-wise training’ (similar to [20]), and ‘weighted average patch aggregation,’ whereby the relative importance of each patch was weighted by training on a subnetwork. The overall loss function was optimized in an end-to-end manner. The authors reported state-of-the-art prediction accuracies on the major synthetic-distortion picture quality databases.

To overcome overfitting problems, several authors have suggested training deep CNN models in two separate stages: a pre-training stage, using a large number of proxy ground truth quality scores, followed
by a stage of regression onto a smaller set of subjective scores. For example, [23] describes a two-stage CNN-based no-reference quality prediction model, whereby local quality scores generated by a full-reference algorithm are used as proxy patch labels, in a first stage of training. In the second stage, the feature vectors obtained from image patches are aggregated using statistical moments, then regressed onto subjective scores. In this instance, the first stage is patch-wise training, while the second stage is image-wise training. Since the local proxy scores reflect the non-stationary characteristics of perceived quality, they are reasonable local regression targets, and training of the CNN model is enabled by the abundant training samples. Following the second stage of training on human ground truth, their model attains highly competitive prediction accuracy on the legacy datasets.

The same authors later developed a two-stage training scheme for no-reference picture quality prediction called the Deep Image Quality Assessor (DIQA) [24]. The training process of that model was separated into an objective training stage, followed by a subjective training stage. Rather than using a sophisticated picture quality predictor to produce proxy scores, they instead efficiently compute PSNR (or MSE) maps. Unlike prior methods, in the first stage, the input images are not broken into patches. Using only convolutional layers, high-dimensional feature maps are obtained, which are then regressed onto high-dimensional objective error maps. The second stage aggregates the high-dimensional feature maps by weighted averaging, then regresses these global features onto ground truth subjective scores. The weighting maps are also learned during training. The reported prediction accuracy of these models is competitive with state-of-the-art models on the legacy databases.

B. CNN-Based Full-Reference Picture Quality Models

While CNNs were first used to model no-reference picture quality, more recently, they have been applied to the reference prediction problem as well.

Liang et al. [25] proposed a dual-path CNN-based full-reference quality prediction model. They generalized the problem by seeking to predict quality using a non-aligned image of a similar scene as reference. Locally normalized distorted and reference image patches are fed into a dual-path CNN model, each using the same parameter values. Then the concatenated learned feature vectors are regressed onto the subjective scores of source distorted images. They report state-of-the-art prediction accuracies in both aligned and non-aligned full-reference scenarios.

Gao et al. deployed a deep CNN model pre-trained on ImageNet. They used it to conduct full-reference picture quality prediction [26] by feeding pairs of reference and distorted pictures into the CNN, where each output layer is used as a feature map. Local similarities between the feature maps obtained from the reference and distorted images are then computed and pooled to arrive at global picture quality scores.
Table I compares the implementations of reported CNN-based no-reference picture quality models. ‘Conv’ indicates convolutional layers, and ’FC’ indicates fully-connected layers.

<table>
<thead>
<tr>
<th>Models</th>
<th>Layer depth</th>
<th>Preprocessing</th>
<th>Feature aggregation or score pooling</th>
</tr>
</thead>
<tbody>
<tr>
<td>[20]</td>
<td>2 Conv &amp; 2 FC</td>
<td>Local normalization</td>
<td>Mean pooling (during testing)</td>
</tr>
<tr>
<td>[21]</td>
<td>14 Conv (4 NiN blocks)</td>
<td>Local normalization</td>
<td>Mean pooling (during testing)</td>
</tr>
<tr>
<td>[22]</td>
<td>10 Conv &amp; 2 FC&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Raw RGB image</td>
<td>Mean or weighted average pooling</td>
</tr>
<tr>
<td>[23]</td>
<td>2 Conv &amp; 6 FC</td>
<td>Local normalization</td>
<td>Mean and standard deviation aggregation</td>
</tr>
<tr>
<td>[24]</td>
<td>8 Conv &amp; 3 FC&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Low-freq. subtraction</td>
<td>Mean or weighted average aggregation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Models</th>
<th>Training targets</th>
<th>1st stage</th>
<th>2nd stage</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>[20]</td>
<td>Subjective scores</td>
<td>N.A.</td>
<td></td>
<td>Patch-wise training</td>
</tr>
<tr>
<td>[21]</td>
<td>Semantic label</td>
<td>Subjective scores</td>
<td></td>
<td>Fine-tuning of pre-trained CNN on ImageNet</td>
</tr>
<tr>
<td>[22]</td>
<td>Subjective scores</td>
<td>N.A.</td>
<td></td>
<td>Weighted average patch aggregation</td>
</tr>
<tr>
<td>[23]</td>
<td>Full-reference scores</td>
<td>Subjective scores</td>
<td></td>
<td>Uses proxy patch labels</td>
</tr>
<tr>
<td>[24]</td>
<td>Objective error map</td>
<td>Subjective scores</td>
<td></td>
<td>Uses proxy patch labels</td>
</tr>
</tbody>
</table>

<sup>a</sup>Patch-wise model in [22]

<sup>b</sup>DIIQA-ERR model in [24]

The deep CNN-based full-reference quality prediction model in [27], called DeepQA, takes the distorted image and an objective error map (e.g. MSE) as inputs. The model is trained to compute a visual sensitivity weight at each coordinate, from measured local spatial characteristics of the distorted image. The authors reported consistent state-of-the-art prediction accuracies compared to other reference quality models, on the synthetic-distortion legacy picture quality databases.

C. Summary of CNN-Based No-Reference Picture Quality Models

Table I compares the implementations of reported CNN-based no-reference picture quality models. Most of the models consist of convolutional layers followed by fully-connected layers. Most of the methods employ local divisive normalization [20]. Several of the models use proxy ground truth scores generated by objective quality prediction models to augment the subjective scores, or alternately, to pre-train the network on a large amount of data before fine-tuning on subjective scores.

In CNN-based schemes, the process of feature aggregation or score pooling determines the form of a loss function. Examples of aggregation and pooling strategies are shown in Fig. 4. The patch-based
algorithms described in [20], [21], did not use aggregation or pooling during training. Instead, each image patch was independently regressed onto the global subjective quality score. The loss function used is

$$L = \frac{1}{N} \sum_{i} \| f(p_i) - S \|,$$  

(1)

where $p_i$ refers to the $i$-th patch obtained, $N$ is the number of patches, $S$ is the ground truth score, and $f(\cdot)$ is a neural network process. The models were trained via a path-wise optimization, and during testing, the outputs of multiple patches composing an input image were averaged to obtain a final predicted subjective score. Conversely, image-wise approaches use aggregation or pooling during training. For example, weighted average pooling methods [22] may be used, where the loss function looks like

$$L' = \| \text{pool}(f(p_1), ..., f(p_N)) - S \|$$  

(2)

where $\text{pool}(\cdot)$ refers an unspecified pooling method (Fig. 4(a)). In [23] and [24] (Figs. 4(b) and (c)), feature aggregation was used. A more complicated model, such as a recurrent neural network (RNN) [4], could also be used for aggregation (Fig. 4(d)). Here, the loss function becomes

$$L'' = \| g(\text{aggr}(f(p_1), ..., f(p_N))) - S \|$$  

(3)
where \( \text{aggr}(\cdot) \) refers to a feature aggregation process, and \( g(\cdot) \) is a regression neural network. (2) and (3) have the advantage that the model can be trained under the same conditions as the actual testing conditions where the image-wise scores are predicted.

**D. Description on Picture Quality Databases**

The choice and consideration of database for training is important for deep learning-based models, since their performance depends highly on the size of the training set. In most picture quality databases, the distorted images are afflicted by only a single type of synthetically introduced distortion, such as JPEG compression, simulated sensor noise, or simulated blur, as exemplified in Fig. 5(a). Since they have played important roles in the development of perceptual picture quality studies, we briefly describe several popular legacy databases.

The LIVE IQA database [11], which was the first successful public-domain picture quality database, and is still the most widely-used, contains 29 reference images and 982 images, each impaired by one of five types of synthetic distortions: JPEG and JPEG2000 (JP2K) compression, white Gaussian noise (WN), Gaussian blur (GB), and Rayleigh fast-fading channel distortion (FF). The differential Mean Opinion Score (DMOS) of each distorted image is provided. The CSIQ database [28] includes 30 reference images and 866 synthetically distorted images of six types: JPEG, JP2K, WN, GB, pink Gaussian noise...
TABLE II

Comparison of IQA databases in terms of numbers of reference images (Ref.), distorted images (Dist.), distortion types, authenticity of distortions, type of subjective scores, and whether distortions are mixed.

<table>
<thead>
<tr>
<th>Database</th>
<th>Ref.</th>
<th>Dist.</th>
<th>Dist. types</th>
<th>Synthetic/Authentic</th>
<th>Score type</th>
<th>Mixtures of distortions</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSIQ [28]</td>
<td>30</td>
<td>866</td>
<td>6</td>
<td>Synthetic</td>
<td>DMOS</td>
<td>N.A.</td>
</tr>
<tr>
<td>TID2013 [12]</td>
<td>25</td>
<td>3,000</td>
<td>24</td>
<td>Synthetic</td>
<td>MOS</td>
<td>N.A.</td>
</tr>
<tr>
<td>LIVE MD [29]</td>
<td>15</td>
<td>405</td>
<td>2</td>
<td>Synthetic</td>
<td>DMOS</td>
<td>✓</td>
</tr>
<tr>
<td>LIVE Challenge [3]</td>
<td>N.A.</td>
<td>1,162</td>
<td>Numerous</td>
<td>Authentic</td>
<td>MOS</td>
<td>✓</td>
</tr>
</tbody>
</table>

(PGN), and global contrast decrements (CTD). The DMOS of the distorted images is also provided. TID2013 [12] contains the largest number of distorted images. It consists of 25 reference images and 3,000 synthetically distorted images with 24 different distortions at five levels of degradation. The database also provides the Mean Opinion Scores (MOS). The LIVE MD database [29], was the first to include multiply (synthetically) distorted images. Images in it are distorted by two types of distortion in two combinations: Gaussian blur followed by JPEG compression, Gaussian blur followed by white and Gaussian noise. It contains 15 reference and 405 distorted images, and the DMOS of each distorted image is provided.

Lastly, the LIVE Challenge Database [3] contains nearly 1,200 unique image contents, captured by a wide variety of mobile camera devices under highly diverse conditions. As such, the images were subjected to numerous types of authentic distortions during the capture process, often in complex combinations of multiple interacting impairments, as shown in Fig. 5(b). The distortions include, for example, low-light blur and noise, motion blur, camera shake, overexposure, underexposure, a variety of color errors, compression errors and many combinations of these and other impairments. There are no reference images in the LIVE Challenge database, since the distorted images are originals, captured by dozens of ordinary photographers. The LIVE Challenge pictures were judged by more than 8,100 human subjects in a tightly monitored crowdsourced study, yielding more than 350,000 human judgments that exhibit excellent internal consistency [3].

A summary of the attributes of these five databases is shown in Table II.

IV. PERFORMANCES OF CNN PICTURE QUALITY MODELS

Since only a few CNN-based picture quality models have been released, we provide the prediction accuracies of baseline models on the five databases as performance references to be compared against. We
selected the well-known CNN models, AlexNet [2], and ResNet50 [30], as the architectures of the baseline models, where each was pre-trained on ImageNet. For each pre-trained architecture, two types of back-end training strategies were tested: using an SVR to regress the extracted features from the CNN model onto subjective scores, and fine-tuning the pre-trained networks to conduct picture quality prediction. We did not test direct training of these models on any of the picture quality databases, since they are not large enough. Very deep networks easily overfit on insufficient training samples, causing significant decreases in testing accuracy (AlexNet has 62 million and ResNet50 has 26 million parameters). Instead, we tested a smaller CNN network as a baseline model of direct training.

In the first approach, the output of 6-th fully connected layer (4096 dimensions) from AlexNet and averaged-pooled features (2048 dimensions) from ResNet50 were used as the input feature vectors to the SVR. From each input image, 25 randomly cropped image patches were used for training and testing. The obtained feature vectors from 25 image patches were averaged to obtain a single global feature vector.

In the second approach, we randomly cropped 100 image patches from each training image to be used for training (except on the TID2013 database, where 30 cropped patches were used due to the large number of distorted images in the database). The image patches inherited the quality scores from the source distorted images which are first normalized into (0,1). This preprocessing enables us to use the same parameter setting on all databases. The basic regression loss (1) was used. To alleviate overfitting, one dropout layer with dropout rate 0.5 was added before the last fully-connected layer. The learning rate was set to $10^{-3}$, and the fine-tuning process iterated for 8 and 6 epochs on AlexNet and ResNet50, respectively. The batch size was fixed at 48 for both models. In the testing stage, the pre-trained models were used to predict quality scores on each of 25 random image crops. These were average pooled to produce the final picture quality scores.

For the direct training approach, we used the following CNN architecture: Conv-48, Conv-48 with stride 2, Conv-64, Conv-64 with stride 2, Conv-64, Conv-64, Conv-128, Conv-128, FC-128, FC-128, and FC-1. Here, “Conv” refers to convolutional layers, “FC” refers to fully-connected layers, and the trailing numbers indicate the number of feature maps (of Conv) or output nodes (of FC). All of the convolutional layers were configured to use $3 \times 3$ filters, using zero-padding to preserve the spatial size. Following the convolutional layers, each feature map is averaged, then fed into the fully-connected layers. The number of parameters in this model is about 0.4 million, which is much lower than AlexNet and ResNet50. This baseline model was trained using the image-wise $L_2$ loss in (3). Each input image was partitioned into $112 \times 112$ patches when training on the LIVE IQA database, while full-sized images were used on the other databases. The data was augmented by supplementing the training set with horizontally flipped replicas of each image. Each minibatch contained patches extracted from five images. The training was
iterated over 80 epochs.

Two performance metrics were used to benchmark the models: Spearman’s rank order correlation coefficient (SRCC), and Pearson’s linear correlation coefficient (PLCC). To evaluate the baseline models,
we randomly divided each database into two subsets of non-overlapping content (distorted or otherwise), 80% for training and 20% for testing. Of course, all of the LIVE Challenge pictures contain different contents. The SRCC and PLCC were averaged after 10 repetitions of this random process.

The performance of all of the exemplar picture quality prediction models on the LIVE IQA database are shown in Fig. 6. The first five are no-reference learning-based models, where the last two of these used deep learning. The next seven are CNN-based no-reference quality prediction models, and the last three are CNN-based full-reference models. The reported SRCC and PLCC scores of the listed models were taken from the original papers. Overall, the CNN-based full-reference models followed by CNN-based no-reference models achieved higher prediction accuracies relative to conventional learning-based models on the legacy databases.

Table III compares the performance of the various picture quality prediction models on all of the reviewed databases. The last five rows show results for the baseline models. The three best models performing no-reference picture quality models are shown in bold. Generally, the existing CNN-based models were able to achieve remarkable prediction accuracies on the legacy databases. However, it is much harder to obtain successful results on the LIVE Challenge database. For example, two models proposed in [24], DIQA-ERR and DIQA-NR, achieved SRCC of 0.701 and 0.627 respectively, which are lower than the results attained by a recent successful SVR-based method, FRIQUEE-ALL [18] which achieved an SRCC of 0.72.

However, the baseline models that were pre-trained on the ImageNet databases achieved standout accuracies on the LIVE Challenge database. This is likely because the real-world ImageNet pictures are not synthetically distorted. Like the LIVE Challenge pictures, any distortions occurred as a natural consequence of photography, without intervention by the database creator. This further suggests that the pre-trained CNNs are, to same degree, already “quality-aware,” meaning that their learned internal features assist the performance of the task (recognition) in the presence of authentic distortions.

The baseline models using the first approach achieved very low accuracies on the legacy databases, since they were not exposed to any synthetic distortions during training. Fine-tuning of the pre-trained baseline models significantly improved both the SRCC and the PLCC in most cases, with one exception of AlexNet on LIVE Challenge, where the performance was slightly lower than using SVR. The fine-tuning brought larger improvement when the target domain (legacy picture quality databases) has larger difference to the source domain (ImageNet). However the fine-tuning models did not achieve superior accuracies on the legacy databases, whereas the directly trained CNN baseline model achieved competitive performance on the legacy databases, but lower accuracies on the LIVE Challenge database. A possible explanation for the above results is that the pre-trained models adapted easily to the authentic distortions.
in LIVE Challenge as a consequence of learning image recognition tasks on real-world pictures. Applying them to databases with synthetic distortions, like LIVE IQA and TID2013, likely failed to exploit what was learned regarding authentic distortions, hence significant retraining would be needed to deal with the synthetic distortions, which could not be solved by the SVR taking the pre-trained CNN outputs as inputs. This may help explain the excellent generalization power of pre-trained models when applied to other real world image tasks: their ability to handle authentic distortions.

V. ENVISIONING THE FUTURE

The size of the training sets is critical to the success of deep neural network models. Current public-domain databases have insufficient size as compared to widely-used image recognition databases. However, constructing large-scale perceptual quality databases is a much more difficult problem than image recognition databases. Creating databases for picture quality assessment requires time-consuming and expensive subjective studies, which must be conducted under controlled laboratory conditions. Even if the number of reference images is small, the required number of subjective tests quickly becomes excessive. Conducting subjective tests using online crowdsourcing is one possible solution (like the LIVE Challenge database), yet even online tests are (probably) prohibitively difficult to scale up to the necessary size, especially while ensuring the aggregate quality of the collected human data. Another possibility would be if a large social media company engages their customers to provide picture quality scores, similar to the Netflix DVD ratings-by-email of a decade ago. Generally, understanding how to successfully create reliable, very large-scale, and authentic picture quality databases remains an open question.

REFERENCES


REFERENCES


