Blind Image Quality Assessment on Real Distorted Images using Deep Belief Nets

Deepti Ghadiyaram
The University of Texas at Austin

Alan C. Bovik
The University of Texas at Austin

***Abstract*—We present a novel natural-scene-statistics-based blind image quality assessment model that is created by training a deep belief net to discover good feature representations that are used to learn a regressor for quality prediction. The proposed deep model has an unsupervised pre-training stage followed by a supervised fine-tuning stage, enabling it to generalize over different distortion types, mixtures, and severities. We evaluated our new model on a recently created database of images afflicted by real distortions, and show that it outperforms current state-of-the-art blind image quality prediction models.**

***Index Terms*—Perceptual quality, deep belief nets, blind image quality assessment, natural scene statistics.**

I. Introduction

Objective blind or no-reference (NR) image quality assessment (IQA) models are algorithms that can automatically predict the perceptual quality of images such that the only information that the algorithm has available is the distorted image whose quality is to be ascertained. Because of the surge in visual media content across the internet, IQA algorithms are fast gaining importance. These algorithms are used for monitoring and controlling multimedia services on wired and wireless networks and devices with an aim to ensure that end users have a satisfactory quality of experience (QoE).

The most efficient NR IQA algorithms to date are founded on natural scene statistical (NSS) models [1] that capture the *naturalness* of images that are not distorted. These models are based on the well-founded observation that good quality real-world photographic images obey certain perceptually relevant statistical laws that are violated by the presence of common image distortions. State-of-the-art NSS-based NR IQA models [2] - [6] exploit these statistical perturbations by first extracting image features and then learning a kernel function that maps these features to ground truth subjective quality scores. We refer readers to [2] for a detailed comparison of the performance of several blind IQA models.

To date, the performance of these techniques has been gauged only on legacy databases such as the LIVE Image Quality Database [7] and the TID2008 database [8], which contain images corrupted by only one of a few synthetically introduced distortions viz., images containing only JPEG-compression artifacts or images corrupted by simulated camera sensor noise. However, in practice, every image captured by a typical mobile digital camera device passes through numerous processing stages, each of which could potentially introduce visual artifacts. Thus, authentically distorted images are likely to be impaired by a broad range of diverse quality "types," mixtures, and distortion severities.

Challenging image dataset: The lack of diversity and realism of distortions in existing, widely-used image quality databases [7], [8] impedes our goal to be able to model and predict the perception of real image distortions. To overcome this limitation and towards being able to design robust, perceptually-aware image assessment models, we designed and created a new image quality database called the **LIVE Blind Image Quality Challenge Database** [9], that contains images afflicted by diverse authentic distortions and genuine artifacts captured using a variety of commercial devices. Each image was collected without artificially introducing any distortions beyond those occurring in each camera device during the capture, processing, and storage processes. The database consists of 1,163 images afflicted by varied artifacts such as low-light noise and blur, motion-induced blur, over and underexposure, compression errors, and so on. See [9] for a more detailed description of this unique and difficult distorted image corpus.

Better image features: Existing blind IQA models that learn feature representations on images that contain only single, unmixed distortions, may not perform as well when applied on images afflicted by mixtures of distortions. The LIVE challenge database has a high percentage of images distorted by multiple processes. Our recently proposed IQA model, FRIQUEE [10] alleviates this problem to some extent by generating more informative features that predict human quality judgments better than state-of-the-art blind IQA methods.

Deeper architecture: Shallow architectures typically consist of one layer of fixed kernel functions and can sometimes be inefficient when matching the features from the training data with ground truth labels. On the other hand, deep architectures such as the Deep Belief Network [11] and the Deep Boltzmann Machine [12] progressively combine lower level inputs into more abstract and higher-level representations and have shown remarkable performance on complex tasks such as digit classification [11], visual object recognition [12], and image denoising [13]. These models, which try to learn a "deep" structure from the input data are motivated in part by the hierarchical organization of human visual cortex.

Here, we combine recently proposed natural-scene-statistics-based perceptual image features with a deep belief network and a regressor to tackle the difficult problem of blind

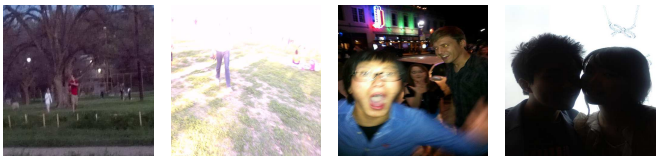

Fig. 1. Sample images from the LIVE Blind Image Quality Challenge Database [9].

image quality assessment on authentically distorted images. Our chief contribution is a robust image quality assessment model that outperforms existing techniques on a wide range of diverse and authentic distortions, as shown by the results of experiments on the LIVE Challenge database [9] (Table I). We are aware of only one other very recent project [14] reporting an effort made in the same spirit as our proposed model. There are, however, several crucial differences that distinguish our approach. First, we utilize a state-of-the-art image database containing only real distortions whereas the authors of [14] tested their model on the legacy databases [7], [8]. Second, we use a different (thinner) architecture with many fewer input units (330 vs. 16689 units in [14]) yielding a much shorter total learning time. Third, under the given problem setting, we study the representative power of image features by training different DBNs individually on features derived from several top-performing IQA models, which has not been studied in [14].

II. Blind Image Quality Assessment

A. Images and Quality Scores

Figure 1 shows a subset of images from the new challenge database [9]. Each of the 1163 images contained in the challenge database is a unique content that has been rated by many thousands of subjects via an online crowdsourcing system that we designed for subjective quality assessment [9]. The study is on-going and we have so far gathered nearly 280,000 opinion scores from over 5,000 unique subjects. The mean opinion score (MOS) of each image is computed by averaging the individual ratings across subjects. These are used as ground truth quality scores.

B. Perceptually Relevant FRIQUEE Features

We recently proposed a new quality assessment model, FRIQUEE [10], to overcome the limitations of existing blind IQA techniques with regard to representing mixtures of distortions, such as those contained in the LIVE Challenge database [9]. FRIQUEE is a natural scene statistical (NSS) model that is based on the hypothesis that different modalities capture distinctive aspects of the loss of the perceived quality of any given image. Figure 2 is a schematic illustration of some of the feature maps that are built into our model. In our framework, a total of 330 statistical features that have been observed to contribute significant information regarding distortion visibility and perceived perceptual quality of an image were selected as input to the learning model.

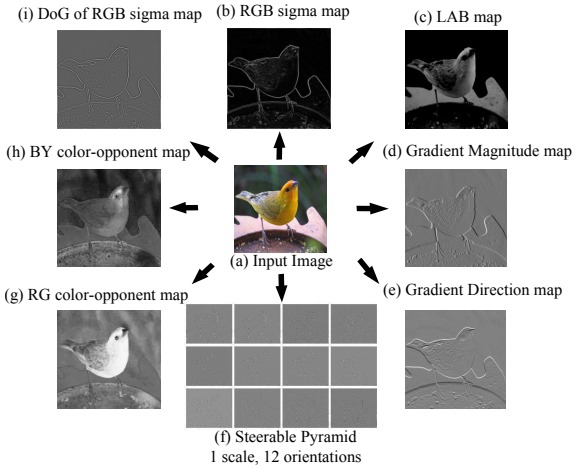

Fig. 2. Given any image (a) FRIQUEE first constructs several feature maps in multiple transform domains (some of them are shown here) and then extracts scene statistics from these maps after performing perceptually significant divisive normalization [15] on them.

C. Our Model

Briefly, a restricted boltzmann machine (RBM) is an undirected graphical model with bipartite symmetrical connections between the stochastic binary units of the visible and hidden layer with no inter-layer connections [11]. The weights on each of the connections in an RBM can be efficiently learned by first performing alternating Gibbs sampling, then employing a contrastive divergence learning technique to update all of the units in a given layer in parallel until the RBM reaches a *conditional equilibrium*. Stacking multiple RBMs and learning multiple weight matrices by treating the hidden activities of one RBM as the visible input data for training a higher-level RBM in a greedy layer-by-layer way results in a hybrid generative model called a deep belief net (DBN).

Hinton *et.al.* [11] proposed an unsupervised pre-training step where greedy layer-by-layer learning was employed to initialize the model parameters of a DBN. This was followed by a supervised fine-tuning phase where labeled data was used to further update these parameters which resulted in a model with superior classification performance than some of the shallower architectures.

Our proposed model is a combination of a deep belief network of three hidden layers (Figure 3) and a regressor. As mentioned earlier, after training the RBM of a layer $l-1$, its hidden activations serve as an input to train the next layer l. Thus, each layer captures strong, high-order correlations between the activities of the units in the layer below. Our DBN model attempts to build more complex representations of the simple statistical features provided as input, by gradually propagating them from one layer to another.

We also employ an unsupervised pre-training strategy to regulate the weight space of the deep network, followed by a supervised fine-tuning step to learn the parameters of the entire model by aligning the output of the network with the ground truth MOS values. These two stages pertaining to DBN are described in detail below, followed by a description of the

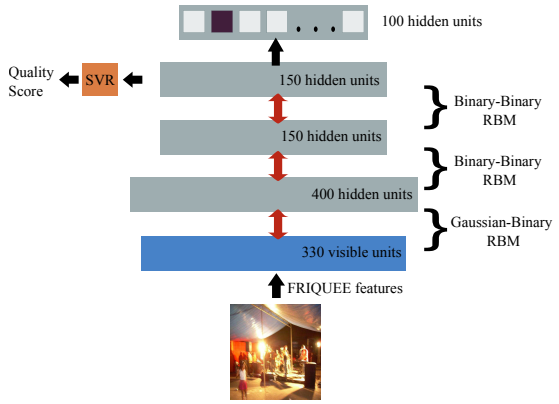

Fig. 3. Configuration of our DBN model. The unit in the topmost layer that is activated determines the class label that is used to by the cross-entropy error function.

final regression step.

1) Unsupervised pre-training: In this phase, each layer is treated as an RBM, the individual layers of the network are trained and the weights are learned greedily, one layer at a time, from the bottom up. Since the input is real-valued continuous data (FRIQUEE features), the first layer is modeled using a Gaussian-Binary RBM (GRBM) [16]. The rest of the layers are modeled as Binary-Binary RBMs (Figure 3).

First Layer: A Gaussian-Binary RBM allows us to model real-valued image features by using a Gaussian distribution to model the observed "visible" input data (**v**) and a conditional Bernoulli distribution to model the "hidden" unit values (**h**). Holding either **h** or **v** fixed, we can sample from the other as follows:

$$P(v_i = x|\mathbf{h}) = \mathcal{N}(c_i + \sigma \sum_j w_{ij} h_j) \tag{1}$$

$$P(h_j = 1|\mathbf{v}) = logistic(b_j + \sum_i w_{ij} \frac{v_i}{\sigma}) \tag{2}$$

Here $\mathcal{N}(.)$ is the Gaussian density function and $logistic(.)$ is the logistic function. The RBM is parameterized by the network weights **W**, the hidden layer bias **b**, the visible layer bias **c**, and by the standard deviation of the visible units σ. We normalize the data to have zero mean and unit variance along each feature dimension and set σ to 1 in (1) and (2) while reconstructing the hidden and visible layer probabilities.

Due to space constraints, we refer readers to [16] for technical details on computing contrastive divergence gradients to learn and evolve weights and biases.

Higher Layers: For the three hidden layers that follow the input layer, we followed the probabilistic model of Binary-Binary RBMs to sample visible and hidden unit values and a contrastive divergence step for updating the weights and biases as explained in [11].

The weights of the RBM connecting the visible and hidden units at every level are initialized from random samples from a uniform distribution $\mathcal{U}(0, 0.1)$. The first layer is trained more gently at a smaller learning rate of 0.001 for the weights as compared with that of 0.1 for the higher layers. We split the images from the challenge database into train and test sets and use only the training data in the pre-training stage. To speed up the learning process, the momentum which is 0.5 in the initial 5 epochs is increased to 0.9 for the rest of the epochs. We stop the training process after 3000 epochs.

2) Supervised fine-tuning phase: In this phase, we pose the problem of predicting the perceptual quality of an image as a classification problem, with MOS values as class labels. Since the MOS values range from $1 - 100$, each ground truth MOS value is represented as a vector of 100 binary units with exactly one unit that is indexed by the MOS value turned on and all the other units turned off. This binary representation serves as a ground truth class label and aligns well with the binary units of an RBM. Consequently, a *fourth hidden layer* with the number of hidden nodes equal to the number of classes (100) is temporarily added at the top, only to fine-tune the system. For a given input (FRIQUEE features computed on an image), the probability of each unit in the top layer is given by

$$p_i = \frac{e^{x_i}}{\sum_j e^{x_j}}, \tag{3}$$

where x_i is the total input received (from all the lower layers) by top unit i. These probabilities are binarized by turning on the unit with maximum p_i and turning off all the other units.

Thus, in the global fine-tuning phase, the weights and biases learned at every level from pre-training are retained and the quality labels of the training data are computed using (3). For every training data sample, the predicted quality label $(\widehat{t_i})$ generated by the top-most hidden layer (of 100 units) is compared with the corresponding ground truth class label t_i to minimize an objective function, which is the multi-class cross-entropy function [17] defined as follows:

$$CE = \sum_i t_i log(\widehat{t_i}) \tag{4}$$

We then employ the conjugate gradient descent technique to adjust each of the weight matrices. The gradients are obtained by backpropagating the error derivatives such that the predicted class labels from the network on the training data align with the ground truth class labels. In our experiments, the fine-tuning phase was run for 100 epochs.

3) Regression: This fine-tuned DBN model and its learned weights and biases can now be used to generate feature representations for any given image. Specifically, the train and test sets constructed in the pre-training phase are considered, the FRIQUEE features of each image are fed to the DBN as input, and the probabilities of the third hidden layer (with 150 hidden nodes) are extracted, thus generating "deep features." These deep features, along with the corresponding MOS values of the training set are used to train a support vector regressor (SVR). Following this, given any test image's deep features as input to the trained SVR, a final quality score may be predicted.

III. Experiments

We determined the number of hidden units in each RBM and the number of layers in the final model by cross-validation. Specifically, we divided the data into disjoint training and

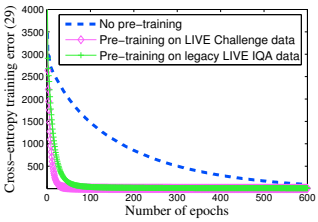

Fig. 4. Cross-entropy training errors (defined in (4)) with and without the pre-training step in our model.

testing sets. Then, for each RBM layer, we varied the number of hidden nodes and trained a series of models using the training data and evaluated the performance of the model on the test data. The model with a configuration of 400-150-150 units per layer yielded the highest prediction accuracy of the regressor and was thus chosen as the optimal model. As mentioned earlier, a fourth hidden layer of 100 units was added during the fine-tuning phase. The input layer has units equal to the number of features each algorithm extracts per image, which is 330 in the case of FRIQUEE.

Comparing different IQA techniques: We extracted the features proposed by several other prominent blind IQA algorithms (whose code was publicly available) on the images of the LIVE Blind Image Quality Challenge Database, fed them to the input units and trained each DBN model under the same train/test setting. To mitigate any bias due to division of data, we repeated this process of randomly splitting the data over 100 iterations and computed Spearman's rank ordered correlation coefficient and Pearson's correlation coefficient between the predicted and the ground truth quality scores at the end of every iteration. We report the median of these correlations across 100 iterations in Table I. A higher value for each of these metrics indicates good performance both in terms of correlation with human opinion as well as the performance of the entire model. From Table I, we conclude that the performance of our proposed model on unseen test data is significantly better than the currently top-performing state-of-the-art methods on the LIVE Challenge database [9].

Without pre-training: Next, using FRIQUEE features as input, we also tried skipping the unsupervised pre-training phase and initialized the weights at each layer to random values drawn from a uniform distribution $\mathcal{U}(0, 0.1)$. The faster drop in the error when the pre-training step was included in the model (Fig. 4) demonstrates that greedy layer-wise unsupervised pre-training on unlabeled images is a crucial step as it overcomes the challenges of deep learning by introducing a useful prior to the supervised fine-tuning training procedure and thus makes it possible to fine-tune the network efficiently.

Comparison with a different model architecture: A different experiment where an SVR is trained directly on different IQA features on the same train/test splits used earlier to report the results in Table I is conducted and again, the median of the correlation values across 100 iterations is reported in Table II. The significantly better correlation values in Table I compared to those in Table II when the same set of features are used is indicative of the ability of a deep belief network to learn and discriminate features belonging to different distortions more effectively than an SVR.

TABLE I
MEDIAN LCC AND MEDIAN SROCC ACROSS 100 TRAIN-TEST COMBINATIONS ON THE LIVE CHALLENGE DATABASE.[9] WHEN THE PROPOSED DBN WAS USED TO GENERATE "DEEP FEATURES".

	LCC	SROCC
FRIQUEE [10]	**0.7051**	**0.6721**
BRISQUE [2]	0.6204	0.6018
DIIVINE [3]	0.5577	0.5094
BLIINDS-II [4]	0.4977	0.4893

TABLE II
MEDIAN LCC AND MEDIAN SROCC ACROSS 100 TRAIN-TEST COMBINATIONS ON THE LIVE CHALLENGE DATABASE [9] WHEN SVM WAS USED.

	LCC	SROCC
FRIQUEE	**0.67**	**0.64**
BRISQUE	0.56	0.53
DIIVINE	0.50	0.48
BLIINDS-II	0.45	0.40

It can be thus be concluded that the combination of recently proposed FRIQUEE features and our deep regression model performs extremely well in comparison with all of the other models on [9]. These results illustrate the importance of perceptually significant features that are representative of authentic distortions as well as the ability of a hierarchical, non-linear model to offer the flexibility to distinctly represent those features that belong to different mixtures of distortions leading to a significant improvement in the final performance of the model.

IV. Conclusions and Future Work

We explored the problem of image quality assessment on a challenging new database of real distorted images [9] by using a deep belief net to derive informative feature representations from a bag of perceptually relevant statistical image features. These derived representations were in turn used to train a regressor that predicts a quality score. We achieved a significant improvement over previous blind IQA methods, underscoring the benefits of perceptually significant features as well as a densely connected deep learning model to generate complex feature representations. We believe our work in this direction is the first substantial effort towards designing blind IQA models for predicting the perceptual quality of images corrupted by complex distortion mixtures. Its success encourages us to explore the feasibility of developing analogous powerful blind *video* quality assessment models using space-time natural video statistics based models [18], [19] and also to practically adapt our model in real-world applications such as monitoring the quality of streamed media content. Going forward, we also believe that accounting for image content when predicting quality [20] may further improve the performance of NR IQA models.

REFERENCES

[1] A.C. Bovik, "Automatic prediction of perceptual image and video quality," *IEEE Proc.*, vol. 101, no. 9, pp. 2008-2024, Sept. 2013.

[2] A. Mittal, A.K. Moorthy, and A.C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695-4708, Dec 2012.

[3] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350-3364, Dec. 2011.

[4] M. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339-3352, Aug. 2012.

[5] P. Ye and D. Doermann, "No-reference image quality assessment using visual codebooks," *Proc. IEEE Int. Conf. Image Process.*, pp. 3129-3138, Jul. 2011.

[6] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp. 305-312, Jun. 2011.

[7] H.R. Sheikh, M.F. Sabir, and A.C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440-3451, Nov 2006.

[8] N Ponomarenko, V Lukin, A Zelensky, K Egiazarian, M Carli, and F Battisti, "TID2008-A database for evaluation of full-reference visual quality assessment metrics," *Adv of Modern Radio Electron.*,vol. 10, no. 4, pp. 30-45, 2009.

[9] D. Ghadiyaram and A.C. Bovik, "Crowdsourced study of subjective image quality," *Asilomar Conf. Signals, Syst. Comput.*, Nov 2014, Invited Paper in this proceedings, in press.

[10] D. Ghadiyaram and A.C. Bovik, "Feature maps driven no-reference image quality prediction of authentically distorted images," *In Proc. SPIE Conf. Human Vision and Electronic Imaging.*, Feb. 2015, unpublished.

[11] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, 18(7):1527-1554, 2006.

[12] R. Salakhutdinov and H. Larochelle, "Efficient learning of deep Boltzmann machines," *Int. Conf. on Artifical Intelligence and Statistics*, pp. 693-700, 2010.

[13] M. Ranzato, J. Susskind, V. Mnih, and G. E. Hinton, "On deep generative models with applications to recognition," *IEEE Conf. on Comp. Vision and Pattern Recog.*, pp. 2857-2864, 2011.

[14] H. Tang, N. Joshi, and A. Kapoor, "Blind Image Quality Assessment using Semi-supervised Rectifier Networks," *Proc. IEEE Conf. on Comp. Vision and Pattern Recog.*, 2014, in press.

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

[16] G. E. Hinton and R. Salakhutdinov, "Using Deep Belief Nets to Learn Covariance Kernels for Gaussian Processes," *Advances in Neural Info. Proc. Sys.*, pp. 1249-1256, 2008.

[17] G. E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, pp. 504-507, 2006.

[18] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circ. and Sys. Video Tech.*, vol. 23, no. 4, pp. 684-694, 2013.

[19] M. Saad, A.C. Bovik, and C. Charrier, "Blind Prediction of Natural Video Quality," *IEEE Trans. on Image Proc.*, vol. PP no. 99, 2014.

[20] C. Li and A.C. Bovik, "Content-partitioned Structural Similarity Index for Image Quality Assessment," *Signal Process. Image Commun.*, vol. 25, no. 7, pp 517-526, Aug. 2010.