Model-Based Blind Image Quality Assessment Using Natural DCT Statistics

Michele A. Saad*, Student Member, IEEE, Alan C. Bovik, Fellow, IEEE, and Christophe Charrier, Member, IEEE

Abstract—We propose an efficient, general-purpose, non-distortion specific, blind/no-reference image quality assessment (NR-IQA) algorithm based on a natural scene statistics model of discrete cosine transform (DCT) coefficients. The algorithm is computationally appealing, given the availability of platforms optimized for DCT computation. We propose a generalized parametric model of the extracted DCT coefficients. The parameters of the model are utilized to predict image quality scores. The resulting algorithm, which we name BLIINDS-II, requires minimal training and adopts a simple probabilistic model for score prediction. When tested on the LIVE IQA database, BLIINDS-II is shown to correlate highly with human visual perception of quality, at a level that is even competitive with the popular SSIM index. (EDICS category: SMR-REP, SMR-HPM)

Index Terms—No-reference image quality assessment, discrete cosine transform, natural scene statistics, generalized Gaussian density.

I. INTRODUCTION

The ubiquity of transmitted digital visual information (in the form of images and video) in every economic sector, and the broad range of applications that rely on it, such as PDAs, high definition televisions, internet video streaming, and video on demand, to name a few, necessitates means to evaluate the visual quality of this information. The various stages of the pipeline through which an image passes, introduce distortions to the image or modify it in one way or another, starting from its capture until its consumption by a viewer. The capture, digitization, compression, storage, transmission, and display processes all introduce modifications to the original image. These modifications, also termed distortions or impairments, may or may not be perceptually visible to the viewer. If they are visible, quantifying how perceptually annoying they are is an important process for improving Quality of Service (QoS) in the applications listed above. Since human raters are generally unavailable or too expensive for these applications, there is a significant need for objective IQA algorithms.

Only recently did full-reference image quality assessment (FR-IQA) methods reach a satisfactory level of performance, as demonstrated by high correlations with human subjective judgements of visual quality. SSIM [1], MS-SSIM [2], VSNR [3], and the VIF index [4] are examples of FR-IQA algorithms, to name a few. These methods require the availability of a reference signal against which to compare the test signal. In many applications, however, the reference signal is not available to perform a comparison against. This strictly limits the application domain of FR-IQA algorithms and points up the need for reliable blind/NR-IQA algorithms. However, no current NR-IQA algorithm exists that has been proven consistently reliable in performance. While the performance of some FR-IQA algorithms has proven satisfactory enough to be deployed in standards, (e.g. the inclusion of the SSIM index in the H.264/MPEG4 Part 10 AVC reference software [5], [1]), and their performance improvement has possibly plateaued, generic NR-IQA algorithms, on the other hand, are regarded as having a long way to go before reaching the same level of performance.

The problem of blindly assessing the visual quality of images, in the absence of a reference, and without assuming a single distortion type, requires dispensing with older ideas of quality such as fidelity, similarity, and metric comparison. Presently, NR-IQA algorithms generally follow one of three trends: 1) Distortion-specific approaches: These employ a specific distortion model to drive an objective algorithm to predict a subjective quality score. These algorithms quantify one or more distortions such as blockiness [6], blur [7], [8], or ringing [9] and score the image accordingly. 2) Training-based approaches: these train a model to predict the image quality score based on a number of features extracted from the image [10], [11], [12]. 3) Natural scene statistics (NSS) approaches: these rely on the hypothesis that images of the natural world (i.e. distortion free images) occupy a small subspace in the space of all possible images and seek to find a distance between the test image and the subspace of natural images [13].

The first approach is distortion-specific, and hence to some degree, application specific. It is important to understand, however, that while distortion modeling is important, it does not necessarily embody perceptual relevance (distortion annoyance), since such factors as masking and contrast sensitivity need to be considered. The second approach is only as reliable as the representativeness of the features used to train the learning model. Moreover, most existing algorithms following this trend rely on a large number of features without providing an interpretable meaning of each individual feature. The third approach is a promising one but relies on extensive statistical modeling and reliable
In this paper, we propose a framework that derives entirely from a simple statistical model of local DCT coefficients. We name our algorithm BLIINDS-II (Blind Image Integrity Notator using DCT Statistics). The new BLIINDS-II index greatly improves upon a preliminary algorithm (BLIINDS-I) [14], which uses no statistical modeling and a different set of sample DCT statistics. BLIINDS-I was a reasonably successful experiment to determine whether DCT statistics could be used for blind IQA. BLIINDS-II fully unfolds this possibility and provides a leap forward in both performance and in the use of an elegant and general underlying statistical model. We derive a generalized NSS-based model of local DCT coefficients, and transform the model parameters into features used for perceptual image quality score prediction. It is observed that the statistics of the DCT features change as the image quality changes. A generalized probabilistic model is obtained for these features, and is used to make probabilistic predictions of visual quality. We show that the method correlates highly with human subjective judgements of quality. We also interpret, analyze, and report how each feature in isolation correlates with human visual perception.

The contributions of our approach are the following: 1) The proposed method inherits the advantages of the NSS approach to IQA. While the goal of IQA research is to produce results in accordance with human visual perception of quality, one can to some degree avoid modeling poorly understood functions of the human visual system (HVS), and resort to models of the natural environment instead. This is motivated by the fact that HVS modeling and NSS modeling can be regarded as dual problems, owing to the widely accepted hypothesis that the HVS has evolutionally adapted to its surrounding visual natural environment [15], [16]. 2) The approach is non-distortion specific; while most NR-IQA algorithms quantify a specific type of distortion, the features used in our algorithm are derived independently of the type of distortion of the image and are effective across multiple distortion types. Consequently, it can be deployed in a wide range of applications. 3) We propose a novel model for the statistics of DCT coefficients. Previous work on reduced-reference RR-IQA has shown that local image wavelet coefficients are Laplacian in nature and tend to be Gaussian distributed when a divisive normalization transform is applied [17]. Our observations have shown that DCT coefficients are symmetrically distributed, but the distribution tails, peakedness and skewness are not Laplacian nor Gaussian. We observe that the distribution shape varies with the perceptual visual quality. To capture such general quantities we propose the use of the generalized Gaussian distribution to model local DCT coefficients. The generalized Gaussian family is symmetric about its mean, and depending on its shape parameter it can range over a wide range of probabilistic behaviors. 4) Since the framework operates entirely in the DCT domain, one can take advantage of the availability of platforms devised for fast computation of DCT transforms. Additionally, many image and video compression algorithms are based on block-based DCT transforms (JPEG, MPEG2, H263, and H264 which relies on a variation of the DCT transform). One can apply our general method to the coefficients already extracted and available, hence reducing computational complexity. 5) The method, as we will show, requires minimal training effort, and relies on a simple probabilistic model for quality score prediction. This leads to further computational gains. 6) Finally, we show that the method correlates highly with human visual perception of quality and yields highly competitive performance, even with respect to state-of-the-art FR-IQA algorithms. The algorithm is evaluated on the LIVE IQA Database, and performance results are reported for 1000 iterations of randomly chosen, completely content-separate train and test sets. We also provide a Matlab implementation of BLIINDS-II, which can be downloaded from the Laboratory of Image and Video Engineering (LIVE) website at http://live.ece.utexas.edu/.

The rest of the paper is organized as follows. In Section 2, we describe the DCT-domain features and the motivation behind the choice of the features. In Section 3, we show how each feature alone correlates with subjective differential-mean-opinion-scores (DMOS). In Section 4, we describe the generalized probabilistic prediction model. We present the results in Section 5, and we conclude in Section 6.

II. OVERVIEW OF THE METHOD

In the context of NSS modeling, undistorted images taken by a camera are referred to as natural scenes, and models built for undistorted natural scenes are referred to as NSS models. Deviations from NSS models, caused by the introduction of distortions to an image, can be used to predict the perceptual quality of the image. The model-based NSS-IQA approach we develop here is a process of feature extraction from the image, followed by statistical modeling of the extracted features. Purely NSS-based IQA approaches require the development of a distance measure between a given distorted test image and the NSS model. This leads to the question of what constitutes appropriate and perceptually meaningful distance measures between distorted image features and NSS models. The Kullback-Leibler (KL) divergence has been used for this purpose [17], but no perceptual justification has been provided for its use.

Our approach relies on the IQA algorithm learning how the NSS model varies across different perceptual levels of image distortion. The algorithm is trained using features derived directly from a generalized parametric statistical model of natural image DCT coefficients against various perceptual levels of image distortion. The learning model is then used to predict perceptual image quality scores.

Unlike much of our prior work on image/video QA [1], [2], [4], [18], [19], we make little direct use of specific perceptual models such as area V1 cortical decompositions [19], masking [1], [2], [4], [17], and motion perception [19].
Yet we consider our approach as perceptually consistent since the NSS models reflect statistical properties of the world that drive perceptual functions of the HVS. This is a consequence of the belief that the HVS is adapted to the statistics of its visual natural environment. In other words, models of natural scenes embody characteristics of the HVS, since the HVS itself has evolutionally adapted to models conforming to natural scenes [15], [16]. A number of HVS characteristics that are intrinsic to, or can be incorporated into NSS models are: 1) visual sensitivity to structural information [1], [2], 2) perceptual masking [15], [17], [19], 3) visual sensitivity to directional information [20], [21], 4) multiscale spatial visual processing [4], [15], [19], and 5) intolerance to flagrantly visible visual distortion [22]. In the following sections we indicate where one of more of these HVS properties are embedded in the model.

The framework of the proposed approach is summarized in Fig. 1. An image entering the IQA “pipeline” is first subjected to local 2-dimensional DCT-transform coefficient computation. This stage of the pipeline consists of partitioning the image into equally sized $n \times n$ blocks, henceforth referred to as local image patches, and computing a local 2-dimensional DCT transform on each of the blocks. The coefficient extraction is performed locally in the spatial domain in accordance with the HVS’s property of local spatial visual processing (i.e. in accordance with the fact that the HVS processes the visual space locally) [15]. As will be seen, this $n \times n$ DCT decomposition may be accomplished across scales. The second stage of the pipeline applies a generalized Gaussian density model to each block of DCT coefficients, as well as for specific partitions within each DCT block.

We next briefly describe the $n \times n$ DCT block partitions that are used. In order to capture directional information from the local image patches, the DCT block is partitioned directionally as shown in Fig. 7 into 3 oriented subregions. A generalized Gaussian fit is obtained for each of the oriented DCT-coefficient subregions. Another configuration for the DCT block partition is shown in Fig. 5. The partition reflects 3 radial frequency subbands in the DCT block. The upper, middle, and lower partitions correspond to the low frequency, mid-frequency, and high frequency DCT subbands respectively. A generalized Gaussian fit is obtained for each of the subregions as well.

The third step of the pipeline computes functions of the derived model parameters. These are the features used to predict image quality scores. These features are derived from the model parameters. In the following sections we define and analyze each model-based feature, and demonstrate how it changes with visual quality, and how well it correlates with human subjective judgement of quality.

The fourth and final stage of the pipeline uses a simple Bayesian model to predict a quality score for the image. The Bayesian approach maximizes the probability that the image has a certain quality score given the extracted features and extracted from the image. The posterior probability that the image has a certain quality score given the extracted features is modeled as a multidimensional generalized Gaussian distribution.

![Fig. 1: High level overview of the BLIINDS-II framework](image)

### A. The Generalized Probabilistic Model

The Laplacian model has often been used to approximate the distribution of DCT image coefficients [23]. The Laplacian model is characterized by a large concentration of values around zero and heavy tails. However, the introduction of distortions to the images changes the distribution of the coefficients, as shown in [24], [4]. Such descriptive terms as heavy tails, peakedness at zero, and skewness, which have often been used to describe distributions, are intrinsically heuristic. In our prior work in [14], we used such sample statistics (kurtosis, entropy, etc.), without image modeling, to create a reasonably successful but preliminary blind IQA algorithm. Instead, here we model image features using a generalized Gaussian family of distributions which encompasses a wide range of observed behavior of distorted DCT coefficients. The generalized Gaussian model has recently been used as a feature in a NSS-based reduced-reference (RR) IQA algorithm [17] and in a simple two-stage NR-IQA algorithm in [12].

The univariate generalized Gaussian density is given by

$$
\alpha e^{-\beta |x-\mu| \gamma},
$$

where $\mu$ is the mean, $\beta$ is the scale parameter, $\gamma$ is the shape parameter, and $\Gamma$ denotes the gamma function given by

$$
\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt,
$$

and $\alpha$ and $\beta$ are normalizing and scaling constants given by

$$
\alpha = \frac{\beta \gamma}{2 \Gamma(1/\gamma)},
$$

$$
\beta = \frac{1}{\sigma \sqrt{\Gamma(3/\gamma)/\Gamma(1/\gamma)}}.
$$

This family of distributions includes the Gaussian distribution ($\beta = 2$) and the Laplacian distribution ($\beta = 1$) [25]. As $\beta \to \infty$ the distribution converges to a uniform distribution. Fig. 2
shows the generalized Gaussian distribution at varying levels of the shape parameter (\(\gamma\)) value.

![Generalized Gaussian Distribution at Varying Gamma Parameter Values](image)

Fig. 2: The generalized Gaussian density at varying levels of the shape parameter \(\gamma\)

A variety of parameter estimation methods have been proposed for this model in the literature. We deploy the reliable method given in [26].

The multivariate version of the generalized Gaussian density is given by

\[
\alpha e^{-\left(\beta(x-\mu)^T \Sigma^{-1}(x-\mu)\right)^\gamma},
\]  

(5)

where \(\Sigma\) is the covariance matrix of the multivariate random variable \(x\), and the remaining parameters are as defined in the univariate case. Fig. 3 shows 2D generalized Gaussian distributions at varying levels of the shape parameter (\(\gamma\)) value. Parameter estimation is treated similarly as in the univariate case once the quantity \((x-\mu)^T \Sigma^{-1}(x-\mu)\) is estimated from the sample data.

![Generalized Gaussian Gamma Parameter Values](image)

Fig. 3: The 2D generalized Gaussian density at varying levels of the shape parameter \(\gamma\)

B. The DCT Feature Domain

The performance of any IQA model is a function of the representativeness of the features that are used for quality score prediction. In other words, the prediction is only as good as the choice of features extracted. It is broadly agreed upon that the HVS is adapted to the statistics of images in its natural surrounding. It has also been shown that natural images exhibit strong spatial structural dependencies [1]. Consequently, we define features representative of image structure, and whose statistics are observed to change with image distortions. The structural information in images is largely contained in edge-information represented by spatial frequencies that constructively and destructively interfere with each other to produce spatial structure in the form of natural scenes. Moreover, visual images are subjected to local spatial frequency decompositions in the visual cortex [4], [15], [27].

For this reason we perform feature extraction in a local frequency-domain, namely the DCT domain. The main motivation behind feature extraction in the DCT domain is the observation that the statistics of DCT coefficients change with varying levels of image distortions. Another advantage of having a framework in the DCT domain is computational convenience: 1) The availability of optimized DCT-specific platforms [28], [29], [30], [31] eases its computation. 2) The availability of fast algorithms for DCT computation [32], [33] presents another computational advantage. For instance, DCTs can be computed efficiently by variable-change transforms from the very computationally efficient FFT algorithm. This has been explained and derived in [34]. 3) Many image and video compression algorithms are based on block-based DCT transforms (JPEG, MPEG2, H263, and H264 which rely on a variation of the DCT transform). Consequently, the model-based method may be applied to the already computed coefficients, resulting in even greater computational efficiency. 4) Finally, and perhaps most importantly, we are able to define simple and naturally-defined model-based DCT features that capture perceptually-relevant image and distortion characteristics in a natural and convenient manner.

We illustrate one instance of how the statistics of DCT coefficients change as an image becomes distorted in Fig. 4, which shows the DCT coefficient histograms of a distortion free image and a Gaussian blur distorted image, respectively. Similar trends in the histogram statistics are observed over many images in the LIVE IQA Database [35], on which we perform our study. Among the observed differences in the
histograms is the degree of peakedness at zero, (blurred images are observed to have a higher histogram peak at zero), and variance (blurred images exhibit reduced variance). We utilize statistical differences, such as these, to develop an NR-IQA index. We describe each of the model-based features used and show how each correlates with human judgements of quality in the following.

III. MODEL-BASED DCT DOMAIN NSS FEATURES

A. The Generalized Gaussian Model Shape Parameter

We deploy a generalized Gaussian model of the non-DC DCT coefficients. In other words, we model the DCT coefficients in an $nxn$ block, omitting the DC coefficient. The generalized Gaussian density in (1) is parametrized by mean $\mu$, scale parameter $\beta$, and shape parameter $\gamma$. The shape parameter $\gamma$ is used as a model-based feature. This feature is computed over all blocks in the image.

The overall shape parameter-based quality feature used is found by computing the lowest 10$^{th}$ percentile average of the local block shape scores ($\gamma$) across the image. The reason for this pooling of the histogram shape features, as opposed to simple averaging, is that percentile pooling has been observed to result in high correlations with subjective perception of quality [18], [36]. Percentile pooling is motivated by the observation that the "worst" distortions in subjective perception of quality [18], [36]. Percentile pooling has been observed to result in high correlations with subjective perception of quality [18], [36]. Percentile pooling is defined as follows:

$$
\zeta = \frac{\sigma_{|X|}}{\mu_{|X|}}
$$

(6)

In Table I we report Spearman rank order correlation coefficient (SROCC) scores between the LIVE IQA Database DMOS scores, and the features obtained by averaging the lowest 10$^{th}$ block scores and by simple (100%) averaging, respectively. Table I also reports linear correlation coefficient (LCC) scores between a fitted function of the features and the DMOS scores. (The blocks were chosen to be of dimension 5x5).

<table>
<thead>
<tr>
<th>LIVE Subset</th>
<th>SROCC 10%</th>
<th>SROCC 100%</th>
<th>LCC 10%</th>
<th>LCC 100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG2000</td>
<td>0.9214</td>
<td>0.7329</td>
<td>0.8040</td>
<td>0.7453</td>
</tr>
<tr>
<td>JPEG</td>
<td>0.7790</td>
<td>0.7295</td>
<td>0.7166</td>
<td>0.8008</td>
</tr>
<tr>
<td>WN</td>
<td>0.9580</td>
<td>0.9233</td>
<td>0.9660</td>
<td>0.9295</td>
</tr>
<tr>
<td>GBLUR</td>
<td>0.9099</td>
<td>0.3298</td>
<td>0.8067</td>
<td>0.3176</td>
</tr>
<tr>
<td>FASTFADING</td>
<td>0.8266</td>
<td>0.6282</td>
<td>0.7734</td>
<td>0.2964</td>
</tr>
</tbody>
</table>

B. The Coefficient of Frequency Variation

The next feature is the coefficient of frequency variation feature:

$$
\zeta = \frac{\sigma_{|X|}}{\mu_{|X|}}
$$

(6)

which we shall show is equivalent to

$$
\zeta = \sqrt{\frac{\Gamma(1/\gamma)\Gamma(3/\gamma)}{\Gamma^2(2/\gamma)}} - 1,
$$

(7)

under our model (1), where $\sigma_{|X|}$ is the variance of the DCT coefficient magnitudes, and $\mu_{|X|}$ is the mean of the absolute value of the coefficients, (i.e. the mean of the coefficient magnitudes).

Let $X$ be a random variable with density function (1). Since $\mu = 0$, (the density of the DCT coefficients is centered around 0), then

$$
\mu_{|X|} = \int_{-\infty}^{+\infty} |x|\alpha e^{-(\beta|x|)^\gamma} dx
$$

$$
= \frac{2\alpha}{\beta^\gamma} \Gamma\left(\frac{2}{\gamma}\right)
$$

(8)

where $\alpha$ and $\beta$ are given by (3) and (4) respectively, and $\sigma$ is the variance of $X$. Substituting for the values of $\alpha$ and $\beta$ yields

$$
\frac{\Gamma(1/\gamma)\Gamma(3/\gamma)}{\Gamma^2(2/\gamma)} = \frac{\sigma^2}{\mu^2_{|X|}}.
$$

(9)

On the other hand,

$$
\sigma^2_{|X|} = \sigma^2 - \mu^2_{|X|},
$$

(10)

Hence,

$$
\zeta = \frac{\sigma_{|X|}}{\mu_{|X|}} = \sqrt{\frac{\Gamma(1/\gamma)\Gamma(3/\gamma)}{\Gamma^2(2/\gamma)}} - 1.
$$

(11)

The feature $\zeta$ is computed for all blocks in the image. The highest 10$^{th}$ percentile average and the mean (100$^{th}$ percentile) of the local block scores across the image are then computed. The motivation behind the percentile pooling strategy is similar to the one for the pooling of the shape parameter feature $\gamma$, in the previous subsection. As shown in Table II, the highest 10$^{th}$ percentile pooling correlates well with human visual perception of quality. Both pooling results (10% and 100%) are used as pooled features, since the difference between these is a compact but rich form of information.

Observe that the correlations are consistently higher when the lowest 10$^{th}$ percentile pooling strategy is adopted. This may be interpreted as further evidence that human sensitivity to distortion in images is not a linear function of the distortion. For instance humans tend to judge poor regions in an image more harshly than good ones, and hence penalize images with even a small number or area of poor regions more heavily [22], [36].
TABLE II: SROCC and LCC correlations (subjective DMOS vs DCT $\zeta$ highest 10th percentile and 100th percentile means (5x5))

<table>
<thead>
<tr>
<th></th>
<th>SROCC</th>
<th>LCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG2000</td>
<td>0.9334</td>
<td>0.9131</td>
</tr>
<tr>
<td>JPEG</td>
<td>0.8070</td>
<td>0.4066</td>
</tr>
<tr>
<td>WN</td>
<td>0.9607</td>
<td>0.9368</td>
</tr>
<tr>
<td>GBBLUR</td>
<td>0.9245</td>
<td>0.8614</td>
</tr>
<tr>
<td>FASTFADING</td>
<td>0.8312</td>
<td>0.8410</td>
</tr>
</tbody>
</table>

TABLE III: SROCC and LCC correlations (subjective DMOS vs product of $\zeta$ and $\gamma$)

<table>
<thead>
<tr>
<th></th>
<th>SROCC</th>
<th>LCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG2000</td>
<td>0.8725</td>
<td>0.8962</td>
</tr>
<tr>
<td>JPEG</td>
<td>0.8217</td>
<td>0.8154</td>
</tr>
<tr>
<td>WN</td>
<td>0.8545</td>
<td>0.8190</td>
</tr>
<tr>
<td>GBBLUR</td>
<td>0.8857</td>
<td>0.1761</td>
</tr>
<tr>
<td>FASTFADING</td>
<td>0.7427</td>
<td>0.0686</td>
</tr>
</tbody>
</table>

In the coefficient of frequency variation $\zeta$, the denominator $\mu_{|X|}$ measures the center of the DCT coefficient magnitude distribution, whereas $\sigma_{|X|}$ measures the spread or energy of the DCT coefficient magnitudes. The ratio $\zeta$ correlates well with human visual impressions of quality as shown in Table II. The high correlation between $\zeta$ and subjective judgement of perceptual quality is an indication of the monotonicity between $\zeta$ and subjective DMOS. Since $\zeta$ is the ratio of the variance $\sigma_{|X|}$ to the mean $\mu_{|X|}$, the effect of an increase (or decrease) of $\sigma_{|X|}$ in the numerator is mediated by the decrease (or increase) of $\mu_{|X|}$ in the denominator of $\zeta$. Indeed two images may have similar perceptual quality even if their respective DCT coefficient magnitude energy ($\sigma_{|X|}$) is very different, depending on where the coefficient magnitude energy distribution is centered ($\mu_{|X|}$).

We have also found that the product of the 10th percentile shape parameter $\gamma$ and the 100th percentile coefficient of frequency variation $\zeta$, as well as the product of the mean shape parameter $\gamma$ and the mean coefficient of frequency variation $\zeta$ improve the performance of our prediction model when applied across all distortion types (JPEG2000, JPEG, white noise, Gaussian blur, and fast fading channel distortions). The prediction model and the challenge of applying one unified prediction model over any type of distortion are discussed in the coming sections (sections IV and VI). The LCC and SROCC correlations between the product of $\zeta$ and $\gamma$ and subjective DMOS are reported in Table III.

C. Energy Subband Ratio Measure

Image distortions often modify the local spectral signatures of an image in unnatural ways. Towards measuring this, define a local DCT energy-subband ratio measure. Consider the 5x5 matrix shown in Fig. 6. Moving along diagonal lines on the figure from the top-left corner of the matrix to the bottom-right corner, the DCT coefficients represent higher radial spatial frequencies in the block. Consequently, we define 3 frequency bands in the block, as depicted in Fig. 5. Let $\Omega_n$ denote the set of coefficients belonging to band $n$, where $n = 1, 2, 3$ (lower, middle, higher). Then define the average energy in frequency band $n$ as the model variance $\sigma_n^2$ corresponding to band $n$:

$$E_n = \sigma_n^2.$$  

This is found by fitting the DCT data histogram in each band to the generalized Gaussian density model (1), then using the $\sigma_n^2$ value from the fit. We then compute the ratio of the difference between the average energy in frequency band $n$ and the average energy up to frequency band $n$, over the sum of these two quantities:

$$R_n = \frac{|E_n - \frac{1}{n-1} \sum_{j<n} E_j|}{E_n + \frac{1}{n-1} \sum_{j<n} E_j}.$$  

$R_n$ is defined for $n = 2, 3$. A large ratio corresponds to a large disparity in the frequency energy between a local frequency band and the average energy in the lower bands. Thus, this feature measures the relative distribution of energies in lower and higher bands, which may be affected by distortions. In the spatial domain, a large ratio roughly corresponds to a uniform frequency (textural) content in the image patch. A low ratio, on the other hand, corresponds to a small frequency disparity between the feature band and the average energy in the lower bands. The mean of $R_2$ and $R_3$ is computed. This feature is computed for all blocks in the image. We compute the highest 10th percentile average and the 100th percentile average (regular mean) of the local block scores across all the image. The reason behind pooling the local feature measures in this manner is similar to the reasoning given for the features listed above.

In Table IV, we report SROCC and LCC scores between the LIVE IQA Database DMOS scores and the feature obtained by averaging the highest 10% block scores and by simple (100%) averaging, respectively. (The blocks in this example were of dimension 5x5). We observe that the correlation is consistently higher when the 10th percentile pooling strategy is adopted.
TABLE IV: SROCC and LCC correlations (subjective DMOS vs energy subband ratio feature pooled according to the highest 10<sup>th</sup> percentile and 100<sup>th</sup> percentile mean (5x5))

<table>
<thead>
<tr>
<th>LIVE Subset</th>
<th>SROCC</th>
<th>LCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG2000</td>
<td>0.9313 0.8745</td>
<td>0.9370 0.8823</td>
</tr>
<tr>
<td>JPEG</td>
<td>0.9493 0.4601</td>
<td>0.8021 0.3075</td>
</tr>
<tr>
<td>WN</td>
<td>0.9754 0.9608</td>
<td>0.9478 0.8748</td>
</tr>
<tr>
<td>GBLUR</td>
<td>0.8850 0.5808</td>
<td>0.8916 0.6797</td>
</tr>
<tr>
<td>FASTFADING</td>
<td>0.8602 0.7558</td>
<td>0.8511 0.7351</td>
</tr>
</tbody>
</table>

TABLE V: SROCC and LCC correlations (subjective DMOS vs oriented ζ variance pooled according to the highest 10<sup>th</sup> percentile and 100<sup>th</sup> percentile mean)

<table>
<thead>
<tr>
<th>LIVE Subset</th>
<th>SROCC</th>
<th>LCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG2000</td>
<td>0.8805 0.8104</td>
<td>0.8001 0.7885</td>
</tr>
<tr>
<td>JPEG</td>
<td>0.2095 0.8233</td>
<td>0.1560 0.7786</td>
</tr>
<tr>
<td>White Noise</td>
<td>0.9214 0.9171</td>
<td>0.9227 0.8445</td>
</tr>
<tr>
<td>Gaussian Blur</td>
<td>0.6963 0.8858</td>
<td>0.6011 0.6895</td>
</tr>
<tr>
<td>Fast Fading</td>
<td>0.7550 0.8258</td>
<td>0.7005 0.7245</td>
</tr>
</tbody>
</table>

D. Orientation Model-Based Feature

Image distortions often modify local orientation energy in an unnatural manner. The HVS, which is highly sensitive to local orientation energy [15] is likely to respond to these changes. To capture directional information in the image that may correlate with changes in human subjective impressions of quality, we model block DCT coefficients along 3 orientations. We demonstrate how oriented DCT coefficients are captured in Fig. 7 below. The 3 shaded areas represent the DCT coefficients along 3 orientations. A generalized Gaussian model is fit to each shaded region in the block, and γ and ζ are obtained from the model histogram fits for each orientation. We then compute the variance of ζ along the 3 orientations. The variance of ζ across the 3 orientations from all the blocks in the image is then averaged (highest 10<sup>th</sup> percentile and 100<sup>th</sup> percentile) to obtain two numbers per image. We report how this feature correlates with subjective DMOS in Table V, again using 5x5 blocks.

![DCT coefficients along 3 orientations](image)

IV. Blind Quality Assessment Across Distortion Types

Tables I - V show that each of the features correlates highly with human visual perception of quality when applied on one specific distortion type (JPEG2000, JPEG, white noise, Gaussian blur, Fast fading channel distortions). A major challenge arises when one assumes no knowledge of the type of distortion affecting an image. In such a case, it becomes necessary to combine a set of features that perform well in predicting image quality blindly and irrespective of the type of distortion in the image. We demonstrate the challenge by showing in Table VI how each feature alone correlates with DMOS on the entire LIVE IQA database of images (i.e. with all distortion types mixed together). The low correlations between each individual feature and subjective DMOS across all distortion types implies the need to combine the features in a manner that enables quality score prediction reliably, with no assumption on the type of distortion in the image. The manner in which we combine the features to predict blind image quality scores with no assumption about the type of distortion affecting the image is discussed in section VI.

V. Multiscale Feature Extraction

It is well understood that images are naturally multiscale [4], [16], and that the early visual system involves decompositions over scales [15]. Successful FR-IQA algorithms have exploited this fact to create natural multiscale measurements of image quality [2], [4]. Towards this end, we implement the BLIINDS-II concept over multiple scales in a simple way. Specifically, the NSS-based DCT features are extracted from 5x5, overlapping blocks in the image. The feature extraction is repeated after lowpass filtering the image and subsampling it by a factor of 2 as shown in Fig. 8. Prior to downsampling, the image is filtered by the rotationally symmetric discrete 3x3 Gaussian filter kernel depicted in Fig. 9. At each scale, the overlap between neighboring blocks is 2 pixels. This defines a multiscale feature extraction approach, which enables BLIINDS-II to deal with changes in the image resolution, with distance from the image display to the observer, and with variations in the acuity of the observer’s visual system.

VI. Prediction Model

We have found that a simple probabilistic predictive model is quite adequate for training the features used in BLIINDS-II. The prediction model is the only element of BLIINDS-II that carries over from BLIINDS-I. The efficacy of this simple predictor points up the power of the NSS-based features we have defined. Let $X_i = [x_1, x_2, ..., x_m]$ be the vector of features extracted from the image, where $i$ is the index of the image being assessed, and $m$ is the number of features extracted (in our case $m = 10$ per scale). Additionally, let $DMOS_i$ be the subjective $DMOS$ associated with the image $i$. We model the distribution of the pair $(X_i, DMOS_i)$. 
The probabilistic model is trained on a subset of the LIVE IQA Database, which includes DMOS scores, to determine the parameters of the probabilistic model by distribution fitting. A multivariate generalized Gaussian model is used to model the data. Parameter estimation of the model only requires the mean and covariance of the empirical data from the test set. The probabilistic model \( P(X, DMOS) \) is designed by distribution fitting to the empirical data of the training set. The training and test sets are completely content independent, in the sense that no two images of the same scene are present in both sets. The probabilistic model is then used to perform prediction by maximizing the quantity \( P(DMOS_i/X_i) \). This is equivalent to maximizing the joint distribution of \( X \) and \( DMOS, P(X, DMOS) \) since \( P(X, DMOS) = P(DMOS/X)p(X) \).

VII. EXPERIMENTS AND RESULTS

BLIINDS-II was rigorously tested on the LIVE IQA Database [35] which contains 29 reference images, each impaired by many levels of 5 distortion types: JPEG2000, JPEG, white noise, Gaussian blur, and fast-fading channel distortions (simulated by JPEG2000 compression followed by channel bit errors.). The total number of distorted images (excluding the 29 reference images) is 779 images.

Multiple train-test sequences were run. In each, the image database was subdivided into distinct training and test sets (completely content-separate). In each train-test sequence, 80% of the LIVE IQA Database content was chosen for training, and the remaining 20% for testing. Specifically, each training set contained images derived from 23 reference images, while each test set contained the images derived from the remaining 6 reference images. 1000 randomly chosen training and test sets were obtained and the prediction of the quality scores was run over the 1000 iterations.

The model based-features were extracted over 3 scales. The total number of features per scale is 10. These 10 features are: 1) the lowest \( 10^{th} \) percentile of the shape parameter \( \gamma \), 2) the mean of the shape parameter \( \gamma \), 3) the highest \( 10^{th} \) percentile of the coefficient of frequency variation \( \zeta \), 4) the mean \( (100^{th} \) percentile) of the coefficient of frequency variation \( \zeta \), 5) the product of the lowest \( 10^{th} \) percentile \( \gamma \) and the highest \( 10^{th} \) percentile \( \zeta \), 6) the product of the mean of \( \gamma \) and the mean of \( \zeta \), 7) the highest \( 10^{th} \) percentile of the energy subband ratio measure \( R_n \), 9) the mean of the energy subband ratio measure, 8) the highest \( 10^{th} \) percentile of the orientation feature (which is the variance of \( \zeta \) across the 3 orientations), and 10) the mean of the orientation feature.

We report quality score prediction results for features extracted at 1 scale only (10 features), at 2 scales (20 features, 10 features per scale), and at 3 scales (30 features, 10 per scale). Linear Correlation Coefficient (LCC) scores (on a fitted function of the predicted DMOS and subjective DMOS scores), as well as SROCC scores between the predicted DMOS scores and the subjective DMOS scores of the LIVE IQA Database are computed for each of the 1000 iterations. The comparison of prediction results for 1 scale, 2 scale, and 3 scale feature extraction is shown in Tables VII and VIII. We observe that no significant gain in performance is obtained beyond the 3rd scale of feature extraction.

For comparison purposes, in addition to the simple probabilistic model described in this section, we also trained a radial basis function-kernel regression SVM, based on the implementation in [37], and perform the prediction utilizing this slightly more complex model as well. We also compared BLIINDS-II to the recent SVM-based NR-IQA algorithm BIQI, in [12], to the full-reference PSNR measure which has long been used for the assessment of the quality of images in the presence of a reference image, and to the state-of-the-art FR-IQA SSIM index.

The SROCC and LCC results are shown in Tables IX - XII. Tables IX and X compare the SROCC and the LCC results between the 3 NR-IQA methods (BLIINDS-II with the probabilistic prediction model, BLIINDS-II with the SVM prediction model, and a recent SVM-based NR-IQA algorithm, BIQI), respectively. Tables XI and XII report the SROCC and LCC results of PSNR and SSIM both of which are full-reference methods that require the presence of a reference image to perform quality score prediction on a test image.

The two prediction models (probabilistic and SVM) used on BLIINDS-II perform very similarly, with slightly higher correlation for the probabilistic prediction model on the individual distortion subsets (JPEG2000, JPEG, white noise, Gaussian blur, and fast-fading channel distortions) than on the entire dataset. The SVM prediction model only slightly outperforms the simple probabilistic prediction on the entire
TABLE VII: Median SROCC correlations for 1000 iterations of train and test sets (subjective DMOS vs predicted DMOS). Comparison for multiple scales of feature extraction.

<table>
<thead>
<tr>
<th>LIVE Subset</th>
<th>BLIINDS-II 1 Scale</th>
<th>BLIINDS-II 2 Scales</th>
<th>BLIINDS-II 3 Scales</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG2000</td>
<td>0.9505</td>
<td>0.9533</td>
<td>0.9506</td>
</tr>
<tr>
<td>JPEG</td>
<td>0.9394</td>
<td>0.9403</td>
<td>0.9419</td>
</tr>
<tr>
<td>White Noise</td>
<td>0.9753</td>
<td>0.9772</td>
<td>0.9783</td>
</tr>
<tr>
<td>GBlur</td>
<td>0.9335</td>
<td>0.9409</td>
<td>0.9435</td>
</tr>
<tr>
<td>Fast Fading</td>
<td>0.9039</td>
<td>0.9164</td>
<td>0.9268</td>
</tr>
<tr>
<td>ALL</td>
<td>0.8925</td>
<td>0.9081</td>
<td>0.9122</td>
</tr>
</tbody>
</table>

TABLE VIII: Median LCC correlations for 1000 iterations of train and test sets (subjective DMOS vs predicted DMOS). Comparison for multiple scales of feature extraction.

<table>
<thead>
<tr>
<th>LIVE Subset</th>
<th>BLIINDS-II 1 Scale</th>
<th>BLIINDS-II 2 Scales</th>
<th>BLIINDS-II 3 Scales</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG2000</td>
<td>0.9550</td>
<td>0.9571</td>
<td>0.9630</td>
</tr>
<tr>
<td>JPEG</td>
<td>0.9664</td>
<td>0.9781</td>
<td>0.9793</td>
</tr>
<tr>
<td>White Noise</td>
<td>0.9804</td>
<td>0.9833</td>
<td>0.9854</td>
</tr>
<tr>
<td>GBlur</td>
<td>0.9300</td>
<td>0.9450</td>
<td>0.9481</td>
</tr>
<tr>
<td>Fast Fading</td>
<td>0.9088</td>
<td>0.9267</td>
<td>0.9436</td>
</tr>
<tr>
<td>ALL</td>
<td>0.8919</td>
<td>0.9091</td>
<td>0.9132</td>
</tr>
</tbody>
</table>

LIVE IQA Database. With either of the prediction models, BLIINDS-II outperforms BIQI [12] and the full-reference PSNR measure. BLIINDS-II also approaches the powerful full-reference SSIM index in performance.

Scatter plots (for each of the distortion sets as well as for the entire LIVE IQA Database), of the predicted DMOS versus the subjective one on the test sets are shown in Fig. 10 - Fig. 15. These exhibit nice properties: a nearly linear relationship against DMOS, tight clustering, and a roughly uniform density along each axis.

VIII. CONCLUSION

We have proposed a generalized Gaussian density model-based, general (non-distortion specific) approach to no-reference/blind image quality assessment using a minimal number of features extracted entirely from the DCT-domain which is also computationally convenient. We have shown that the new BLIINDS-II algorithm can be easily trained and it employs a simple probabilistic model (based on a multivariate generalized Gaussian density) for prediction. The method is shown to correlate highly with human visual perception of quality, and to outperform the full-reference PSNR measure and the recent effective no-reference BIQI index.

IX. ACKNOWLEDGEMENTS

This research was supported by the CSOSG ANR-08-SECU-007-04 grant, and by Intel and Cisco, Inc., under the VAWN program.

REFERENCES


[29] J. Huan, M. Parris, J. Lee, and R.F. DeMarra, “Scalable FPGA-based ar-
Fig. 15: Predicted Versus Subjective DMOS on the entire LIVE IQA database


