# Learning a Blind Measure of Perceptual Image Quality

Huixuan Tang
University of Toronto
hxtang@dgp.toronto.edu

Neel Joshi
Microsoft Research
neel@microsoft.com

Ashish Kapoor
Microsoft Research
akapoor@microsoft.com

## Abstract

*It is often desirable to evaluate an image based on its quality. For many computer vision applications, a perceptually meaningful measure is the most relevant for evaluation; however, most commonly used measure do not map well to human judgements of image quality. A further complication of many existing image measure is that they require a reference image, which is often not available in practice. In this paper, we present a "blind" image quality measure, where potentially neither the groundtruth image nor the degradation process are known. Our method uses a set of novel low-level image features in a machine learning framework to learn a mapping from these features to subjective image quality scores. The image quality features stem from natural image measure and texture statistics. Experiments on a standard image quality benchmark dataset shows that our method outperforms the current state of art.*

## 1. Introduction

In numerous computer vision, computer graphics, and image processing applications it is necessary to evaluate image quality. The measurement of "quality" cannot be easily defined, as it often depends on context and personal preferences. However, when restricted to low-level aspects, image quality as perceived by human observers is a measurable and consistent property [23], even when comparing images with different content and degradation types.

In the signal and image processing literature, the most common measure for judging image quality are straightforward measure such as PSNR (Peak-Signal-To-Noise) [25], yet, it is well known that PSNR does not correlate well with perceptual quality. Furthermore many measure require a reference image for comparison, making them useful only in limited situations, such as in synthetic experiments.

In most practical cases, a reference image is not available, and image quality assessment is more difficult. Recovering a reference image or its properties (either explicitly or implicitly) for image quality assessment is equivalent to the general "blind image enhancement" problem, which is ill-conditioned even if the degradation process is known.

Recent work has sought to break these limitations by developing more perceptually meaningful reference-based measure [27] and ones that do not require a reference [20].

While existing methods have shown some promise, they still do not predict human quality judgements very accurately. One of the largest difficulties in computing a perceptually relevant score is the variability in how different types of image degradation processes affect an image's structure and statistics. As a result the scores from exiting methods often are biased by the type of degradation, making it difficult to compare quality between images with different or unknown degradation processes, *e.g*. comparing a blurry image to a noisy image.

The Blind Image Quality Index (BIQI) addressed this problem by using distortion-specific image quality measure as well as a distortion-type classifier [17]. They learn a mapping for images under five different types of distortion (noise, blur, JPEG, JPEG2000, fast fading) over a range of distortion amount. Given a new image they have to identify the distortion type and then measure it. Although BIQI performs impressively in its specific setup, its utility is somewhat limited. It requires an accurate classification of the distortion type, which is itself a difficult problem. Consequently it does not address the bias very well, as shown in Fig. 1(a). And it assumes that only one distortion type dominates, which is often not the case in practice. This hard classification can prohibit computing a meaningful quality measure. Furthermore, it is not trivial to extend BIQI method to handle additional distortion types, as a dual problem needs to be solved to both reliably recognize and measure the distortion.

We propose a learning based blind image quality measure (LBIQ) that is more perceptually correlated and less biased by distortion type, as shown in Fig. 1(b). Our measure addresses the above limitations by designing novel low-level image quality features that measure aspects of image structure and statistics that are useful for discriminating degraded and un-degraded images. Instead of using a small number of measure, LBIQ achieves good performance by

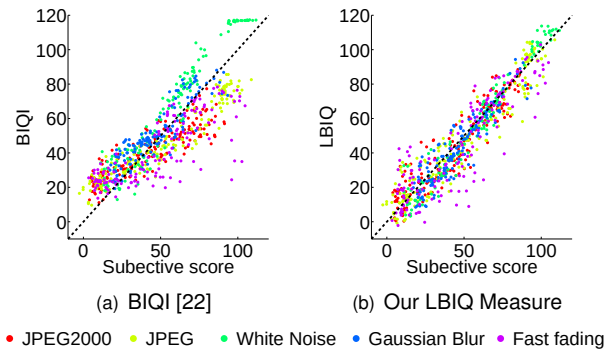



Figure 1. Computational v.s. subjective scores on the LIVE image quality assessment database [24]. Ideally the plot would scatter on the dashed line. Due to the variability of how different types of distortion degrades image structure and statistics, BIQI is biased by the type of degradation, yet, our LBIQ measure minimizes this bias by incorporating numerous image quality features in a learning framework. Besides, the rich set of feature we use ensures LBIQ to be more tightly correlated to the subjective score.

combining and incorporating numerous image quality features with a regression algorithm. The algorithm is able to correlate the underlying structure of distorted images with perceptual image quality without the need to provide a reference image.

The key contributions of our method include: (1) several novel low-level features for measuring image quality and (2) an algorithm to combine these features in order to learn a perceptually relevant image measure. To the best of our knowledge, our method is the first perceptually accurate image quality measure that does not require a reference image nor knowledge of the image degradation process and provides scores that are not biased by the degradation type. Our experimental results show that our LBIQ measure significantly outperforms state of art blind image quality assessment methods.

## 2. Related work

There are numerous cases where it is desirable to computationally evaluate image quality. For most applications, a perceptual accurate measure is the most relevant measure for evaluation, as a human observer is the final consumer of the image. Depending on the application, the measurement of "image quality" conveys many different aspects from how much is the image degraded by a specific distortion type to how "realistic" or "beautiful" an image looks.

### 2.1. Low-level quality assessment

A common use of an image quality measure is to judge the accuracy of an image compression or rendering algorithm against some reference "ground-truth" solution. In these cases, the relevant evaluation measure are known as "full-reference" measure. Since the reference image is known, a quality index can be computed from an image dissimilarity measure such as PSNR [25]. As direct measure often do not correlate well to image quality as perceived by human observers, numerous researchers have extended image dissimilarity measure to be more perceptually meaningful [27, 14]. Further improvements have been made by modulating and pooling local reference scores according to the gradient of the reference image [13].

Unfortunately, there are numerous cases where a reference image is unavailable, such as when judging the quality of a denoising algorithm on a real-world dataset, where the underlying noise-free image is unknowable. In this case, one would need a "no-reference" or "blind" measure. In a "no-reference" measure, while the ground-truth image is unknown, some assumptions about the underlying image structure or content is usually made. Several different approaches have been taken along these lines. One approach is to model artifacts for specific types of image degradation and then measure the amount of these artifacts present in an image to determine a quality measure. Another approach is to correlate natural image statistics with subjective measure, *e.g.* the BLIINDS measure [20] predicted perceptual image quality by linear regression on DCT statistics.

### 2.2. High-level quality assessment

The definition of "image quality" can be as high-level as measuring "realism" or "beauty" of an image. In this regime, digital forensics algorithms aims at differentiating real photographs from photo-realistic synthesized images [5, 16] and doctored photographs [6]. Evaluating the aesthetic quality of photographs [4, 12, 15] is also a very related area of work. The primary distinction is that aesthetic quality appears to be a much more subjective and personalized measure than that of low-level image quality [23]. Although these final goals slightly differ from image quality assessment, the underlying goals are quite similar, as synthesized images can be viewed as being (slightly) distorted images, and our work is in some sense a subset of aesthetic quality. We utilize some of the concepts in this related area of work to develop our low-level image features, but explicitly focus on an objective evaluation of how distortions affect perceptual image quality.

### 2.3. Learning-based quality assessment

To achieve good generalization across image content and distortion types and to have a perceptually meaningful result, we learn a measure that combines a number of low-level features to map to image quality scores from human observers. This concept has been also used by other recent works in evaluating image quality, especially in the case of high-level image measure such as digital forensics and photograph aesthetics.

There has also been some work involving using learning techniques for the low-level aspect of image quality, *e.g*. [2] used SVM regression to assess the quality of color images given a known reference image.

Jung and Leger [11] used artificial neural networks to compute a blind image quality index, but only apply this framework in situations with a specific type of distortion. BIQI [17] computed a more generic quality measure with SVM and addressed the bias across distortion by identifying distortion type first. While our LBIQ measure shares some similarities with this work, in that we too learn an image-quality measure from image features, our contribution in not only from the learning framework we use, but also from our design of a *rich* set of low-level features. Our feature set allows any distortion type to find a considerably large set of supporting features for evaluating its quality, which enables our algorithm to have good generalization power across images and distortion types regardless of image content and the potential presence of multiple distortions types in one image.

## 3. Image quality features

The key contributions of our work are our analysis and development of novel low-level features for measuring image quality and the use of these features to learn an image measure that correlates with the perception of human observers. In such learning based paradigms, the strength of the results heavily weighs on having a comprehensive set of discriminant features for the desired task. Given previous work, we believe that the task of image quality assessment benefits from a thorough analysis of current measure and features and this analysis leads to a better understanding of the area and the development of novel features.

Thus our methodology for selecting and designing relevant features included a lengthy process of extracting different features across a set of images and distortions and observing trends and correlations that reflected the change in perceptual quality. For this process, we tested each feature on the LIVE Image Quality dataset [24]. This dataset includes 29 reference images, each containing around 30 images that covers 5 distortion types: JPEG2000, JPEG, white noise, blur, and analog transmission loss of JPEG2000 encoded images (also known as fast fading). The perceptual score of each image is computed by collecting evaluations of about 23 trained human subjects, removing outlier subjects and scores, and finally compensating for the bias across reference images and subjects.

The design of our features relies on several key observations regarding image quality:

- A good objective function (or image prior) for image enhancement is a good measure of image quality;

Figure 2. Reference image used for extracting features that are visualized in subsequent figures in this paper.

- Texture statistics are a good indicator of distortion artifacts;
- Noise and blur are two fundamental degradation process that occur in a variety of distortion types, and can be directly measured.

In the following sections we present the image quality features we used to learn our LBIQ measure. For each feature, we illustrate its behavior on images of the same reference image (Fig. 2) of similar subjective quality but with different distortion types.

### 3.1. Natural image statistics

We assume that an image that is likely to be a "natural" image is also of high quality and thus investigate numerous objective functions used in the literature of image enhancement. Among them, we have found that high-frequency responses of images are an effective facility for many image enhancement problems. Although they appear in different forms from image gradients [3], DCT coefficients [28] to field of experts responses [19], the statistics of these responses behave similarly.

In our work, we use complex pyramid wavelet transform due to its reconstruction properties, pyramid representation, and translational invariance property[18].

Under this setup, natural images are most commonly described as images whose real or imaginary coefficients follow a zero-peak, heavy-tailed distribution. As shown in Fig. 3(a), noise smooths the distribution of the wavelet coefficients, while blur compresses the distribution towards zero by reducing contrast of the image. The extent of blur in compressed images is not as significant as with Gaussian blurred images of the same quality because the former also suffers from compression artifacts, which is not conveyed in the distribution of real coefficients.

In our test data, we found it more effective to represent the coefficients by magnitude and phase instead. Similar to real coefficient distributions, we found blur compresses the magnitude distribution and noise changes its shape (Fig. 3(b)). The phase distribution of the coefficients

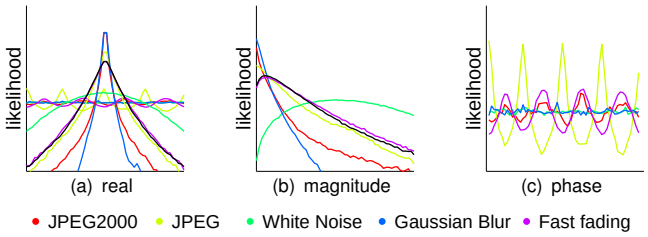



Figure 3. Marginal histograms of wavelet coefficients.

shows a distinct capability to discriminate compression artifacts by showing a oscillating pattern (Fig. 3(c)), resulting from quantization error produced by compression.

The distribution can be compactly described by analytical models. It has been found that the real and imaginary coefficients distribution can be modeled by a generalized Gaussian [3]:

$$p(x;\gamma,\lambda) = \frac{\gamma\lambda^{1/\gamma}}{2\Gamma(1/\gamma)} \exp\left(-\lambda|x|^{\gamma}\right), \qquad (1)$$

and the magnitude distribution can be modeled with a Weibull distribution [7]:

$$p(x;\gamma,\lambda) = \begin{cases} \lambda\gamma x^{\gamma-1}\exp\left(-\lambda|x|^{\gamma}\right) & x \geq 0 \\ 0 & x < 0 \end{cases}. \qquad (2)$$

With these two models, we can evaluate the maximal likelihood of an image as a natural image with a MAP (Maximum A-Posteriori) estimate. In our implementation, we estimated the generalized Gaussian and Weibull parameters with MLE(Maximal Likelihood Estimation). Both the estimated model parameters and the likelihood achieved with these parameters are used as features, in order to convey both prior distribution of parameters and the likelihood of the image under the most likely parameters.

The cross-scale distribution of wavelet coefficient magnitude is also a meaningful feature, as high quality image often show self-similarity across scales [8]. Accordingly, the coarse scale and fine scale coefficients are statistically correlated. The behavior of a distortion on this joint distribution is similar to what occurs to the marginal distributions (Fig. 4(b)); the only difference is that the extent of degradation is larger in the finer scale than the coarser scale.

### 3.2. Distortion texture statistics

When the distortion becomes severe, the likelihood of being a natural image is so low that it is difficult to discriminate the difference in quality using a natural image model. However, a distortion-specific texture typically arises. For instance, JPEG images often present an 8x8 block texture, and JPEG2000 images of high compression rates often suffer from ringing around corners and edges. Therefore, the prominence of these textures is a good measure that complements that of natural image prior based features. We observed in the test data that the cross-scale distribution of coefficient phase a good indicator of distortion-induce local image texture, as shown in Fig. 4(b).

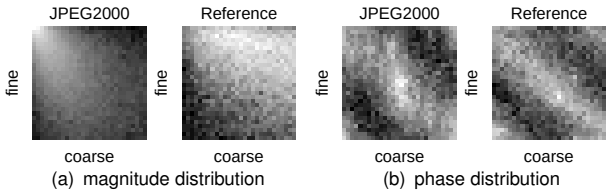



Figure 4. Cross-scale wavelet coefficient distribution for JPEG2000 distorted images of bad quality.

### 3.3. Blur/noise statistics

Although each distortion types has a distinctive way to degrade image structure, we found blur and noise fundamental to various distortion types. In the following we characterize these two degradation processes with three existing techniques.

**Patch PCA singularity** Due to the redundancy of natural images in content, the intrinsic dimensionality of local patches of an image is much lower than its actual dimension. Therefore, we perform principal component analysis on these patches and use the singular values as an indicator of the intrinsic dimensionality of the patch manifold. The singular values are then a meaningful measure of smoothness vs. randomness in a patch. Increases in image blur will squeeze the values to zero as it goes to less significant eigenvectors. In comparison, noise increases evenly in each eigenvectors and results in a more uniform distribution in singular values(Fig. 5(a)).

**Two-color prior based blur statistics** Joshi *et al*. show that the "two-color model", i.e., assuming that all local colors are a linear combination of two colors, is a good model for natural, sharp images [10]. Thus we use the method described by Joshi *et al*. to fit each local patch with a two color model and recover a primary and secondary color layer, an alpha layer, and a residual image. The alpha layer is a good indication of blur as a more peaked alpha distribution indicates a sharper image (i.e., more pixels in the image are exactly equal to the primary or secondary and are not in-

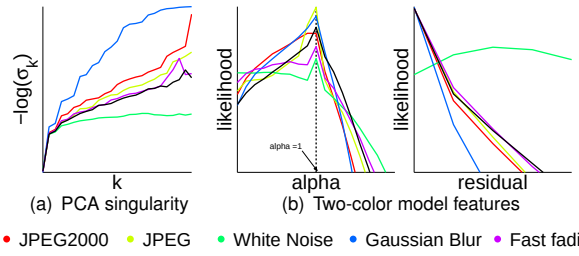



Figure 5. Singularities of $5 \times 5$ patches and Two-color model coefficient histograms.

(a) $M_1$: marginal distribution

| feature | # dimension |
|---|---|
| negative log histogram of magnitude | 720 |
| negative log histogram of real | 720 |
| negative log histogram of phase | 720 |
| MLE estimates of GGD parameter/likelihood of real | 36 |
| MLE estimates of WBD parameter/likelihood of magnitude | 36 |

(b) $M_2$: cross-scale joint distribution

| feature | # dimension |
|---|---|
| negative log histogram of phase | 7200 |
| negative log histogram of magnitude | 7200 |

(c) $M_3$: blur/noise statistics

| feature | # dimension |
|---|---|
| Patch PCA singular values | 25 |
| negative log histogram of alpha value | 10 |
| negative log histogram of residual | 20 |
| step edge based blur/noise estimation | 2 |

Table 1. List of features for each kernel machine

between) and residual image is a good measure of noise artifacts, since independently distributed colored noise lies outside the color model. Therefore, we use the log distribution of the alpha and residual image as image quality features. As shown in Fig. 5(b), blurry images have more transparent pixels and therefore have a less peaked distribution of alpha values in comparison to the reference image, while noisy images are not well modeled by the two-color model and thus have larger residual.

**Direct blur kernel and noise estimation** As a final measure of blur and noise level, we run the blind kernel estimation method of Joshi *et al*. [9] to compute a spatially invariant blur kernel and noise level estimate. This method makes predictions based on edges in the images and tries to infer a kernel that would produce the observed edges from assumed underlying step edges. We use the maximum of the covariance of the blur kernel as a feature and the reported standard deviation of the noise as a separate feature.

## 4. Learning algorithm

While each feature is carefully designed and motivated by the physical properties of image distortions, we cannot expect that each individual feature would work well across all the distortion types. Consequently, we propose exploiting the complementary properties of all the features by combining different predictors to build an estimator that predicts a perceptual image quality measure. In particular, our LBIQ measure consists of an ensemble of regressors trained on three different groups of features (summarized in Table 1), whose outputs are then further combined to produce the final LBIQ score.

As many of our features are negative log histograms, the dimensionality of the features is extremely high. Therefore, we first perform principal component analysis (PCA) for each group of feature in order to reduce to a lower dimension, which is selected by cross validation. These low-dimensional projections are then used to train a $\epsilon$-SVM regression model [21] for each group of feature. Formally, if we denote the low dimension projection of $j^{th}$ feature coefficients/histogram for image $i$ as $x_i^j$, we then solve the following optimization problem:

$$\mathbf{w}^j = \arg\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 + C\sum_i \xi_i + C\sum_i \xi_i^* \quad (3)$$

s.t.

$$\sum_n w_n k(x_i^j, x_n^j) + b^j - y_i \leq \epsilon + \xi_i, \xi_i \geq 0 \quad (4)$$

$$y_i - \sum_n w_n k(x_i^j, x_n^j) - b^j \leq \epsilon + \xi_i, \xi_i, \xi_i^* \geq 0 \quad (5)$$

Here $y_i$ is the subjective image quality of the $i-$th image and $k(\cdot,\cdot)$ is the kernel function. In our implementation, we use radial basis function(RBF) as a kernel:

$$k(x, x_i) = \exp(-\gamma|x - x_i|^2). \quad (6)$$

Once this optimization is performed, the image quality for a test image can be computed as:

$$\bar{y}^j = \sum_i k(x^j, x_i^j) w_i^j + b^j \quad (7)$$

We combine the results of our three individual SVM regression outputs using a weighted linear combination of the the kernel SVM outputs:

$$LBIQ = \sum_j \mu_j \cdot \bar{y}^j. \quad (8)$$

The weights of the linear combination are learned by minimizing prediction error on the validation set:

$$\mu^* = \arg\min \sum_i (LBIQ_i - y_i)^2 \quad (9)$$

This is a least squares minimization problem, and the unique, global optimum can be easily found by solving a small linear system. We experimented training a kernel regression on the concatenation of all features, but found that our combination of multiple SVMs to be more effective and far more efficient.

## 5. Implementation details

Although there is some redundancy in our feature sets, we didn't prone the features as we found the combination of all features outperform using a subset of it. Besides, we projected all features into the log space as we found this would allow the RBF kernel to produce a more meaningful distance metric.

The complex steerable pyramid we use has 3 scales and 4 orientations. When computing histograms of wavelet coefficients, we discretize each coefficient into 60 bins for marginal histograms and 30 bins for joint histograms. In computing the two-color prior based features, we perform color clustering on local $5 \times 5$ patches. The resulting alpha values and residuals are discretized into 20 bins. We

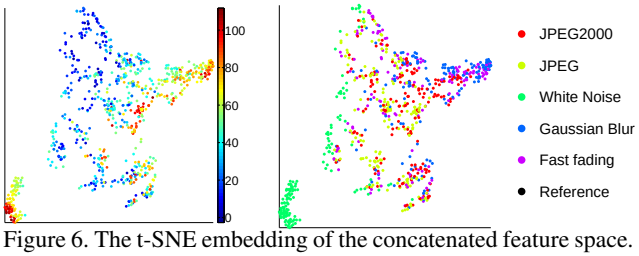
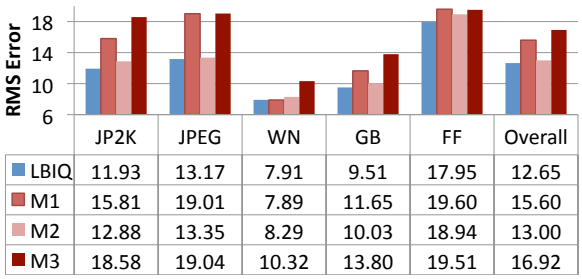



Figure 6. The t-SNE embedding of the concatenated feature space. In both images, each point corresponds to an image. In left image, color encodes subjective quality score; in right image, color encodes distortion type. This embedding verifies that our designed feature set naturally clusters images degraded in a similar way and by a similar extent.




| | JP2K | JPEG | WN | GB | FF | Overall |
|---|---|---|---|---|---|---|
| LBIQ | 11.93 | 13.17 | 7.91 | 9.51 | 17.95 | 12.65 |
| M1 | 15.81 | 19.01 | 7.89 | 11.65 | 19.60 | 15.60 |
| M2 | 12.88 | 13.35 | 8.29 | 10.03 | 18.94 | 13.00 |
| M3 | 18.58 | 19.04 | 10.32 | 13.80 | 19.51 | 16.92 |

Figure 7. Median root mean square error between prediction and subjective score (lower means better). Much of the strength of our LBIQ measure comes from the second kernel machine which trains on cross-scale histogram and statistics of wavelet coefficients. However, the combination of all three machines decreases the Median RMSE up to 1.

actually only used the first 10 bins in the alpha histogram, noticing that the over-saturated alpha values does not help prediction.

We used LIBSVM [1] to perform the regression. Both the dimension of PCA and the SVM parameters are selected by cross validation. In particular, we select PCA dimension for each SVM among 20, 40, 60, and 80, and tune the SVM parameters with a 2D grid search in the log space followed by a simplex search based refinement.

## 6. Experimental results

We perform experiments to explore (1) how well do our features capture variations due to different distortions type, (2) how well does the learned predictive model perform on the task of assessing perceptual image quality, and (3) the failure modes. All experiments are tested on the LIVE image quality assessment database [24].

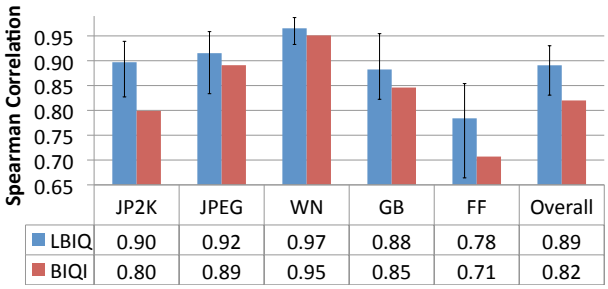
We split the dataset into a training set of 10 reference images, a validation set of 5 reference images and a test set of 14 reference images. Because we are interested in gaining generalization power across images, we made sure that these three sets do not overlap in reference images. The actual number of images in each set is spanned by the 5 types of distortions and 5 to 6 distortion levels per type.

**Local neighborhood embedding of features** We investigate how well the features are able to capture the variations due to distortion type by performing non-linear dimensionality reduction using the t-SNE [26] algorithm to recover a local embedding using the feature representation of the images. Figure 6 shows the embedding of the LIVE dataset with color encoding the subjective quality score(left) and distortion type(right) of each image. We observe that this embedding tends to cluster together blurry images regardless of the source of blur (Gaussian, JPEG2000, *etc*.) while separating them from noisy images. This means we can predict the quality of an image from all relevant distortion types, rather than referring to a specific subset of the training data as BIQI. Further, adjacent images in the t-SNE embedding are typically close in quality, also guaranteeing the exemplar to be expressive enough for prediction.

**Regression performance** We first examine regression performance on individual machines. Fig. 7 shows the root-mean-square prediction error for all distortion types by the three kernel machines we trained (lower values are better). We found that the second SVM, which was trained on cross-scale wavelet coefficients, was best performing because it is both effective in images of good quality with the self-similarity feature and in images with severe distortions with the phase based texture feature. However, the combination of all SVMs takes advantage of the strength of all three machines, and improve the performance on challenging data such as JPEG2000 and fast fading images.

Next, we explore predictive power of the learned SVM regression model. Since many applications require a reliable order of images based on quality, we used Spearman order correlation coefficient as the final performance measure (higher means better). We compared the performance of our LBIQ measure with BIQI [17] as the state of art blind image quality assessment method. We conducted 150 rounds of performance evaluation. In each round we randomly partitioned the dataset into three content-independent sets for training, validation, and testing. As shown in Fig. 8, our method significantly out-performs BIQI for every distortion type without explicitly estimating the distortion types.

We also performed a more thorough comparison to numerous full-reference measure including PNSR, multi-scale SSIM and VIF and blind measure including BIQI and BLIINDS [20]. As shown in Fig. 9, our LBIQ measure is very competitive compared to state of art blind quality indices such as BIQI and BLIINDS. It also outperforms PNSR, the most common full-reference measure. Although it is not as good as that of SSIM and VIF, we believe the performance of LBIQ is comparable to these two contemporary




| | JP2K | JPEG | WN | GB | FF | Overall |
|---|---|---|---|---|---|---|
| LBIQ | 0.90 | 0.92 | 0.97 | 0.88 | 0.78 | 0.89 |
| BIQI | 0.80 | 0.89 | 0.95 | 0.85 | 0.71 | 0.82 |

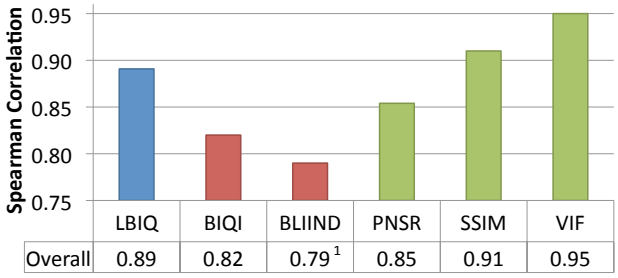
Figure 8. Median spearman order correlation between prediction and subjective score (higher the better) on each distortion type and the entire dataset. Our LBIQ measure significantly outperforms BIQI in all distortion types (as reported in [17]). The error bar provides the minimal and maximal spearman correlations achieved through 150 test runs.

reference-based methods yet the no-reference requirement of LBIQ makes the problem much harder.

**Success and failure modes** Finally we evaluate our method on individual subsets of the same reference images to examine the success and failure modes of our algorithm. We found that the poorly performing datasets typically include periodic textures, such as with roof tiles or water, that are difficult to discriminate from JPEG artifacts (see Fig. 10(a)). In comparison, the reference image achieving highest correlation (Fig. 10(b)) is composed of smooth areas of sky and water and texture areas of stones and grass, as well as step edges: all well conveyed in our natural image model.

## 7. Conclusions

In this work, we dived into the design of low level features for image quality assessment by looking into features derived from natural image statistics, texture features and blur/noise estimation. After in-depth analysis, we found that the magnitude and phase of high-frequency filter responses encapsulate much more information than the conventional real-imaginary (i.e. odd and even filter) representations as they can capture compression artifacts with the phase features. Also, using this representation to model the cross-scale wavelet coefficient distributions renders features very well correlated with perceptual image quality. Our use of direct blur/noise measurements also produces useful features for image quality assessment. Neighborhood embedding of our proposed features well clusters images of similar quality and relevant distortion type, indicating good potential to generalize our learned measure to new images and distortion types. To take advantage of the strength of all features, we used kernel SVM to combine different features. Experiments on the LIVE image quality benchmark dataset shows that our method significantly outperforms state of art no-reference image assessment algorithm in all aspects.




| | LBIQ | BIQI | BLIIND | PNSR | SSIM | VIF |
|---|---|---|---|---|---|---|
| Overall | 0.89 | 0.82 | 0.79[1] | 0.85 | 0.91 | 0.95 |

Figure 9. Median spearman order correlation between prediction and subjective score on the entire dataset through 150 test runs. Our method (blue) perform much better than state of art blind quality indices (red) and comparably to current reference-based quality indices (green) although we do not require a reference image.

For future work, we believe our algorithm can be improved by using a more elegant learning framework such as boosting or image specific weighting to combine the features. Besides, observing that the strength of the RBF kernel relies on dense sample of image, we also consider doing a larger user study to enrich the training set we use. We also think it worthwhile to apply our image quality features and learning algorithm to create a new full-reference measure and, given that LBIQ eliminates distortion-specific bias so well, expect it to become the new frontier in reference-based image quality assessment .

## References

[1] C. C. Chang and C. J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[2] C. Charrier, G. Lebrun, and O. Lezoray. A machine learning-based color image quality metric. In *Color in Graphics, Imaging and Vision (CGIV)*, pages 251–256, 2006.

[1] The performance evaluation of BLIINDS was obtained from a specific test run [20]. In supplemental material we provide test results of LBIQ and BIQI on this specific setting, which show that this is actually a favorable setting to both our method and BIQI. Therefore, potentially the average performance of BLIINDS is lower than 0.79.

(a) Spearman correlation = 0.89 (b) Spearmen correlation = 0.97

Figure 10. Test reference images of the worst and best performance.

[3] T. S. Cho, N. Joshi, C. L. Zitnick, S. B. Kang, R. Szeliski, and W. T. Freeman. A content-aware image prior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision (ECCV)*, volume 3, pages 288–301, 2006.

[5] A. L. F. Cutzu, R. Hammoud. Estimating the photorealism of images: Distinguishing paintings from photographs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 305, 2003.

[6] H. Farid and S. Lyu. Higher-order wavelet statistics and their application to digital forensics. In *IEEE Workshop on Statistical Analysis in Computer Vision (in conjunction with CVPR)*, 2003.

[7] S. Ghebreab, H. S. Scholte, V. A. F. Lamme, and A. W. M. Smeulders. A biologically plausible model for rapid natural scene identification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1–9, 2009.

[8] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[9] N. Joshi, R. Szeliski, and D. Kriegman. Psf estimation using sharp edge prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[10] N. Joshi, C. L. Zitnick, R. Szeliski, and D. Kriegman. Image deblurring and denoising using color priors. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[11] M. Jung, D. Léger, and M. Gazalet. Univariant assessment of the quality of images. *Journal of Electronic Imaging*, 11(3):354–364, 2002.

[12] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 419–426, 2006.

[13] C. Li and A. C. Bovik. Content-partitioned structural similarity index for image quality assessment. *Image Communication*, 25(7):517–526, 2010.

[14] J. Lubin. Sarnoff jnd vision model: Algorithm description and testing. Technical report, Sarnoff Corporation, 1997.

[15] Y. Luo and X. Tang. Photo and video quality evaluation: focusing on the subject. In *European Conference on Computer Vision (ECCV)*, pages 386–399, 2008.

[16] S. Lyu and H. Farid. How realistic is photorealistic? *IEEE Transactions on Signal Processing*, 53(2):845–850, 2005.

[17] A. K. Moorthy and A. C. Bovik. A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters*, 17(5):513–516, 2010.

[18] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40:49–71, 2000.

[19] S. Roth and M. J. Black. Fields of experts: A framework for learning image priors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–867, 2005.

[20] M. A. Saad, A. C. Bovik, and C. Charrier. A dct statistics-based blind image quality index. *IEEE Signal Processing Letters*, 17(6):583–586, 2010.

[21] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000.

[22] H. R. Sheikh, A. C. Bovik, and G. D. Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transaction on Image Processing*, 14:2117–2128, 2005.

[23] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006.

[24] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. *LIVE image quality assessment database release 2*. `http://live.ece.utexas.edu/research/quality`.

[25] P. C. Teo and D. J. Heeger. Perceptual image distortion. *IEEE International Conference Image Processing*, 2:982–986, 1994.

[26] L. J. P. van der Maaten and G. E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2009.

[27] H. R. S. Z. Wang, A. C. Bovik and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 3(4):600–612, 2004.

[28] D. Zoran and Y. Weiss. Scale invariance and noise in natural images. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2209–2216, 2009.