Signal Processing: Image Communication 61 (2018) 54–72

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Perceptual quality evaluation of synthetic pictures distorted by compression and transmission

Debarati Kundu *,[1], Lark Kwon Choi [1], Alan C. Bovik [1], Brian L. Evans [1]

Department of Electrical and Computer Engineering, Wireless Networking and Communications Group, The University of Texas at Austin, TX, United States

ARTICLE INFO

Keywords:
Image quality assessment
Natural scene statistics
Subjective study
Synthetic scenes

ABSTRACT

Measuring visual quality, as perceived by human observers, is becoming increasingly important in a large number of applications where humans are the ultimate consumers of visual information. Many natural image databases have been developed that contain human subjective ratings of the images. Subjective quality evaluation data is less available for synthetic images, such as those commonly encountered in graphics novels, online games or internet ads. A wide variety of powerful full-reference, reduced-reference and no-reference Image Quality Assessment (IQA) algorithms have been proposed for natural images, but their performance has not been evaluated on synthetic images. In this paper we (1) conduct a series of subjective tests on a new publicly available Embedded Signal Processing Laboratory (ESPL) Synthetic Image Database, which contains 500 distorted images (20 distorted images for each of the 25 original images) in 1920×1080 resolution, and (2) evaluate the performance of more than 50 publicly available IQA algorithms on the new database. The synthetic images in the database were processed by post acquisition distortions, including those arising from compression and transmission. We collected 26,000 individual ratings from 64 human subjects which can be used to evaluate full-reference, reduced-reference, and no-reference IQA algorithm performance. We find that IQA models based on scene statistics models can successfully predict the perceptual quality of synthetic scenes. The database is available at: http://signal.ece.utexas.edu/%7Ebevans/synthetic/.

1. Introduction

Recent years have seen tremendous growth in the acquisition, transmission, and storage of both natural and synthetic digital pictures [3]. In addition to pictures captured by optical cameras, picture traffic also often includes synthetic scenes, such as animations, cartoons, comics, games, and internet ads. In all these cases, humans are the final consumers of the visual data and the ultimate goal is to provide a satisfactory quality-of-experience (QoE) [4]. The visual quality of synthetic scenes can be degraded both by the rendering process (e.g. video gaming on standalone devices) and by post acquisition processes such as wireless transmission. Methods of evaluating visual quality play important roles in perceptually optimized design of display devices, rendering engines, and compression standards as well as for maintaining a satisfactory QoE in streaming applications under challenging network conditions.

Although a subjective study with human observers is the most reliable way to gauge perceptual quality of pictures, human studies are time consuming and rarely feasible. The ground-truth data obtained from human observers can be used to benchmark objective IQA algorithms that aim to automate the process of visual quality assessment. Some of the largest natural image databases are the LIVE Image Quality Database [5], the Tampere Image Database 2013 [6], LIVE Challenge Database [7], the Categorical Image Quality Database [8] and EPFL JPEG XR codec [9]. Recently Cadík et al. [10] developed a synthetic image database of computer graphics generated imagery afflicted by distortions such as noise, aliasing, changes in brightness, light leakage and tone mapping artifacts. Some IQA approaches for synthetic images have also been proposed [11,12].

To automate perceptual quality evaluation, two broad categories of objective IQA algorithms are available: reference and blind or no-reference methods, based on the availability (or not) of a reference image. Reference methods may have access to either the complete reference image or some statistical features extracted from it. The

* Corresponding author.
E-mail address: debarati@utexas.edu (D. Kundu).

[1] This paper is an expanded version of [1] and [2]. D. Kundu and L.K. Choi conducted this research as Ph.D. students at UT Austin and now both of them work at Qualcomm. B.L. Evans is also with the Embedded Signal Processing Laboratory (ESPL) at UT Austin. A.C. Bovik is also with the Laboratory for Image and Video Engineering (LIVE) at UT Austin.

https://doi.org/10.1016/j.image.2017.11.004
Received 2 May 2017; Received in revised form 19 November 2017; Accepted 19 November 2017
Available online 23 November 2017
0923-5965/

former defines full-reference (FR) IQA algorithms, while the latter defines reduced-reference (RR) IQA algorithms. The performance of several publicly available state-of-the-art FR-IQA algorithms has been evaluated on popular natural image databases [13–15]. Cadík et al. [10] evaluated the performance of six FR-IQA algorithms and demonstrated that they were sensitive to brightness and contrast changes, could not distinguish between plausible and implausible shading, and failed to localize distortions precisely.

When information about the reference image is not available, no-reference (NR) IQA models are more suitable. Many NR metrics rely on machine learning approaches using features expressive of statistical regularities possessed by pristine images, commonly called natural scene statistics (NSS) models [16,17]. NSS models of good quality natural images hold reliably well irrespective of image content. In NR-IQA model design, it is often assumed that distortions tend to deviate from these statistical regularities. NR-IQA algorithms have not yet been studied in the context of images generated using computer graphics. Herzog et al. [18] propose an NR-IQA metric for quantifying rendering distortions based on machine learning. Their features were chosen heuristically, instead of being based on properties of pristine synthetic images.

With the advent of more powerful Graphics Processing Units, the degree of realism of graphical images [19] has vastly narrowed between natural scenes and high quality synthetic scenes. This does not imply that they share identical statistical properties. In our earlier work [20], we created a database of photorealistic synthetic images and modeled the distribution of mean-subtracted-contrast-normalized (MSCN) pixels [21] obtained from the image intensities using Generalized Gaussian and Symmetric α-Stable distributions. Irrespective of the content, we discovered that the scene statistics of the photorealistic graphics images show substantial similarity to those of natural images.

Here we present the results of a series of subjective tests conducted on distorted synthetic images [2]. The study included 25 high definition reference images, from which 500 images were created by adding controlled amounts of different levels of five commonly encountered artifacts: interpolation, blur and additive noise (processing artifacts), JPEG blocking (compression artifact) and fast-fading effects (transmission artifacts). Every image was evaluated by 64 observers under controlled laboratory conditions in a single stimulus experiment, where the observers rated visual quality on a continuous quality scale. Differential Mean Opinion Scores (DMOS) are supplied which augment the ESPL Synthetic Image Database [20] containing unannotated pristine and distorted images.

When creating the new database, we considered processing, transmission and compression artifacts on synthetic images that have not been considered in any previous subjective study to the best of our knowledge. In [10,18], Cadík et al. mainly focused on computer graphics generated artifacts. However, with the advent of mobile cloud gaming and animations, compression and transmission artifacts commonly occur on synthetic scenes in addition to processing and rendering artifacts.

We also evaluate the performance of more than 50 state-of-the-art FR, RR and NR IQA algorithms on the synthetic scenes and compare them to the subjective test results. The performances of the algorithms are compared and the leading models are subjected to a statistical significance analysis. We hypothesize that with some modifications, NSS based NR-IQA metrics could be successfully applied to photorealistic graphics images. Here we take a first step towards evaluating scene statistics based NR-IQA methods on synthetic scenes, expressed both in the spatial as well as transform domains. Top performing NSS-based NR-IQA algorithms show a high degree of correlation with human perception of distorted image quality on synthetic scenes, which is a promising development in regards to the successful automatic prediction of the perceptual quality of computer graphics generated imagery for which no 'ground truth' information is available.

The remainder of this paper is organized as follows. Section 2 describes the subjective study: methods employed in generating the synthetic scenes and the subjective testing framework. Section 3 outlines the statistics of the pristine and distorted synthetic scenes considered in the database. The quantitative performances of many objective IQA performance on the ESPL Synthetic Image Database are detailed in Section 4, and the results are discussed in Section 5. Section 6 concludes the paper.

2. Human subjective study

2.1. The image database

2.1.1. Source images

A total of 25 synthetic images were chosen. These high quality color images from the Internet are 1920 × 1080 pixels in size. Some of the images are from multiplayer role playing games (such as War of Warcraft), first person shooter games (such as Counter Strike), motorcycle and car racing games, and games with more realistic content (such as FIFA). Single frames were also collected from animated movies: The Lion King, the Tinkerbell series, Avatar, Beauty and the Beast, Monster series, Ratatouille, the Cars series, etc.[1] We incorporated natural and non-photorealistic renderings of human figures and human-made objects, renderings of fantasy figures such as fairies and monsters, close-up shots, wide angle shots, images containing both high and low degrees of color saturation, and background textures without a foreground object.

Fig. 1 shows the 25 reference images. The source complexities of the images were analyzed using the quantitative metrics of scene complexity and colorfulness in [22]. Fig. 2 shows a scatter plot of spatial information vs. colorfulness computed on the images in the ESPL Synthetic Image Database and three other publicly available image quality assessment databases (Cadík's [10], LIVE [5] and TID [6] databases). The scatter plots from the ESPL database, shown in Fig. 2(a) indicate that spatial information and colorfulness in ESPL span a similar range of scene complexity as in the other natural image databases as shown in Fig. 2(c) and (d). In Fig. 2(b), Cadík's Synthetic Image database shows a larger range but sparsely covers the range.

2.1.2. Distortion simulations

Distortions of synthetic images are often more varied than those affecting natural images. This is because distortions of synthetic images arise from two sources: firstly, the image might have artifacts from the rendering process, display and other processing steps, such as tone mapping and contrast amplification, and secondly, distortions might be introduced due to encoding at a low bit-rate or transmission over a network, such as JPEG block artifacts and transmission noise. Other distortions may arise, such as unnaturalness of shading, which can be evaluated only given access to both the rendered 2D scene, and the information provided by the 3D depth buffer. The ESPL database does not contain these other kinds of distortions, focusing instead on transmission and compression artifacts. Since we did not have access to the proprietary 3D models and the lighting information that were used to render the scenes, we introduced distortions on the rendered image themselves.

Three categories of processing artifacts are considered: interpolation (which arises frequently in texture maps, and causes jaggedness of crisp edges), blurring and additive Gaussian noise. Blur artifacts often appear in synthetic images when simulating objects in high motion or depth-of-field. Depth-of-field blur can be synthesized by placing sharper foreground objects on a uniformly blurred background. Evaluation of images with blur is an important component of image quality prediction. Although blur in computer graphics generated imagery is often intentionally introduced, it is not always, and may arise from other sources, such as a loss of resolution during transmission or inadequate resolution

[2] All images are copyright of their rightful owners, and the authors do not claim ownership. No copyright infringement is intended. The database is to be used strictly for non-profit educational purposes.

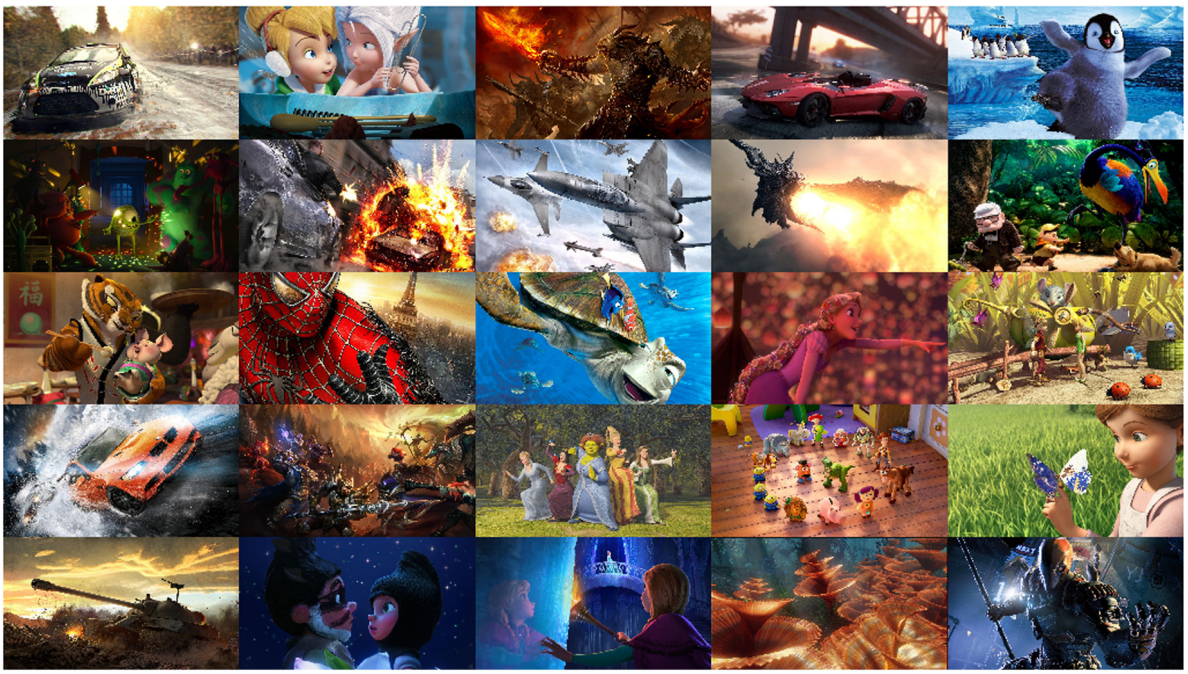

Fig. 1. Sources images in the ESPL database.

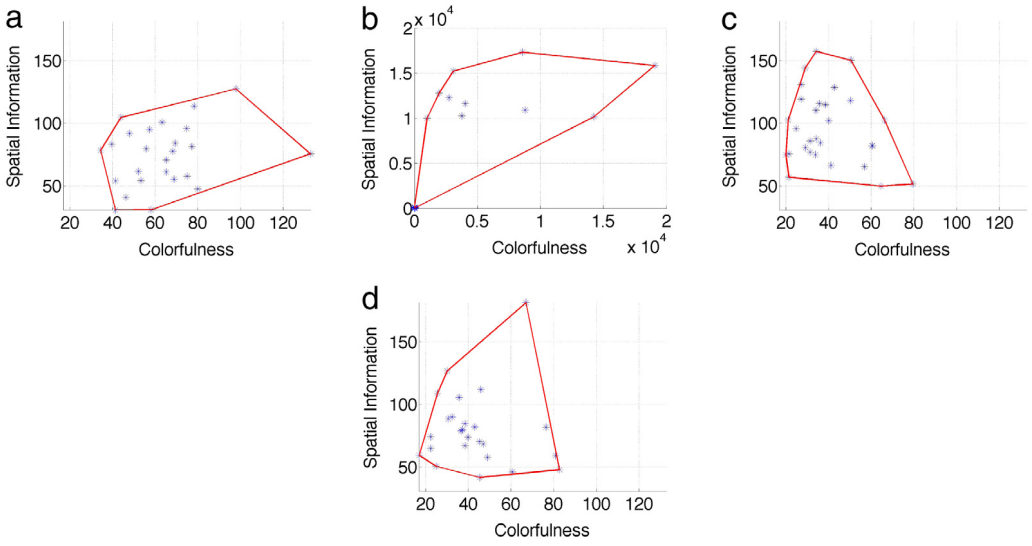

Fig. 2. Spatial Information vs. Colorfulness scatter plots for the source images in the following databases (a) ESPL Synthetic, (b) Cadík's Synthetic Image [10], (c) LIVE [5], (d) TID 2013[6]. Red lines indicate the convex hull of the points in the scatter plot, which approximates the range of scene complexity.

during rendering. This also supplies a way to study the performance of IQA algorithms that were originally intended for use on natural scenes, when applying them on synthetic scenes, since most existing natural image databases [5,6] only contain uniformly blurred images.

In computer graphics, noise may arise in images generated by, for example progressive stochastic ray tracing employing Monte Carlo methods. The visibility of the noise decreases as more samples are employed, but in the case of progressive renderers, high levels of noise may be objectionable to the user early on [23]. Noise may also occur during subsequent transmission or processing of a synthetic image, for example, in mobile viewing scenarios. Our database includes noise of varying degrees for each source image in order to simulate different degrees of noise visibility at, for example different stages of rendering. Since no assumption was made with regards to the distribution of the noise, we decided to simulate Gaussian noise distributed uniformly across the image.

Given the proliferation of mobile applications where rendered 2D images are streamed from the server to 'dumb' clients, we chose to study the effect of compression and transmission artifacts on computer graphics generated images. Thus in the ESPL database, JPEG compression and Rayleigh fast-fading wireless channel artifacts are considered. For each artifact type, the intensity of distortion was varied to create four distorted versions of the same pristine image, so that they range from barely noticeable to a high degree of visual impairment. The following paragraphs briefly describe the types of distortions considered, and the methodology used in their generation. Examples of distortions are shown in Fig. 3.

(1) *Interpolation*: The original images were downsampled using integer downsampling factors ranging from 3 to 6, which are upsampled back using nearest neighbor interpolation. This can be used to simulate jagged edges caused by 'aliasing' in rendering. While bilinear and trilinear interpolation is often used, these methods eliminate jagged

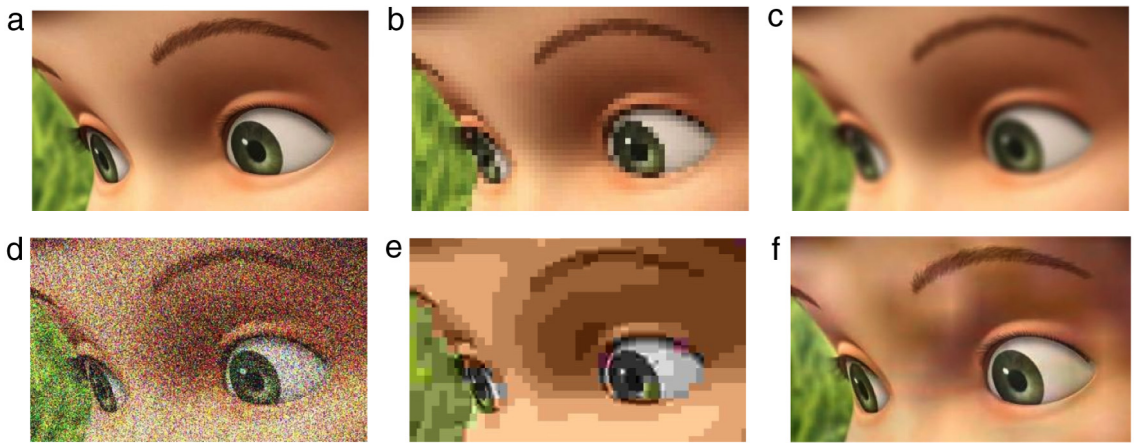

Fig. 3. Cropped parts of (a) original image and images with (b) Interpolation, (c) Gaussian blur, (d) Gaussian noise (e) JPEG compression and (d) Fast Fading artifacts.

edges, hence to retain a higher degree of jaggedness and perceptual separation of these pictures, simple nearest neighbor (zeroth order) interpolation was used.

(2) *Gaussian Blur*: The RGB color channels were filtered using a circularly symmetric 2D Gaussian blur kernel with standard deviation ranging from 1.25 to 3.5 pixels. The same kernel was employed on each of the color channels at every pixel location. Natural photographic images often suffer from severe blur as a consequence of lens defocus and/or motion of the camera. In computer graphics, generating the degree of blur (motion blur or depth-of-field blur) is generally controlled. For this reason, serious blur conditions (e.g. those in the LIVE IQA database [5]) were avoided.

(3) *Gaussian Noise*: Zero mean white Gaussian noise was added to the RGB components of the images (same noise variance were used for all the color channels). The noise standard deviation ranged from 0.071 to 0.316 pixels, using the *imnoise* MATLAB function. Noise can occur in the generation of synthetic images when using random sampling based rendering methods, such as Monte Carlo [24]. When creating the current database, high levels of noise were not simulated since it is unusual and would likely result in de-rendering.

(4) *JPEG compression*: The MATLAB *imwrite* function was used to compress the reference images into JPEG format. The bits-per-pixel (bpp) ranged from 0.0445 to 0.1843. Higher bpp images were not considered, in order to better simulate synthetic picture transmission under restricted bandwidth conditions. Blockiness in images arises from independent coding of spatially correlated adjacent blocks [25]. This can occur in both JPEG still picture compression.

(5) *Simulated Fast Fading Channel*: The reference images were compressed into JPEG2000 bitstreams (with wireless error resilience features enabled and 64 × 64 tiles) and then transmitted over a simulated Rayleigh-fading channel. The images were degraded using the executables provided with the OpenJPEG 2000 encoder [26]. The errors were introduced at the bitstream level. The default functionality was retained for error concealment. The decoder had the ability to check errors with "Termination consistency check". The Rayleigh-fading channel assumption is a simplified model of a realistic wireless channel suffering from multipath effects and Doppler shifts. The signal-to-noise ratio (SNR) was varied at the receiver from 14 to 17 dB to introduce different degrees of transmission errors. SNRs greater than 17 dB did not introduce perceptible distortions due to the error resilience feature of the JPEG2000 codec. We chose to use JPEG2000 in this part of the database to match the fast-fading distortions in the LIVE IQA Database, recognizing that future authors may wish to compare IQA model performance between the different classes of contents.

2.2. Testing methodology

Since the number of images (525) was prohibitively high for a double stimulus setup, a single stimulus continuous evaluation testing procedure with hidden reference [27] was used.

Every image in the database was viewed by each subject, over three sessions of one hour each, with each session separated by roughly 24 h. Each session was divided into two sub-sessions of 25 min with a break of five minutes to reduce visual fatigue and eye strain. The 64 subjects who participated in the test were graduate and undergraduate students at The University of Texas at Austin (Fall 2014), with ages ranging from 18–30 years, mostly without prior experience participating in subjective tests or image quality assessment. The gender ratio of the subjects was roughly 1:1.

Before the start of each session, the study procedure was explained to each subject and verbal confirmation of normal or corrected normal vision was obtained. Subjects viewed approximately 175 test images during each session which were randomly ordered using a random number generator, and randomized for each subject. In order to familiarize themselves with the testing setup, each testing session was preceded by a short training session comprising of around 10 images, which had different content but the same type of distortions as the test images.

2.2.1. Subjective testing display

The user interface for the study was designed on two identical PCs in MATLAB, using the Psychology Toolbox [28]. Both PCs used identical NVIDIA Quadro NVS 285 GPUs and were interfaced to identical Dell 24 inch U2412M displays, which were roughly of the same age with identical display settings. The monitors had 16:10 aspect ratio, 1000:1 static contrast ratio. Any additional digital processing by the monitor was turned off. It was found that the peak luminance of the monitors is 339 cd/m^2, minimum black level is 0.04 cd/m^2 and color gamut is 71% NTSC, 74.3% Adobe RGB, 95.8% sRGB. Each image was displayed on the screen for 12 s and the experiment was carried out under normal office illumination conditions. The ambient lighting was measured using a 200,000 Lux Docooler Digital LCD Pocket Light Meter and was found to be 540lux. Subjects viewed the images from about 2–2.25 times the display height.

The screen resolution was set at 1920 × 1200 pixels, but the images were displayed at their normal resolution (1920 × 1080) without any distortion introduced by interpolation. The pixels per degree was found to be 43.63, assuming a viewing distance of 0.66 m. The top and bottom portions of the display were mid gray. At the end of the image display duration, a continuous quality scale was displayed on the screen, where the default location of the slider was at the center of the scale. It was marked with five Likert-like qualitative adjectives: "Bad", "Poor", "Fair", "Good", and "Excellent" placed at equal distances along the

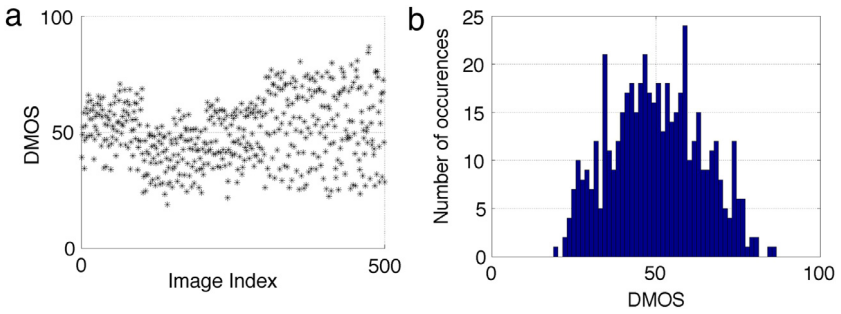

Fig. 4. (a) Scatter Plot and (b) Histogram of DMOS scores obtained on the test images. The DMOS scores span a wide perceptual quality range.

scale. After the subject entered a rating for the image, the location of the slider along the scale was converted into an integer numerical score lying between [0,100]. The subject could take as much time as needed to decide the score, but there was no provision for changing the score once entered or viewing the image again. The next image was automatically displayed once the score was recorded.

2.3. Processing of raw scores

The subject rejection procedure outlined in ITU-R BT.500-13 [27] was used to discard scores from unreliable subjects. The kurtosis of the scores was first used to determine whether the scores assigned by a subject follows a normal distribution. If the kurtosis fell between the values of 2 and 4, the scores were considered to be distributed normally. For the normally distributed scores, a subject was rejected whenever more than 5% of the scores assigned by the subject fell outside the range of two standard deviations from the mean scores, otherwise the subject was considered rejected whenever more than 5% of the scores assigned by her fell outside the range of 4.47 standard deviations from the mean scores. Of the 64 subjects, 12 were treated as outliers and ratings from the remaining 52 subjects were used to calculate the final DMOS.

The differences between the scores on the test images and those on the corresponding reference images were calculated for each subject to account for the preference of certain subjects for certain images. Since any reference image and its distorted version were shown in the same testing session, it is assumed that the quality scale used by the subject remained the same for any single session. The difference scores for the reference images were 0 and were not taken into consideration in subsequent processing steps. The difference scores were normalized by subtracting the mean and dividing by the standard deviation to obtain Z-scores [29]. Z-scores were assumed to be distributed as a standard normal distribution and 99.9% of the scores fell in the range of [−3,3]. The scores were rescaled to lie in the range [0,100]. The DMOS score for each test image was calculated as the mean of the rescaled Z-scores from the 52 subjects remaining after outlier rejection.

One major goal of an IQA database is that the images should span a wide range of visual quality. To illustrate this, the scatter plot and histogram of the DMOS scores of the test images are shown in Fig. 4. The ESPL Synthetic Image Database DMOS scores span the range from 18 to 87. Assuming that the Z-scores assigned by a subject comes from a standard normal distribution, 99% of the Z-scores should lie in the interval [−3,3], which translates to DMOS scores in the range of [0,100]. [18,87] on the DMOS scale corresponds to mean Z-scores in the range of [−1.92,2.22], which covers approximately 96% of the area of the standard normal distribution. The standard deviation of the DMOS scores was 13.89, and the standard error was 0.6212 across the distorted images. Standard error is computed by dividing the standard deviation of the DMOS scores (on a scale of [0,100]) by the square root of the number of distorted images (in this case, 500).

In order to evaluate the degree of consensus among the subjects in judging quality, the subjects were divided into two groups, the DMOS scores for all the images were calculated using the ratings obtained from each group, and the rank correlation was measured between the two sets of DMOS scores. The mean of the Pearson's linear correlation coefficient thus obtained was found to be 0.9813 over 50 such randomized splits. This shows a high level of agreement among the users when evaluating the quality of the images.

3. Synthetic scene statistics

In this section we discuss synthetic scene statistics and the way in which the presence of compression and transmission distortions affects them. For this purpose, we leverage the models employed by past NSS studies. The different statistical properties of natural and synthetic scenes is a well-researched topic. In [19], the authors analyzed the distributions of the first and higher order wavelet coefficients of natural and photorealistic images obtained by a decomposition using Quadrature Mirror Filters and found subtle differences that could be used to distinguish these two classes of images.

In [20] we use spatial domain scene statistics extracted from the Mean-Subtracted-Contrast-Normalized (MSCN) coefficients [21] computed on the luminance component of the images. We collected 221 synthetic images of highly diverse picture content [20]. The skewness and excess kurtosis values computed from the empirical distributions of the MSCN coefficients of the synthetic images were compared with those of natural images obtained from the Berkeley segmentation dataset [30], as shown in Fig. 5.

The empirical histogram skewness values are mostly clustered around the zero value, with some showing small amounts of positive shifts. This shows that a symmetric non-skewed distribution should be able to model the variation in most of the images. However, when compared to the natural images, some of the synthetic images tend to show a higher degree of excess kurtosis. This is common if the images show large textureless regions, and/or abrupt changes of contrast, e.g., those occurring across sharp boundaries. This is a common feature of cartoon images. In this case, most of the MSCN coefficients are equal to or are near zero, hence, a sharp spike is observed near the origin of the MSCN histogram. When modeling these type of images, the Symmetric α-Stable distribution with small values of α was found to be a better model than the GGD models.

The next step is to estimate the GGD mean μ, scale α, and shape β parameters from the sample histograms. This was done by the method of maximum likelihood estimation [31]. In order to understand how much these parameters differ for natural and synthetic images being considered, we plotted the histograms of the scale and shape parameters. Fig. 6 shows the histogram of the GGD shape parameter β. A substantial overlap in the distribution of β was found among natural and synthetic images, suggesting that it is a poor discriminator between computer generated imagery and natural images. In fact, a natural scene and a highly non realistic synthetic scene may have very similar distributions of MSCN coefficients. For natural images, β tends to cluster around 2, which corresponds to the shape parameter of a Gaussian distribution. For synthetic images, the peak of the distributions usually occurs for

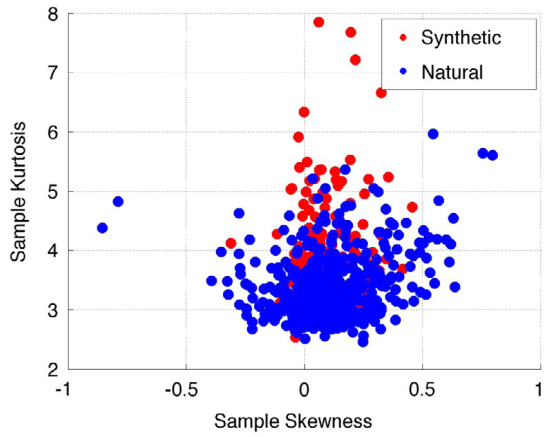

Fig. 5. Scatter plot of skewness (X-axis) and kurtosis (Y-axis) of 221 synthetic and 500 natural images. Note that while most the synthetic images show zero or very small skewness values, some of them might exhibit high excess kurtosis, indicating heavily peaked distribution of the MSCN coefficients.

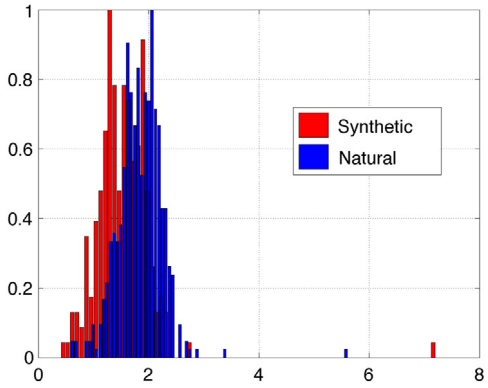

Fig. 6. Normalized histogram of the shape parameter β obtained over 221 synthetic and 500 natural images. Note how β for natural images tends to cluster around 2, indicating a Gaussian-like distribution of the MSCN coefficients. The synthetic images show more variability in the value of β.

$\beta < 2$, which means that more leptokurtic GGDs are needed to model the MSCN coefficients.

In order to quantify the extent to which the probability models fit the empirical distributions, we used the mean-squared error, and the J-divergence. Given two probability distributions, the J-divergence between them is defined as the arithmetic mean of the two possible Kullback–Leibler distances (provided they exist). We also performed χ^2 tests at the 1% confidence level. The null hypothesis was assumed to be the distribution which we were trying to fit to the empirical spatial domain data, and for all the cases, the null hypothesis was accepted. The chi-square values in all cases were found to be smaller than the upper cut off of the χ^2 distribution, $\chi^2_{(0.01)} = 6.635$ with degree of freedom $= 1$, which indicates that the values were generated from the fitted distributions, instead of by chance. Table 1 show the values of the mean square error, J-divergence, and Pearson's χ^2 values for the distributions, calculated on the synthetic image database.

Table 1
Mean square error, J-Divergence, and Pearson's χ^2 values for the distributions fitted to the histogram of the MSCN coefficients of an image for all the considered parametric families, averaged over the entire database.

	MSE	J	χ^2
GGD	**0.00257**	**0.0386**	0.00252
SαS	0.00264	0.0474	0.00174

In [32] it was found that the presence of distortions alters the model statistics of the MSCN coefficients from a Gaussian-like signature. We have found similar behavior of synthetic scenes afflicted by transmission and compression errors. Fig. 7 shows several exemplar plots. This has led us to the hypothesis that, like natural scenes, scene-statistics based approaches can be used to evaluate the distortions present in synthetic images [1].

4. Results

This section outlines the results of evaluating the performance of more than 50 publicly available state-of-the-art objective IQA algorithms on the ESPL Synthetic Image Database. Table 2 lists the algorithms considered. An interested reader may visit the references for more details on the algorithms. The performance metrics and the methods of statistical evaluation are also provided.

4.1. Correlation measures

The performance of the objective IQA algorithms mentioned in Table 2 were evaluated using two correlation measures: the Spearman Rank Order Correlation Coefficient (SROCC) and the Pearson Linear Correlation Coefficient (PLCC) after non-linear regression on the objective IQA scores using a five-parameter monotonic logistic function, following the procedure outlined in [29]. Throughout this paper, all of the results were computed following a non-linear logistic fit of the objective IQA scores.

4.2. Root mean square error

The accuracy of the quality predictions delivered by the IQA algorithms was quantified using the Root Mean Square Error (RMSE) [5] between the DMOS scores and the objective IQA scores (after non-linear regression), and are also tabulated in Tables 4, 6 and 9 which

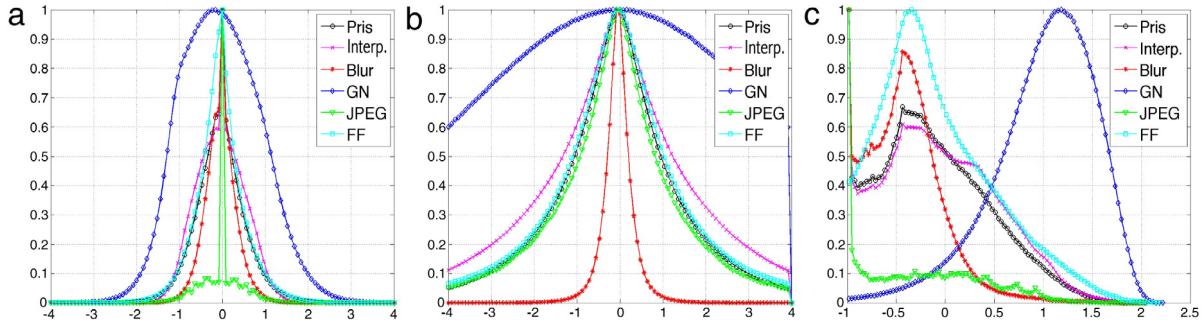

Fig. 7. Histograms of (a) MSCN pixels, (b) Steerable Pyramid Wavelet Coefficients and (c) Curvelet Coefficients of pristine and distorted image patches obtained from the ESPL Synthetic Image Database. The figure shows how distortions change the statistics of pristine images. The legends Pris, Interp., Blur, GN, JPEG, FF refer to pristine images, images with interpolation distortion, blur distortion, additive white Gaussian noise, JPEG compression and simulated transmission over a Rayleigh fast-fading wireless channel, respectively.

Table 2
List of Image Quality Assessment algorithms evaluated in this study.

Category of IQA	Method		Algorithm
Full Reference	Mean Square Error		Peak Signal-to-Noise Ratio Metric based on Singular Value Decomposition (MSVD) [33]
	Structural Similarity based		Structural Similarity Index (SSIM) [34] Multi-scale Structural Similarity Index (MS-SSIM) [35] Quarternion Structural Similarity Index (QSSIM) [36] Universal Quality Index (UQI) [37]
	Human Visual System model based		Visual Difference Predictor (VDP) [38] High Dynamic Range VDP (HDR-VDP-2) [39] Noise Quality Measure (NQM) [40] Weighted Signal-to-Noise ratio (WSNR) [41] Peak Signal-to-Noise ratio-Human Visual System (PHVS) [42] Peak Signal-to-Noise ratio-Human Visual System-A (PHA) [43] Peak Signal-to-Noise ratio-Human Visual System(modified) (PHVSM) [44] Peak Signal-to-Noise ratio-Human Visual System(modified)-A (PHMA) [43]
	Information Theory based		Information Fidelity Criterion (IFC) [45] Visual Information Fidelity (VIF) [46] Information Content Weighted SSIM (IW-SSIM) [47]
	Feature Similarity based		Feature Similarity Index (FSIM) [48] Gradient Magnitude Similarity Deviation (GMSD) [49] Gradient Similarity Measure (GSM) [50] Riesz-transform based Feature Similarity Metric (RFSIM) [51]
	Visual Saliency based		Visual Saliency-Induced Index (VSI) [52] Spectral Residual Based Similarity (SR-SIM) [53]
	Strategy based		Most Apparent Distortion algorithm (MAD) [8] Visual Signal-to-Noise ratio (VSNR) [54]
Reduced Reference	Natural Scene Statistics based		Reduced-Reference Image Quality Assessment (Wavelet Domain) (RRIQA) [55] RRIQA with Divisive Normalization (RRDNT) [56] Reduced-Reference Entropic Differences (RRED) [57]
	Image Feature based		RRIQA with Weibull Statistics [58] RRIQA with Sub-Image Similarity [59] RRIQA with Edge-Pattern map [60]
No-Reference	Artifact based	Blur	Local Phase Coherence based sharpness index (LPCM) [61] Metric based on Cumulative Probability of Blur Detection (CPBD) [62] Metric based on Just-Noticeable Blur (JNBM) [63] Spectral and Spatial Measure of Local Perceived Sharpness (S_3) [64] Fast Wavelet-Based Image Sharpness Estimation (FISH) [65]
		Blocking	NRIQA of JPEG compressed images (JPEG-NR) [66] NRIQA of JPEG compressed images via Quality Relevance Map (NJQA) [67]
		Noise	Noise-level Estimation using weak textured patches (NLWT) [68] Fast Noise Variance Estimation (FNVE) [69]
	Learning based	Spatial Domain	Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) [32] Natural Image Quality Evaluator (NIQE) [21] Derivative Statistics-based QUality Evaluator (DESIQUE) [70] HDR Image GRADient based Evaluator (HIGRADE-1 and HIGRADE-2) [71][a] Gradient Magnitude and Laplacian of Gaussian based NR-IQA (GM-LOG) [72]
		Transform Domain	Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE) [73] Complex-DIIVINE (C-DIIVINE) [74] Blind Image Quality Index (BIQI) [75] BLind Image Integrity Notator using DCT Statistics-II (BLIINDS-II) [76] General Regression Neural Network IQA (GRNN) [77] NR-IQA based on Curvelets (CurveletQA) [78] NR-IQA based on Anisotropy (Anisotropy) [79] COdebook Representation for No-Reference Image Assessment (CORNIA) [80] Topic Model based IQA (TM-IQA) [81]

[a] HIGRADE-1 (L) and HIGRADE-2 (L) are the corresponding algorithm using features extracted from the L-channel only when the image is represented in the LAB color space.

also includes the reduced $\tilde{\chi}^2$ statistic between the algorithm scores and the DMOS for the various algorithms, after logistic function fitting. The reduced $\tilde{\chi}^2$ statistic indicates whether the difference between the DMOS scores and the objective IQA scores (after non-linear regression) is well-modeled as following a normal distribution.

4.3. Outlier ratio

The prediction consistency of the objective IQA algorithms was evaluated by measuring the outlier ratio (OR) [82]. Let Q'_j be the objective IQA algorithm score obtained for image j on the ESPL Synthetic Image Database after the logistic fit. Let $Z'_j = \{z_{ij}\}, i = 1, 2, \ldots, M$ be the Z-scores obtained for image j for M observers and σ_j be the corresponding standard deviation. An image is defined as an outlier if $Q'_j - Z'_j > 2\sigma_j$. The outlier ratio is given by the ratio of the number of outliers to the total number of images (expressed as %). These values are also given in Tables 4, 6 and 9.

4.4. Statistical significance and hypothesis testing

To understand whether the differences in performances of the compared algorithms are statistically significant (based on the number of sample points used), we used two variance-based F-tests, following the similar procedures in [29].

(1) *Hypothesis testing based on individual quality scores*: The 'optimal' or 'null' mode obtained from the subjective study predicts the quality

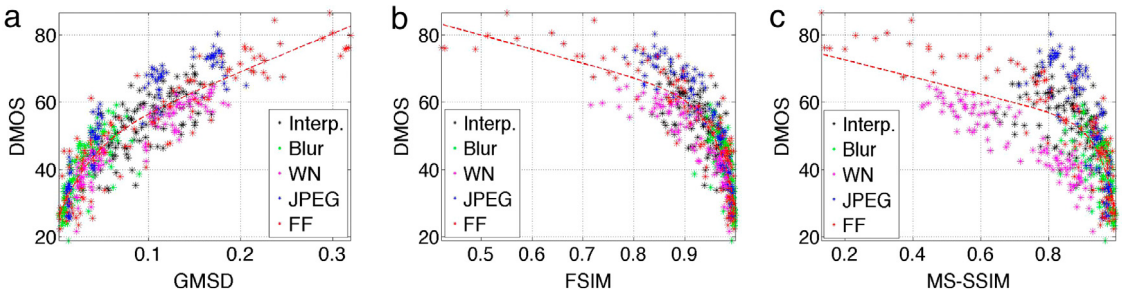

Fig. 8. Scatter plots of predicted IQA scores vs. DMOS for some selected full-reference IQA Algorithms.

of an image using the DMOS score obtained by averaging the Z-scores obtained from all of the subjects. However, subjects show individual variations when assigning subjective ratings to an image. The baseline residual of the null model comprises the differences between the individual ratings assigned by the different subjects and the averaged DMOS, which cannot be taken into account by any objective IQA algorithm. A similar quantity was defined for each objective IQA algorithm. The F-test assumes that the residuals are independent samples drawn from a Gaussian distribution. The F-ratio is the ratio of the variance of the model residual to that of the null residual at the 95% significance level.

$$\text{Null residual (individual ratings)} = z'_{ij} - DMOS_j, \quad i = 1, 2, \ldots, M \text{ and } j = 1, 2, \ldots, N \tag{1}$$

$$\text{Model residual (individual ratings)} = z'_{ij} - Q'_j, \quad i = 1, 2, \ldots, M \text{ and } j = 1, 2, \ldots, N \tag{2}$$

(2) *Hypothesis testing based on DMOS scores*: This hypothesis test was used to determine the statistical superiority (or inferiority), if any of one objective IQA algorithm over another by using the residual error between the quality predictions by an objective IQA algorithm and the DMOS scores obtained from the human subjects. This F-test also assumes Gaussianity of the residuals.

$$\text{Model residual (averaged ratings)} = Q'_j - DMOS_j, \quad j = 1, 2, \ldots, N \tag{3}$$

Following a procedure similar to the prior one outlined for individual quality scores, an F-test was performed on the ratio of the variances of the model residuals of the two candidate objective IQA algorithms at the 95% significance level.

5. Discussion of IQA algorithm performance

This section observes trends and draws conclusions from the experimental results on the IQA algorithms discussed in Section 4. Figs. 8 and 9 show the scatter plot between the predicted IQA scores and DMOS for selected full-reference and no-reference algorithms, respectively. The dotted lines indicate the non-linear regression fit to the IQA scores, as explained in Section 4. For the better-performing IQA models, the DMOS scores and the IQA algorithm predictions are more closely clustered.

5.1. Discussion of results for FR-IQA algorithms

This part of our study aims at benchmarking the performance of different categories of IQA algorithms when applied to specific distortion categories. We evaluated the performance of 27 state-of-art FR-IQA algorithms on the ESPL Synthetic Image Database, where the source codes were provided by the authors [84]. Among these, the single-scale algorithms were evaluated on images rescaled by a factor dependent on the image dimension and the viewing distance [83]. We isolated those distortion categories where the FR-IQA algorithms perform worse and gained insights into those factors that enable certain types of FR-IQA algorithms to perform well, such as using color information instead of only luminance and/or efficient pooling strategies.

In Table 3, PSNR (row 26) is outperformed by nearly all of the other objective IQA algorithms (except for SSIM on row 27 and MSVD on row 28), but it performs reasonably well on additive noise and fast-fading artifacts since it captures high-frequency distortions. The SSIM and MS-SSIM IQA algorithms, which perform exceedingly well on the LIVE database [5], achieve less impressive performance on our database, primarily due to the low correlation with human judgments on certain classes of distortions, such as interpolation, which has not been included in any of previous IQA database. However, SSIM is a single-scale measure; hence the image scale and viewing distance may affect performance. Moreover, SSIM performs optimally well on images satisfying NSS models [85], while the ESPL database contains synthetic images. Based on the rule-of-thumb proposed in [83], if the SSIM index is computed on downsampled images (SSIM-D in row 14), a better degree of correlation is achieved against the human ground truth subjective data. Indeed, if the scale is chosen appropriately, SSIM-D outperforms MS-SSIM (in row 23).

Almost all of the existing IQA algorithms fail to accurately predict the subjective ratings of the interpolation artifact. MAD [8] achieved the highest correlation against DMOS for this type of artifact. MAD advocates multiple strategies for determining the overall image quality, based on whether the distortions are near-threshold or supra-threshold. Low down-sampling factors result in near-threshold artifacts, which might appear almost imperceptible, especially at a normal viewing distance. Although both interpolation and JPEG compression lead to blocking artifacts, the algorithms which perform exceedingly well on JPEG compression distortion (such as FSIM [48]) show much-less impressive performance on interpolation artifacts. This is because the two types of blocking artifacts deviate the statistics of the pristine scenes in different ways (Fig. 7). In the future, we plan to study the effects of varying display sizes on error visibility for interpolated images, which could prove valuable for display designers. Blurred images also led to a lower degree of correlation with human scores compared to other categories. Thus our subjective test reveals a significant performance gap for certain distortion categories between synthetic and natural images, topics which future researchers may address.

Several recently proposed FR-IQA algorithms, such as GMSD [49], FSIM [48], VSI [52], SR-SIM [53] and MAD [8] correlate rather well with human subjectivity. GMSD uses the standard deviation of the gradient map as a pooling strategy. FSIM takes into account image gradient magnitude and phase congruency (a dimensionless measure of significance of local structure), then uses it as a pooling strategy. VSI and SR-SIM use more sophisticated pooling strategies based on visual fixations. Hence, irrespective of whether the image is natural or synthetic, IQA algorithms that use more efficient pooling strategies by taking into account localized distortions perform better than other IQA algorithms, as suggested by [13]. This shows that finding "salient" image regions can improve the performance of IQA algorithms. Some of the IQA algorithms which model different aspects of the human visual system (HVS), such as NQM, VSNR, PSNR-HVSM, perform worse than

Table 3
Spearman's Rank Ordered Correlation Coefficient (SROCC) and Pearson's Linear Correlation Coefficient (PLCC) between the algorithm scores and the DMOS for various FR-IQA Algorithms along with algorithm computation time (on a Macintosh laptop having 8 GB RAM, 2.9 GHz clock, Intel Core i7 CPU). PSNR is Peak Signal-to-Noise Ratio. Table is sorted in descending order of SROCC for the "Overall" category. The numbers within parentheses in the "Overall" category show the confidence intervals on correlation values, computed by bootstrapping using 100 samples. Bold values indicate the best performing algorithm for that category. SSIM-D computes SSIM on images downsampled by a factor determined by the image dimensions and the viewing distance [83].

	IQA	Interp.		Blur		Additive Noise		JPEG Blocking		Fast Fading		Overall (Confidence Interval)		Time (s)
		SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	
1	GMSD	0.727	0.743	0.827	0.838	**0.923**	0.925	0.918	0.954	0.922	0.915	**0.892** (0.877,0.905)	**0.890** (0.871, 0.905)	0.014
2	SR-SIM	0.752	0.772	0.823	0.729	0.916	0.878	0.925	0.832	0.920	0.913	0.880 (0.853, 0.902)	0.873 (0.834, 0.891)	0.042
3	FSIMc	0.694	0.697	0.802	0.808	0.902	0.917	0.938	0.874	0.911	0.907	0.877 (0.855, 0.896)	0.874 (0.850, 0.891)	0.133
4	FSIM	0.692	0.697	0.801	0.809	0.902	0.917	**0.940**	**0.965**	0.907	0.902	0.876 (0.857, 0.898)	0.872 (0.854, 0.892)	0.165
5	VSI	0.692	0.663	0.811	0.814	0.914	0.883	0.880	0.844	0.923	0.917	0.872 (0.856, 0.897)	0.873 (0.855, 0.889)	0.114
6	MAD	**0.788**	**0.806**	0.813	0.815	0.909	0.915	0.933	0.950	**0.927**	0.917	0.863 (0.834, 0.880)	0.869 (0.846, 0.889)	1.257
7	PHA	0.716	0.717	0.781	0.772	0.842	0.883	0.898	0.927	0.905	0.900	0.863 (0.844, 0.884)	0.861 (0.840, 0.879)	0.458
8	PHMA	0.737	0.755	0.823	0.822	0.852	0.889	0.924	0.953	0.911	0.904	0.853 (0.822, 0.878)	0.859 (0.837, 0.881)	0.234
9	PHVS	0.717	0.718	0.778	0.771	0.876	0.885	0.896	0.926	0.903	0.897	0.853 (0.832, 0.874)	0.846 (0.822, 0.863)	0.195
10	GSM	0.676	0.630	0.780	0.655	0.919	**0.927**	0.903	0.881	0.921	0.678	0.839 (0.811, 0.866)	0.627 (0.584, 0.697)	0.054
11	PHVSM	0.736	0.748	**0.839**	**0.840**	0.854	0.874	0.925	0.954	0.905	0.902	0.833 (0.808, 0.857)	0.838 (0.813, 0.862)	0.207
12	IW-SSIM	0.761	0.793	0.823	0.836	0.902	0.921	0.933	0.959	0.925	**0.922**	0.827 (0.796, 0.849)	0.831 (0.790, 0.847)	0.663
13	RFSIM	0.706	0.717	0.763	0.766	0.906	0.912	0.907	0.930	0.891	0.886	0.825 (0.794, 0.846)	0.826 (0.796, 0.850)	0.218
14	SSIM-D	0.688	0.681	0.772	0.777	0.915	0.922	0.904	0.943	0.914	0.906	0.796 (0.758, 0.823)	0.801 (0.775, 0.833)	0.052
15	IFC	0.728	0.722	0.792	0.789	0.837	0.845	0.913	0.922	0.850	0.858	0.791 (0.757, 0.829)	0.786 (0.742, 0.814)	1.199
16	NQM	0.751	0.767	0.831	0.837	0.879	0.893	0.919	0.936	0.859	0.854	0.789 (0.760, 0.818)	0.796 (0.761, 0.822)	0.107
17	QSSIM	0.697	0.693	0.774	0.647	0.913	0.925	0.905	0.940	0.918	0.915	0.786 (0.758, 0.815)	0.793 (0.753, 0.812)	0.104
18	UQI	0.707	0.704	0.780	0.678	0.816	0.824	0.869	0.889	0.848	0.848	0.767 (0.718, 0.791)	0.776 (0.748, 0.818)	0.040
19	CIELAB	0.575	0.572	0.623	0.627	0.840	0.870	0.910	0.925	0.875	0.878	0.758 (0.716, 0.795)	0.772 (0.736, 0.812)	0.116
20	VIF	0.716	0.737	0.788	0.802	0.874	0.903	0.901	0.925	0.761	0.778	0.755 (0.710, 0.799)	0.748 (0.705, 0.782)	6.337
21	WSNR	0.627	0.638	0.773	0.777	0.821	0.825	0.886	0.911	0.839	0.845	0.744 (0.705, 0.780)	0.745 (0.700, 0.775)	0.048
22	HDR-VDP	0.662	0.699	0.766	0.795	0.854	0.861	0.791	0.790	0.856	0.863	0.712 (0.666, 0.753)	0.738 (0.698, 0.768)	2.245
23	MS-SSIM	0.623	0.635	0.646	0.650	0.908	0.924	0.871	0.891	0.903	0.900	0.699 (0.660, 0.742)	0.712 (0.678, 0.764)	0.276
24	VIFP	0.651	0.661	0.624	0.623	0.895	0.912	0.878	0.887	0.791	0.802	0.693 (0.655, 0.729)	0.695 (0.655, 0.730)	0.244
25	VSNR	0.607	0.619	0.611	0.600	0.848	0.889	0.756	0.771	0.884	0.882	0.690 (0.639, 0.734)	0.696 (0.652, 0.741)	0.237
26	PSNR	0.565	0.591	0.481	0.492	0.864	0.897	0.695	0.702	0.846	0.858	0.590 (0.529, 0.632)	0.603 (0.556, 0.645)	0.149
27	SSIM	0.463	0.476	0.440	0.455	0.909	0.927	0.633	0.653	0.797	0.815	0.542 (0.482, 0.590)	0.531 (0.481, 0.592)	0.570
28	MSVD	0.165	0.160	0.403	0.397	0.415	0.423	0.652	0.630	0.363	0.400	0.261 (0.176, 0.341)	0.253 (0.167, 0.321)	2.272

Table 4
Root-mean-square error (RMSE), reduced $\tilde{\chi}^2$ statistic between the algorithm scores and the DMOS for various FR-IQA algorithms (after logistic function fitting) and outlier ratio (expressed in percentage) for each distortion category. The bold values indicate the best performing algorithm for that category.

	IQA	Interp.			Blur			Additive Noise			JPEG Blocking			Fast Fading			Overall		
		RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR
1	GMSD	5.675	3.202	**0.0**	6.400	0.632	**0.0**	4.789	1.746	**0.0**	8.756	4.411	**0.0**	12.355	2.632	1.0	10.689	2.065	**0.0**
2	SR-SIM	6.935	1.230	**0.0**	7.065	2.159	1.0	4.641	1.048	**0.0**	7.463	0.801	2.0	12.549	3.105	14.0	10.808	3.539	7.6
3	FSIMc	7.308	2.886	**0.0**	6.856	0.768	**0.0**	5.301	1.376	**0.0**	8.093	1.461	**0.0**	8.166	2.626	6.0	9.182	3.043	4.2
4	FSIM	6.876	2.964	**0.0**	5.885	1.094	**0.0**	5.835	1.782	**0.0**	7.285	1.559	**0.0**	9.382	2.195	5.0	9.373	2.265	4.6
5	VSI	5.441	0.860	1.0	5.128	1.141	**0.0**	3.904	2.757	**0.0**	6.999	0.657	**0.0**	9.227	1.883	13.0	7.725	1.014	5.6
6	MAD	6.225	1.682	**0.0**	6.012	1.492	**0.0**	4.113	1.020	**0.0**	7.264	0.509	**0.0**	8.122	2.979	**0.0**	8.145	6.005	0.4
7	PHA	6.261	1.164	**0.0**	5.950	2.803	1.0	4.665	4.098	**0.0**	5.960	0.281	**0.0**	7.589	2.138	1.0	7.483	2.957	0.4
8	PHMA	5.981	2.862	**0.0**	**5.069**	1.439	**0.0**	5.016	3.620	**0.0**	4.756	1.733	**0.0**	7.481	2.077	**0.0**	8.111	3.507	1.0
9	PHVS	**5.298**	1.164	**0.0**	5.594	3.014	1.0	4.143	2.923	**0.0**	5.521	0.621	**0.0**	**6.918**	0.829	**0.0**	6.886	2.016	0.6
10	GSM	6.402	1.127	**0.0**	5.506	2.548	1.0	4.000	2.546	**0.0**	7.670	0.405	**0.0**	10.626	2.612	17.0	9.214	1.534	7.6
11	PHVSM	6.157	2.863	**0.0**	5.094	1.431	**0.0**	**3.791**	3.704	**0.0**	**4.740**	1.576	**0.0**	7.087	1.669	1.0	**6.335**	3.284	0.4
12	IW-SSIM	6.402	4.254	**0.0**	5.491	1.511	**0.0**	3.989	1.383	**0.0**	7.721	1.821	1.0	10.765	3.006	3.0	9.283	2.109	1.0
13	RFSIM	8.607	1.263	**0.0**	5.424	1.340	**0.0**	4.704	0.893	**0.0**	8.455	0.601	1.0	11.731	4.082	**0.0**	10.437	2.434	2.6
14	SSIM-D	7.213	2.718	**0.0**	8.213	1.025	**0.0**	4.462	1.403	**0.0**	11.477	1.486	3.0	8.847	1.946	4.0	11.171	4.429	7.6
15	IFC	6.344	1.422	2.0	6.866	1.314	**0.0**	6.206	0.638	**0.0**	9.522	0.608	**0.0**	9.612	1.632	3.0	8.818	1.729	6.8
16	NQM	7.409	1.118	**0.0**	7.021	1.040	**0.0**	5.375	3.101	**0.0**	6.946	1.146	**0.0**	8.859	1.064	**0.0**	10.415	1.934	2.4
17	QSSIM	8.813	3.107	**0.0**	8.578	3.258	**0.0**	9.142	0.694	**0.0**	12.267	0.665	**0.0**	16.062	2.783	1.0	13.426	5.622	5.4
18	UQI	6.697	1.550	**0.0**	7.307	1.928	1.0	4.150	0.318	**0.0**	7.883	1.379	7.0	10.196	1.177	3.0	10.017	3.893	2.6
19	CIELAB	6.447	0.234	**0.0**	5.872	1.052	**0.0**	5.362	4.063	**0.0**	9.675	0.590	**0.0**	8.779	0.651	1.0	9.357	2.711	3.0
20	VIF	7.038	1.417	**0.0**	7.497	1.560	**0.0**	5.123	4.230	**0.0**	10.647	1.648	**0.0**	8.230	2.038	7.0	10.015	6.305	4.2
21	WSNR	5.742	1.580	1.0	5.200	0.095	**0.0**	4.660	0.910	1.0	6.363	1.019	1.0	9.028	0.974	4.0	8.567	1.058	4.8
22	HDR-VDP	5.980	1.766	**0.0**	5.322	1.515	**0.0**	4.846	0.493	**0.0**	4.785	1.263	5.0	7.316	1.598	3.0	7.370	0.667	4.6
23	MS-SSIM	6.535	3.464	**0.0**	5.880	1.601	**0.0**	4.503	0.417	**0.0**	8.727	2.223	**0.0**	10.012	2.142	5.0	10.247	6.758	8.4
24	VIFP	6.093	2.373	**0.0**	5.693	1.058	1.0	4.448	3.016	**0.0**	6.001	2.211	1.0	10.858	1.500	9.0	9.209	2.751	5.4
25	VSNR	6.899	0.544	1.0	7.103	0.201	1.0	4.072	0.267	**0.0**	7.201	0.392	6.0	9.955	2.400	1.0	11.006	4.417	6.8
26	PSNR	6.681	1.753	**0.0**	6.822	2.059	1.0	5.591	6.533	**0.0**	9.249	1.316	8.0	13.111	2.197	1.0	12.697	1.682	9.2
27	SSIM	7.325	1.278	2.0	7.278	1.727	**0.0**	5.005	1.156	**0.0**	6.006	0.237	11.0	8.183	2.069	6.0	8.818	1.167	11.2
28	MSVD	6.260	1.880	2.0	5.934	0.603	1.0	4.697	2.038	2.0	5.936	2.877	17.0	7.554	0.880	27.0	7.168	3.128	16.8

Table 5
Spearman's Rank Ordered Correlation Coefficient (SROCC) and Pearson's Linear Correlation Coefficient (PLCC) between the algorithm scores and the DMOS for various RR-IQA algorithms along with algorithm computation time (on a Macintosh laptop having 8 GB RAM, 2.9 GHz clock, Intel Core i7 CPU). PSNR is Peak Signal-to-Noise Ratio. Table was sorted in descending order of SROCC for the "Overall" category. The numbers within parentheses in the "Overall" category show the confidence intervals on correlation values, computed by bootstrapping using 100 samples. Bold values indicate the best performing algorithm for that category.

	IQA	Interp.		Blur		Additive Noise		JPEG Blocking		Fast Fading		Overall (Confidence Interval)		Time (s)
		SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	
1	RRED	**0.691**	**0.694**	**0.813**	**0.815**	0.908	0.923	0.878	0.892	0.798	0.802	**0.658** (0.593,0.702)	**0.666** (0.611, 0.706)	5.380
2	RRSIS	0.381	0.471	0.772	0.805	0.888	0.900	**0.938**	**0.955**	0.838	**0.853**	0.624 (0.537, 0.676)	0.635 (0.584, 0.686)	3.290
3	RRDNT	0.478	0.508	0.643	0.657	0.918	**0.928**	0.703	0.745	0.657	0.677	0.394 (0.311, 0.488)	0.406 (0.335, 0.487)	15.100
4	RREdge	0.424	0.489	0.578	0.589	0.842	0.871	0.747	0.809	0.690	0.707	0.351 (0.261, 0.420)	0.359 (0.297, 0.421)	2.290
5	RRIQA	0.206	0.243	0.613	0.628	0.822	0.840	0.621	0.686	0.669	0.738	0.349 (0.264, 0.429)	0.348 (0.268, 0.416)	5.920
6	RRWeibull	0.401	0.302	0.789	0.793	**0.918**	0.919	0.860	0.869	**0.844**	0.842	0.299 (0.203, 0.385)	0.400 (0.337, 0.463)	7.100

Table 6
Root-mean-square error (RMSE), reduced $\tilde{\chi}^2$ statistic between the algorithm scores and the DMOS for various RR-IQA algorithms (after logistic function fitting) and outlier ratio (expressed in percentage) for each distortion category. The bold values indicate the best performing algorithm for that category.

	IQA	Interp.			Blur			Additive Noise			JPEG Blocking			Fast Fading			Overall		
		RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR
1	RRED	**6.490**	3.579	**0.0**	**5.486**	2.816	**0.0**	4.061	0.611	**0.0**	7.173	0.670	**0.0**	9.885	1.553	6.8	**10.264**	6.322	**6.1**
2	RRSIS	7.887	1.069	**0.0**	5.818	1.742	**0.0**	5.206	2.342	**0.0**	**4.798**	0.883	**0.0**	**8.645**	2.373	3.4	10.621	2.606	7.7
3	RRDNT	7.823	0.762	**0.0**	7.057	0.559	**0.0**	**3.854**	1.045	**0.0**	11.184	2.179	8.0	12.578	1.201	12.5	12.566	1.826	14.3
4	RREdge	7.876	0.860	**0.0**	7.593	2.045	**0.0**	5.026	0.464	**0.0**	9.873	1.885	5.0	12.713	2.972	9.0	12.952	2.503	17.0
5	RRIQA	8.772	0.428	2.3	7.288	1.655	**0.0**	5.768	1.361	**0.0**	12.226	2.473	14.8	11.951	2.017	11.4	12.894	1.390	16.1
6	RRWeibull	8.544	0.200	**0.0**	6.049	0.869	**0.0**	4.350	4.098	**0.0**	11.321	1.330	10.2	8.933	0.967	**2.3**	12.650	4.955	15.5

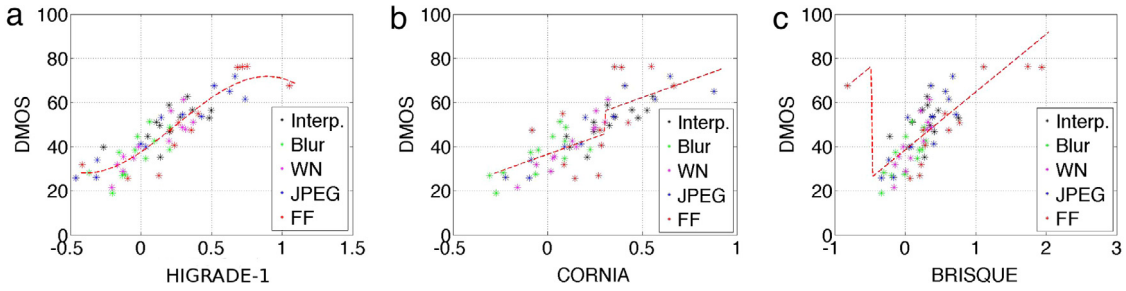

Fig. 9. Scatter plots of predicted IQA scores vs. DMOS for some selected no-reference IQA algorithms.

Table 7
Mean classification accuracy (in percentage) of selected learning based NR-IQA algorithms (described in Table 2) across 100 train-test (4:1) combinations on the ESPL Synthetic Image Database.

IQA	Alias	Blur	GN	JPEG	FF	All
GM-LOG	**99.8**	**96.2**	**100.0**	**96.8**	92.5	**97.1**
C-DIIVINE	91.7	95.3	**100.0**	95.5	**93.3**	95.2
BRISQUE	90.3	95.6	**100.0**	92.8	87.2	93.2
DESIQUE	90.7	87.3	**100.0**	89.1	85.3	90.5
BIQI	89.3	87.9	94.0	92.4	83.0	89.3
HIGRADE-1	78.5	83.4	**100.0**	90.2	87.7	88.0
BLIINDS-II	86.2	84.6	**100.0**	81.1	81.8	86.7
CurveletQA	87.0	87.0	**100.0**	81.2	69.5	84.9
DIIVINE	21.8	74.7	80.8	45.1	51.7	54.8

the top performing signal driven IQA algorithms. Significant progress has been made towards understanding low-level processing. However, on synthetic scenes, higher level cognitive factors, such as predicted gaze direction might be highly relevant to the perception of synthetic scenes. The authors of [86] have taken an important step in that direction.

Table 4 shows the RMSE, reduced $\tilde{\chi}^2$ statistic between scores predicted by the algorithms and the DMOS for various FR-IQA algorithms (after logistic function fitting) and outlier ratio. The top performing algorithm (GMSD) produced a zero outlier ratio, meaning that all of the predicted scores lie within two times the standard deviation of the DMOS scores. The "Additive Noise" distortion category was generally the one where the algorithms achieved the best performance in terms of the outlier ratio, while the images with "Fast fading" artifacts caused the highest percentages of outlying predictions by the IQA algorithms. The RMSE scores were computed after ensuring that both the DMOS and the algorithm scores lie in the same scale of [0,100]. Since the absolute RMSE values are more affected by the presence of outliers, a low value of RMSE may not always correspond to a low value of outlier ratio. The smallest RMSE values occurred in the "Additive Noise" category, while the "Fast Fading" artifact produced the maximum variation of RMSE. Overall, we found that the RMSEs lie within similar ranges as in other natural image IQA subjective datasets, such as [5].

5.2. Discussion of results for RR-IQA algorithms

RR-IQA algorithms generally achieve lower degrees of correlation against human subjective scores as compared to state-of-the-art FR-IQA algorithms, as shown in Table 5. Among the NSS based RR-IQA algorithms, RRED achieves the best overall performance (also the best RR-IQA algorithm). RRED also achieved the best performance on the interpolation distortion category, since it captures changes in wavelet coefficient statistics of images with interpolation artifact. Overall, the NSS based RR-IQA algorithms performed better than edge-map based RR-IQA algorithms, primarily due to the poor performance of the latter on the interpolation artifact category. However, RRSIS is an edge based algorithm that shows good performance for the JPEG compression artifact. Since this algorithm also uses only six features extracted from the edge detection procedure LoG (a widely used computation in many image processing algorithm), it is a promising tool for analyzing image compression algorithms while allowing for rapid and accurate evaluation of image quality. As per Table 6, the RMSE and outlier ratios of the best performing RR-IQA algorithms was worse than that of the best-performing FR-algorithms.

5.3. Discussion of results for NR-IQA algorithms

Next we discuss the performance of NR-IQA algorithms in predicting the type of distortion in the test image and also the quality score. Many NR-IQA algorithms operate in two steps: prediction of the type of distortion present in the test image, then using the features of the detected class to map the extracted image features to a quality score. We conducted an experiment where the features employed in different learning based NR-IQA algorithms were used to classify different types of distortions. We used a support vector machine classifier (SVC) in LibSVM [87]. The training set had 80% of the reference images (and their corresponding distorted versions), while the test set had the remaining 20% of the reference images (and their corresponding distorted versions). The process was repeated 100 times with random train-test splits to eliminate any bias due to varying spatial content. Table 7 highlights the results. Algorithms like GM-LOG, C-DIIVINE, BRISQUE and DESIQUE achieved good performances on distortion identification. Gaussian Noise was the easiest to detect among all the distortion categories for most of the learning based NR-IQA algorithms.

Table 8 compares the performances of 26 NR-IQA algorithms which comprise both learning based methods and artifact based methods in terms of SROCC and PLCC. For rows 1–9 (learning based methods), after the feature extraction step, a mapping is obtained from the feature space to the DMOS scores using a regression method, which provides a measure of the perceptual quality. We used a support vector machine regressor (SVR), specifically LibSVM [87] to implement ϵ-SVR with the radial basis function kernel, where γ was by default the inverse of the number of features. The training set had 80% of the reference images (and their corresponding distorted versions) and the test set had the remaining 20% of the reference images (and their corresponding distorted versions). The process was repeated 100 times to eliminate any bias due to varying spatial content.

Tables 7 and 8 show that C-DIIVINE, BRISQUE, GM-LOG, HIGRADE-1, and DESIQUE features perform well in classifying distortions and deducing the mapping between the feature space and DMOS scores. Similar conclusions were drawn while evaluating these algorithms on natural image databases. The reader may look at the references for these algorithms for more details on the results on natural image databases.

Fig. 12 shows box plots of the distribution of SROCC values for each of the 100 trials of random train-test splits for some NR-IQA algorithms. This enables us to study the robustness of performance of the algorithms with variations of the choice of the training set. CORNIA, C-DIIVINE, and BRISQUE show smaller variation in the degree of correlation with human perception.

Table 8
Median Spearman's Rank Ordered Correlation Coefficient (SROCC) and Pearson's Linear Correlation Coefficient (PLCC) between algorithm scores and DMOS for various NR-IQA algorithms (described in Table 2) along with algorithm computation time needed (on a Macintosh laptop having 8 GB RAM, 2.9 GHz clock, Intel Core i7 CPU) across 100 train-test (4:1) combinations on the ESPL Synthetic Image Database (50 trials for CORNIA in row 2). For the distortion specific NR-IQA algorithms, the non-italicized entries are NR-IQA algorithms meant for particular distortion categories. Italicized algorithms indicate the values obtained when the mentioned NR-IQA algorithms were applied for distortion categories other than what they were originally intended for. For these algorithms, we have copied the correlation values on the distortion class for which the algorithm was originally meant for, to the "Overall" column. The numbers within parentheses in the "Overall" category show the confidence intervals on correlation values, obtained by considering the maximum and minimum values of the correlations obtained over 100 trials for the learning based NR-IQA algorithms. Table was sorted in descending order of SROCC for the "Overall" category. Bold values indicate the best performing algorithm for that category.

	IQA	Interp.		Blur		GN		JPEG		FF		Overall (Confidence Interval)		Time (s)
		SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	
1	HIGRADE-1 (L)	0.605	0.646	0.612	0.640	0.858	0.904	0.901	0.927	**0.774**	**0.833**	**0.813** (0.562, 0.918)	**0.819**(0.626, 0.911)	2.134
2	CORNIA	**0.808**	**0.823**	**0.775**	**0.801**	0.793	0.821	0.898	0.918	0.706	0.763	0.810 (0.687, 0.875)	0.807(0.682, 0.880)	84.330
3	C-DIIVINE	0.702	0.760	0.730	0.769	0.847	0.896	0.841	0.879	0.738	0.802	0.798 (0.691, 0.916)	0.808(0.712, 0.912)	65.720
4	BRISQUE	0.631	0.643	0.720	0.782	0.840	0.902	0.898	0.935	0.717	0.740	0.789 (0.663, 0.897)	0.795(0.690, 0.895)	0.590
5	GM-LOG	0.680	0.711	0.653	0.694	0.853	0.906	0.912	0.944	0.701	0.746	0.787 (0.627, 0.893)	0.791(0.594, 0.892)	0.590
6	HIGRADE-1	0.580	0.647	0.474	0.508	0.871	0.920	0.922	0.942	0.726	0.758	0.774 (0.552, 0.893)	0.786(0.569, 0.887)	4.641
7	DESIQUE	0.595	0.678	0.590	0.617	**0.886**	**0.922**	**0.934**	**0.955**	0.714	0.737	0.773 (0.570, 0.909)	0.781(0.588, 0.901)	2.250
8	HIGRADE-2	0.510	0.584	0.565	0.576	0.857	0.906	0.865	0.879	0.728	0.762	0.743 (0.387, 0.888)	0.744(0.406, 0.877)	42.693
9	CurveletQA	0.658	0.695	0.695	0.753	0.880	0.916	0.854	0.880	0.553	0.595	0.731 (0.460, 0.872)	0.734(0.490, 0.863)	20.130
10	HIGRADE-2 (L)	0.509	0.563	0.488	0.529	0.859	0.906	0.874	0.909	0.668	0.729	0.689 (0.489, 0.876)	0.714(0.538, 0.881)	14.893
11	BIQI	0.665	0.733	0.732	0.764	0.837	0.903	0.735	0.769	0.538	0.593	0.676 (0.338, 0.849)	0.676(0.414, 0.858)	0.330
12	GRNN	0.537	0.592	0.371	0.409	0.811	0.896	0.738	0.790	0.408	0.551	0.602 (0.422, 0.777)	0.643(0.422, 0.802)	2.480
13	BLIINDS-II	0.388	0.444	0.499	0.556	0.794	0.839	0.680	0.754	0.548	0.608	0.596 (0.333, 0.834)	0.622(0.382, 0.835)	81.790
14	Anisotropy	0.364	0.354	0.357	0.400	0.835	0.871	0.385	0.449	0.392	0.439	0.470 (0.379, 0.513)	0.431(0.391, 0.483)	10.780
15	NIQE	0.428	0.496	0.425	0.528	0.740	0.511	0.732	0.834	0.606	0.623	0.377 (0.144, 0.600)	0.395(0.181, 0.601)	3.240
16	DIIVINE	0.421	0.523	0.441	0.490	0.484	0.537	0.444	0.489	0.439	0.513	0.372 (0.080, 0.700)	0.404(0.121, 0.705)	118.040
17	TMIQA	0.367	0.376	0.437	0.353	0.741	0.681	0.159	0.227	0.411	0.469	0.220 (0.097, 0.300)	0.311(0.223, 0.387)	0.120
18	*LPCM*	*0.415*	*0.444*	**0.836**	**0.847**	*0.623*	*0.621*	*0.211*	*0.231*	*0.108*	*0.237*	**0.836**(0.791, 0.890)	**0.847**(0.792, 0.885)	11.570
19	*CPBDM*	*0.676*	*0.720*	0.757	0.766	*0.746*	*0.815*	*0.765*	*0.749*	*0.347*	*0.405*	0.757 (0.678, 0.808)	0.766(0.669, 0.830)	3.500
20	*FISH*	*0.222*	*0.305*	0.705	0.716	*0.823*	*0.870*	*0.196*	*0.252*	*0.432*	*0.472*	0.705 (0.548, 0.787)	0.716(0.631, 0.793)	0.250
21	S_3	*0.409*	*0.449*	0.700	0.756	*0.747*	*0.786*	*0.151*	*0.189*	*0.402*	*0.450*	0.700 (0.554, 0.792)	0.756(0.692, 0.818)	308.150
22	*JNBM*	*0.598*	*0.635*	0.506	0.528	*0.756*	*0.816*	*0.536*	*0.512*	*0.448*	*0.455*	0.506 (0.327, 0.627)	0.528(0.336, 0.676)	7.520
23	*NLWT*	*0.324*	*0.334*	*0.024*	*0.141*	**0.872**	**0.888**	*0.000*	*0.187*	*0.559*	*0.589*	**0.872**(0.821, 0.905)	**0.888**(0.847, 0.928)	10.410
24	*FNVE*	*0.320*	*0.332*	*0.463*	*0.553*	0.863	0.887	*0.517*	*0.543*	*0.461*	*0.459*	0.863 (0.817, 0.894)	0.887(0.838, 0.915)	0.030
25	*JPEG-NR*	*0.540*	*0.570*	*0.593*	*0.650*	*0.748*	*0.865*	**0.928**	**0.954**	*0.464*	*0.607*	**0.928**(0.878, 0.952)	**0.954**(0.940, 0.969)	0.110
26	*NJQA*	*0.373*	*0.406*	*0.333*	*0.367*	*0.878*	*0.808*	0.743	0.819	*0.420*	*0.437*	0.743 (0.649, 0.854)	0.819(0.732, 0.869)	192.590

Table 9
Root-mean-square error (RMSE), reduced $\tilde{\chi}^2$ statistic between the algorithm scores and the DMOS for various NR-IQA algorithms (after logistic function fitting) and outlier ratio (expressed in percentage) for each distortion category. The bold values indicate the best performing algorithm for that category.

	IQA	Interp.			Blur			Additive Noise			JPEG Blocking			Fast Fading			Overall		
		RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR
1	HIGRADE-1 (L)	6.981	0.006	**0.000**	8.326	0.665	**0.000**	4.690	3.608	**0.000**	14.908	0.172	20.000	13.615	0.202	17.500	9.209	4.099	3.000
2	CORNIA	**0.112**	0.057	**0.000**	**0.131**	0.719	**0.000**	**0.136**	0.051	**0.000**	**0.151**	0.662	**0.000**	**0.262**	0.445	**0.000**	**0.190**	8.032	**0.000**
3	C-DIIVINE	5.897	0.084	**0.000**	8.290	0.306	**0.000**	5.067	0.131	**0.000**	14.858	0.429	20.000	14.269	0.019	20.000	9.631	4.586	4.000
4	BRISQUE	6.747	0.007	**0.000**	6.804	0.017	**0.000**	5.087	1.105	**0.000**	15.202	0.005	25.000	14.214	0.017	20.000	9.231	2.427	4.000
5	GM-LOG	6.182	0.000	**0.000**	7.858	0.166	**0.000**	4.856	2.231	**0.000**	14.953	0.006	20.000	14.846	0.009	20.000	9.579	1.390	5.000
6	HIGRADE-1	6.803	0.124	**0.000**	8.256	0.076	**0.000**	4.720	0.689	**0.000**	14.985	0.012	25.000	13.532	0.030	10.000	9.933	10.419	6.000
7	DESIQUE	6.799	0.107	**0.000**	7.993	0.025	**0.000**	4.527	3.408	**0.000**	15.207	0.010	25.000	14.205	0.462	20.000	9.799	1.119	5.000
8	HIGRADE-2	7.287	0.201	**0.000**	8.207	0.009	**0.000**	4.956	0.401	**0.000**	15.200	0.003	25.000	13.386	1.144	15.000	10.870	2.906	8.500
9	CurveletQA	6.535	0.215	**0.000**	7.136	0.069	**0.000**	4.735	0.466	**0.000**	15.152	0.004	25.000	15.279	0.434	25.000	11.272	6.938	9.000
10	HIGRADE-2 (L)	7.480	0.155	**0.000**	8.250	0.280	**0.000**	4.912	5.519	**0.000**	15.204	0.002	25.000	14.095	0.923	20.000	10.836	14.526	8.000
11	BIQI	6.177	0.520	**0.000**	8.216	0.970	**0.000**	4.915	0.002	**0.000**	14.838	0.143	20.000	14.514	0.893	25.000	10.741	3.509	9.000
12	GRNN	6.725	0.296	**0.000**	8.318	1.415	**0.000**	5.089	0.778	**0.000**	15.065	0.004	25.000	15.193	0.772	20.000	11.336	4.263	9.500
13	BLIINDS-II	7.546	0.884	**0.000**	7.884	0.686	**0.000**	5.826	0.000	**0.000**	15.312	0.002	25.000	14.689	0.009	20.000	11.060	6.710	9.000
14	Anisotropy	8.496	0.406	**0.000**	9.113	0.934	1.000	2.956	2.626	1.000	9.561	1.618	16.000	14.354	1.308	27.000	10.846	3.328	12.800
15	NIQE	7.683	0.030	**0.000**	8.095	0.234	**0.000**	8.582	0.346	**0.000**	10.994	0.002	5.000	12.394	1.493	10.000	12.490	2.538	14.000
16	DIIVINE	7.682	0.000	**0.000**	8.133	0.028	**0.000**	8.172	0.126	**0.000**	14.874	0.004	20.000	14.724	0.172	25.000	12.632	5.402	14.000
17	TMIQA	14.342	1.373	1.000	10.219	0.478	2.000	5.275	2.338	**0.000**	6.478	4.082	27.000	10.586	1.102	22.000	13.245	2.466	15.200
18	*LPCM*	–	–	–	**4.968**	1.019	**0.000**	–	–	–	–	–	–	–	–	–	**4.968**	1.019	**0.000**
19	*CPBDM*	–	–	–	6.485	0.440	**0.000**	–	–	–	–	–	–	–	–	–	6.485	0.440	**0.000**
20	*FISH*	–	–	–	6.603	0.324	**0.000**	–	–	–	–	–	–	–	–	–	6.603	0.324	**0.000**
21	S_3	–	–	–	6.339	0.162	**0.000**	–	–	–	–	–	–	–	–	–	6.339	0.162	**0.000**
22	*JNBM*	–	–	–	7.952	0.360	1.000	–	–	–	–	–	–	–	–	–	7.952	0.360	1.000
23	*NLWT*	–	–	–	–	–	–	4.611	3.620	**0.000**	–	–	–	–	–	–	**4.611**	3.620	**0.000**
24	*FNVE*	–	–	–	–	–	–	4.626	6.129	**0.000**	–	–	–	–	–	–	4.626	6.129	**0.000**
25	*JPEG-NR*	–	–	–	–	–	–	–	–	–	6.949	1.088	**0.000**	–	–	–	6.949	1.088	**0.000**
26	*NJQA*	–	–	–	–	–	–	–	–	–	9.279	1.453	8.000	–	–	–	9.279	1.453	8.000

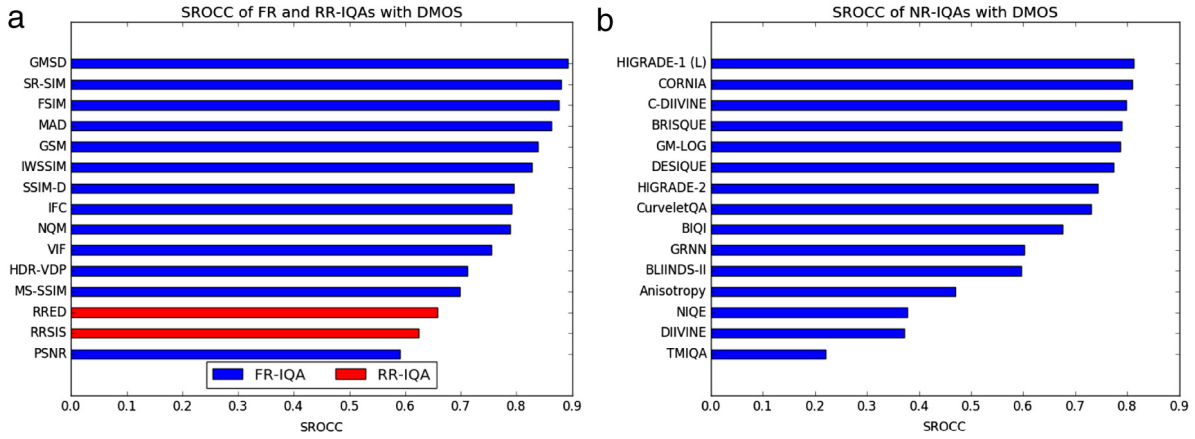

Fig. 10. Bar plots of SROCC of selected (a) FR and RR-IQA (b) NR-IQA algorithms with DMOS for images in the ESPL Synthetic Image Database.

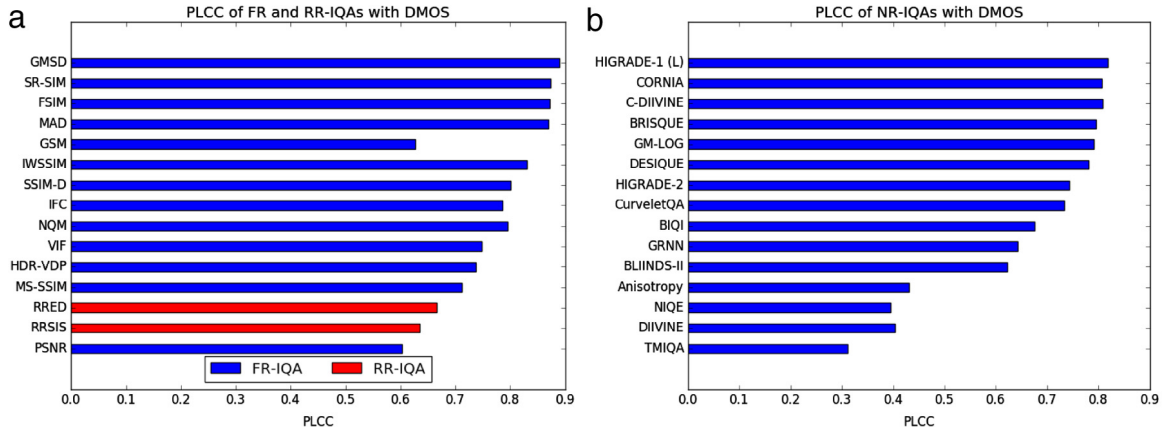

Fig. 11. Bar plots of PLCC of selected (a) FR and RR-IQA (b) NR-IQA algorithms with DMOS for images in the ESPL Synthetic Image Database.

Table 10
Results of the F-test performed on the residuals between model predictions and DMOS scores. Each cell in the table is a codeword consisting of 5 symbols that correspond to "Interpolation"', "Blur", "Gaussian Noise", "JPEG Blocking", "Fast Fading" and "Overall" distortions. "1"("0") indicates that the performance of the row IQA model is superior (inferior) to that of the column IQA model. - indicates that the statistical performance of the row IQA is indistinguishable from that of the column IQA. The matrix is symmetric.

	GMSD	FSIM	MS-SSIM	PSNR	RRED	HIGRADE-1	CORNIA	BRISQUE	DESIQUE	DIIVINE
GMSD	- - - - - -	- - - 1 - 1	- - - - - 1	- - - 1 - 1	0 0 - - - 1	1 1 - 1 - 1	- - 1 1 - 1	- - - 1 - -	- 1 - 1 - 1	1 1 1 1 - 1
FSIM	- - - 0 - 0	- - - - - -	- - - 0 - -	- - - - - 1	0 0 - - - 1	- - - 1 - -	- - - - - -	- - - 1 - -	- - - 1 - -	- 1 - 1 - 1
MS-SSIM	- - - - - 0	- - - 1 - -	- - - - - -	- 1 - 1 - 1	0 0 - - - -	1 1 - 1 - -	- - - 1 - -	- - - 1 - -	- 1 - 1 - -	1 1 1 1 - 1
PSNR	- - - 0 - 0	- - - - - 0	- 0 - 0 - 0	- - - - - -	0 0 - - - -	- - - 1 - -	- - - - - 0	- - - 1 - 0	- - - 1 - -	- - 1 1 - -
RRED	1 1 - - - 0	1 1 - - - 0	1 1 - - - -	1 1 - - - -	- - - - - -	1 1 - 1 - -	1 1 - - - -	1 1 - 1 - 0	1 1 - 1 - -	1 1 1 1 - -
HIGRADE-1	0 0 - 0 - 0	- - - 0 - -	0 0 - 0 - -	- - - 0 - -	0 0 - 0 - -	- - - - - -	0 - - 0 - -	- - - - - 0	- - - - - -	- - 1 - - 1
CORNIA	- - 0 0 - 0	- - - - - -	- - - 0 - -	- - - - - 1	0 0 - - - -	1 - - 1 - -	- - - - - -	- - - 1 - -	- - - 1 - -	1 - - 1 - 1
BRISQUE	- - - 0 - -	- - - 0 - -	- - - 0 - -	- - - 0 - 1	0 0 - 0 - 1	- - - - - 1	- - - 0 - -	- - - - - -	- - - - - 1	- - - - - 1
DESIQUE	- 0 - 0 - 0	- - - 0 - -	- 0 - 0 - -	- - - 0 - -	0 0 - 0 - -	- - - - - -	- - - 0 - -	- - - - - 0	- - - - - -	- - - - - 1
DIIVINE	0 0 0 0 - 0	- 0 - 0 - 0	0 0 0 0 - 0	- - 0 0 - -	0 0 0 0 - -	- - 0 - - 0	0 - - 0 - 0	- - - - - 0	- - - - - 0	- - - - - -

Compared to learning based models, NIQE and TMIQA use unsupervised learning models and are not trained on any corpus of distorted images or human scores. Instead, they are created by measuring the NSS of pristine natural pictures. As such, these models perform worse on synthetic images in spite of showing competitive performance on natural images. This might occur due to the higher amount of variability in the distribution of the MSCN coefficients of synthetic images as compared to natural scenes [20]. The performance of artifact based NR-IQA algorithms is shown in rows 18–22 (blur), 23–24 (noise) and 25–26 (JPEG blocking). To the best of our knowledge, we did not find any artifact based NR-IQA algorithm meant only for images having interpolation or fast-fading artifacts. For Gaussian Noise and JPEG blocking, the learning based NR-IQA algorithms perform better than artifact based NR-IQA algorithms. For the blur artifact, the LPCM algorithm outperformed the learning based NR-IQA algorithms. LPCM is based on the concept of local phase coherence, following the idea that the phases of complex wavelet coefficients follow a highly predictable pattern in scale space in the vicinity of sharp features. It is a powerful tool when assessing blur severity, whereas learning based NR-IQA algorithms are based on statistics that generalize better over a wider array of distortions. On the present set of synthetic images, the local phase coherence based features outperformed the generic learning based NR-IQA algorithms on the blur distortion.

Table 9 shows that the "Interpolation", "Blur", and "Additive Noise" distortion categories resulted in lower values of RMSE and outlier ratio as compared to the "JPEG blocking" and "Fast Fading" distortions. This in turn has leads to higher RMSE and outlier ratio values overall for the database. In general, as compared to learning based NR-IQA algorithms,

Table 11
Variance (Var.) and Gaussianity (Gauss.) of the residuals between individual scores and NR-IQA algorithm predictions. ✓ indicates that the residuals follow a Gaussian distribution.

IQA	Interp.		Blur		GN		JPEG		FF		All	
	Var.	Gauss.	Var.	Gauss.	Var.	Gauss.	Var.	Gauss.	Var.	Gauss.	Var.	Gauss.
GMSD	106.40	✓	122.92	✓	111.90	✓	103.94	✓	172.92	✓	133.47	✓
FSIM	130.65	✓	128.74	✓	125.80	✓	143.81	✓	194.77	✓	156.45	✓
MS-SSIM	106.82	✓	119.33	✗	120.28	✓	101.66	✓	189.93	✓	171.53	✓
PSNR	125.14	✓	161.65	✓	119.93	✓	135.76	✓	189.99	✓	216.02	✓
RRED	95.70	✓	107.18	✓	120.27	✓	116.31	✓	169.66	✓	193.03	✓
HIGRADE-1	162.68	✓	178.64	✓	115.16	✓	337.30	✓	218.88	✓	187.16	✓
CORNIA	116.20	✓	144.61	✓	140.40	✓	152.78	✓	223.88	✓	172.02	✓
BRISQUE	129.21	✓	147.86	✓	133.83	✓	330.32	✓	186.70	✓	150.14	✓
DESIQUE	122.27	✓	167.63	✓	126.76	✓	327.28	✓	271.96	✓	187.14	✓
DIIVINE	163.11	✓	189.27	✓	167.15	✓	338.06	✓	276.63	✓	243.71	✓
Null Model	93.13	✓	101.71	✓	99.80	✓	88.17	✓	98.83	✓	96.24	✓
Number of samples	832		832		832		832		832		4160	
Threshold F-ratio	1.11		1.09		1.10		1.10		1.12		1.04	

Table 12
Variance (Var.) and Gaussianity (Gauss.) of the residuals between DMOS values and NR-IQA algorithm predictions. ✓ indicates that the residuals follow a Gaussian distribution.

IQA	Interp.		Blur		GN		JPEG		FF		All	
	Var.	Gauss.	Var.	Gauss.	Var.	Gauss.	Var.	Gauss.	Var.	Gauss.	Var.	Gauss.
GMSD	14.14	✗	22.59	✓	12.89	✓	16.80	✓	78.93	✓	37.70	✓
FSIM	39.98	✓	28.80	✓	27.70	✓	59.28	✗	102.21	✓	60.96	✓
MS-SSIM	14.60	✓	18.77	✗	21.82	✓	14.37	✗	97.06	✓	76.23	✓
PSNR	34.11	✓	63.86	✓	21.45	✓	50.70	✓	97.12	✓	121.27	✓
RRED	2.74	✓	5.82	✓	21.81	✓	29.98	✓	75.46	✗	97.99	✓
HIGRADE-1	74.10	✗	81.96	✓	16.37	✓	265.42	✗	127.90	✗	92.05	✓
CORNIA	24.58	✗	45.70	✓	43.25	✓	68.83	✓	133.23	✓	76.73	✗
BRISQUE	38.44	✗	49.16	✓	36.26	✓	257.99	✗	93.62	✓	54.58	✓
DESIQUE	31.05	✓	70.23	✓	28.73	✓	254.75	✗	184.44	✓	92.03	✗
DIIVINE	74.56	✓	93.28	✓	71.75	✗	266.23	✗	189.42	✓	149.31	✓
Null Model	2.40	✓	2.40	✓	2.40	✗	2.40	✗	2.40	✓	1.45	✓
Number of samples	16		16		16		16		16		80	
Threshold F-ratio	2.40		2.40		2.40		2.40		2.40		1.45	

distortion-specific NR-IQA algorithms delivered lower RMSE and outlier ratio values.

In order to help visualize the results, bar plots of the SROCC and PLCC values of selected IQA algorithms against DMOS are shown in Fig. 10 and 11 respectively. Overall, the FR and NR-IQA algorithms performed better than the RR-IQA algorithms. Although the FR-IQA algorithms performed better overall than the NR-IQA algorithms, the latter also produced promising results. For the "Interpolation" distortion category, the NR-IQA algorithm CORNIA yielded better correlation against human opinions than did the best performing FR-IQA algorithm (MAD).

It is interesting to observe that the best NR-IQA algorithms were able to obtain higher levels of correlation against DMOS than some of the FR-IQA algorithms. While this was not generally the case, the NR-IQA algorithms benefit by training on the corpus of synthetic images (both pristine and those inflicted with distortions) and the human subjective judgments of them, hence were able to learn the statistical properties of the different distortion categories. The tested FR-IQA algorithms are not learning based, instead relying on measurements of changes in the local image structure.

5.4. Determination of statistical significance

Results of statistical significance tests are summarized in Tables 10–12. For this purpose, ten representative IQA algorithms were selected. These were chosen as the best-performing models on the ESPL database or because they were very widely used, like SSIM. For the learning based methods, the statistical significance tests were carried out over multiple training-test splits, using 60 test images each time, and similar results were obtained. Tables summarize the results obtained for one such representative trial. For the F-test based on quality scores provided by individual human observers, the variance of the residuals obtained from the null-model and by the ten selected IQA algorithms, along with the number of samples considered in each category and the threshold F-ratio at the 95% significance are shown in Table 11. In addition, we also show whether the assumptions of Gaussianity of the residuals hold. None of the 10 IQA algorithms tested was found to be statistically indistinguishable from the null-model corresponding to human judgment in any of the distortion categories. Similar conclusions were reached in [29]. GMSD achieved the lowest variance of the residuals for the overall database among the ten IQA algorithms.

For the F-test based on the DMOS scores, Table 12 summarizes the variances of the residuals obtained from the ten selected IQA algorithms, along with the number of samples considered in each category, and the threshold F-ratio at the 95% significance level. For some of the cases, it was found that the assumption of Gaussianity of the residuals did not hold.

To determine whether the IQA algorithms were significantly different from each other, the F-statistic, as in [5,29], was used to determine the statistical significance between the variances of the residuals after a non-linear logistic mapping between the two IQA algorithms, at the 95% confidence level. Table 10 shows the results for the ten selected IQA algorithms and all distortions. Overall, the FR-IQA algorithms were found to be statistically superior to the NR-IQA algorithms. This conclusion is most pronounced for the "JPEG blocking" and "Interpolation" artifacts.

5.5. Computational complexity

All of the IQA algorithms were profiled using the original source codes provided by the respective authors. FR-IQA metrics like SR-SIM and GMSD achieved a high degree of correlation with human perception yet are computationally inexpensive. As expected, the learning based NR-IQA algorithms (like CORNIA, BRISQUE, DESIQUE, C-DIIVINE, BLIINDS-II) achieved performance comparable to the best performing FR-IQA algorithms, but they are computationally more intensive since

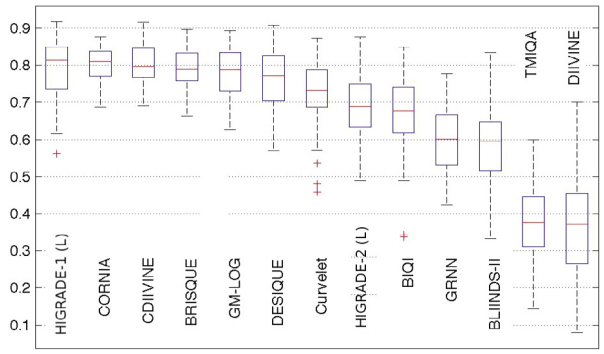

Fig. 12. Box plot of SROCC of learning based NR-IQA algorithms on images in the ESPL Synthetic Image Database for 4:1 train-test splits over 100 trials. For each box, the median is the line dividing the central box, the edges of the box represent the 25th and 75th percentiles, the whiskers span the most extreme non-outlier data points, and the outliers are plotted individually.

many image features must be computed. RRED shows intermediate performance between FR-IQA and NR-IQA algorithms both in terms of correlation with human judgment and time complexity.

6. Conclusion

We presented the new publicly available ESPL Synthetic Database comprising pristine synthetic source images and images containing five different types of distortions, annotated by 26,000 quality scores from 52 subjects. We evaluated the performance of more than 50 state-of-the-art IQA algorithms.

For FR-IQA algorithms, we observed the importance of saliency based spatial pooling and different strategies for evaluating the quality of the image based on whether the artifacts are subthreshold or suprathreshold. GMSD offered the best trade-off between performance and run-time complexity. RR-IQA algorithms perform worse than FR-IQA and NR-IQA algorithms. RRED was the best performing RR-IQA algorithm. For NR-IQA, deviations in statistical regularity caused by distortions can be used to successfully evaluate the quality of synthetic images. Scene statistics based algorithms take more time to run. Algorithms such as GMSD, SR-SIM, GM-LOG, and HIGRADE-1 (L) show high correlation with human perception and reasonable runtime.

We also found that for synthetic images, interpolation distortion is the most challenging category to conduct IQA on, but scene statistics based NR-IQA algorithms perform well on this artifact. Future databases may involve subjective quality evaluation of images interpolated using more sophisticated schemes such as bilinear or bicubic interpolation.

A future avenue of work would be to introduce localized blur, such as motion blur or depth-of-field blur. Also, we would like to study more specialized distortions, such as Perlin noise used in texture synthesis. Lastly, we would like to conduct a study on computer graphics generated video content with associated spatio-temporal distortions.

Acknowledgments

The authors would like to thank the Office of the Vice President for Research, The University of Texas at Austin, for providing the funds used to conduct the subjective experiment. They would also like to thank Prof. Donald Fussell, Department of Computer Science, The University of Texas at Austin, for helpful discussions regarding the type of artifacts arising in computer graphics generated imagery.

References

[1] D. Kundu, B.L. Evans, No-reference synthetic image quality assessment using scene statistics, in: Proc. Asilomar Conf. Signals, Systems and Computers, 2015. http://users.ece.utexas.edu/%7Ebevans/papers/2015/imagequalitynoref/index.html.

[2] D. Kundu, B.L. Evans, Proc. International Conference on Image Processing, 2015. http://users.ece.utexas.edu/%7Ebevans/papers/2015/imagequality/index.html.

[3] Cisco Systems, Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 20142019, White Paper, 2015. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.pdf.

[4] L.K. Choi, Y. Liao, A.C. Bovik, Video QoE metrics for the compute continuum, IEEE Commun. Soc. Multimed. Tech. Comm (MMTC) E-Lett. 8 (5) (2013) 26–29.

[5] H.R. Sheikh, M.F. Sabir, A.C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Trans. Image Process. 15 (11) (2006) 3440–3451.

[6] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C.-C.J. Kuo, Image database TID2013: Peculiarities, results and perspectives, Signal Process., Image Commun. 1 (2015) 57–77.

[7] D. Ghadiyaram, A.C. Bovik, Massive online crowdsourced study of subjective and objective picture quality, IEEE Trans. Image Process. 25 (1) (2016) 372–387 [Online]. Available: http://dx.doi.org/10.1109/TIP.2015.2500021.

[8] E.C. Larson, D.M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, J. Electr. Imaging 19 (1) (2010) 011006.

[9] F. De Simone, L. Goldmann, V. Baroncini, T. Ebrahimi, Subjective evaluation of jpeg XR image compression, in: Proc. SPIE Applications of Digital Image Processing XXXII, 7443, 2009.

[10] M. Čadík, R. Herzog, R. Mantiuk, K. Myszkowski, H.-P. Seidel, New measurements reveal weaknesses of image quality metrics in evaluating graphics artifacts, ACM Trans. Graphics 31 (6) (2012) 1–10.

[11] J.A. Ferwerda, P. Shirley, S.N. Pattanaik, D.P. Greenberg, A model of visual masking for computer graphics, in: SIGGRAPH, ACM, 1997, pp. 143–152.

[12] G. Ramanarayanan, J. Ferwerda, B. Walter, K. Bala, Visual equivalence: towards a new standard for image fidelity, in: SIGGRAPH, ACM, 2007.

[13] L. Zhang, L. Zhang, X. Mou, D. Zhang, A comprehensive evaluation of full reference image quality assessment algorithms, in: Proc. IEEE International Conference on Image Processing, 2012, pp. 1477–1480.

[14] P. Mohammadi, A. Ebrahimi-Moghadam, S. Shirani, Subjective and objective quality assessment of image: A survey. arXiv preprint arXiv:1406.7799 (2014).

[15] R. Piórkowski, R. Mantiuk, Using full reference image quality metrics to detect game engine artefacts, in: Proceedings of the ACM SIGGRAPH Symposium on Applied Perception, SAP '15, ACM, New York, NY, USA, 2015, pp. 83–90 [Online]. Available: http://doi.acm.org/10.1145/2804408.2804414.

[16] W.S. Geisler, Visual perception and the statistical properties of natural scenes, Ann. Rev. Psychol. 59 (1) (2008) 167–192.

[17] E.P. Simoncelli, B.A. Olshausen, Natural image statistics and neural representation, Ann. Rev. Neurosci. 24 (2001) 1193–1216.

[18] R. Herzog, M. Čadík, T.O. Aydin, K.I. Kim, K. Myszkowski, H.-P. Seidel, NoRM: No-reference image quality metric for realistic image synthesis, Comput. Graph. Forum 31 (2) (2012) 545–554.

[19] S. Lyu, H. Farid, How realistic is photorealistic?, IEEE Signal Process. Lett. 53 (2) (2005) 845–850.

[20] D. Kundu, B.L. Evans, Spatial domain synthetic scene statistics, in: Proc. Asilomar Conf. Signals, Systems and Computers, 2014.

[21] A. Mittal, R. Soundararajan, A.C. Bovik, Making a 'completely blind' image quality analyzer, IEEE Signal Process. Lett. 20 (3) (2013) 209–212.

[22] S. Winkler, Analysis of public image and video databases for quality assessment, IEEE J. Sel. Topics Signal Process. 6 (6) (2012) 616–625.

[23] K. Schwenk, A. Kuijper, J. Behr, D.W. Fellner, Practical noise reduction for progressive stochastic ray tracing with perceptual control, IEEE Comput. Graph. Appl. 32 (6) (2012) 46–55.

[24] P. Sen, S. Darabi, On filtering the noise from the random parameters in Monte Carlo rendering, ACM Trans. Graph. 31 (3) (2012) 18:1–18:15 [Online]. Available: http://doi.acm.org/10.1145/2167076.2167083.

[25] A. Unterweger, Compression artifacts in modern video coding and state-of-the-art means of compensation, Multimedia Netw. Coding (2012) 28.

[26] OpenJPEG 2000, http://www.openjpeg.org/.

[27] ITU-R BT.500-13 methodology for the subjective assessment of the quality of television pictures [Online]. Available: https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-13-201201-I!!PDF-E.pdf.

[28] M. Kleiner, D. Brainard, D. Pelli, C. Broussard, T. Wolf, D. Niehorster, The psychology toolbox, http://psychtoolbox.org/.

[29] K. Seshadrinathan, R. Soundararajan, A.C. Bovik, L.K. Cormack, Study of subjective and objective quality assessment of video, IEEE Trans. Image Process. 19 (6) (2010) 1427–1441.

[30] D. Martin, C. Fowlkes, D. Tal, J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: Proc. Int'l Conf. Comp. Vision, vol. 2, July 2001, pp. 416–423.

[31] M. Do, M. Vetterli, Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance, IEEE Trans. Image Process. 11 (2) (2002) 146–158.

[32] A. Mittal, A.K. Moorthy, A.C. Bovik, No-Reference image quality assessment in the spatial domain, IEEE Trans. Image Process. 21 (12) (2012) 4695–4708.

[33] A. Shnayderman, A. Gusev, A.M. Eskicioglu, A multidimensional image quality measure using singular value decomposition, Proc. SPIE 5294 (2004) 82–92.

[34] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.

[35] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in: Proc. Asilomar Conf. Signals, Systems and Computers, vol. 2, Nov 2003, pp. 1398–1402.

[36] A. Kolaman, O. Yadid-Pecht, Quaternion structural similarity: a new quality index for color images, IEEE Trans. Image Process. 21 (4) (2012) 1526–1536.

[37] Z. Wang, A.C. Bovik, A universal image quality index, IEEE Signal Process. Lett. 9 (3) (2002) 81–84.

[38] S. Daly, The visible differences predictor: An algorithm for the assessment of image fidelity, in: A.B. Watson (Ed.), Digital Images and Human Vision, MIT Press, 1993, pp. 179–206.

[39] R. Mantiuk, K.J. Kim, A.G. Rempel, W. Heidrich, HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions, in: SIGGRAPH, ACM, 2011, pp. 1–14.

[40] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, A.C. Bovik, Image quality assessment based on a degradation model, IEEE Trans. Image Process. 9 (2000) 636–650.

[41] T. Mitsa, K.L. Varkur, Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms, in: Proc. International Conference on Acoustics Speech and Signal Processing, vol. 5, April 1993, pp. 301–304.

[42] K. Egiazarian, J. Astola, V. Lukin, F. Battisti, M. Carli, New full-reference quality metrics based on hvs, in: Proc. Int. Work. Video Process. and Quality Metrics, 2006.

[43] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, M. Carli, Modified image visual quality metrics for contrast change and mean shift accounting, in: Proc. Int. Conf. Experience of Designing and Appl. of CAD Systems in Microelectronics, 2011, pp. 305–311.

[44] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, V. Lukin, On between-coefficient contrast masking of DCT basis functions, in: Proc. Int. Work. Video Process. and Quality Metrics, 2007.

[45] H.R. Sheikh, A.C. Bovik, G. de Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, IEEE Trans. Image Process. 14 (12) (2005) 2117–2128.

[46] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444.

[47] Z. Wang, Q. Li, Information content weighting for perceptual image quality assessment, IEEE Trans. Image Process. 20 (5) (2011) 1185–1198.

[48] L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: A feature similarity index for image quality assessment, IEEE Trans. Image Process. 20 (8) (2011) 2378–2386.

[49] W. Xue, L. Zhang, X. Mou, A.C. Bovik, Gradient magnitude similarity deviation: A highly efficient perceptual image quality index, IEEE Trans. Image Process. 23 (2) (2014) 684–695.

[50] A. Liu, W. Lin, M. Narwaria, Image quality assessment based on gradient similarity, IEEE Trans. Image Process. 21 (4) (2012) 1500–1512.

[51] L. Zhang, D. Zhang, X. Mou, RFSIM: A feature based image quality assessment metric using Riesz transforms, in: Proc. International Conference on Image Processing, 2010, pp. 321–324.

[52] L. Zhang, Y. Shen, H. Li, VSI: A visual saliency-induced index for perceptual image quality assessment, IEEE Trans. Image Process. 23 (10) (2014) 4270–4281.

[53] L. Zhang, H. Li, SR-SIM: A fast and high performance IQA index based on spectral residual, in: Proc. International Conference on Image Processing, 2012, pp. 1473–1476.

[54] D.M. Chandler, S.S. Hemami, VSNR: A wavelet-based visual signal-to-noise ratio for natural images, IEEE Trans. Image Process. 16 (9) (2007) 2284–2298.

[55] Z. Wang, E.P. Simoncelli, Reduced-reference image quality assessment using a wavelet-domain natural image statistic model, Proc. SPIE 5666 (2005) 149–159.

[56] Q. Li, Z. Wang, Reduced-reference image quality assessment using divisive normalization-based image representation, IEEE J. Sel. Topics Signal Process. 3 (2) (2009) 202–211.

[57] R. Soundararajan, A. Bovik, RRED indices: Reduced reference entropic differencing for image quality assessment, IEEE Trans. Image Process. 21 (2) (2012) 517–526.

[58] W. Xue, X. Mou, Reduced reference image quality assessment based on Weibull statistics, in: Proc. IEEE International Conference on Quality of Multimedia Experience, 2010, pp. 1–6.

[59] X. Mou, W. Xue, L. Zhang, Reduced reference image quality assessment via sub-image similarity based redundancy measurement, Proc. SPIE 8291 (2012) 82911S–82911S–7.

[60] M. Zhang, W. Xue, X. Mou, Reduced reference image quality assessment based on statistics of edge, Proc. SPIE 7876 (2011) 787611–787611–7.

[61] R. Hassen, Z. Wang, M. Salam, Image sharpness assessment based on local phase coherence, IEEE Trans. Image Process. 22 (7) (2013) 2798–2810.

[62] N.D. Narvekar, L.J. Karam, A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection, in: Proc. IEEE International Conference on Quality of Multimedia Experience, 2009, pp. 87–91.

[63] R. Ferzli, L.J. Karam, A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB), IEEE Trans. Image Process. 18 (4) (2009) 717–728.

[64] C. Vu, T. Phan, D.M. Chandler, S3: A spectral and spatial measure of local perceived sharpness in natural images, IEEE Trans. Image Process. 21 (3) (2011).

[65] P. Vu, D. Chandler, A fast wavelet-based algorithm for global and local image sharpness estimation, IEEE Signal Process. Lett. 19 (7) (2012) 423–426.

[66] Z. Wang, H.R. Sheikh, A.C. Bovik, No-reference perceptual quality assessment of JPEG compressed images, in: Proc. IEEE International Conference on Image Processing, vol. 1, 2002, pp. I–477–I–480.

[67] S. Golestaneh, D. Chandler, No-reference quality assessment of jpeg images via a quality relevance map, IEEE Signal Process. Lett. 21 (2) (2014) 155–158.

[68] X. Liu, M. Tanaka, M. Okutomi, Single-Image noise level estimation for blind denoising, IEEE Trans. Image Process. 22 (12) (2013) 5226–5237.

[69] J. Immerkr, Fast noise variance estimation, Comput. Vis. Image Underst. 64 (2) (1996) 300–302.

[70] Y. Zhang, D.M. Chandler, No-reference image quality assessment based on log-derivative statistics of natural scenes, J. Electronic Imaging 22 (4) (2013).

[71] D. Kundu, D. Ghadiyaram, A.C. Bovik, B.L. Evans, No-reference quality assessment of tone-mapped HDR pictures, IEEE Trans. Image Process. 26 (6) (2017) 2957–2971.

[72] W. Xue, X. Mou, L. Zhang, A.C. Bovik, X. Feng, Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features, IEEE Trans. Image Process. 23 (11) (2014) 4850–4862.

[73] A.K. Moorthy, A.C. Bovik, Blind image quality assessment: From natural scene statistics to perceptual quality, IEEE Trans. Image Process. 20 (12) (2011) 3350–3364.

[74] Y. Zhang, A.K. Moorthy, D.M. Chandler, A.C. Bovik, C-DIIVINE: No-reference image quality assessment based on local magnitude and phase statistics of natural scenes, Signal Process., Image Commun. 29 (7) (2014) 725–747.

[75] A.K. Moorthy, A.C. Bovik, A two-step framework for constructing blind image quality indices, IEEE Signal Process. Lett. 7 (5) (2010).

[76] M.A. Saad, A.C. Bovik, C. Charrier, Blind image quality assessment: A natural scene statistics approach in the DCT domain, IEEE Trans. Image Process. 21 (8) (2012) 3339–3352.

[77] C. Li, A.C. Bovik, X. Wu, Blind image quality assessment using a general regression neural network, IEEE Trans. Neural Netw. 22 (5) (2011) 793–799.

[78] L. Liu, H. Dong, H. Huang, A.C. Bovik, No-reference image quality assessment in curvelet domain, Signal Process., Image Commun. 29 (4) (2014).

[79] S. Gabarda, G. Cristbal, Blind image quality assessment through anisotropy, J. Opt. Soc. Amer. 24 (12) (2007).

[80] Peng Ye, Jayant Kumar, Le Kang, David Doermann, Unsupervised feature learning framework for no-reference image quality assessment, in: Proc. CVPR, 2012, pp. 1098–1105.

[81] A. Mittal, G.S. Muralidhar, J. Ghosh, A.C. Bovik, Blind image quality assessment without human training using latent quality factors, IEEE Signal Process. Lett. 19 (2) (2012) 75–78.

[82] Final report from the video quality experts group on the validation of objective models of video quality assessment, 2003. ftp://vqeg.its.bldrdoc.gov/Documents/Meetings/Hillsboro_VQEG_Mar_03/VQEGIIDraftReportv2a.pdf.

[83] H.R.S. Zhou Wang, AlanC. Bovik, E.P. Simoncelli, The ssim Index for Image Quality Assessment, Feb 2003. https://ece.uwaterloo.ca/%7Ez70wang/research/ssim/.

[84] M. Gaubatz, Metrix mux visual quality assessment package. http://foulard.ece.cornell.edu/gaubatz/metrix_mux/.

[85] K. Seshadrinathan, A.C. Bovik, Unifying analysis of full reference image quality assessment, in: 2008 15th IEEE International Conference on Image Processing, Oct 2008, pp. 1200–1203.

[86] M. Čadík, R. Herzog, R. Mantiuk, R. Mantiuk, K. Myszkowski, H.-P. Seidel, Learning to predict localized distortions in rendered images, in: Computer Graphics Forum, 32, Wiley Online Library, 2013, pp. 401–410.

[87] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Trans. on Intelligent Systems and Technology 2 (2011) Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.