Large-Scale Crowdsourced Study for Tone-Mapped HDR Pictures

Debarati Kundu, Deepti Ghadiyaram, *Student Member, IEEE*,
Alan C. Bovik, *Fellow, IEEE*, and Brian L. Evans, *Fellow, IEEE*

***Abstract*— Measuring digital picture quality, as perceived by human observers, is increasingly important in many applications in which humans are the ultimate consumers of visual information. Standard dynamic range (SDR) images provide 8 b/color/pixel. High dynamic range (HDR) images, usually created from multiple exposures of the same scene, can provide 16 or 32 b/color/pixel, but need to be tonemapped to SDR for display on standard monitors. Multiexposure fusion (MEF) techniques bypass HDR creation by fusing an exposure stack directly to SDR images to achieve aesthetically pleasing luminance and color distributions. Many HDR and MEF databases have a relatively small number of images and human opinion scores, obtained under stringently controlled conditions, thereby limiting realistic viewing. Moreover, many of these databases are intended to compare tone-mapping algorithms, rather than being specialized for developing and comparing image quality assessment models. To overcome these challenges, we conducted a massively crowdsourced online subjective study. The primary contributions described in this paper are: 1) the new ESPL-LIVE HDR Image Database that we created containing diverse images obtained by tone-mapping operators and MEF algorithms, with and without post-processing; 2) a large-scale subjective study that we conducted using a crowdsourced platform to gather more than 300 000 opinion scores on 1811 images from over 5000 unique observers; and 3) a detailed study of the correlation performance of the state-of-the-art no-reference image quality assessment algorithms against human opinion scores of these images. The database is available at http://signal.ece.utexas.edu/%7Edebarati/HDRDatabase.zip.**

***Index Terms*— Image quality assessment, high dynamic range, subjective study, crowdsourcing.**

I. Introduction

THERE has been significant growth in the acquisition, processing and transmission of pictures and videos in recent years. While most pictures are still Standard Dynamic Range (SDR) images represented by 8 bits/color/pixel obtained by taking photographs at a fixed exposure, there is a growing interest in the acquisition/creation and display of high dynamic range images and other types of pictures created by multiple exposure fusion. These images allow for more pleasing representation and better use of the available luminance and color ranges in real scenes, ranging from direct sunlight to faint starlight [1]. Several video-on-demand services can stream and home HDR monitors can display HDR content, while smart phones and digital SLR cameras can create aesthetically pleasing images by fusing a multiply-exposed stack of images.

HDR images, commonly represented by 16 or 32 bits/color/pixel, typically are obtained by blending a stack of SDR images at varying exposure levels, thereby allowing a range of intensity levels on the order of 10,000 to 1. HDR rendering also finds use in computer graphics, where lighting calculations are performed over a wider dynamic range. This results in better contrast variation thereby leading to a higher degree of detail preservation. However, in order to visualize these images on standard display devices designed for SDR images, they must be tonemapped to SDR [2]. In addition to tone-mapped SDR images, images are also created by multi-exposure fusion, where a stack of SDR images taken at varying exposure levels are fused to create an SDR image that is more visually informative than the input images. This bypasses the intermediate step of creating an HDR irradiance map. HDR images may also be post-processed (color saturation, color temperature, detail enhancement, etc.) for aesthetic purposes.

Subjective quality evaluation of images produced by TMO or MEF algorithms is of considerable interest given the ongoing rollout of HDR products and standards. A subjective study using human observers is the most reliable way, although this process is time consuming and expensive. However it provides the necessary ground-truth data to benchmark objective image quality assessment (IQA) algorithms that automate the process of visual quality assessment. Some of the earliest psychophysical experiements on HDR images were carried out in [3]–[10]. Many existing HDR IQA databases suffer from the limitations of a relatively small number of images and human subjective scores. The subjective scores have typically been obtained by experiments conducted under stringently controlled conditions. In addition, most of these studies either asked the subjects to rank multiple versions of the same HDR scene created using different processing algorithms, or used a two-alternative forced choice method of subjective evaluation. These approaches severely restrict the number of source images that can be considered, the type of processing

Manuscript received May 25, 2016; revised December 19, 2016, April 19, 2017, and May 21, 2017; accepted May 29, 2017. Date of publication June 8, 2017; date of current version July 18, 2017. This work was supported in part by The University of Texas at Austin, Special Research Grant, Vice President for Research and in part by the National Science Foundation under Grant 1116656. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rahul Vanam. *(Corresponding author: Debarati Kundu.)*

D. Kundu was with Embedded Signal Processing Laboratory, The University of Texas at Austin, Austin, TX 78701 USA. She is now with Qualcomm Research, Bengluru 560066, India (e-mail: debarati@utexas.edu).

D. Ghadiyaram and A. C. Bovik are with the Laboratory for Image and Video Engineering, The University of Texas at Austin, Austin, TX 78701 USA (e-mail: deepti@cs.utexas.edu; bovik@ece.utexas.edu).

B. L. Evans is with the Embedded Signal Processing Laboratory, The University of Texas at Austin, Austin, TX 78701 USA (e-mail: bevans@ece.utexas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2713945

algorithms examined and the number of subjects participating in the experiments. Moreover, many of these studies have been directed towards comparing the results of optimized tone-mapping algorithms, rather than for creating and comparing IQA models to access tone-mapped HDR images.

Krasula *et al.* [11] showed that scores assigned by human subjects to different tone-mapped images may differ, based on whether the source HDR images were also shown to them. Ashikhmin and Goyal found significant differences between rank-based evaluations of different TMOs, depending on whether the subjects were also shown the corresponding real physical scenes. In real-world scenarios, viewers of tone-mapped HDR content do not view or have access to a reference image. In such situations, no-reference (blind) IQA models are required. However, to be able to design blind IQA models, it is necessary to have data from large-scale subjective studies, ideally without showing the subjects reference HDR images. Here, our aim is to design a subjective study that mimics realistic viewing conditions, whereby human subjects provide absolute ratings of HDR processed images on a Single Stimulus Continuous Quality Scale (SSCQS) instead of forcing them to rank images that were output by different algorithms. No reference images are introduced in our experiments.

Unlike subjective data collected in a strictly controlled laboratory setting, the subjective data should reflect the visual quality perceived by consumers using diverse display devices under varying viewing conditions. Similar studies are needed that study the subjective quality of tone-mapped HDR videos [13], [14].

Automatic objective IQA models may be classified as full-reference (FR) or as no-reference (NR).[1] FR-IQA algorithms designed for tonemapping applications [1], [15], [16] compare a tonemapped SDR image with a corresponding HDR irradiance map. However, in many applications the reference 32-bit irradiance map is not available for comparison (such as the huge traffic of HDR-processed pictures shared on social media or on photo-sharing platforms, like Picasa). Hence, FR evaluation of these images is not a practical goal. Again if multi-exposure fusion (MEF) of an exposure stack is used, it is impractical to compare the processed output with a "reference," since among the multiple images in the exposure stack, there is no single identifiable "reference." Hence we focus here on the important NR-IQA aspect of HDR, recognizing that there are likewise many HDR applications where FR-IQA might be useful.

NR-IQA is practical in many applications. Some of the most successful NR-IQA algorithms for SDR images have been developed using Natural Scene Statistics models [17]. NSS models are based on the observation that pristine real-world optical images obey certain statistical principles ('naturalness') that are violated by the presence of distortions ('unnaturalness'). NR-IQA algorithms extract NSS features, then usually train a kernel function to map the features to ground-truth human subjective scores using a supervised learning framework. It is important that these algorithms are trained on a large number of HDR-processed images that are sufficiently representative of photos captured and processed in practice. It is also important to collect a large number of subjective evaluations per image to accommodate variations of perceived quality among human observers on each image.

Present legacy HDR databases are limited in the following ways. First, the small number of images considered may not represent the diversity of HDR images captured in practice. Second, a small number of human subject scores may not adequately capture the variability of user perception in a large population of human subjects. Third, most HDR-processed images in these databases have been annotated by a rank relative to other images instead of being given a raw quality score, thereby making it difficult to map the extracted statistical features to quantifiable human judgments.

In order to address these limitations, we conducted a large-scale crowdsourced subjective study on a large corpus of HDR-processed images to obtain a very large number of subjective opinion scores. Following are the contributions of the paper:

1) We created the new ESPL-LIVE HDR Image Database, comprising 1,811 HDR-processed images created from 605 high quality source HDR scenes. The images were obtained using eleven HDR processing algorithms involving both tonemapping and multi-exposure fusion. We also considered post-processing artifacts of HDR image creation, which typically occur in commercial HDR systems.
2) We conducted subjective experiments on more than 5,000 observers using Amazon's online crowdsourcing platform, Mechanical Turk.
3) We studied variations in the perceived quality of the images with respect to different viewing conditions, demographics, and user familiarity with HDR image processing.
4) We analyzed the performance of several state-of-the-art NR-IQA algorithms (usually studied in the context of SDR images afflicted by commonly occuring artifacts such as blur, additive noise, compression and so on) on the ESPL-LIVE HDR Image Database.

Thus we have designed an IQA database containing a set of images to be evaluated by a pool of human observers. The tone-mapping, multi-exposure or post-processing parameters were not fine tuned to make the images appear aesthetically pleasing; instead we designed a set of images that span the quality spectrum. We do not propose any modification to the compared algorithms to make them correlate better with human perception, although this is an interesting line of inquiry. Indeed, the provided ground truth human subjective scores may be used by future researchers to design HDR-processing algorithms that correlate more highly with human perception.

The remainder of the paper is organized as follows. Section II outlines related previous work on subjective image quality evaluation of HDR images. Details of the source HDR images used and the different processing algorithms deployed are described in Section III. Section IV explains the subjective study setup: a small-scale laboratory sub-

[1]Setting aside reduced-reference models here, which also require a reference.

jective study (to obtain 'gold standard' ratings), and the large-scale crowdsourced subjective study. The raw quality scores obtained from the subjects is analyzed in Section V. Section VI evaluates the performance of several state-of-the-art objective NR-IQA algorithms on the new ESPL-LIVE HDR Image Database and discusses the results. The limitations of the current study have been discussed in Section VII. Section VIII concludes the paper.

II. Related Work

Existing HDR IQA databases have been used to study two typical HDR processing methods: tonemapping and multi-exposure fusion. Yeganeh and Wang [1] carried out a subjective study using 15 reference natural HDR images and 8 tone-mapped SDR images generated using different algorithms. The SDR images were quality ranked from 1 (best) to 8 (worst) by 209 subjects. Ma *et al.* [18] conducted a subjective experiment using 17 reference HDR images and 8 images created using different multi-exposure fusion algorithms. A total of 25 subjects participated in their study.

HDR compression artifacts were subjectively evaluated in [19]–[22]. Narwaria *et al.* [19] and Hanhart *et al.* [20] used still HDR images and distorted versions of them obtained by a combination of different TMOs and JPEG compression at different bit rates. Hanhart *et al.* [21] conducted a subjective experiment using 240 images obtained by tonemapping 20 HDR images with a display adaptive tone-mapping algorithm and compressing them using different profiles of the JPEG XT [23] compression algorithm. Liu *et al.* [24] considered 192 images created from 6 source HDR images impaired by four types of distortions (JPEG/JPEG2K compression, white noise, and Gaussian blur) assessed by 25 participants.

Crowdsourcing for IQA is relatively new. Analyses of the best practices for using crowdsourcing as a method of large scale collection of data may be found in [25] and [26]. One of the earliest crowdsourced subjective experiments [27] gathered ratings from 40 subjects on 116 JPEG compressed SDR images. Ghadiyaram and Bovik [28] developed the LIVE In the Wild Image Quality Challenge Database comprising 1,162 images containing diverse, authentic, real world distortions assessed by more than 8,100 unique subjects. Crowdsourcing of HDR images was used in [20] and [29] to evaluate privacy and compression artifacts in HDR images, respectively. To the best of our knowledge, crowdsourcing has not been used before to conduct subjective quality evaluation of HDR-processed images such as tone-mapping and multi-exposure fusion artifacts at a large scale.

III. ESPL-LIVE HDR Database

This section describes the types of source images, the method of capturing them and the processing algorithms used to generate the processed HDR images in the ESPL-LIVE HDR Database.

A. Source Content

The source images in the new database are real-world HDR scenes of nature, lakes, snow, forests, cities, man-made structures, historical architectures etc. The images were shot both during the day and the night and include both indoor and outdoor scenes. Figure 1 shows some sample images from the new database. The high dynamic range images used in the database were obtained by combining photographs of the same scene shot at multiple exposures using a modern digital SLR camera. The auto-bracketing feature of modern SLR cameras allows multiple photos of the same scene to be captured at several exposure settings with one depression of the shutter release. The new database contains 518 daytime photos and 87 night-time photos. In addition, 444 of the images were taken outdoors while 161 of them are indoor pictures.

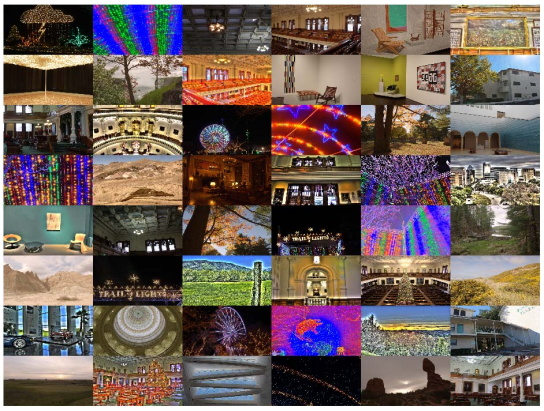
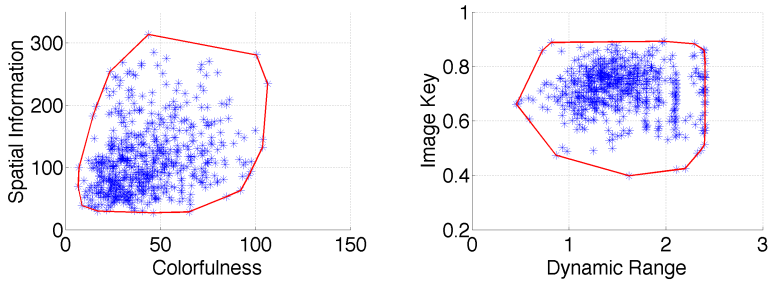


Fig. 1. Sample images from the ESPL-LIVE HDR Image Quality Database. The images include pictures taken during day and night under different illumination conditions. Both indoor and outdoor photos are included, along with scenes containing both natural and man-made objects.

A total of 106 images were obtained from the HDR Photographic Survey [30]. These images were captured with a Nikon D2x using a selection of lenses. Most of the images were obtained with a Nikon 17-55mm f/2.8 ED-IF AF-S DX Zoom-Nikkor lens. The D2x is a professional digital SLR with a 12.4 Megapixel CMOS sensor. The auto-bracketing function allowed for nine exposures to be made at one stop increments in exposure time at a fixed aperture. Capturing them at 5 frames/s allowed nine-exposure HDR sequences covering a nine-stop exposure range to be made in less than two seconds given sufficient light, a feature that is helpful for subjects that might tend to move. These images have a resolution of 4288×2848.

The rest of the images were captured using a Canon Rebel T5 and Nikon D5300 digital SLR camera, with an 18 Megapixel CMOS sensor. An 18-55mm standard zoom lens was used. The auto-bracketing function allowed three exposures to be captured on each scene. The exact range of exposures varied from scene to scene depending on the subject and the available lighting conditions. Under low light conditions, a tripod was used to prevent inadvertent camera shakes. These images have a resolution of 5184×3456. All images were saved in raw electronic format (NEF for Nikon and CR2 for Canon cameras).

In order to minimize the degree of ghosting artifacts arising from moving objects, care was taken to ensure that no high motion objects were present in the scenes. Photomatix was used to process the multiply-exposed stack of images obtained from the DSLRs, to obtain floating point irradiance maps



(a) (b)

Fig. 2. Scatter plots of (a) Spatial Information vs. Colorfulness and (b) Dynamic Range vs. Image Key for the source images in the ESPL-LIVE HDR Database. Red lines indicate the convex hull of the points in the scatter plot, which illustrates the range of scene complexities.

stored in OpenEXR format. Since it is a copyrighted software, we do not have access to the algorithm used to create the OpenEXR files. Correction for small amounts of intra-frame motion (caused by minor movements in a scene) is a common technique in HDR image processing, so that setting was turned on.

B. Source Complexity

The source complexity of the image database was evaluated using four measures: *spatial information* [31], which gives an indication of the richness of the edge distribution in the image; *colorfulness* [31], which quantifies color saturation; *pixel-based dynamic range* [32]; and *image key* [32], which indicates the average image brightness. These quantities are computed on the full-resolution images obtained from the DSLRs. Since for HDR images the scenes are captured at multiple exposures, the scene complexity was determined from the middle exposure image. Figure 2 shows scatter plots between the measured spatial information and colorfulness and that between the dynamic range and image key of the source scenes. As may be observed, the database contains a wide and rich range of scene content according to these measures.

C. HDR Processing Algorithms

Legacy subjective image quality assessment databases usually divide images into distortion categories (such as "Blur", "JPEG Compression", and "Color Saturation"). However, our new database makes no such attempt, although the TMO/MEF algorithms are indeed regarded as sources of distortion. Indeed, it is practically infeasible to superimpose such artificial classification schemes onto realistic HDR images. Depending on the scene and the type of processing algorithm considered, the image could be impaired by a complex interplay of multiple luminance, structural or chromatic artifacts that are hard to categorize. Furthermore, many commercial HDR processing programs postprocess images to modify the local contrast and color saturation, thereby creating a wider perceptual gamut.

Prior to fusing the exposure stack, the bracketed photos need to be registered to correct small misalignments due to camera movement between the shots. Even if the camera is held fixed (as with a tripod), the scene may contain moving objects. Since the merging process assumes that the pixels in the bracketed stack are aligned perfectly, the moving objects may result in ghosting or blurring artifacts, depending on whether the amount of motion is high or low (respectively) [33]. If the trailing 'ghosts' of the moving objects are not removed, viewers may be annoyed by the artifacts. Hence, in this section, we outline the HDR algorithms the we used to create images instead of discretely defining distortion categories. Figure 3 shows the distribution of the algorithms considered in our database. In order to show the difference between the type of artifacts that arise from using HDR processing algorithms and other commonly occurring artifacts, we show the same source scene processed by two TMO operators along with a JPEG compressed version of it in Fig. 4.

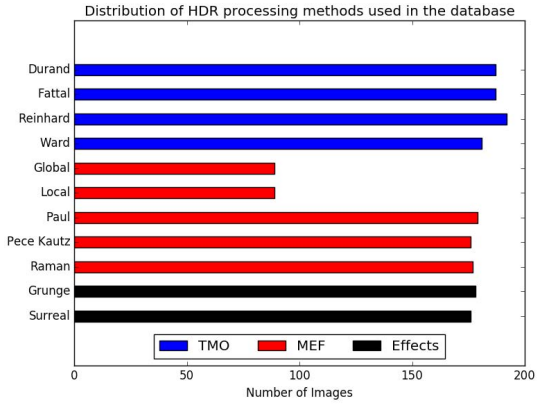


Fig. 3. Bar chart showing the number of images in the database created by each of the different HDR algorithms. 'TMO', 'MEF', and 'Effects' denote Tone-Mapping Operators, Multi-Exposure Fusion Algorithms and Post Processing respectively.

Most of the algorithms were obtained from the HDR Toolbox [41] implemented in MATLAB. The remaining source code was provided by the authors of the algorithms. Instead of generating the best-quality HDR-processed images, they were generated in such a way that they span the entire quality scale so as to present the subjects with a wide range of stimulus. The final images displayed to the subjects had resolutions of 960×540 for landscape orientation and 304×540 for portrait orientation (both downsampled from the original resolution using *imresize* in MATLAB using bicubic interpolation). This was done to ensure that the images fit comfortably within smaller displays and so that the subjects would not encounter delays when loading the images over low bandwidth

(a) (b) (c)

Fig. 4. Image of the same scene tone-mapped and JPEG compressed. (a) Method 1 (Durand TMO [34]) and (b) Method 2 (Fattal TMO [35]) show two different TMO and a JPEG compressed image is shown in (c) . Tone-mapping operators primarily manipulate the contrast of the scene whereas JPEG compression leads to annoying blockiness artifact.

TABLE I

TABLE OF THE NUMBER OF IMAGES IN THE DATABASE CREATED BY EACH HDR ALGORITHM. 'TMO', 'MEF', AND 'EFFECTS' DENOTE TONE-MAPPING OPERATORS, MULTI-EXPOSURE FUSION ALGORITHMS AND POST PROCESSING RESPECTIVELY

Type	Algorithm	Number of images
TMO	Durand [34]	187
TMO	Fattal [35]	187
TMO	Reinhard [36]	192
TMO	Ward [37]	181
MEF	Global energy weighting	89
MEF	Local energy weighting	89
MEF	Paul [38]	179
MEF	Pece Kautz [39]	176
MEF	Raman [40]	177
Effects	Grunge	178
Effects	Surreal	176
	Total	1811

internet connections. Figure 3 and Table I show the number of images in the database created by each of the different HDR algorithms. In total, 1,811 images were used in the subjective experiment. The following sections briefly describe the algorithms used to generate the images in the new database.

1) Images Generated by Tone Mapping Operators: The process of generating well-exposed SDR scenes involves estimating the scene radiance map, followed by tone-mapping it to the displayable gamut of the SDR displays. Some of the earliest algorithms for estimating the radiance map of a natural scene in the HDR format were proposed in [42]–[44] using photographs taken with conventional digital cameras. Given multiple photographs of the same scene taken at different degrees of exposures, the algorithms first recover the camera response function (up to a scale factor) and use it to fuse multiply exposed images into a single HDR radiance map whose pixel values are proportional to the true radiance values of the scene. It is presumed that each scene is static and that the associated series of images were captured by deliberately changing the exposure in quick succession so that lighting changes can be safely ignored.

Once the radiance map is obtained, it is tonemapped to a lower gamut (8 bit/color/pixel) of the SDR display. These algorithms try to replicate the local-adaptation behavior of the human visual system. The human eye adapts to the vast range of real-world illuminations by changing its sensitivity to be responsive at different illumination levels in a highly localized fashion, thereby making it possible to see details in both bright and dark regions [45]. Tone-mapping algorithms compute either a spatially varying transfer function or shrink image gradients to fit within the available dynamic range [46].

On every scene, the raw exposure stack was registered and combined into a 32-bit floating point irradiance map (in OpenEXR format) using Photomatix software with minimal processing. Apart from capturing photographs of the same scene at multiple exposures, some OpenEXR images were also obtained from [47]. The tonemapped images were created by using four representative TMOs proposed by Durand and Dorsey [34], Fattal *et al.* [35], Reinhard *et al.* [36], and Larson *et al.* [37]. The resulting image was downsampled to resolution 960×540 for landscape orientation and 304×540 for portrait images.

2) Images Generated by Multi-Exposure Fusion: The bracketed stack of images, after being downsampled to the display resolution, was first registered using a SIFT based image alignment method [41], and then the aligned images were cropped so that every pixel was visible in every image of the stack, thus avoiding "black border" artifacts. The multiply exposed images were then blended using a MEF algorithm, which can broadly be expressed as [18]

$$Y(i) = \sum_{k=1}^{K} W_k(i) X_k(i) \tag{1}$$

Here K is the number of bracketed images, Y is the fused output image, and $X_k(i)$ and $W_k(i)$ indicate luminance or color either in the spatial domain or coefficients in a transform domain, and the weight at the i-th pixel in the k-th exposure image, respectively. W_k is a relative spatial weight on the images captured at different exposure levels based on a measurement of perceptual information content. Different MEF algorithms differ in the ways that the weights are captured, but they all have an end goal of maintaining details in both underexposed and overexposed regions. These methods bypass the intermediate step of creating an HDR irradiance map by instead creating an SDR image that can be directly displayed on standard displays.

The five algorithms that we used to create multi-exposure fused images are: local and global energy weighting methods, Raman's method based on bilateral filtering [40], the

TABLE II

TABLE SHOWING THE PARAMETERS USED FOR THE PHOTOMATIX POST-PROCESSING EFFECTS: "SURREAL" AND "GRUNGE"

Parameter	"Surreal"	"Grunge"
Strength	50	100
Color Saturation	45	80
Tone compression	1.4	7.1
Detail contrast	10	4.6
Smooth highlights	0	0
White point	10	1.8
Black point	0.097	0.02
Gamma	2.00	0.8
Temperature	0	0

multi-exposure fusion method by Pece *et al.* [39] that also deghosts and Paul *et. al*'s method [38] based on blending the luminance component in the gradient domain. The methods were chosen in order to cover a spectrum of representative MEF algorithms based on a range of processing techniques and computational complexity.

3) Post Processed Images: Many HDR images created by professional and amateur photographers are post-processed in order to convey different 'feels' of a scene. This can drastically alter the final look of the image. We also included post-processed HDR images in the database for subjective evaluation, since these types of effects are not represented in any existing HDR quality databases. In our implementation, we first created an irradiance map using Photomatix and tonemapped it using their default tone-mapping algorithm, followed by post-processing using two commonly used effects: "Surreal" and "Grunge" as determined by the choice of different parameter settings on color saturation, color temperature and detail contrast preservation. Here our goal was to provide samples of post-processing artifacts that often arise among the community of amateur and professional photographers. We chose the Photomatix platform because of its popularity. To constrain the number of images used in the crowdsourcing platform to a reasonable value, we included only two types of special effects. We provide the parameter values used to achieve the "Surreal" and "Grunge" effects in Table II.

IV. SUBJECTIVE STUDY SETUP

Crowdsourced subjective image quality assessment studies provide a wider range of challenges as compared to a traditional subjective study in a laboratory study, primarily due to the lack of control over the precise experimental setup. To validate the subjective results we obtained in the crowdsourced study, we also conducted a separate small-scale controlled laboratory subjective test using a small subset of the HDR images (mentioned in the database link) as a control group to obtain 'gold standard' subjective quality scores. This section describes the setup of the laboratory and online subjective experiments, the methods used to check the consistency of ratings, and the techniques used to analyze the raw scores. In addition, we also studied the dependency of the subjective scores on various demographic factors such as age and gender and various viewing parameters.

A. Laboratory Subjective Evaluation

We conducted a smaller, separate subjective study under controlled conditions to serve as a validation of the crowdsourced study. Fifteen graduate students comprised of five women and ten men in the age group of roughly 20-30 years participated in the laboratory subjective study conducted in the Department of Electrical and Computer Engineering at The University of Texas at Austin in Spring 2016. Most of the subjects did not have any prior experience of participating in a visual subjective test. A single stimulus testing procedure [48] was used. The subjects viewed a total of 38 images of a range of qualities produced by a variety of HDR algorithms. Each testing session entailed viewing 27 images and was preceded by a short training phase, where the subject was shown 11 exemplar images. The training phase was provided in order to familiarize a subject with the experimental setup and hence, the scores entered by the subject during this phase were not considered. On average, each subject required roughly 15 minutes to complete the task.

The user interface for the study was designed on a PC with NVIDIA Quadro NVS 285 GPU using the MATLAB Psychology Toolbox [49] and the images were displayed on a Dell 24-inch U2412M monitor. Each image was displayed on the screen for 12 seconds. The subjects viewed the images from about 2 - 2.25 times of the display height. 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 around 540 lux.

The screen resolution was set at 1920×1200 pixels, but the images were displayed at their normal resolution (1920×1080) without introducing any distortion by interpolation. The top and bottom portions of the display were set to gray color. At the end of each image's display interval, a continuous quality scale was displayed on the screen, where the default initial location of the slider was at the center of the scale. The scale was marked with five Likert adjectives: "Bad", "Poor," "Fair," "Good," and "Excellent". After the subject entered a rating for an image, the location of the slider along the scale was converted into an integer score lying between [0,100]. The subject was allowed to 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 once the rating bar was presented. The next image was automatically displayed once the score for the current image was recorded.

Regarding subject rejection, 3 of the 15 subjects were found to be outliers following the standard ITU BT.500-13 recommendation [48]; hence the mean opinion score (MOS) of each image was calculated using the scores of the remaining 12 subjects. In order to account for variability among subjects, the raw subjective scores were converted to Z-scores [50] before calculating MOS. Based on the MOS scores, five images were chosen as gold standard exemplars spanning the quality scale.

B. Challenges to Crowdsourcing

There has recently been growing interest in using online crowdsourcing platforms such as Amazon Mechanical Turk (AMT) [51], Microworkers [52], and Crowdflower [53] to collect large-scale human data from a diverse and distributed global population. The registered 'requesters' advertise their

tasks to registered 'workers' who can choose to provide their inputs for data-collection in return for monetary compensation. The following salient features should be kept in mind while designing a crowdsourced subjective experiment:

- While the reach of these online platforms to a large number of potential subjects does help the requesters collect a large number of image ratings in a much shorter time than via standard laboratory experiments, the requesters have limited control over the experimental setup, e.g., display devices used by workers, their distance from the display, and illumination conditions in the workers' viewing environment. Since these factors may have a significant effect on the image ratings provided by the users, information regarding these factors was collected from the users at the end of each viewing session by asking them to complete a short survey. We gathered information from them on their familiarity with HDR photography, the devices used to capture HDR content and the softwares used to process HDR images. Further details are in the next section.
- The time spent by a subject on a subjective experiment via a crowdsourcing platform differs from a laboratory experiment. In the latter setup, the goal is make the subject evaluate each and every image in the dataset; hence the study may last for a couple of hours which may be broken into multiple shorter sessions to avoid subject fatigue. However, in a crowdsourced setting, since it is difficult to induce workers to participate in time-consuming activities [54], the online tasks need to be segmented into smaller chunks. Hence, each image in the database was viewed and evaluated by a subset of the participating workers.

C. Instructions, Training, and Testing

The subjects were instructed to focus on image quality rather than image aesthetics. Care was taken to provide a wide variety of images that they are likely to encounter in real life. We relied on the human subjective judgments of the experimenters to select the images. On AMT, requesters present the tasks as Human Intelligence Tasks (HITs). The workers are shown an instructions page explaining the details of the study along with the monetary reimbursement offered. If the worker is interested in participating, she has to click the "Accept HIT" button to begin the actual task. At the end of the task, the worker submits her results to the requester by clicking on the "Submit Results" button.

1) Interface Used: Apart from the instructions, the workers were also shown some representative images in the database along with a screenshot of the interface to be used to rate the images. Once the worker accepted the HIT, she was presented with a rating interface, as shown in Figure 5, containing the image to be evaluated and a slider below it. A single stimulus quality evaluation [48] method was used in the experiment. The subjects entered the ratings by dragging a horizontal slider bar along a continuous scale marked at equal intervals "bad," "poor," "fair," "good," and "excellent," to aid the subject in entering her judgment. Once she decides on the rating, she changes the slider position accordingly. Upon pressing the

Fig. 5. Rating Screen for Amazon Mechanical Task HIT shown to the subjects.

"Next Image" button, the position of the slider was converted to an integer valued quality score between [1-100] and the next image was presented. Unlike the laboratory experiments where the subjects were shown each image for a fixed amount of time, on the crowdsourced platform, the subjects could view each image for as long as they desired.

2) Training and Testing Phase: Following a similar procedure as the laboratory experiment, before the testing phase, each participant was shown a set of 11 training images to familiarize them with the user interface, to get a sense of the range of image qualities and the types of processing artifacts that they might encounter during the actual testing phase. The training set of images was the same for all participants.

The testing phase experienced by each subject involved viewing 49 images selected randomly from the corpus of 1,811 images in the database, and presented in a randomized order for each subject. The testing phase was followed by a short survey. On average, the subjects required 9 minutes to complete the task of evaluating a total of 60 images and they were paid US$0.45 for their participation.

D. Subject Reliability and Rejection Strategies

Although AMT makes it possible to gather subjective evaluations from a large number of subjects in a relatively short period of time, stringent subject rejection strategies were implemented in order to ensure high quality reliable ratings. Following are the subject rejection methods that we used:

- *Intrinsic metric*: Only those workers on AMT having AMT confidence values greater than 0.75 (on a [0,1] scale) were allowed to participate in the study. Although this number may not take into account the performance of the subject on previous visual tasks, a higher confidence number indicates a more reliable subject. To avoid bias, we only allowed unique participants. Hence if the same worker selected the task again, she was not allowed to proceed beyond the instructions page.
- *Using corrective lens*: If any worker wore corrective lenses in their day-to-day life, they were instructed to wear them during the entire duration of the study. At the end of the task, they were asked whether they normally wore corrective lenses and whether they were wearing them during the task. If a certain worker, who was supposed to be wearing lenses, reported that she was not using them during the study, her scores were rejected.

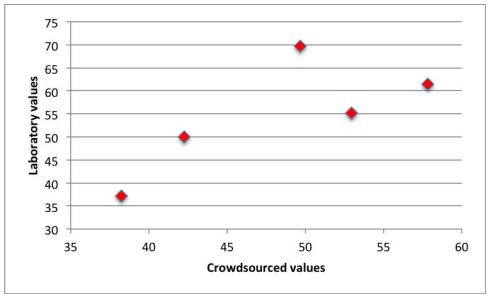


Fig. 6. Scatter plot of the MOS scores of the five 'gold standard' images obtained from the laboratory vs. the ones obtained from the crowdsourcing experiment.

- *Repeated images*: From among the 49 test images, 5 were randomly chosen and presented twice to each subject during the testing phase. If the difference between the two scores provided by the worker to the same image exceeded a certain threshold for at least 3 of the 5 repeated images, the scores from that worker were rejected. During the initial phase of the study, the average standard deviation of the scores obtained from about 400 workers was found to be 17 (rounding up to the nearest integer). A value of 1.5 times the average standard deviation was used as the threshold for rejecting subjects. This method eliminated inattentive or otherwise disengaged subjects who were providing arbitrary scores to the images.
- *Gold standard images*: As described earlier, 5 of the remaining 44 images were chosen from the laboratory subjective study. These images, referred to as "gold standard" set were used to provide a control. The median value of Pearson's linear correlation coefficient (PLCC) between the scores provided by each subject to these five images in the crowdsourced study, and the corresponding MOS calculated from the laboratory subjective test, after applying non-linear regression was found to be **0.9465**,[2] the root-mean-square-error between the subject scores and the ground truth MOS values was 5.4710, and the PLCC without non-linear regression was 0.7514. This high degree of agreement between the ground truth data obtained from the laboratory settings and from the online platform strongly suggests a high degree of reliability of the scores obtained by crowdsourcing. Figure 6 shows the scatter plot of the MOS scores of the five 'gold standard' images obtained from the laboratory vs. the ones obtained from the crowdsourcing experiment.

E. Subject-Consistency Analysis

While a variety of measurements of intersubject agreement are available [55], these generally cannot be applied here.

[2]Unless otherwise mentioned, all correlation values between the IQA algorithm scores and/or human ground truth values were computed following non-linear logistic regression as outlined in [50]. The logistic regression helps measure the degree of monotonicity between the two sets of scores when the relationship between them is non-linear or their scales differ. Correlations measured after logistic regression might be higher than without, if the relationship between the variables is not very linear.

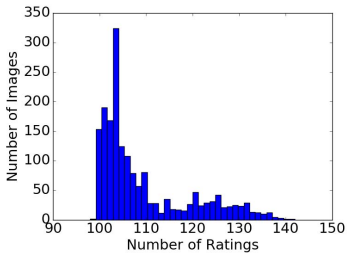


Fig. 7. Distribution of number of ratings per image.

In our study, each subject was exposed to only a very small percentage of the images in the database. Moreover, no two subjects viewed the same set of images. Therefore, we utilized the following methods to analyze the degree of consistency of the scores obtained from the many subjects:

- *Inter-subject consistency*: For each image, the ratings were randomly divided into two disjoint equal sized subsets and MOS values were computed on each of them. This procedure was repeated over 25 random splits. The median Spearman's Rank Order Correlation Coefficient (SROCC) between the MOS between the two sets was found to be **0.9700** $\pm$ **0.0013**, while the Pearson Linear Correlation Coefficient (PLCC) was **0.9721** $\pm$ **0.0011**. The corresponding root-mean-square-error was 2.3713.
- *Consistency with 'gold standard' images*: Pearson's linear correlation coefficient was measured between the individual opinion scores and the MOS values of the gold standard images. A median PLCC value of **0.8743** was obtained over all subjects. The corresponding median root-mean-square-error was found to be 7.7703.

The high values of these measures indicate good consistency between the scores obtained from the subjects on each image.

V. Analysis of Subjective Scores

We gathered 327,720 ratings of picture quality from 5,462 unique participants. Of these, the scores from 388 subjects were eliminated following the rejection criterion based on their performance on the "gold standard" images, and/or for not following the instruction of wearing corrective lenses when they were supposed to. The images were evaluated by an average of 110 observers. Figure 7 plots the histogram of the number of ratings per image.

The MOS was computed by averaging the Z-scores using the method outlined in [56]. The range of MOS values spans [16.941 - 68.502]. Figure 8 shows a histogram of the MOS scores for every image obtained from the Z-scores. The average standard deviation of all of the subjective scores was found to be 21.131.

We also gathered demographic information about the subjects, such as age and gender, as shown in Figure 9. Since familiarity of the subjects with HDR photography might affect the quality scores provided by them, the subjects were also requested to provide information regarding the same. Figure 10 summarizes the levels of awareness of the subjects about HDR photography, the type of optical devices used by them to

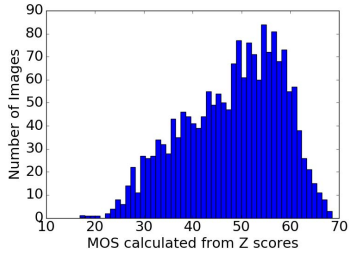


Fig. 8. Histogram of MOS obtained from the human subjects. The range of the MOS values spans [16.941 - 68.502].

capture HDR content (if they indeed knew about HDR), and their familiarity with image processing software such as Adobe Photoshop or Photomatix. This last question was included in the survey because some of the images were created by adding special post-processing effects following HDR fusion.

The subjects were instructed to work on the HIT only from personal computers instead of smartphones or tablets. The type of display devices used and the distance from the screen can affect the visual quality of the image. The subjects also provided information on these aspects. Figures 10(b) and (c), respectively, show the types and distribution of displays used by the subjects, and their estimated distances from the screen while completing each HIT.

A. Variation of Subjective Scores With Number of Subjects

While the MOS scores that are used in our study were computed on all of the subjects, we conducted the following procedure to study the effects of subject count on MOS. To do this, we randomly selected five images from the database (these are shown in Fig. 11), each of which had more than 100 subjective scores associated with it. We then randomly sampled 20, 40, 60, 80 and 100 of the subject scores of each image and recomputed the MOS again on each of these reduced subsets. Figure 12 shows that these computed MOS values remained relatively constant with respect to the number of subjects viewing the images, although the standard error increased noticeably below 40 subjects. The confidence intervals were calculated based on the variation of the MOS scores over the 25 trials, which gives a rough indication of the number of subjective evaluations that are needed to obtain reliable MOS scores (depending on the accuracy needed in an application).

B. Variation of Subjective Scores

Here we summarize observations on how the perceptual quality judgments of the subjects were affected by parameters such as age, gender, display devices used when participating in the subjective study, distance from the display, and their familiarity with HDR image processing. Figure 11 shows five representative images on which the effects of the above mentioned factors on the subjective scores was studied. At a confidence interval of 95%, one-way (Analysis of variance) ANOVA test was performed in order to find out whether these factors affect the MOS scores.

1) Age: Data from subjects who used a laptop during the study and were sitting about 15 - 30 inches away from the screen was used to isolate the effects of age on the perceived quality of the images while holding other factors relatively fixed. These display settings were selected because most of the subjects participated in the experiment using their laptops and reported to be sitting at about 15 - 30 inches away from the screen, thereby providing us with sufficient number of samples to study the effect of age on perceived quality. The individual ratings on the images shown in Fig. 11 were grouped according to three age categories: '20-30', '30-40' and '>40' and the MOS was computed for each group, as shown in Fig. 13.

2) Gender: Data from subjects between 20 - 30 years of age, who used a laptop during the study and were sitting about 15 - 30 inches away from the screen was used to isolate the effects of gender on perceived quality of the images while keeping the other factors relatively constant. These display settings were selected for the same reasons as above. The individual ratings on the images shown in Fig. 11 were grouped according to their gender and the MOS was computed for each group, as shown in Fig. 14.

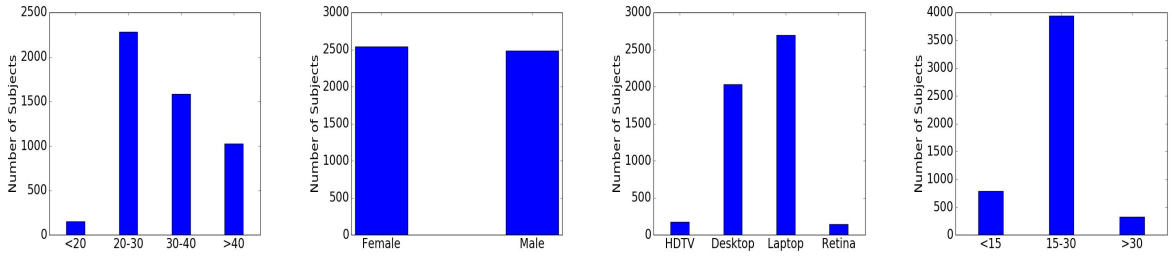
3) HDR Awareness: One of the questions asked of the subjects was whether they were familiar with HDR images. Figure 10 shows the distribution of the answers of the subjects to various HDR related questions. The individual ratings on the images shown in Figure 11 were grouped according to whether the users were familiar with HDR imaging. The MOS was computed for each group, as shown in Fig. 15. We hypothesize that since the subjects were not shown the original HDR irradiance map on an HDR compatible display and were not informed at the beginning of the experiment that they would be evaluating HDR processed content, they judged the artifacts more or less similarly.

4) Display Device Used: The subjects were asked to report the type of display device they used to participate in this study. The individual ratings on the images shown in Fig. 11 were grouped according to whether the users were using a desktop or a laptop computer and the MOS was computed for each group, as shown in Fig. 16. The type of display device used by a subject may impact their perception of quality. Studies of the effects on perceived quality of the type of display, the screen resolution, and the interplay between the display technology and the HDR processing algorithms used are topics of interesting future study.

5) Distance From Display: The subjects were asked to report how far they were sitting from the display while participating in this study. The individual ratings on the images shown in Fig. 11 were grouped according to three distances: '<15,' '15-30,' and '>30' inches from the display and the MOS was computed for each group, as shown in Fig. 17.

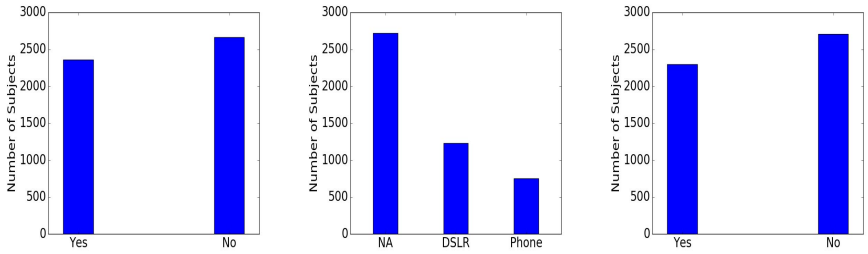
C. Discussion of HDR Processing Algorithms

In order to study the relationship between image quality and the type of HDR processing algorithms, we divided the images into three categories: high quality (having raw scores above 70), medium quality (raw scores lying between 40 and 70) and low quality (raw scores less than 40). For the high quality algorithms, most the images were found to be processed by the TMOs outlined in [36] and [37] and by MEF algorithms like [38] and [40]. A previous study



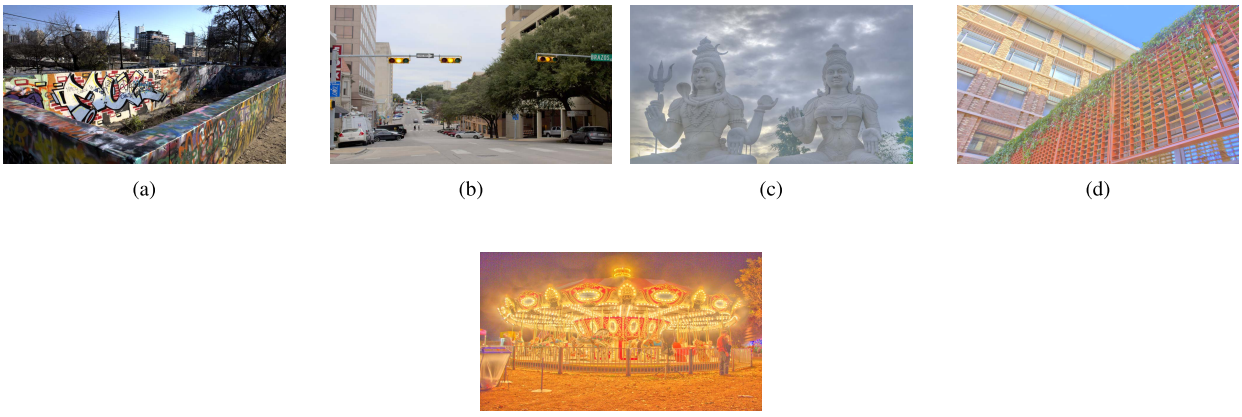
(a) (b) (c) (d)

Fig. 9. Demographics of the participating human subjects by (a) age (b) gender and display (c) different categories of display devices used by the workers to participate in the study and (d) approximate distance in inches between the subject and the viewing screen.



(a) (b) (c)

Fig. 10. HDR awareness of the subjects (a) Number of subjects aware of HDR images (b) The types of devices they used to capture HDR content where 'NA' indicates subjects who are not familiar with HDR and (c) Number of subjects familiar with image processing software such as Photoshop or Photomatix.



(a) (b) (c) (d)

(e)

Fig. 11. Sample images from HDR database used to illustrate the effect of increasing the number of participants on the calculated MOS. The caption of each image gives the MOS values and the associated 95% confidence intervals. (a) MOS = 62.43 ± 2.04. (b) MOS = 52.90 ± 2.17. (c) MOS = 42.33 ± 2.77. (d) MOS = 40.23 ± 2.42. (e) MOS = 31.07 ± 2.82.

by Drago *et al.* [5] showed that histogram adjustment based methods produce high contrast images and preserves spatial details relatively well. Ledda *et al.* [3] found that [36] performs very well when the subjects were asked to rank the outputs of different tone-mapping operators. For the medium quality algorithms, all of the TMO and MEF algorithms were found to perform about equally well. The TMO outlined in [34] using bilateral filters was found to yield lower quality images. This agrees with the results of many previous studies, such as [3] and [12], that bilateral/trilateral filtering yields outputs less similar to real scenes. Thus we find that many of the hypotheses developed in previous laboratory studies of TMOs agree with the MOS collected in our crowdsourced experiment.

VI. Evaluation of IQA Algorithms

We also tested the performance of some of the state-of-the-art NR-IQA algorithms on the new database to demonstrate and study the usefulness of the database and the capabilities and limitations of current models when evaluating HDR processing artifacts. Table III outlines the features extracted by the various NSS based NR-IQA algorithms evaluated on the database. The algorithms HIGRADE-1 and HIGRADE-2 are two recently proposed gradient scene-statistics based

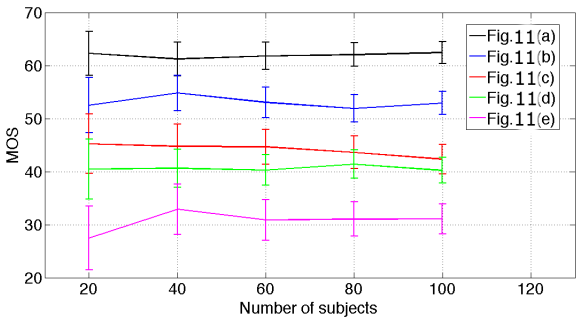


Fig. 12. MOS plotted against the number of workers who viewed and rated the images shown in Fig 11 along with the 95% confidence intervals.

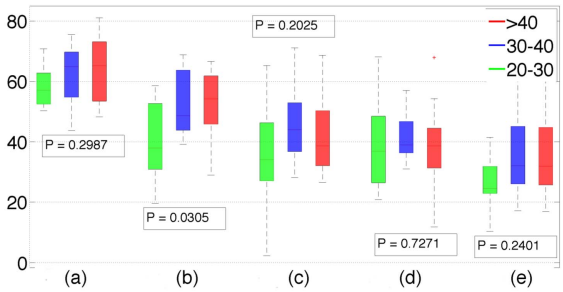


Fig. 13. Individual Z-scores obtained from subjects of different ages who rated the images shown in Fig 11. The letter below each column indicates which image in Fig. 11 was rated. For each vertical column, the median is the center of the central box, while the upper and lower edges of each box represent the 25th and 75th percentiles, and the whiskers span the most extreme non-outlier data points. P-values obtained from the one-way ANOVA tests have been shown for each figure. Other than the image in Fig. 11(b), the P-values indicate that there is no statistical evidence to reject the null hypothesis that people from different age groups rated the images the same.

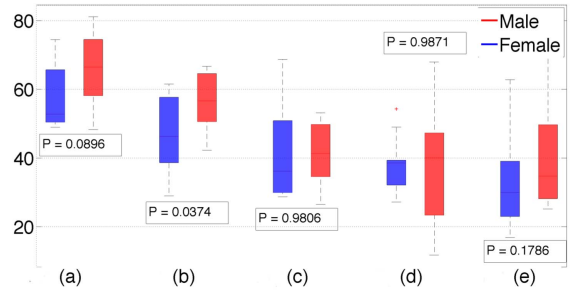


Fig. 14. Individual Z-scores obtained from subjects of different genders who rated the images shown in Fig 11. The letter below each column indicates which image in Fig. 11 was rated. For each vertical column, the median is the center of the central box, while the upper and lower edges of each box represent the 25th and 75th percentiles, and the whiskers span the most extreme non-outlier data points. P-values obtained from the one-way ANOVA tests have been shown for each figure. Other than the image in Fig. 11(b), the P-values indicate that there is no statistical evidence to reject the null hypothesis that people of different genders rated the images the same.

NR-IQA algorithms defined in the LAB color space [58]. HIGRADE-1 (L) and HIGRADE-2 (L) are versions of these algorithms that only operate on the luminance channel (L). Although there are no clear-cut distortion categories that can be defined on this database, results are summarized individually for each class of HDR processing algorithms.

The performances of the algorithms were evaluated by measuring correlations with subjective scores (after non-linear regression). Once the features were extracted, a mapping was 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)

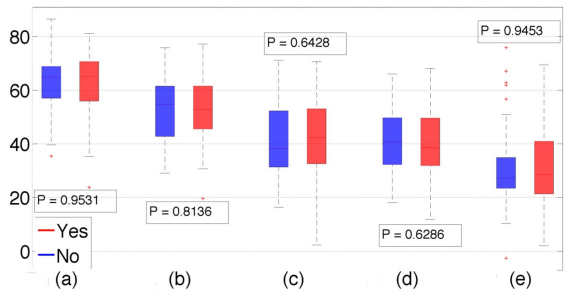


Fig. 15. Individual Z-scores obtained from subjects familiar with or not familiar with HDR imaging who rated the images shown in Fig 11. The letter below each column indicates which image in Fig. 11 was rated. For each vertical column, the median is the center of the central box, while the upper and lower edges of each box represent the 25th and 75th percentiles, and the whiskers span the most extreme non-outlier data points. P-values obtained from the one-way ANOVA tests have been shown for each figure. The P-values indicate that there is no statistical evidence to reject the null hypothesis that people rated the images the same depending on their familiarity with HDR images.

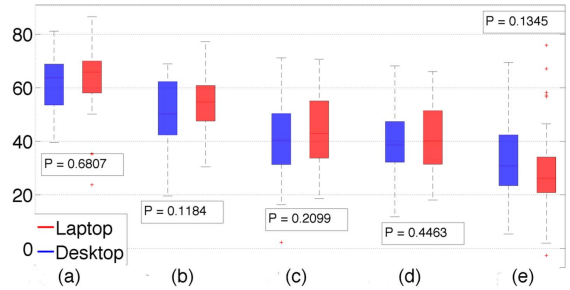


Fig. 16. Individual Z-scores obtained from subjects using different display devices who rated the images shown in Fig 11. The letter below each column indicates which image in Fig. 11 was rated. For each vertical column, the median is the center of the central box, while the upper and lower edges of each box represent the 25th and 75th percentiles, and the whiskers span the most extreme non-outlier data points. P-values obtained from the one-way ANOVA tests have been shown for each figure. The P-values indicate that there is no statistical evidence to reject the null hypothesis that people rated the images the same based on the type of their display devices.

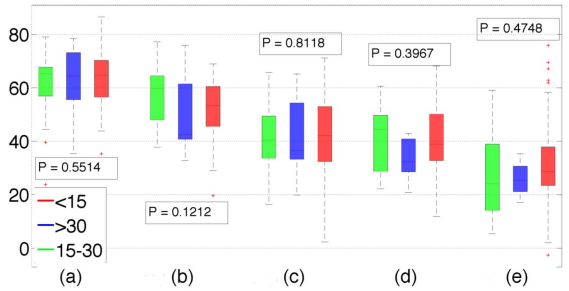


Fig. 17. Individual Z-scores obtained from subjects viewing the images at different distances (expressed in inches) who rated the images shown in Fig 11. The letter below each column indicates which image in Fig. 11 was rated. For each vertical column, the median is the center of the central box, while the upper and lower edges of each box represent the 25th and 75th percentiles, and the whiskers span the most extreme non-outlier data points. P-values obtained from the one-way ANOVA tests have been shown for each figure. The P-values indicate that there is no statistical evidence to reject the null hypothesis that people rated the images the same based on their distance from the display devices.

TABLE III

LIST OF NR-IQA ALGORITHMS EVALUATED IN THIS STUDY

	Algorithm	Feature
1	Derivative Statistics-based QUality Evaluator (DESIQUE) [57]	Pointwise and pairwise log-derivative statistics in spatial and frequency domain
2	Gradient Magnitude NSS based IQA (HIGRADE-1) [58]	Pointwise and pairwise log-derivative statistics of pixels and gradient magnitude
3	Gradient Coherency NSS based IQA (HIGRADE-2) [58]	Pointwise and pairwise log-derivative statistics of pixels and gradient coherency
4	Gradient Magnitude and Laplacian of Gaussian based NR-IQA (GM-LOG) [59]	Joint statistics of Gradient Magnitude and Laplacian of Gaussian
5	NR-IQA based on Curvelets (CurveletQA) [60]	Log-histograms and energy distribution of orientation and scale of curvelet coefficients,
6	NR-IQA for Contrast Distorted Images (Contrast QA) [61]	Sample mean, standard deviation, skewness, kurtosis and entropy features
7	Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE)	Real-valued wavelet coefficients modeled using the Gaussian Scale Mixture
8	BLind Image Integrity Notator using DCT Statistics-II (BLIINDS-II) [62]	DCT coefficients modeled using Generalized Gaussian Distribution
9	Complex-DIIVINE (C-DIIVINE) [63]	Complex-valued wavelet coefficients modeled using the Gaussian Scale Mixture and wrapped Cauchy distribution
10	Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) [64]	Pointwise and pairwise statistics in spatial domain

TABLE IV

MEDIAN SPEARMAN'S RANK ORDERED CORRELATION COEFFICIENT (SROCC) AND PEARSON'S LINEAR CORRELATION COEFFICIENT (PLCC) BETWEEN THE ALGORITHM SCORES FOR VARIOUS IQA ALGORITHMS AND THE MOS SCORES ON THE ESPL-LIVE HDR DATABASE. THE TABLE WAS SORTED IN DESCENDING ORDER OF SROCC OF THE 'OVERALL CATEGORY'. THE BOLD VALUES INDICATE THE BEST PERFORMING ALGORITHM

	IQA	Tone Mapping		Multi-Exposure Fusion		Post Processing		Overall	
		SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
1	HIGRADE-1	0.728	0.764	**0.711**	**0.705**	**0.616**	**0.643**	**0.719** (0.671, 0.766)	**0.718**(0.652, 0.776)
2	HIGRADE-2	**0.752**	**0.777**	0.706	0.690	0.529	0.552	0.711 (0.639, 0.792)	0.704(0.645, 0.788)
3	HIGRADE-2 (L)	0.703	0.737	0.662	0.661	0.465	0.515	0.662 (0.575, 0.730)	0.663(0.571, 0.730)
4	HIGRADE-1 (L)	0.672	0.702	0.634	0.637	0.551	0.582	0.661 (0.595, 0.732)	0.658(0.590, 0.738)
5	DESIQUE	0.542	0.553	0.572	0.584	0.529	0.563	0.570 (0.481, 0.657)	0.568(0.467, 0.650)
6	GM-LOG	0.549	0.562	0.545	0.541	0.578	0.599	0.556 (0.448, 0.638)	0.557(0.465, 0.639)
7	CurveletQA	0.584	0.623	0.517	0.535	0.481	0.506	0.547 (0.458, 0.610)	0.560(0.447, 0.631)
8	ContrastQA	0.683	0.717	0.432	0.433	0.401	0.413	0.506 (0.405, 0.631)	0.521(0.402, 0.632)
9	DIIVINE	0.523	0.530	0.453	0.472	0.392	0.447	0.482 (0.326, 0.578)	0.484(0.331, 0.583)
10	BLIINDS-II	0.412	0.442	0.446	0.459	0.486	0.510	0.444 (0.310, 0.519)	0.454(0.326, 0.545)
11	C-DIIVINE	0.453	0.453	0.423	0.460	0.432	0.470	0.434 (0.265, 0.551)	0.444(0.277, 0.538)
12	BRISQUE	0.340	0.370	0.494	0.516	0.468	0.483	0.418 (0.300, 0.500)	0.444(0.313, 0.528)

TABLE V

ROOT-MEAN-SQUARE ERROR (RMSE), REDUCED $\tilde{\chi}^2$ STATISTIC BETWEEN THE ALGORITHM SCORES AND THE MOS FOR VARIOUS NR-IQA ALGORITHMS (AFTER LOGISTIC FUNCTION FITTING) AND OUTLIER RATIO (EXPRESSED IN PERCENTAGE) FOR EACH DISTORTION CATEGORY ON THE ESPL-LIVE HDR DATABASE. THE BOLD VALUES INDICATE THE BEST PERFORMING ALGORITHM FOR THAT CATEGORY

	IQA	Tone Mapping			Multi-Exposure Fusion			Post Processing			Overall		
		RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR
1	HIGRADE-1	6.711	9.908	**0.000**	**6.884**	21.155	**0.000**	**6.884**	2.376	**0.000**	**7.033**	13.918	**0.275**
2	HIGRADE-2	**6.643**	3.576	**0.000**	6.988	5.983	**0.000**	7.457	6.660	**0.000**	7.231	16.495	0.277
3	HIGRADE-2 (L)	7.070	5.327	**0.000**	7.178	13.882	**0.000**	7.742	3.227	**0.000**	7.607	13.879	0.551
4	HIGRADE-1 (L)	7.434	8.624	0.662	7.484	5.263	**0.000**	7.308	3.131	**0.000**	7.628	12.558	0.552
5	DESIQUE	8.577	12.079	0.683	7.862	11.588	0.687	7.402	1.851	**0.000**	8.296	19.614	0.829
6	GM-LOG	8.632	5.002	1.170	8.028	15.027	0.702	7.420	0.851	**0.000**	8.357	20.659	0.829
7	CurveletQA	8.177	17.408	0.694	8.054	10.754	0.714	7.922	2.892	**0.000**	8.511	15.253	0.829
8	ContrastQA	7.248	8.681	**0.000**	8.562	21.640	0.755	8.183	6.501	**0.000**	8.556	33.433	0.829
9	DIIVINE	8.805	10.025	0.791	8.371	5.663	0.667	7.979	2.659	**0.000**	8.821	12.115	0.829
10	BLIINDS-II	9.330	7.565	0.697	8.517	19.979	0.752	7.818	1.976	**0.000**	8.975	21.948	0.828
11	C-DIIVINE	9.167	15.338	1.356	8.485	8.374	0.671	7.852	1.428	**0.000**	8.983	12.305	0.966
12	BRISQUE	9.535	16.712	1.356	8.227	5.681	0.685	7.894	7.146	**0.000**	9.049	17.259	0.831

(LibSVM [65]) to implement ϵ-SVR with the radial basis function kernel, where the kernel parameter is by default the inverse of the number of features.

We randomly split the data into disjoint training and testing sets at a 4:1 ratio and the split was randomized over 100 trials. Care was taken to ensure that the same source scene did not appear during both training and testing to prevent artificial inflation of the results. The Spearman's rank ordered correlation coefficient (SROCC) and Pearson's linear correlation coefficient (PLCC) values between the predicted and the ground truth quality scores were computed at each iteration and the median values of the correlations were found. The results indicate that there is significant room for improvement among current NR-IQA algorithms when predicting HDR artifacts. The results are summarized in Table IV.

Table V shows the root-mean-squared-errors (RMSE), reduced $\tilde{\chi}^2$ statistic between scores predicted by the algorithms and MOS (after logistic regression) and the outlier ratios (expressed in percentage). The top performing algorithms yielded lower values of RMSE, reduced $\tilde{\chi}^2$ statistic and outlier ratio.

Many of the tonemapping and multi-exposure fusion algorithms modify the gradients of the component images of the exposure stack. We found that algorithms that take into

TABLE VI

ESPL STUDY: VARIANCE OF THE RESIDUALS BETWEEN INDIVIDUAL SUBJECTIVE SCORES AND NR-IQA ALGORITHM PREDICTIONS

	TMO	MEF	PP	Overall
Number of samples	140	149	70	359
Threshold F-ratio	1.32	1.31	1.49	1.19
HIGRADE-1	**43.39**	54.52	**48.23**	50.95
HIGRADE-2	45.12	**45.24**	53.73	**49.67**
DESIQUE	63.78	69.10	62.42	70.65
BRISQUE	96.17	87.27	59.58	87.03
GM-LOG	80.23	83.72	55.70	79.29
C-DIIVINE	81.29	79.20	69.73	81.23
DIIVINE	88.70	76.79	64.41	84.09
BLIINDS-II	77.27	87.84	67.90	85.26
Curvelet	63.66	78.44	64.53	72.57
ContrastQA	54.90	80.05	66.02	71.59

TABLE VII

ESPL STUDY: VARIANCE OF THE RESIDUALS BETWEEN INDIVIDUAL SUBJECTIVE SCORES AND NR-IQA ALGORITHM PREDICTIONS

	TMO	MEF	PP	Overall
Number of samples	15297	16371	7588	39256
Threshold F-ratio	1.03	1.03	1.04	1.02
HIGRADE-1	**185.10**	199.55	**239.37**	250.56
HIGRADE-2	186.84	**190.87**	244.31	**245.66**
DESIQUE	206.32	213.87	251.15	250.54
BRISQUE	238.99	231.84	248.27	254.71
GM-LOG	222.10	229.32	246.53	254.33
C-DIIVINE	224.45	223.14	258.63	250.40
DIIVINE	230.35	222.06	253.34	251.70
BLIINDS-II	219.63	233.38	257.58	250.95
Curvelet	205.46	223.40	252.82	251.02
ContrastQA	196.93	224.25	255.57	252.79
Null Model	142.01	145.96	190.19	229.88

account variations of the gradient of the images achieved a higher degree of correlation with the human ground truth subjective data. Both grayscale and color versions of the proposed models were found to exhibit good correlations with human judgment compared to other state-of-the-art NR-IQA algorithms. However, as expected, algorithms that use all three LAB color channels performed better than models that only extract feature on the L-channel, especially on post-processing artifacts that modify the color-saturation and/or color temperature of the images.

A. Determination of Statistical Significance

Ten representative NR-IQA algorithms were studied in regards to determining the significance of their relative performances. Following the methods outlined in [50], statistical significance tests were carried out over multiple 4:1 train-test splits and similar results were obtained. We show the results obtained for one such representative trial. Results are summarized in Tables VI, VII, VIII and IX. For the F-Test based on MOS, the variance of the residuals obtained from the null-model and the ten selected IQA algorithms, along with the number of samples considered in each category and the threshold F-ratio at 95% significance are shown in Table VI, while Table VII shows the corresponding result considering the individual scores obtained from the human subjects. None of the IQA algorithms tested was found to be statistically equivalent to the null-model corresponding to human judgment. HIGRADE-1 shows the least variance of the residuals for the overall database.

To determine whether the performance of the IQA algorithms are significantly different from each other, the F-statistic, as in [50], 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 interval. Table VIII shows the results for the ten IQA algorithms and all HDR processing methods when MOS scores were considered, while Table IX shows the corresponding result considering the individual scores obtained from human subjects. Both of these indicate that while most of the models produced similar results, HIGRADE-1 and HIGRADE-2 were found to be statistically superior overall relative to the other NR-IQA algorithms.

VII. LIMITATIONS AND DISCUSSION

The overarching goal of this study has been to leverage crowdsourcing tools to collect human opinion data on a large scale, while allowing for a wide range of displays and viewing conditions that are closer to free-viewing scenarios than can be obtained under strictly controlled laboratory settings. By design, the human observers rated the images in the most natural way, using a Single Stimulus Continuous Quality Scale, instead of ranking their qualities. Our study is directed at the evaluation of NR models only, hence a limitation of this new resource is it cannot be used to evaluate the relative performances of FR-IQA algorithms that measure the signal fidelity between original HDR irradiance maps and images obtained by processing them with different TMO operators. To create the image database, we generated images using eleven widely used tonemapping and other HDR processing algorithms that deploy fundamental methods. This has an advantage of generality, but does not supply a resource for directly comparing the efficacies of the latest TMO models against human subjective judgments. Future interesting directions for crowdsourced studies might include large numbers of images generated by emergent HDR-processing algorithms or wider classes of distortions, and multiply distorted images [66], such as those arising from compressing or transmitting HDR-processed images.

While crowdsourcing human opinion scores requires considerable care, we applied the experiences gained on an even larger study [28] to produce a dataset exhibiting excellent internal consistency. Nevertheless, it would be quite useful to conduct dual experiments whereby a large number of the same images are evaluated under both laboratory and crowdsourcing conditions, to determine the differences in human responses under these very different conditions. In a laboratory study, HDR displays could be used to display floating point HDR irradiance maps to the human subjects, against the wide diversity of displays used in a corresponding crowdsourced experiment. This could lead to insightful comparisons between the perceptions of HDR and HDR-processed images.

The kinds of distortions caused by TMOs also depend on the color gamut and peak luminance of each display. Hence, such a study might be used to determine which TMOs are the most efficacious. Optimizing the parameters of TMO models using crowdsourcing is an interesting direction of open research.

TABLE VIII

RESULTS OF THE F-TEST PERFORMED ON THE RESIDUALS BETWEEN MODEL PREDICTIONS AND MOS SCORES ON ESPL-LIVE HDR DATABASE. EACH CELL IN THE TABLE IS A CODEWORD CONSISTING OF 4 SYMBOLS THAT CORRESPOND TO "TONE MAPPING OPERATORS"', "MULTI-EXPOSURE FUSION", "POST PROCESSING", AND "OVERALL" PROCESSING ALGORITHMS. "1"("0") INDICATES THAT THE PERFORMANCE OF THE ROW IQA ALGORITHM IS SUPERIOR(INFERIOR) TO THAT OF THE COLUMN IQA ALGORITHM. - INDICATES THAT THE STATISTICAL PERFORMANCE OF THE ROW IQA IS EQUIVALENT TO THAT OF THE COLUMN IQA. THE MATRIX IS SYMMETRIC

	HIGRADE-1	HIGRADE-2	DESIQUE	BRISQUE	GM-LOG	C-DIIVINE	DIIVINE	BLIINDS-II	Curvelet	ContrastQA
HIGRADE-1	- - - -	- - - -	1 - - 1	1 1 - 1	1 1 - 1	1 1 - 1	1 1 - 1	1 1 - 1	1 1 - 1	- 1 - 1
HIGRADE-2	- - - -	- - - -	1 1 - 1	1 1 - 1	1 1 - 1	1 1 - 1	1 1 - 1	1 1 - 1	1 1 - 1	- 1 - 1
DESIQUE	0 - - 0	0 0 - 0	- - - -	1 - - 1	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -
BRISQUE	0 0 - 0	0 0 - 0	0 - - 0	- - - -	- - - -	- - - -	- - - -	- - - -	0 - - -	0 - - -
GM-LOG	0 0 - 0	0 0 - 0	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -	0 - - -
C-DIIVINE	0 0 - 0	0 0 - 0	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -	0 - - -
DIIVINE	0 0 - 0	0 0 - 0	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -	0 - - -
BLIINDS-II	0 0 - 0	0 0 - 0	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -	0 - - -
Curvelet	0 0 - 0	0 0 - 0	- - - -	1 - - -	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -
ContrastQA	- 0 - 0	- 0 - 0	- - - -	1 - - -	1 - - -	1 - - -	1 - - -	1 - - -	- - - -	- - - -

TABLE IX

RESULTS OF THE F-TEST PERFORMED ON THE RESIDUALS BETWEEN MODEL PREDICTIONS AND INDIVIDUAL QUALITY SCORES ON ESPL-LIVE HDR DATABASE. EACH CELL IN THE TABLE IS A CODEWORD CONSISTING OF 4 SYMBOLS THAT CORRESPOND TO "TONE MAPPING OPERATORS"', "MULTI-EXPOSURE FUSION", "POST PROCESSING", AND "OVERALL" PROCESSING ALGORITHMS. "1"("0") INDICATES THAT THE PERFORMANCE OF THE ROW IQA ALGORITHM IS SUPERIOR(INFERIOR) TO THAT OF THE COLUMN IQA ALGORITHM. - INDICATES THAT THE STATISTICAL PERFORMANCE OF THE ROW IQA IS EQUIVALENT TO THAT OF THE COLUMN IQA. THE MATRIX IS SYMMETRIC

	HIGRADE-1	HIGRADE-2	DESIQUE	BRISQUE	GM-LOG	C-DIIVINE	DIIVINE	BLIINDS-II	Curvelet	ContrastQA
HIGRADE-1	- - - -	- 0 - -	1 1 1 -	1 1 - -	1 1 - -	1 1 1 -	1 1 1 -	1 1 1 -	1 1 1 -	1 1 1 -
HIGRADE-2	- 1 - -	- - - -	1 1 - -	1 1 - 1	1 1 - 1	1 1 1 -	1 1 - 1	1 1 1 1	1 1 - 1	1 1 1 1
DESIQUE	0 0 0 -	0 0 - -	- - - -	1 1 - -	1 1 - -	1 1 - -	1 1 - -	1 1 - -	- 1 - -	0 1 - -
BRISQUE	0 0 - -	0 0 - 0	0 0 - -	- - - -	0 - - -	0 0 - -	0 0 - -	0 - - -	0 0 - -	0 0 - -
GM-LOG	0 0 - -	0 0 - 0	0 0 - -	1 - - -	- - - -	- - 1 -	1 0 - -	- - - -	0 - - -	0 - - -
C-DIIVINE	0 0 0 -	0 0 0 -	0 0 - -	1 1 - -	- - 0 -	- - - -	- - - -	- 1 - -	0 - - -	0 - - -
DIIVINE	0 0 0 -	0 0 - 0	0 0 - -	1 1 - -	0 1 - -	- - - -	- - - -	0 1 - -	0 - - -	0 - - -
BLIINDS-II	0 0 0 -	0 0 0 0	0 0 - -	1 - - -	- - - -	- 0 - -	1 0 - -	- - - -	0 0 - -	0 0 - -
Curvelet	0 0 0 -	0 0 - 0	- 0 - -	1 1 - -	1 - - -	1 - - -	1 - - -	1 1 - -	- - - -	0 - - -
ContrastQA	0 0 0 -	0 0 0 0	1 0 - -	1 1 - -	1 - - -	1 - - -	1 - - -	1 1 - -	1 - - -	- - - -

While we chose to keep the instructions that were given to the subjects simple and performance-directed, we did solicit feedback from them regarding the viewing distance and display type. Given the heterogeneity of display devices and viewing conditions in the crowdsourced experiment, it was not possible (nor desirable) to have the subjects evaluate images on calibrated displays or to measure the ambient lighting. As a complement to the current study, future experiments might target more knowledgeable pools of subjects, to obtain answers to detailed questions regarding the effects of viewing conditions (such as the ambient illumination), and the display settings (e.g., the model, age, resolution, and size of the monitor). Another variation might involve individual human subjects evaluating the same images on multiple displays to better account for how perception varies across displays. The results of the current study cannot answer these questions.

In the present study, only those AMT workers having AMT confidence values greater than 0.75 (on a [0,1] scale) were allowed to participate, and strict subject rejection criteria were imposed. In future crowdsourced studies, it may be useful to collect more background data, such as the average time spent by each subject on each image, the overall time spent by her rating all of the images, and so on.

VIII. CONCLUSION

We have described the new ESPL-LIVE HDR Image Quality Database of more than 300,000 human judgments garnered from more than 5,000 unique subjects. We outlined variable sources of HDR images, the algorithms used to process them and our crowdsourced subjective study framework, which allowed the images to be evaluated by thousands of observers over the Internet.

We also studied current NR-IQA algorithms in light of the collected subjective data as predictors of the perceptual quality of HDR-processed images. To the best of our knowledge, this is the largest and most comprehensive study of HDR-processed image quality conducted to date. It is our hope that the new database will prove to be a valuable resource, allowing researchers to develop improved IQA models of HDR processed images, and tonemapped HDR quality prediction algorithms that can be used for a variety of processing tasks, such as perceptually optimizing HDR processing algorithms for tonemapping and multi-exposure fusion (and for assessing the results).

REFERENCES

[1] H. Yeganeh and Z. Wang, "Objective quality assessment of tone-mapped images," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 657–667, Feb. 2013.

[2] G. Eilertsen, R. K. Mantiuk, and J. Unger, "A comparative review of tone-mapping algorithms for high dynamic range video," *Comput. Graph. Forum*, vol. 36, no. 2, pp. 565–592, 2017.

[3] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen, "Evaluation of tone mapping operators using a high dynamic range display," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 640–648, Jul. 2005. [Online]. Available: http://doi.acm.org/10.1145/1073204.1073242

[4] A. Yoshida, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "Analysis of reproducing real-world appearance on displays of varying dynamic range," *Comput. Graph. Forum*, vol. 25, no. 3, pp. 415–426, 2006.

[5] F. Drago, W. L. Martens, K. Myszkowski, and H.-P. Seidel, "Perceptual evaluation of tone mapping operators," in *Proc. ACM SIGGRAPH Sketches Amp Appl.*, New York, NY, USA, 2003, p. 1. [Online]. Available: http://doi.acm.org/10.1145/965400.965487

[6] M. Čadík, M. Wimmer, L. Neumann, and A. Artusi, "Evaluation of HDR tone mapping methods using essential perceptual attributes," *Comput. Graph.*, vol. 32, no. 3, pp. 330–349, Jun. 2008. [Online]. Available: https://www.cg.tuwien.ac.at/research/publications/2008/CADIK-2008-EHD/

[7] P. B. Delahunt, X. Zhang, and D. H. Brainard, "Perceptual image quality: Effects of tone characteristics," *J. Electron. Imag.*, vol. 14, no. 2, p. 023003, 2005.

[8] J. Kuang, R. Heckaman, and M. D. Fairchild, "Evaluation of HDR tone-mapping algorithms using a high-dynamic-range display to emulate real scenes," *J. Soc. Inf. Disp.*, vol. 18, no. 7, pp. 461–468, 2010.

[9] M Klíma *et al.*, "Deimos-an open source image database," *Radioengineering*, vol. 20, no. 4, pp. 1016–1023, 2011.

[10] I. Sprow, D. Kuepper, Z. Baranczuk, and P. Zolliker, "Image quality assessment using a high dynamic range display," in *Proc. AIC*, 2013, p. 307310.

[11] L. Krasula, M. Narwaria, K. Fliegel, and P. Le Callet, "Influence of HDR reference on observers preference in tone-mapped images evaluation," in *Proc. 7th Int. Workshop Quality Multimedia Exper. (QoMEX)*, Jul. 2015, pp. 1–6.

[12] M. Ashikhmin and J. Goyal, "A reality check for tone-mapping operators," *ACM Trans. Appl. Perception (TAP)*, vol. 3, no. 4, pp. 399–411, 2006.

[13] J. Petit and R. K. Mantiuk, "Assessment of video tone-mapping: Are cameras' S-shaped tone-curves good enough?" *J. Vis. Comun. Image Represent.*, vol. 24, no. 7, pp. 1020–1030, Oct. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.jvcir.2013.06.014

[14] G. Eilertsen, R. Wanat, R. K. Mantiuk, and J. Unger, "Evaluation of tone mapping operators for HDR-video," in *Comput. Graph. Forum*, vol. 32, no. 7, pp. 275–284, 2013.

[15] H. Z. Nafchi, A. Shahkolaei, R. F. Moghaddam, and M. Cheriet, "FSITM: A feature similarity index for tone-mapped images," *IEEE Signal Process. Lett.*, vol. 22, no. 8, pp. 1026–1029, Aug. 2015.

[16] H. R. Nasrinpour and N. D. Bruce, "Saliency weighted quality assessment of tone-mapped images," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 4947–4951.

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

[18] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, Nov. 2015.

[19] M. Narwaria, M. Perreira Da Silva, P. Le Callet, and R. Pepion, "Tone mapping-based high-dynamic-range image compression: Study of optimization criterion and perceptual quality," *Opt. Eng.*, vol. 52, no. 10, p. 102008, Oct. 2013.

[20] P. Hanhart, P. Korshunov, and T. Ebrahimi, "Crowdsourcing evaluation of high dynamic range image compression," in *Proc. SPIE Opt. Eng. Appl.*, 2014, p. 92170D.

[21] P. Hanhart, M. V. Bernardo, M. Pereira, A. M. G. Pinheiro, and T. Ebrahimi, "Benchmarking of objective quality metrics for HDR image quality assessment," *EURASIP J. Image Video Process.*, vol. 2015, no. 1, 2015, Art. no. 39. [Online]. Available: http://dx.doi.org/10.1186/s13640-015-0091-4

[22] C. Mantel, S. C. Ferchiu, and S. Forchhammer, "Comparing subjective and objective quality assessment of HDR images compressed with JPEG-XT," in *Proc. IEEE 16th Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2014, pp. 1–6.

[23] T. Richter, "On the standardization of the JPEG XT image compression," in *Proc. Picture Coding Symp. (PCS)*, Dec. 2013, pp. 37–40.

[24] M. Liu, G. Zhai, S. Tan, Z. Zhang, K. Gu, and X. Yang, "HDR2014—A high dynamic range image quality database," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2014, pp. 1–6.

[25] T. Hoßfeld *et al.*, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 541–558, Feb. 2014.

[26] M. Hirth, T. Hossfeld, M. Mellia, C. Schwartz, and F. Lehrieder, "Crowdsourced network measurements," *Comput. Netw.*, vol. 90, pp. 85–98, Oct. 2015. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2015.07.003

[27] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 3097–3100.

[28] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2016. [Online]. Available: http://dx.doi.org/10.1109/TIP.2015.2500021

[29] P. Korshunov, H. Nemoto, A. Skodras, and T. Ebrahimi, "Crowdsourcing-based evaluation of privacy in HDR images," in *Proc. SPIE Photon. Eur.*, 2014, pp. 913802-1–913802-11.

[30] M. D. Fairchild, "The HDR photographic survey," in *Proc. Color Imag. Conf.*, 2007, pp. 233–238.

[31] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 616–625, Oct. 2012.

[32] V. Hulusic, G. Valenzise, E. Provenzi, K. Debattista, and F. Dufaux, "Perceived dynamic range of HDR images," in *Proc. IEEE Int. Conf. Qual. Multimedia Exper.*, Jun. 2016, pp. 1–6.

[33] J. Hu, O. Gallo, K. Pulli, and X. Sun, "HDR deghosting: How to deal with saturation?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1163–1170.

[34] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," in *Proc. ACM SIGGRAPH*, 2002, pp. 257–266. [Online]. Available: http://doi.acm.org/10.1145/566570.566574

[35] R. Fattal, D. Lischinski, and M. Werman, "Gradient domain high dynamic range compression," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 249–256, Jul. 2002.

[36] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 267–276, Jul. 2002.

[37] G. W. Larson, H. Rushmeier, and C. Piatko, "A visibility matching tone reproduction operator for high dynamic range scenes," *IEEE Trans. Vis. Comput. Graph.*, vol. 3, no. 4, pp. 291–306, Oct. 1997.

[38] S. Paul, I. Sevcenco, and P. Agathoklis, "Multi-exposure and multi-focus image fusion in gradient domain," *J. Circuits, Syst. Comput.*, 2016. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/48782-multi-exposure-and-multi-focus-image-fusion-in-gradient-domain

[39] F. Pece and J. Kautz, "Bitmap movement detection: HDR for dynamic scenes," in *Proc. Conf. Vis. Media Product. (CVMP)*, 2010, pp. 1–8. [Online]. Available: http://nbn-resolving.de/urn:nbn:de:0009-6-36506

[40] S. Raman and S. Chaudhuri, "Bilateral filter based compositing for variable exposure photography," in *Eurographics-Short Papers*, P. Alliez and M. Magnor, Eds. The Eurographics Association, 2009.

[41] F. Banterle. *HDR Toolbox for Matlab*, accessed on Jan. 2016. [Online]. Available: https://github.com/banterle/HDR_Toolbox

[42] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proc. ACM SIGGRAPH*, New York, NY, USA, 1997, pp. 369–378. [Online]. Available: http://dx.doi.org/10.1145/258734.258884

[43] Mann, Picard, S. Mann, and R. W. Picard, "On being 'undigital,' with digit. Cameras: Extending dynamic range by combining differently exposed pictures," *Proc. IS&T*, 1995, pp. 442–448.

[44] S. Nayar and T. Mitsunaga, "High dynamic range imaging: Spatially varying pixel exposures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Jun. 2000, pp. 472–479.

[45] P. Ledda, L. P. Santos, and A. Chalmers, "A local model of eye adaptation for high dynamic range images," in *Proc. 3rd Int. Conf. Comput. Graph., Virtual Reality, Visualisation Interact. Africa*, 2004, pp. 151–160.

[46] R. Szeliski, *Computer Vision: Algorithms and Applications*, 1st ed. New York, NY, USA: Springer-Verlag, 2010.

[47] R. Mantiuk. (2015). *Hdr Image Gallery*. [Online]. Available: http://pfstools.sourceforge.net/hdr_gallery.html

[48] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R BT.500-13. [Online]. Available: https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-13-201201-I!!P% DF-E.pdf

[49] M. Kleiner, D. Brainard, D. Pelli, C. Broussard, T. Wolf, and D. Niehorster. *The Psychology Toolbox*. [Online]. Available: http://psychtoolbox.org/

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

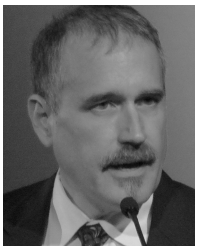
[51] *Amazon Mechanical Turk*, accessed on Jan. 2016. [Online]. Available: https://www.mturk.com

[52] *Microworkers*, accessed on Jan. 2016. [Online]. Available: https://microworkers.com/

[53] *Crowdflower*, accessed on Jan. 2016. [Online]. Available: https://crowdflower.com/
[54] T. Schulze, S. Seedorf, D. Geiger, N. Kaufmann, and M. Schader, "Exploring task properties in crowdsourcing—An empirical study on mechanical turk," in *Proc. 19th Eur. Conf. Inf. Syst. (ECIS)*, Helsinki, Finland, Jun. 2011, p. 122. [Online]. Available: http://aisel.aisnet.org/ecis2011/122
[55] *Inter-Rater Reliability—Wikipedia, the Free Encyclopedia*. [Online]. Available: https://en.wikipedia.org/wiki/Inter-rater_reliability
[56] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.
[57] Y. Zhang and D. M. Chandler, "No-reference image quality assessment based on log-derivative statistics of natural scenes," *J. Electron. Imag.*, vol. 22, no. 4, p. 043025, 2013.
[58] D. Kundu, "Subjective and objective quality evaluation of synthetic and high dynamic range images," Ph.D. dissertation, Dept. Elect. Comput. Eng. Univ. Texas, Austin, TX, USA, May 2016. [Online]. Available: http://users.ece.utexas.edu/ bevans/students/phd/debarati_kundu/
[59] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, Nov. 2014.
[60] L. Liu, H. Dong, H. Huang, and A. C. Bovik, "No-reference image quality assessment in curvelet domain," *Signal Process. Image Commun.*, vol. 29, no. 4, pp. 494–505, Apr. 2014.
[61] Y. Fang, K. Ma, Z. Wang, W. Lin, Z. Fang, and G. Zhai, "No-reference quality assessment of contrast-distorted images based on natural scene statistics," *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 838–842, Jul. 2015.
[62] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.
[63] Y. Zhang, A. K. Moorthy, D. M. Chandler, and A. C. Bovik, "C-DIIVINE: No-reference image quality assessment based on local magnitude and phase statistics of natural scenes," *Signal Process., Image Commun.*, vol. 29, no. 7, pp. 725–747, Aug. 2014.
[64] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.
[65] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, 2011. [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm
[66] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2012, pp. 1693–1697.

Debarati Kundu received the B. Eng. degree in electronics and telecommunications engineering from Jadavpur University, Kolkata, India, in 2010, and the M.Sc. and Ph.D. degrees the Department of Electrical and Computer Engineering, The University of Texas at Austin (UT Austin). She joined the Department of Electrical and Computer Engineering, UT Austin in 2010. She is currently a Senior Engineer with Qualcomm Research Bangalore, India. Her research interests include image and video quality assessment, computer graphics, computer vision, machine learning, and prototyping of real-time systems. She was a recipient of the RGM Advisors Research Award for best poster at Graduate and Industry Networking, UT Austin, in 2016, the Top 10% Paper Award at the IEEE International Conference on Image Processing in 2015, and the Qualcomm Roberto Padovani Fellowship 2014 awarded to the top 1% of the interns.

Deepti Ghadiyaram received the B.Tech. degree in computer science from the International Institute of Information Technology, Hyderabad, in 2009, and the M.S. degree from The University of Texas at Austin (UT Austin), in 2013. She is currently pursuing the Ph.D. degree with the Laboratory for Image and Video Engineering, UT Austin. Her research interests broadly include image and video processing, particularly perceptual image and video quality assessment, computer vision, and machine learning. She was a recipient of the Microelectronics and Computer Development Fellowship from 2013 to 2014 and the Graduate Student Fellowship offered to the top 1% of the students by the Department of Computer Science for the academic years 2013–2016.



Alan C. Bovik (F'96) is currently a Cockrell Family Regents Endowed Chair Professor with The University of Texas at Austin. His books include *The Handbook of Image and Video Processing*, *Modern Image Quality Assessment*, and The Essential Guides to Image and Video Processing. He is a fellow of the Optical Society of America and SPIE. He has received many major international awards, including the 2017 Edwin H. Land Medal from the Optical Society of America, the 2015 Primetime Emmy Award for Outstanding Achievement in Engineering Development from the Academy of Television Arts and Sciences, and the Society Award from the IEEE Signal Processing Society. He cofounded and was the longest serving Editor-in-Chief of the IEEE TRANSACTIONS ON IMAGE PROCESSING and created the IEEE International Conference on Image Processing in Austin, Texas, 1994.

Brian L. Evans (F'09) has authored more than 240 refereed conference and journal papers, and graduated 27 Ph.D. and 11 M.S. students. His research bridges digital signal processing theory and embedded real-time implementation. Applications include image/video acquisition/display and cellular/smart grid communications. He holds the Engineering Foundation Professorship at UT Austin. He received Top 10% Paper Awards at the 2012 IEEE Multimedia Signal Processing Workshop and the 2015 IEEE International Conference on Image Processing, and the Best Paper Award at the 2013 IEEE International Symposiam on Power Line Communications and Its Applications. He received three teaching awards at UT Austin, and a 1997 U.S. National Science Foundation CAREER Award.