CROWDSOURCED STUDY OF SUBJECTIVE IMAGE QUALITY

Deepti Ghadiyaram and Alan C. Bovik

Laboratory of Image and Video Engineering (LIVE) at The University of Texas at Austin

ABSTRACT

We designed and created a new image quality database that models diverse *authentic* image distortions and artifacts that affect images that are captured using modern mobile devices. We also designed and implemented a new online crowdsourcing system, which we are using to conduct a very large-scale, on-going, multi-month image quality assessment (IQA) subjective study, wherein a wide range of diverse observers record their judgments of image quality. Our database currently consists of over 320,000 opinion scores on 1,163 authentically distorted images evaluated by over 7000 human observers. The new database will soon be made freely available for download and we envision that the fruits of our efforts will provide researchers with a valuable tool to benchmark and improve the performance of objective IQA algorithms.

Index Terms— image quality, quality assessment, crowdsourcing, human study.

1. INTRODUCTION

The field of visual media is witnessing an explosive growth in recent years with significant advances in technology made by camera and mobile device manufacturers, and by the synergistic development of very large photo-centric social networking websites such as Pinterest and Instagram, which allow consumers to efficiently capture, store, and share high-resolution images with their friends or the community at large. By 2015, the estimated annual volume of photographs taken on mobile devices is expected to surpass 100 billion in the United States [1]. Every captured image passes through numerous processing stages, each of which could potentially introduce visual artifacts and compromise an end user's quality of experience. Moreover, the capture process is fraught with delicate variables such as lighting, exposure, aperture, noise sensitivity, lens limitations, and the unsure hands and eyes of many amateur photographers. Each of these factors could also potentially introduce annoying artifacts thereby perturbing an image's perceived visual quality. Thus, finding effective and efficient ways to identify and predict the perceptual quality of images is a pressing concern [1].

A major advance in modern image quality assessment has been the development of statistical models that capture the "naturalness" of images that are not distorted [1]. Pristine images, i.e., images with no apparent distortions, obey certain perceptually relevant statistical laws that are violated by the presence of common distortions. The state-of-the-art objective blind image quality assessment models [2, 3, 4] are designed to exploit these statistical perturbations, to accurately predict the perceptual quality of images.

On the other hand, given that the final receivers of these images are humans, the best way to understand and predict the effect of distortions on a typical person's viewing experience is to capture opinions from a large sample of human subjects. While these *subjective* scores are vital for understanding human perception of image quality, they are also crucial for designing and evaluating reliable IQA models that are consistent with subjective human evaluations, regardless of the type and severity of the distortions. Models that lead to IQA algorithms that produce quality predictions that correlate highly with mean opinion scores (MOS) obtained on images from subjective human studies, have the potential to motivate the design of solutions aimed at delivering maximum quality image content to customers across wired and wireless networks.

Collecting a large collection of subjective opinions is time-consuming and cumbersome. Nevertheless, several valuable image quality studies have been conducted that have supported the development of IQA algorithms in the past. The human opinion scores in most of these datasets were collected by conducting subjective studies in a fixed laboratory setup where images were displayed on a single device having a fixed display resolution which the subjects viewed from a fixed distance. In addition to this, almost all of these datasets suffer from one or more of the following problems: (1) a small database size, (2) a lack of diversity and realism of the distortions, (3) an insufficient number of subjective judgments, (4) a limited variability of the image content, (5) limited or no public availability of the database, and (6) a lack of fine-grained, continuous scale ratings.

These limitations motivated us to design and create a new image quality database that models authentic distortions captured using a wide variety of commercial devices and which includes highly diverse and genuine artifacts. By contrast with most databases where the distorted images are *derived*

Fig. 1. Sample images from the LIVE Blind Image Quality Challenge Database. These images include pictures of faces, people, animals, close-up shots, wide-angle shots, nature scenes, man-made objects, images with distinct foreground/background configurations, and images without any specific object of interest

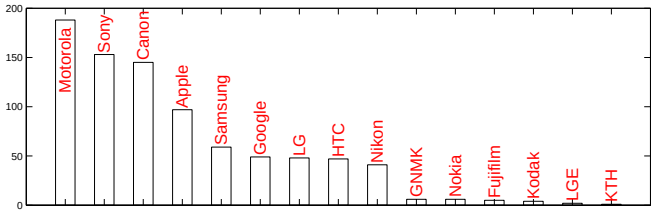

Fig. 2. Distribution of different manufacturers of the cameras that were used to capture a sample of images contained in our database.

from a set of high-quality source images by simulating image impairments [5, 6], we chose to gather *naturally distorted* images representing a broad range of diversity of quality "types," mixtures, and distortion severities. In other words, each image was collected without artificially introducing any distortions beyond those occurring during the capture, processing, and storage processes in each user's device. Here, we summarize the content and characteristics of the resulting database which we have dubbed the **LIVE Blind Authentic Image Quality Challenge Database**. It consists of 1,163 images afflicted by varied artifacts such as low-light noise and blur, motion-induced blur, over and underexposure, compression errors, and so on. We also describe a new online crowdsourcing system for collecting subjective quality assessment scores that we designed and implemented using Amazon's Mechanical Turk (AMT) [7], which we are using to conduct a very large-scale, on-going, multi-month IQA subjective study. We then present several critical factors involved in crowdsourcing IQA such as the design of the online study, subject rejection, task remuneration, and so on.

To the best of our knowledge, we are aware of only one other project [8] reporting efforts made in the same spirit as our work that we report here. However, the authors of [8] tested their crowdsourcing system only on 116 JPEG compressed images from the legacy LIVE Image Quality Database [6] and gathered opinion scores from only forty subjects. By contrast, we have so far collected over 320,000 human opinion scores on 1,163 naturally distorted images from over 7,000 distinct subjects and we plan to collect more than 350,000 subjective judgments overall, making it the world's largest, most comprehensive study of perceptual image quality ever conducted.

2. DETAILS OF THE SUBJECTIVE STUDY

2.1. LIVE Blind Authentic Image Quality Challenge Database

Figure 1 shows a subset of images used in the study. All of the images in the database were captured using different digital cameras including mobile devices as presented in Fig. 2. These images include pictures of faces, people, animals, close-up shots, wide-angle shots, nature scenes, man-made objects, images with distinct foreground/background configurations, and images without any specific object of interest. Some images contain high luminance and/or color activity, while some are mostly smooth. Since these images are naturally distorted as opposed to being artificially calculated post-acquisition from pristine reference images, they often contain mixtures of multiple image distortions that can occur in real-world applications and reflect a broad range of image impairments.

2.2. Subjective Test Methodology

Crowdsourcing systems like Amazon Mechanical Turk (AMT) [7] have emerged as effective human-powered platforms making it feasible to gather a large number of opinions from a diverse distributed populace over the web. On AMT, "requesters" broadcast their task to a selected pool of registered "workers" in the form of an open call for data collection. Workers who select the task are motivated primarily by the monetary compensation offered by the requesters and also by the enjoyment they experience through participation.

2.2.1. Instructions, Training, and Testing

Crowdsourcing has been extensively explored in several object identification tasks [9] to gather segmented objects and their labels. However, the task of labelling objects is often more clearly defined and fairly straightforward to perform, in contrast to our challenging, highly subtle and *subjective* task of gathering opinion scores on the perceived quality of the presented images. The varied level of experience of the workers with respect to understanding the concept of image quality and their geographical diversity made it extremely important that detailed instructions be provided to assist them in

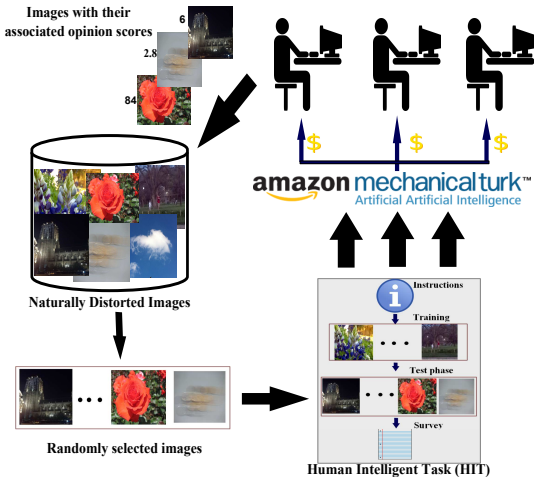

Fig. 3. Illustrating how our system packages the task of rating images as a HIT and disperses it on Mechanical Turk.

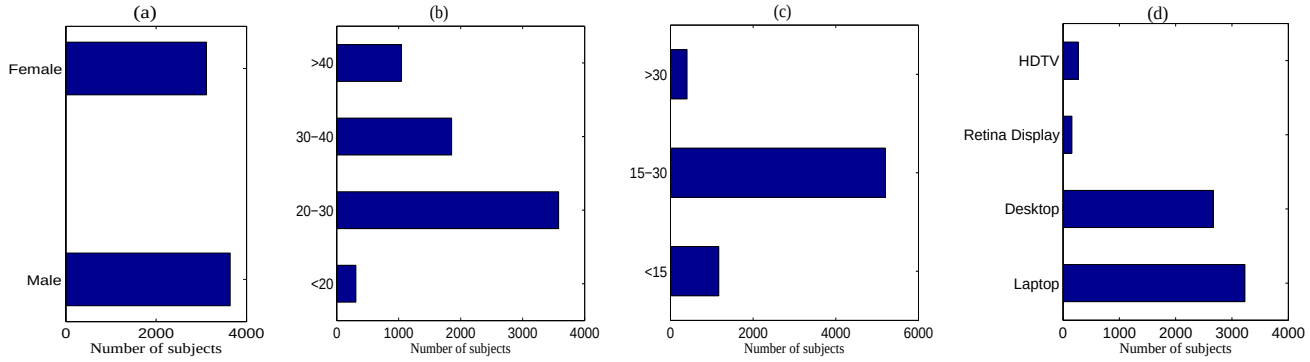

Fig. 4. Demographics of the participants so far (a) gender (b) age (c) approximate distance between the subject and the viewing screen (d) different categories of display devices used by the workers to participate in the study.

understanding how to undertake the task without biasing their perceptual scores. Thus, every unique participating subject on AMT that selects our **HIT** (**H**uman **I**ntelligent **T**ask) is first provided with detailed instructions.

After reading the instructions, if a worker accepts the task, a rating interface is presented that contains a slider by which opinion scores can be interactively provided. We adopted a single stimulus continuous procedure [10] to obtain quality ratings on images where subjects report their quality judgments by dragging the slider located below the image on the rating interface. This continuous rating bar is divided into five equal portions, which are labelled "bad," "poor," "fair," "good," and "excellent". After the subject moves the slider to rate an image and presses the "Next Image" button, the position of the slider is converted to an integer quality score in the range $1 - 100$, and then the next image is presented.

Before the actual study begins, each participant is first presented with a fixed set of 7 training images that were selected by us as being reasonably representative of the approximate range of image qualities and distortion types that might be encountered during the study. We call this the **training phase**. This phase is to help a worker get adjusted to the rating process and the task at hand and thus, we do not include the ratings obtained in the training phase in our database. Next, in the **testing phase**, the subject is presented with 43 images in a random order where the randomization is different for each subject. This is followed by a quick survey session which involves the subject answering a few questions. Each HIT involves rating a total of 50 images and the subject receives a remuneration of 30 cents (if she is not rejected, as discussed in Section 2.2.3) for the task. Fig. 3 illustrates how we package the task of rating images as a HIT and effectively disperse it online via AMT to gather thousands of human opinion scores.

2.2.2. Subjects

Figures 4 (a) and (b) illustrate the demographic details of a random sample of the subjects. Most of them reported in the final survey that they are inexperienced with image quality assessment but do get annoyed by image impairments they come across on the Internet. As we couldn't test the subjects for vision problems, in the instructions, we requested them to wear corrective lenses during the study if they do so in their

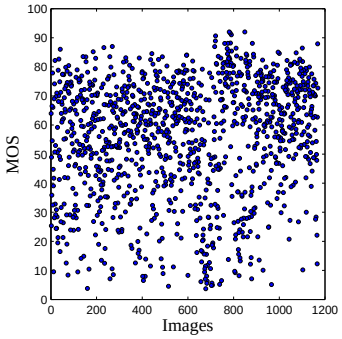

Fig. 5. Scatter plot of the MOS scores obtained so far on all the images in the database.

day-to-day life. Later in the survey, the subjects are asked if they usually wear corrective lenses and whether they wore the lenses while participating in the study. The ratings given by those subjects who were not wearing their corrective lenses they were otherwise supposed to wear are rejected. Figures 4 (c) and (d) illustrate the distribution of the broad classes of different display devices and the distances from which workers (thus far) have viewed the images indicating the diverse testing conditions that exist during the study.

2.2.3. Subject Rejection Techniques

Crowdsourcing has empowered us to efficiently collect large amounts of ratings. However, it raises interesting issues such as how to deal with noisy ratings and address the reliability of the AMT workers.

1. **Intrinsic metric**: To reduce the noise in our ratings, only those workers on AMT having confidence values greater than 75% are allowed to select our task. Also, in order to not bias the ratings due to a single worker picking up our HIT multiple times, we impose a restriction that each worker can select our task only once.
2. **Gold standard data**: 5 of each group of 43 test images are fixed across all the HITs and are drawn from the LIVE Multiply Distorted Image Quality Database [11] to supply a control. These images along with their corresponding MOS from the database are treated as a gold standard. We then reject a subject when at least three of their five ratings on these gold standard images differ by more than a threshold from the corresponding gold standard opinion scores.
   The mean of the correlation values computed between the MOS obtained from the workers on the gold standard images and the corresponding ground truth MOS from the database [11] was found to be **0.985**. This high value indicates a high degree of reliability of the scores that are being collected via Mechanical Turk, reaffirming the efficacy of our approach of gathering opinion scores.
3. **Repeated images**: 5 of the remaining 38 test images are presented twice randomly to each subject in the testing phase. If the difference between the two ratings that a subject provides to the same image each time it is presented exceeds another threshold on at least 3 of the 5 images, then that subject is rejected.

Table 1. Median lcc and Median srocc across 100 train-test combinations on the live challenge database and on the LIVE IQA database (indicated in *italics*)

	LCC	SROCC	LCC	SROCC
FRIQUEE [12]	**0.67**	**0.64**	*0.96*	*0.95*
BRISQUE [2]	0.56	0.53	*0.94*	*0.94*
DIIVINE [3]	0.50	0.48	*0.93*	*0.92*
BLIINDS-II [4]	0.45	0.40	*0.92*	*0.91*

The study is still on-going and the database currently comprises a total of about 320,000 ratings obtained from more than 7,000 unique subjects. The MOS values after subject rejection are computed for each image by averaging the individual opinion scores from multiple workers. MOS is representative of the perceived viewing experience of each image. The MOS values observed to date have ranged between $[3.71 - 92.02]$. Figure 5 depicts a scatter plot of the MOS computed from the individual scores we have collected thus far.

2.3. Performance of Objective IQA Algorithms

We also evaluated the performance of a few leading blind (no-reference) IQA algorithms in regards to their ability to reliably predict the visual quality of the images in our growing database. Specifically, Table 1 presents the median across 100 disjoint train and test splits of linear correlation coefficient (LCC) and Spearman Rank Ordered Correlation Coefficient (SROCC) of a few blind IQA algorithms (whose code was publicly available) on the new LIVE Blind Authentic Image Quality Challenge Database. To further highlight the challenges that the authentic distortions present in our database pose to the top-performing algorithms, we also present the median correlation values when the algorithms are tested on the standard benchmark database [6]. It can be observed that all of the top-performing models, when trained and tested on the legacy LIVE IQA database which comprises of singly distorted images, perform remarkably well when compared to their performance on our difficult database of mixtures of distortions. The new model FRIQUEE [12] that is designed by drawing insights from the scene statistics of authentic distortions, combines a bag of perceptually relevant features with a deep belief net and yields better performance than the state-of-the-art models. This indicates that the existing blind IQA algorithms have significant room for improvement towards being able to accurately predict the quality of images suffering from diverse uncontrolled real world distortions. We hope that this database would encourage the quality assessment community to design robust learning engines that would

push the boundaries of achievable prediction power on authentically distorted images.

3. FUTURE WORK

With an end goal to collect more than 350,000 subjective judgments overall, we believe that our study is the world's largest, most comprehensive online study of perceptual image quality ever conducted. Of course, digital videos (moving pictures) are also being captured with increasing frequency by both professional and casual users. In the increasingly mobile environment, these spatial-temporal signals will be subject to an even larger variety of distortions [1] arising from a multiplicity of natural and artifical processes [13]. Predicting, monitoring, and controlling the perceptual effects of these distortions will require the development of powerful blind video quality assessment models, such as [14], and new VQA databases representative of human opinions of modern, authentic videos captured by current mobile video camera devices and exhibiting contemporary distortions. Current legacy VQA databases, such as [15, 16] are useful tools but are limited in regard to content diversity, number of subjects, and distortion realism and variability. Therefore, we plan to conduct large-scale crowdsourced video quality studies in the future, mirroring the effort described here, and building on our expertise in conducting the current study.

4. REFERENCES

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

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

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

[4] M.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.

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

[6] 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.

[7] "Amazon Mechanical Turk," [Online]. Available: http://mturk.com.

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

[9] B. Russell, A. Torralba, K. Murphy, and W.T. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Computer Vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[10] M.H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, 2004.

[11] D. Jayaraman, A. Mittal, A.K. Moorthy, and A.C. Bovik, "Objective quality assessment of multiply distorted images," *Asilomar Conf. Signals, Syst. Comput.*, pp. 1693–1697, Nov 2012.

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

[13] A. C. Bovik, *The Essential Guide to Video Processing*, Elsevier Academic Press, 2nd edition, 2009.

[14] M. Saad, A.C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, March 2014.

[15] A.K. Moorthy, K. Seshadrinathan, R. Soundararajan, and A.C. Bovik, "Wireless video quality assessment: A study of subjective scores and objective algorithms," *IEEE Trans. Circ. Syst. Video Technol.*, vol. 20, no. 4, pp. 587–599, April 2010.

[16] 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, June 2010.