QUALITY ASSESSMENT OF THUMBNAIL AND BILLBOARD IMAGES ON MOBILE DEVICES

Zeina Sinno[1], Anush Moorthy[2], Jan De Cock[2], Zhi Li[2], Alan Bovik[3]

ABSTRACT

Objective image quality assessment (IQA) research entails developing algorithms that predict human judgments of picture quality. Validating performance entails evaluating algorithms under conditions similar to where they are deployed. Hence, creating image quality databases representative of target use cases is an important endeavor. Here we present a database that relates to quality assessment of billboard images commonly displayed on mobile devices. Billboard images are a subset of thumbnail images, that extend across a display screen, representing things like album covers, banners, or frames or artwork. We conducted a subjective study of the quality of billboard images distorted by processes like compression, scaling and chroma-subsampling, and compared high-performance quality prediction models on the images and subjective data.

Index Terms— Subjective Study, Image Quality Assessment, User Interface, Mobile Devices.

I. INTRODUCTION

Over the past few decades, many researchers have worked on the design of algorithms that automatically assess the quality of images in agreement with human quality judgments. These algorithms are typically classified based on the amount of information that the algorithm has access to. Full reference image quality assessment algorithms (FR IQA), compare a distorted version of a picture to its pristine reference; while no-reference algorithms evaluate the quality of the distorted image without the need for such comparison. Reduced reference algorithms use additional, but incomplete side-channel information regarding the source picture [1].

Objective algorithms in each of the above categories are validated by comparing the computed quality predictions against ground truth human scores, which are obtained by conducting subjective studies. Notable subjective databases include the LIVE IQA database [2], [3], the TID 2013 databases [4], the CSIQ database [5], and the recent LIVE Challenge database [6]. The distortions studied in many databases are mainly JPEG and JPEG 2000 compression, linear blur, simulated packet-loss, noise of several variants (white noise, impulse noise, quantization noise etc.), as well as visual aberrations [6], such as contrast changes. The LIVE Challenge database is much larger and less restrictive, but is limited to no-reference studies. While these databases are wide-ranging in content and distortion, they fail to cover a use case that is of importance thumbnail-sized images viewed on mobile devices. Here we attempt to fill this gap for the case of billboard images.

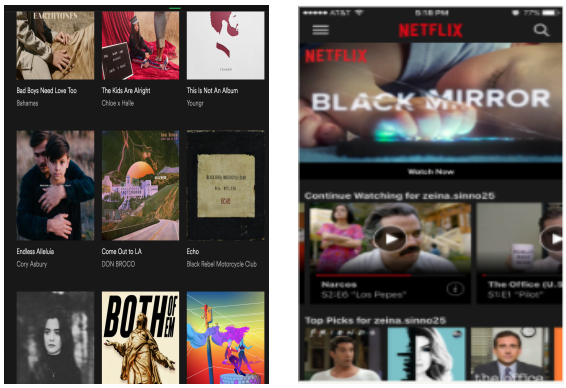
We consider the use case of a visual interface for an audio/video streaming service displayed on small-screen devices such as a mobile phone. In this use case, the user is presented with many content options, typically represented by thumbnail images representing the art associated with the selection in question. For instance, on a music streaming site, such art could correspond to images of album covers. Such visual representations are appealing and are a standard form of representation across multiple platforms. Typically, such images are stored on servers and are transmitted to the client when the application is accessed.

While these kinds of billboard images can be used to create very appealing visual interfaces, it is often the case that multiple images must be transmitted to the end-user to populate the screen as quickly as possible. Further, billboard images often contain stylized text, which needs to be rendered in a manner that is true to the artistic intent should allow for easy reading even on the smallest of screens. These two goals imply conflicting requirements: the need to compress the image as much as possible to enable rapid transmission and a dense population, while also delivering high quality, high resolution versions of the images.

Fig. 1(a) shows a representative interface of such a service [7], where a dense population is seen. Such interfaces may be present elsewhere as well, as depicted in Fig. 1(b), which shows the display page of a video streaming service provider [8], which includes images in thumbnail and billboard formats on the interface. These images have unique characteristics such as text and graphics overlays, as well as the presence of gradients, which are rarely encountered in existing picture quality databases that are used to train and validate image quality assessment algorithms. Images in such traditional IQA databases consist of mostly natural

1: The author is at the Laboratory for Image and Video Engineering (LIVE) at the University of Texas at Austin and was working for Netflix when this work was performed (email: zeina@utexas.edu).
2: The author is at Netflix.
3: The author is at the Laboratory for Image and Video Engineering (LIVE) at the University of Texas at Austin.

images, as seen in Fig. 2, which differ in important ways with the content seen in Figs. 1(a) and (b).



(a) (b)

Fig. 1. Screenshots of mobile audio and video streaming services (a) Spotify and (b) Netflix.

To address the use case depicted in the interfaces in Figs. 1(a) and (b), we conducted an extensive subjective study to capture the perceptual quality of such content. The study detailed here targets billboard images, which are thumbnail images arranged in a landscape orientation, spanning the width of the screen, that may be overlayed with text, gradients and other graphic components. The test was designed to replicate the viewing conditions of users interacting with an audio/video streaming service. The source images that we used are all billboard artwork at Netflix [8]. The distortions considered represent the most common kinds of distortions that billboard images are subjected to when presented by a streaming service.

Specifically we describe the components of the new database, and how the study was conducted, and we examine how several leading IQA models perform.

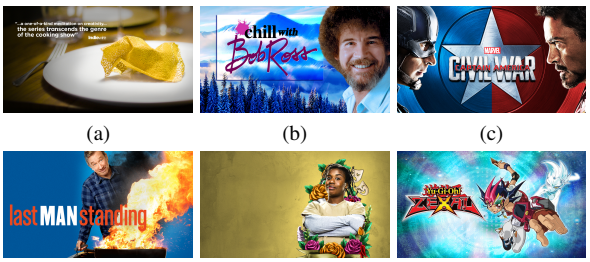
Fig. 2. Sample images from the LIVE IQA database [2], [3], which consists only of natural images.

II. DETAILS OF THE EXPERIMENT

In this section we detail the subjective study that we conducted on billboard images viewed on mobile devices. We first describe the content selection, followed by a description of the distortions studied, and finally detail the test methodology.

II-A. Source Image Content

The source content of the database was selected to represent typical billboard images viewed on audio and video streaming services. Twenty-one high resolution contents were selected to span a wide range of visual categories such as animated content, contents with (blended) graphic overlays, faces, and to include semi-transparent gradient overlays. A few contents were selected containing the Netflix logo, since it has some saturated red which exhibit artifacts when subjected to chroma subsampling. Fig. 3 shows some of the content used in the study.



(a) (b) (c)

(d) (e) (f)

Fig. 3. Several samples of content in the new database.

The images, which were all originally of resolutions 2048×1152, or 2560×1440, were downscaled to 1080×608, so their width matched the pixel density of typical mobile displays (when held in portrait mode). These 1080×608 images form the reference set.

II-B. Distortion Types

The reference images were then subjected to a combination of distortions using [9], that are typically encountered in streaming app scenarios:

- Chroma subsampling: 4:4:4 and 4:2:0.
- Spatial subsampling: by factors of 1, (no subsampling), $\sqrt{2}$, 2, and 4.
- JPEG compression: using quality factors 100, 78, 56, and 34 in the compressor.

All images were displayed at 1080×608 resolution, hence all of the downsampled images were upsampled back to the reference resolution as would be displayed on the device. We applied the following imagemagick [9] commands to obtain the resulting images:

- magick *SrcImg.jpg* -resize 1080x608 *RefImg.jpg*
- magick *RefImg.jpg* -resize *hxw* -compress JPEG -quality f_q -sampling-factor f_c *DistortedImg.jpg*
- magick *DistortedImg.jpg* -resize 1080x608 *DistortedImg.jpg*

where *SrcImg.jpg*, *RefImg.jpg*, and *DistortedImg.jpg* are source, reference and the distorted images respectively, h and w are the dimensions to which the images were downsampled, defined as $w = \frac{1080}{f_s}$ and $h = \frac{608}{f_s}$ where f_s is the imagemagick spatial subsampling factor, f_c is the chroma subsampling factor and f_q is the quality factor.

The above combinations resulted in 32 types of multiply distorted images (2 chroma subsamplings $\times$ 4 spatial downsamplings $\times$ 4 quality levels), yielding a total of 672 distorted images (21 contents $\times$ 32 distortion combinations). The distortion parameters were chosen so that distortion severities were perceptually separable, varying from nearly imperceptible to severe.

II-C. Test Methodology

A double-stimulus study was conducted to gauge human opinions on the distorted image samples. Our goal was that even very subtle artifacts be judged, to ensure that picture quality algorithms could be trained and evaluated in regards to their ability to distinguish even minor distortions. After all, each billboard image might be viewed multiple times by many millions of viewers. Further, we attempt to mimic the interface of a generic streaming application, whereby the billboard would span the width of the display when held horizontally (landscape mode). Hence, we presented the double stimulus images in the way depicted in Fig. 4. In each presentation, one of the images is the reference, while the other is a distorted version of the same content. The subjects were not told which image was the reference. Instead they were asked to rate which image was better than the other, and by how much (on a continuous rating bar) as depicted in Fig. 4. Further presentation order was randomized so that (top/bottom) reference-distorted and distorted-reference pairs were equally likely. Finally, the distorted content order was randomized, but so that the same content was never presented consecutively (to reduce memory effects), and the content ordering was randomized across users.

Equipment and Display

The subjective study was conducted on two LG Nexus 5 mobile devices, so that two subjects could concurrently participate in the study. Auto brightness was disabled and the two devices were calibrated to have the same mid-level of brightness, to ensure that all the subjects had the same viewing conditions. LG Nexus 5 devices are Android based and have a display resolution of 1920×1080, which is ubiquitous on mobile devices. They were mounted in portrait orientation on a stand [10], and the users were positioned at a distance of three screen widths as in [11]. Although the users were told to try to maintain their viewing distance, they were allowed to adjust their position if they felt the need to do so.

Fig. 4. Screenshot of the mobile interface used in the study.

Human Subjects, Training and Testing

The recruited participants were mostly Netflix employees, spanning different departments: engineering, marketing, languages, statistics etc. A total of 122 participants took part in the study. The majority did not have knowledge of image processing of the image quality assessment problem. We did not test the subjects for vision problems, but a verbal confirmation of (corrected-to-) normal vision was obtained.

Each subject participated in one session, lasting about thirty minutes. The subjects were provided with written instructions explaining the task. Next, each subject was given a demonstration of the experimental procedure. During a session, the subject viewed a random subset of eight different contents (from amongst the 21). Thirteen different distortion levels were randomly selected for each session, and each session also included a reference-reference control pair so that a total of fourteen distorted-reference pairs were displayed for the subject to rate. A short training session preceeded the actual study, where six reference-distorted pairs of images were presented to each subject where distorted images approximately spanned the entire range of picture qualities. The images shown in the training sessions were different from those used in the actual experiment.

Thus, each subject viewed a total of 118 pairs [8 contents $\times$ (13 distortion levels + 1 reference pair) + 6 training pairs] in each session. A black screen was displayed for three seconds between successive pair presentations. Finally, at the halfway point, a screen was shown to the subjects indicating that they had completed half of the test, and that they could take a break if they chose to.

II-D. Processing of the results

The position of the slider after the subject ranked the image was converted to a scalar value between 0 and 100, where 0 represents the worst quality and 100 represents the highest possible quality. The reference images were

anchors, having an assumed score of 100. We found that the subject scores on the reference-reference pairs follow a binary distribution with $p \approx 0.5$, indicating that no bias could be attributed to the location (top/bottom) of the images as they were being evaluated.

We used the guidelines of BT. 500-13 (Annex 2, section 2.3) [11] for subject rejection. Two subjects were outliers, hence their scores were removed for all further analysis. Finally, we computed the mean opinion scores (MOS) of the images.

III. ANALYSIS OF THE RESULTS

First, we study the individual effects of the three considered distortion types (JPEG compression, spatial subsampling, and chroma subsampling) on reported quality.

III-A. Impact of the individual distortions

Figure 5(a) plots the distribution of the MOS while varying the chroma subsampling between 4:4:4 and 4:2:0 at a fixed JPEG quality factor $f_q = 56$ and a fixed spatial subsampling factor $f_s = 1$. Chroma subsampling is very difficult to perceive in the quality range 80-95, but easier to perceive (in the quality range 60-75). Confusion often occurs when there are conspicuous red objects, e.g, in Fig. 3(c) and (d), owing to the heavy subsampling on the red channel. Figure 5(b) plots the distribution of MOS against the JPEG quality factor f_q for fixed chroma subsampling (4:4:4) and fixed spatial subsampling factor $f_s = 2$. The plots shown that these values of f_q produce clear perceptual separation, except for $f_q \in \{78, 100\}$ and $f_q \in \{34, 56\}$, which is expected at the quality extremes. Figs 3(b) and (f) show examples of such content.

Finally, Fig. 5(c) plots MOS against the spatial subsampling factor f_s, for fixed chroma subsampling (4:4:4) and quality factors $f_q = 76$. The downsampling artifacts are clearly perceptually separable.

III-B. Objective Algorithm Performance

We also evaluated the performances of several leading objective picture quality predictors on the collected database. We computed Pearson's Linear Correlation Coefficient (LCC) and Spearman's rank correlation coefficient (SROCC) to assess the following IQA model performances on the new dataset: Peak-Signal to-Noise Ratio (PSNR), PSNR-HVS [12], Structural Similarity Index (SSIM) [3], Multi-Scale Structural Similarity Index (MS-SSIM) [13], Visual Information Fidelity (VIF) [14], Additive Distortion Metric (ADM) [15], and the Video Multi-method Assessment Fusion (VMAF) version 1.2.0 [16]. LCC was computed after applying a non-linear map as prescribed in [17]. We also computed the root mean squared error (RMSE) between the obtained objective scores after mapping the subjective scores.

All of the models were computed on the luma channel only, at the source resolution of the reference images. The results are tabulated in Table I. As may be seen, VMAF and ADM were the top performers.

Given its good performance, we decided to analyze VMAF, towards better understanding when it fails or when it could be improved. Figure 6 plots the VMAF scores against the MOS. While the scatter plot is nicely linear, as also reflected by the correlation numbers, of interest are outlying failures that deviated the most from the general linear trend. Based on our analysis of these we can make the following observations:

1) VMAF tends to overestimate quality when chroma subsampling distortions are present. This is not unexpected since the VMAF is computed only on the luminance channel. This suggests that using features computed on the chroma signal when trying and applying VMAF may be advisable.
2) VMAF tends to underestimate quality on certain contents containing gradient overlays when the dominant distortion is spatial subsampling. This may occur because of banding (false contouring) effects that occur on the gradients.
3) Likewise, on contents with both gradient overlays and heavy compression artifacts, VMAF poorly predicts the picture quality, perhaps for similar reasons.

IV. CONCLUSION

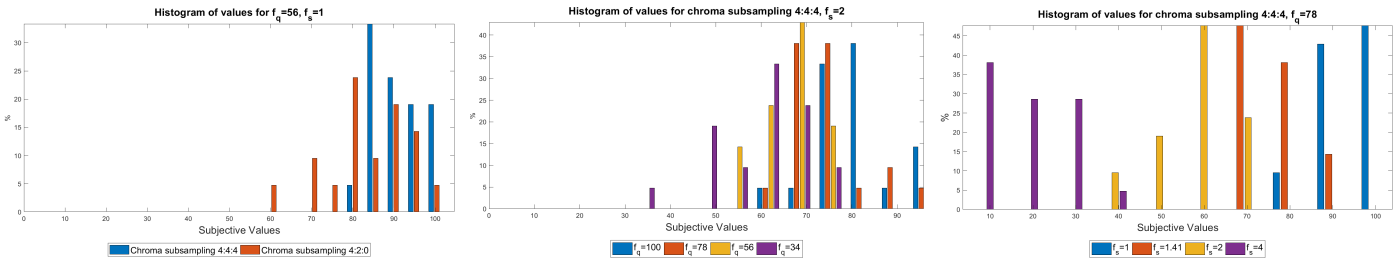
We have taken a first step towards understanding the quality assessment of thumbnail images on mobile devices, like those supplied by audio/video streaming services. We conducted a subjective study where viewers rated the subjective quality of billboard-style thumbnail images that were impaired by chroma subsampling, compression and spatial subsampling artifacts. We observed that spatial subsampling, followed by compression, were the main factors affecting perceptual quality. There are other types of interesting signals, such as boxshot images, which are displayed in a portrait mode at lower resolutions, usually with a dense text component. These characteristics can render both compression and objective quality prediction more challenging. Broader studies encompassing these and other kinds of thumbnail images could prove fruitful for advancing this new subfield of picture quality research.

V. ACKNOWLEDGMENT

The authors would like to thank Dr. Ioannis Katsavounidis and other members of the Video Algorithms team at Netflix for their valuable feedback in designing the study.

VI. REFERENCES

[1] L. He, F. Gao, W. Hou, and L. Hao, "Objective image quality assessment: a survey," *Intern. J. of Comp. Math.*, vol. 91, no. 11, pp. 2374–2388, 2014.



(a) Histogram of MOS plotted against chroma subsampling.

(b) Histogram of MOS plotted against JPEG quality factor f_q.

(c) Histogram of MOS plotted against spatial subsampling factor f_s.

Fig. 5. Distribution of the MOS as a function of different distortions

Table I. PLCC and SROCC results computed between the subjective scores and predicted objective scores

	PSNR	PSNR HVS [12]	SSIM [3]	MS SSIM [13]	VIF [14]	ADM [15]	VMAF [16]
PLCC	0.72	0.86	0.82	0.83	0.88	0.94	0.95
SROCC	0.71	0.85	0.81	0.83	0.86	0.93	0.94
RMSE	20.79	15.42	17.31	16.51	14.31	10.09	9.19

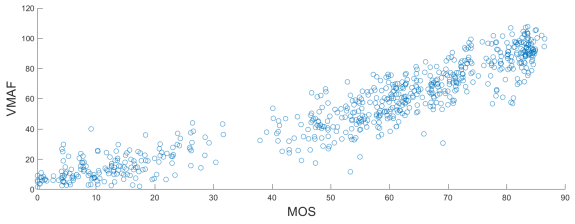


Fig. 6. Scatter plot of predicted VMAF scores against MOS.

[2] H. Sheikh, C. L. Wang, Z., and A. C. Bovik, "Live image quality assesment database release 2," http://live.ece.utexas.edu/research/quality/subjective.htm, accessed on Jan. 8 2018.

[3] Z. Wang, A. C. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. on Imag. Proc.*, vol. 13, no. 4, pp. 600–612, 2004.

[4] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, "Color image database TID2013: Peculiarities and preliminary results," in *Visual Information Processing (EUVIP), 2013 4th European Workshop on*. IEEE, 2013, pp. 106–111.

[5] E. Larson and D. Chandler, "Categorical image quality CSIQ database 2009." [Online]. Available: http://vision. okstate. edu/csiq

[6] 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, 2016.

[7] "Spotify," https://www.spotify.com/, accessed on Jan. 8 2018.

[8] "Netflix," https://www.netflix.com, accessed on Jan. 8 2018.

[9] "Imagemagick," https://www.imagemagick.org/, accessed on Jan. 8 2018.

[10] "Cell phone stand, lamicall s1 dock : Cradle, holder, stand for switch, all android smartphone, iphone 6 6s 7 8 x plus 5 5s 5c charging, accessories desk - black," http://a.co/bDukbam, accessed on Jan. 8 2018.

[11] "Methodology for the subjective assessment of the quality of television pictures." ITU-R Rec. BT. 500-13, 2012.

[12] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on hvs," in *Proc. of the Sec. Int. Worksh. Vid. Process. Qual. Met.*, vol. 4, 2006.

[13] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Asilomar Conf. Sign., Sys. Comp.*, vol. 2. IEEE, 2003, pp. 1398–1402.

[14] H. R. Sheikh, A. C. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on image processing*, vol. 14, no. 12, pp. 2117–2128, 2005.

[15] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Trans. Mult.*, vol. 13, no. 5, pp. 935–949, 2011.

[16] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," Jun 2016, accessed on Jan. 8 2018. [Online]. Available: https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652

[17] "Final report from the video quality experts group on the validation of objective models of video quality assessment." Video Quality Expert Group (VQEG), 2000.