OBJECTIVE QUALITY ASSESSMENT OF MULTIPLY DISTORTED IMAGES

Dinesh Jayaraman, Anish Mittal, Anush K. Moorthy and Alan C. Bovik

Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX, USA

ABSTRACT

Subjective studies have been conducted in the past to obtain human judgments of visual quality on distorted images in order, among other things, to benchmark objective image quality assessment (IQA) algorithms. Existing subjective studies primarily have records of human ratings on images that were corrupted by only one of many possible distortions. However, the majority of images that are available for consumption are corrupted by multiple distortions. Towards broadening the corpora of records of human responses to visual distortions, we recently conducted a study on two types of multiply distorted images to obtain human judgments of the visual quality of such images. Further, we compared the performance of several existing objective image quality measures on the new database and analyze the effects of multiple distortions on commonly used quality-determinant features and on human ratings.

Index Terms— Image quality assessment, Subjective study, Multiple distortions

1. INTRODUCTION

With the explosion of camera usage, visual traffic over networks and increasing efforts to improve bandwidth usage, it is important to be able to monitor the quality of visual content that may be corrupted by multiple distortions. There is an increasing demand to develop quality assessment algorithms which can supervise this vast amount of data and ensure its perceptually optimized delivery.

The performance of existing objective IQA models such as [1, 2, 3, 4] has previously been measured on databases such as [5, 6] containing images corrupted by one of several possible distortions. However, images available to consumers usually reaches them after several stages - acquisition, compression, transmission and reception, in which process, they may suffer multiple distortions. Hence it is important to study the performance of IQA algorithms applied to multi-distorted images. The relatively few existing studies [7][8] of image quality in multiple distortion scenarios have used databases that are limited by the content and size of the data used. To bridge this gap, we have conducted a large study involving 37 subjects and 8880 human judgments on 15 pristine reference images and 405 multiply distorted images of two types.

This work was supported under NSF grant number IIS-1116656 and by Intel and Cisco under the VAWN program.

2. DETAILS OF THE EXPERIMENT

A subjective study was conducted in two parts to obtain human judgements on images corrupted under two multiple distortion scenarios: 1) image storage where images are first blurred and then compressed by a JPEG encoder. 2) camera image acquisition process where images are first blurred due to narrow depth of field or other defocus and then corrupted by white Gaussian noise to simulate sensor noise. We analyzed the performance of full-reference and no-reference algorithms to gauge their performance on our dataset and demonstrate how multiple distortions affect the quality judgments of humans and objective algorithms.

2.1. Compilation of image dataset

Images in the study dataset were derived from 15 high-quality pristine images, chosen to span a wide range of content, colors, illumination levels and foreground/background configurations. Distorted images were generated from each of these images as follows:

2.1.1. Single distortions

- **Blur**: Gaussian kernels (standard deviation σ_G) were used for blurring with a square kernel window of side 3σ (rounded off) using the Matlab *fspecial* and *imfilter* commands. Each of the R, G and B planes of the image was blurred using the same kernel.
- **JPEG**: The Matlab *imwrite* command was used to generate JPEG compressed images by varying the Q parameter (whose range is from 0 to 100) which parametrizes the DCT quantization matrix.
- **Noise**: Noise generated from a standard normal pdf of variance σ_N^2 was added to each of the three planes R,G and B using the $imnoise$ Matlab function.

Three different values each were used for each distortion parameter: σ_G = 3.2, 3.9, 4.6 pixels, Q = 27, 18, 12 and σ_N^2 = 0.002, 0.008 and 0.032 were chosen. They were selected to keep the distorted images perceptually separable from each

other and from the references, and to keep the distortions within a realistic range.

2.1.2. *Multiple Distortions*

Four levels of blur, JPEG compression and noise - the 0 level (no distortion) and levels 1, 2 and 3 with above mentioned values were considered.

- **Blur followed by JPEG**: Each of the four blurred images was compressed using the JPEG encoder bank. 16 images $I_{ij}^{k,1}$ were generated from the k-th reference image R^k, $0 \leq i, j \leq 3$ where i denotes the degree of blur and j the degree of JPEG compression.
- **Blur followed by Noise**: Noise at each of the four levels was added to each of the four images generated by the blurring stage. Therefore, 16 images $I_{ij}^{k,2}$ were generated from R^k, where i denotes the degree of blur and j the degree of noise.

In all, 15 reference images were used to generate 225 images for each part of the study of which 90 are singly distorted (45 of each type) and 135 are multiply distorted. Both parts of the study were conducted under identical conditions with separate sets of subjects. To confirm that the human scores from both parts of the study may be analyzed together, the same blurred images were used in both parts of study, hence there are 405 images in all.[1] The dataset compilation is schematically represented in Fig 1.

2.2. Study conditions

2.2.1. *Equipment and Display Configuration*

Images of size 1280×720 were displayed on an LCD monitor with 73.4 ppi resolution, calibrated in accordance with the recommendations in [10]. The study was conducted in a workspace environment under normal indoor illumination levels. The subjects viewed the monitor from a distance approximately equal to 4 times the screen height. The Matlab Psychometric Toolbox [11] was used to render images on the screen and acquire human ratings.

2.2.2. *Study design*

The study was conducted using a single stimulus (SS) with hidden reference [12] method with numerical non-categorical assessment [10]. Each image was presented for a duration of 8 seconds after which the rating acquisition screen was displayed containing a slider with a continuous scale from 0 to 100. Semantic labels 'Bad', 'Poor', 'Fair', 'Good' and 'Excellent' were marked at equal distances along the scale to guide the subjects.

[1]It was verified that the human scores on these blurred images from the two parts of the study did not differ significantly. This was taken as evidence that the scores from the two parts of the study could be combined for analysis without the need for realignment as in [9]

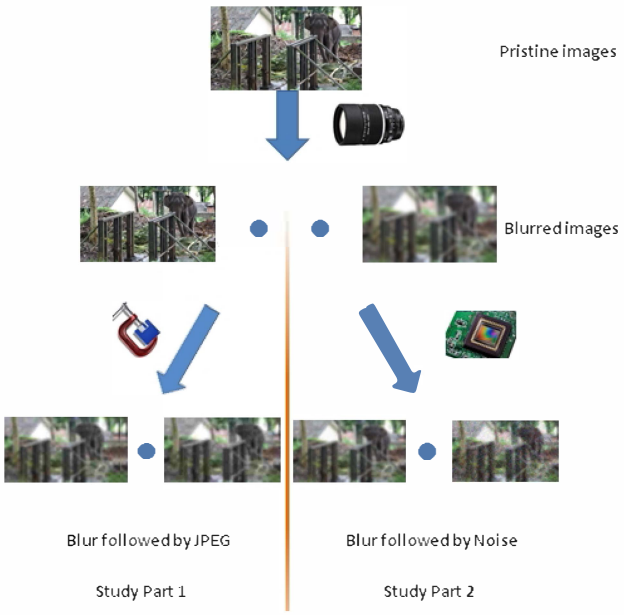


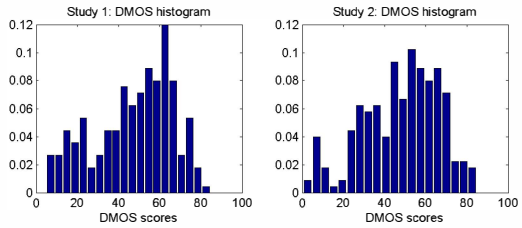
Fig. 1. Schematic of the image dataset compilation

At the beginning of each session, the subject was briefed based on recommendations in [10] and then asked to rate 6 training images, carefully selected to approximately span the range of image qualities in the test dataset. In the test session that followed, subjects were asked to rate images from the test dataset. The test images were presented in random order, different for each subject, to eliminate memory effects on the mean scores. No two images derived from the same reference were shown one after the other. Subjects were not informed of the presence or location of reference images in the test image sequence. The rating procedure was thus completely blind to the reference images.

Ratings for each part of the study were acquired during the test phase from each subject for all the 240 images, of which 15 were reference. To minimize subject fatigue[10], each subject therefore rated images in two sessions lasting no longer than 30 minutes each.Analysis from part 1 showed that the difference mean opinion (DMOS) scores did not change significantly if the reference image was displayed in only one of the sessions. Each reference was presented only once to each subject in part 2 and the acquired ratings were used for difference score computations over both sessions.

2.2.3. *Subjects*

Subjects for the study were mostly graduate students at The University of Texas at Austin (UT) who volunteered to take the study. A majority of the subjects were male, and between 23 and 30 years old. Subjects were allowed to continue to wear corrective lenses if they thought it necessary to do so. The study was conducted over a period of four weeks and ratings were acquired from each subject in two sessions as



(a) Part 1 (b) Part 2

Fig. 2. Distribution of DMOS scores

described above. A total of 19 and 18 subjects participated in the first and second parts of the study respectively.

2.3. Processing of scores

To simplify our notation, we drop the superscript l labelling the study part with the understanding that the rest of our discussion is applicable to each of the two parts of the study. Further, we used three labels for each test image: i, j and k , which we now collapse into one subscript $y = i, j, k$. Then let s_{xyz} be the score assigned by subject x to image I_y in session z. Further let $I_{y_{ref}}$ be the reference image corresponding to I_y, which is displayed in one/both sessions $z = 1, 2$ and N_{zm} be the number of images rated by subject x in session z.

- **Difference Scores**: As a first step, the raw rating assigned to an image was subtracted from the rating assigned to its reference image in that session to form the difference score $d_{xyz} = s_{xyz} - s_{xy_{ref}z}$. In this scheme, all reference images are assigned the same difference score of zero. This eliminates rating biases associated with image content.

- **Mean Human Score**: We then computed a Difference Mean Opinion Score(DMOS) for each image I_y as $\frac{1}{N_X}\sum_x d_{xym}$, where N_X is the number of subjects. The distributions of DMOS scores for parts I and II of the study are shown in Fig 2. We found that analysis done using Z scores in place of DMOS scores as in [13] did not offer any new insights. Hence, we only report analysis using DMOS scores here.

- **Subject Screening**: This mean score may be easily contaminated by outliers such as inattentive subjects. To prevent this, we follow a procedure recommended in [10] to screen subjects. No outlier subjects were detected in either part of the study.

3. RESULTS

3.1. Algorithm Performance Evaluation

In this section, we analyze the performance of a variety of existing full-reference image quality algorithms [14], and a recently proposed state-of-the-art no-reference algorithm called BRISQUE [15] which uses features from [16] on our database. BRISQUE is a learning-based algorithm. We tested its performance on our multi-distortion dataset after training on two separate training sets: the LIVE single distortion dataset[5] (BRISQUE-1), and the multi-distortion dataset itself, using 4:1 train-test splits (BRISQUE-2). To compare algorithms, we ran 1000 iterations with random 4:1 train-test splits. We report the median performance over all iterations. We report Spearman Rank Ordered Correlation Coefficient (SROCC) and Pearson Linear Correlation Coefficient (LCC) as a measure of the correlation of IQA algorithms with DMOS. To obtain the LCC, a four-parameter monotonic function is used to map IQA scores to DMOS: $h(u) = \frac{\beta_1}{1+\exp(\beta_2(u-\beta_3))} + \beta_4$ where $h(u)$ is the predicted human score when the algorithm returns the value u. We then compute the LCC between h and DMOS. Results on five classes of images and on the complete dataset are shown in Tables 1 and 2. The first three columns contain scores on single-distortion images from the dataset. The highest score in each column is highlighted.

	Blur	JPEG	Noise	Study 1	Study 2	Overall
PSNR	0.5000	0.0909	0.8000	0.6634	0.7077	0.6954
MS-SSIM	0.7579	0.4643	**0.8892**	0.8350	0.8559	0.8454
VIF	0.7857	**0.6667**	0.8524	0.8795	0.8749	0.8874
IFC	0.8182	0.6264	0.8364	0.8914	0.8716	0.8888
NQM	0.8462	0.5000	0.7619	0.8936	**0.8982**	0.9020
VSNR	0.6685	0.3571	0.8041	0.7761	0.7575	0.7844
WSNR	0.6190	0.6000	0.7940	0.7692	0.7488	0.7768
BRISQUE-1	0.8000	0.2909	0.7972	0.7925	0.2139	0.4231
BRISQUE-2	**0.8818**	0.6364	0.8799	**0.9214**	0.8934	**0.9111**

Table 1. SROCC of IQA scores with DMOS

	Blur	JPEG	Noise	Study 1	Study 2	Overall
PSNR	0.5661	0.4161	0.9235	0.7461	0.7864	0.7637
MS-SSIM	0.8683	0.6090	0.9567	0.8785	0.8951	0.8825
VIF	0.9079	0.7907	0.9533	0.9214	0.8930	0.9083
IFC	**0.9151**	0.8140	0.9403	0.9271	0.8997	0.9137
NQM	0.8643	0.6367	0.8984	0.9179	0.9126	0.9160
VSNR	0.7684	0.5304	0.9404	0.8372	0.8090	0.8326
WSNR	0.6887	0.6759	0.9310	0.8457	0.8108	0.8408
BRISQUE-1	0.8412	0.5803	0.9349	0.8687	0.3776	0.5001
BRISQUE-2	0.8918	**0.8143**	**0.9614**	**0.9462**	**0.9226**	**0.9349**

Table 2. LCC of IQA scores with DMOS

MS-SSIM tends to perform poorer on our dataset than do IFC, VIF and NQM, while it performs on par with VIF in [5]. Performance on JPEG distorted images in particular was generally very poor. Also, several algorithms that perform poorly on JPEG images score well on images afflicted with blur followed by JPEG (part 1) because human scores on such images in our study are mainly dominated by the blur component. These observations indicate that there might have been insufficient perceptual separation of JPEG levels used to generate test data for part 1. The NR-IQA BRISQUE-2 outperforms all full reference IQA measures on our dataset. However, it should be noted that BRISQUE-2 has the advantage of being trained on multi-distorted images.

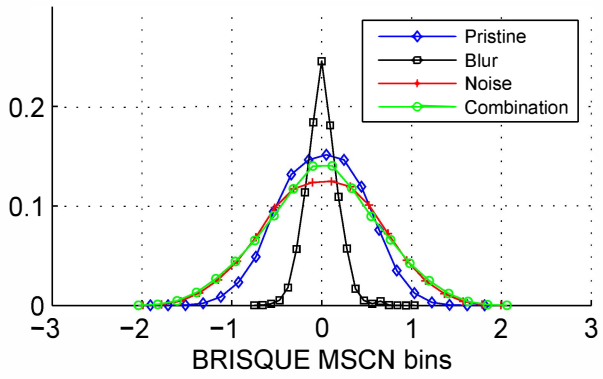


Fig. 3. Effect of blur, noise and blur followed by noise on normalized luminance histogram used by BRISQUE

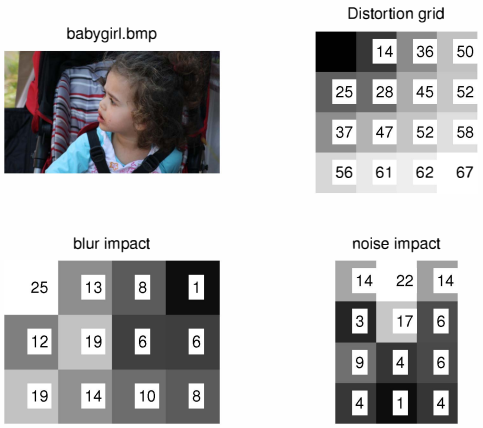


Fig. 4. Distortion grid analysis of the *baby girl* image: (top left) the pristine image, (top right) distortion grid score arrangement on part 2 of the study, (bottom left) blur impact grid and (bottom right) noise impact grid. (see text for details)

3.2. Impact of multiple distortions on quality features

Comparison of BRISQUE-1 and BRISQUE-2 shows that the single-distortion-trained BRISQUE-1 performs particularly poorly on images afflicted with blur followed by noise (part 2 of the study), while it does well on the blur and noise single distortion classes. Similar to other previously proposed successful algorithms such as SSIM, BRISQUE uses the statistics of pixel intensities subtracted from local means and normalized by local contrasts. Broadly speaking, noise and blur tend to widen and narrow the distribution of this feature respectively. Thus this distribution for an image afflicted with both noise and blur resembles that for pristine images as shown in Fig 3. It is conceivable therefore that BRISQUE models trained on singly distorted images would overestimate the quality of multi-distorted images. This evidence indicates that the behavior of quality-determinant features such as these in multi-distortion scenarios warrants further study. The vastly improved performance of BRISQUE-2 over BRISQUE-1 on part 2 shows how a multi-distortion database might provide valuable training data for image quality algorithms.

3.3. Towards understanding interactions of distortions

In this section we analyze the impact of interaction of distortions on the DMOS scores in our dataset. To do this, for each part of our dataset we arrange the scores for all the distorted variants of each image in a distortion grid, where every row represents a constant value of distortion 1 (blur for both parts) and every column represents a constant value of distortion 2 (JPEG and noise for the two parts respectively). Such a grid is demonstrated in the top-right of Figure 4. This distortion grid itself demonstrates the expected behavior of low DMOS scores near the top-left (close to pristine) and subsequently increasing DMOS scores near the bottom-right (most distorted).

If we now form a new grid by subtracting every row from its preceding row, we can analyze the incremental impact of varying distortion 1 at fixed distortion 2 (impact matrix 1). Similarly, by subtracting every column from its preceding column, we can analyze the incremental impact of varying distortion 2 at fixed distortion 1 (impact matrix 2). Examples of blur and noise impact grids are presented in Fig 4. From these grids, we note that the patterns of incremental noise impact at fixed blur vary as a function of the blur and vice versa.

To demonstrate this, we plot every column of impact matrix 1, and similarly, every row of impact matrix 2. These plots are shown in Fig 5 for the *baby girl* image. A general trend that can be observed from these plots is that the impact on DMOS of increment in distortion A is generally lower at higher levels of distortion B. For instance, in the bottom plot in Fig 5, the plots nearly line up one below the other in the order of increasing blur levels. Thus the impact of visual masking of one distortion by another is demonstrated in our analysis.

Another interesting observation is regarding the shape of the plots themselves. We expect that the shapes of these plots must be a function of the distortion parameters that we have chosen for our various distortion levels. An interesting observation that emerges from the data is that the shape of distortion incremental impact plots in the presence of a second distortion itself varies as a function of the level of the other distortion as in Fig 5. For instance, while noise increment 2 has the largest impact of any for blur levels 0 (pristine) and 1, it has the smallest impact of any for blur levels 2 and 3, in the case that is shown here. Similar trends are observed for other images in the dataset. This suggests complex interactions among distortions in determining the perceptual quality of visual content.

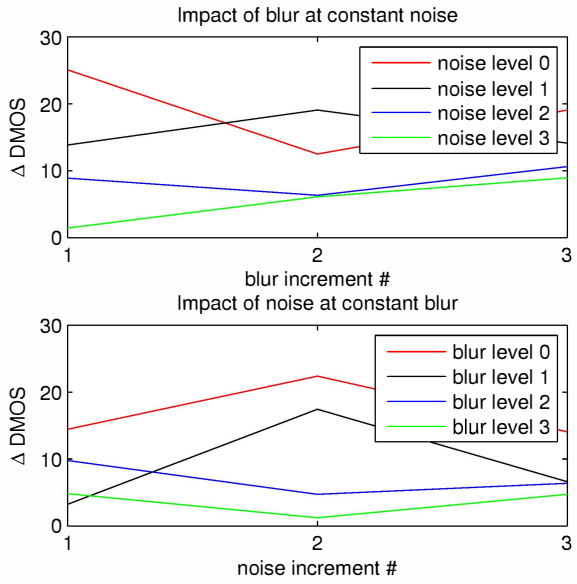


Fig. 5. Plots of distortion impact: (top) impact of blur increments at constant noise level, (bottom) impact of noise increments at constant blur level

4. CONCLUDING REMARKS

We found that correlation scores of objective FR IQA algorithms with human judgments are lower compared to [5] which indicates that the new database is more challenging. This is attributed to the presence of multi-distorted images and individual distortions severities deliberately kept within a small range to resemble images that are available for consumption. The NR-IQA BRISQUE trained on multi-distorted images outperforms all full reference measures on our database. We proposed a method to understand the impact of interaction of distortions on human ratings and demonstrated that some interesting trends emerge from the data. To understand these trends in further detail requires investigation into the nature of the non-linearity of DMOS scores as a function of perceived quality. For instance, human scores are often collected, as in our experiment, on a fixed scale with fixed end points, and it is frequently observed in practice that subjects tend to be more conservative with score increments or decrements near the end points, which may lead to a sigmoidal nonlinearity in the ratings computed from this data. Investigations in this direction will help us better understand interactions of distortions, and incorporate these into our models for image quality. With our database, we have taken a step towards investigating image quality in more realistic settings. Future work in this direction will involve building databases of human scores on more realistic images to pave the way for understanding the problem of perceptual quality assessment on real visual content.

5. REFERENCES

[1] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multi-Scale Structural Similarity for Image Quality Assessment," in *Proceed Asilomar Conf Signals, Syst. and Comput.*, 2003.

[2] H.R. Sheikh, A.C. Bovik, and G. de Veciana, "An Information Fidelity Criterion for Image Quality Assessment using Natural Scene Statistics," *IEEE Trans Image Process.*, 2005.

[3] H.R. Sheikh and A.C. Bovik, "A Visual Information Fidelity Approach to Video Quality Assessment," in *Workshop Video Process. and Quality Metr. for Consumer Elect.*, 2005.

[4] M.A. Saad, A.C. Bovik, and C. Charrier, "A DCT Statistics-Based Blind Image Quality Index," *IEEE Sig. Process. Lett.*, 2010.

[5] H.R. Sheikh, Z. Wang, L.K. Cormack, and A.C. Bovik, "LIVE Image Quality Assessment Database," *http://live.ece.utexas.edu/research/quality.*

[6] N. Ponomarenko, F. Battisti, K. Egiazarian, J. Astola, and V. Lukin, "Metrics performance comparison for Color Image Database," in *Internat'l workshop on video process. and quality metr. for consumer elect.*, 2009.

[7] Vishwakumara Kayargadde and Jean-Bernard Martens, "Perceptual characterization of images degraded by blur and noise: model," *J. Opt. Soc. Am. A*, vol. 13, no. 6, pp. 1178–1188, Jun 1996.

[8] Damon M. Chandler, Kenny H. Lim, and Sheila S. Hemami, "Effects of spatial correlations and global precedence on the visual fidelity of distorted images," 2006, vol. 6057, p. 60570F, SPIE.

[9] 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.*, 2006.

[10] International Telecommunication Union, "BT-500-11: Methodology for the Subjective Assessment of the Quality of Television Pictures," Tech. Rep.

[11] D.H. Brainard, "The Psychophysics Toolbox," *Spatial Vision*, vol. 10, 1997.

[12] M. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," in *SPIE Video Comm. and Image Process. Conf.*

[13] 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. Cir and Syst Video Tech*, 2010.

[14] M. Gaubatz, "Metrix MUX Visual Quality Assessment Package," *http://foulard.ece.cornell.edu/gaubatz/metrix-mux.*

[15] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," submitted to TIP 2012.

[16] A. Mittal, A.K. Moorthy, and A.C. Bovik, "Blind/Referenceless Image Spatial Quality Evaluator," in *Proceed Asilomar Conf Signals, Syst. and Comput.*, 2011.