https://doi.org/10.2352/ISSN.2470-1173.2020.9.IQSP-066

Perceptual Quality Assessment of Enhanced Images Using a Crowd-Sourcing Framework

Muhammad Irshad*[1] *, Alessandro R. Silva*[2] *, Sana Alamgeer*[1] *, Mylène C.Q. Farias*[1] *;
[1] ***Department of Electrical Engineering,*** [2] ***Department of Computer Science,***
University of Brasilia, Brazil.

Abstract
In this work, we present a psychophysical study, in which, we analyzed the perceptual quality of images enhanced with several types of enhancement algorithms, including color, sharpness, histogram, and contrast enhancements. To estimate and compare the qualities of enhanced images, we performed a psychophysical experiment with 35 source images, obtained from publicly available databases. More specifically, we used images from the Challenge Database, the CSIQ database, and the TID2013 database. To generate the test sequences, we used 12 different image enhancement algorithms, generating a dataset with a total of 455 images. We used a Double Stimulus Continuous Quality Scale (DSCQS) experimental methodology, with a between-subjects approach where each subject scored a subset of the total database to avoid fatigue. Given the high number of test images, we designed a crowd-sourcing interface to perform an online psychophysical experiment. This type of interface has the advantage of making it possible to collect data from many participants. We also performed an experiment in a controlled laboratory environment and compared its results with the crowd-sourcing results. Since there are very few quality enhancement databases available in the literature, this works represents a contribution to the area of image quality.

Keywords: Enhancement; Perceptual Quality Assessment; Crowd-Sourcing Framework, Subjective Quality Assessment.

Introduction
Image enhancement is frequently used to improve or restore the visual quality of images and videos. Currently, there are several image enhancement algorithms, but there is not yet a performance metric that is able to estimate the performance of these methods. Since the final consumers of the resulting enhanced visual content are human viewers, the performance of these algorithms should be measured by estimating the visual quality of the enhanced images, taking into consideration the human visual system [20].

Image quality can be estimated using subjective (psychophysical experiments) and objective (quality metrics) methods [9, 21]. Subjective methods are simply psychophysical experiments where participants rate one or more aspects of a set of processed images. Most often, these experiments are performed in a controlled environment (e.g. a laboratory), following standard recommendations for the environment conditions and experimental methodologies [6]. It worth pointing out that although data (subjective scores) collected in psychophysical experiments are considered as ground-truth, these experiments are time-consuming and expensive. Objective quality methods, on the other hand, are algorithms (implemented in hardware or software) that automatically estimate the quality of an image [14, 10]. These methods are designed and tested using subjective scores as ground-truth.

The area of image and video quality has achieved great progress in the last decades [2]. But, although the performance accuracy of quality metrics has improved, there are still many challenges in this area. Among them is the design of objective quality metrics for enhanced contents. Since most of the quality metrics have been designed to capture visual distortions, they are not able to quantify the changes in quality introduced by enhancement algorithms. Therefore, currently, there is a need for quality metrics that can automatically estimate the quality of enhanced images and videos. It is worth pointing out that developing quality metrics for enhanced images is a challenge due to the lack of quality databases containing enhanced images and their respective (ground-truth) subjective quality scores.

In this paper, our goal is to introduce a quality database for enhanced images. Up to our knowledge, currently, there is only one image enhancement quality database that can be used for research in image quality [19]. However, this database contains images of low resolution that were processed manually, using a professional graphics editing software (Adobe Photoshop) to produce the best possible enhanced images. In our database, we used images of a higher resolution, which are enhanced with twelve different image enhancement algorithms. Our goal was to produce a set of images that were like consumer applications contents. Also, we performed a crowd-sourcing subjective experiment to obtain quality scores for all database images. With this experiment, we were able to obtain a large and diverse pool of participants.

Database Content Generation
Figure 1 shows a block diagram of the strategy used to generate the database. Our first step was to choose 35 original (source - SRC) images. These images were taken from three image quality databases, to allow for future comparisons of enhanced and degraded images. More specifically, we took 5 SRC images from the CSIQ database [18], 5 original images from the TID2013 database [16], and 25 original images from the ChallengeDB database [3]. Table 6 (in the Appendix) shows a list of the SRC images, along with their names in the corresponding databases. These chosen source contents are diverse, in terms of spatial activity, semantic content, and color distribution. In Figure 2, the first row (SRCs) shows examples of SRC images taken from the (a-b) TID2013, (c-d) CSIQ, and (e-f) ChallengeDB databases.

Our next step consists of choosing the enhancement algo-

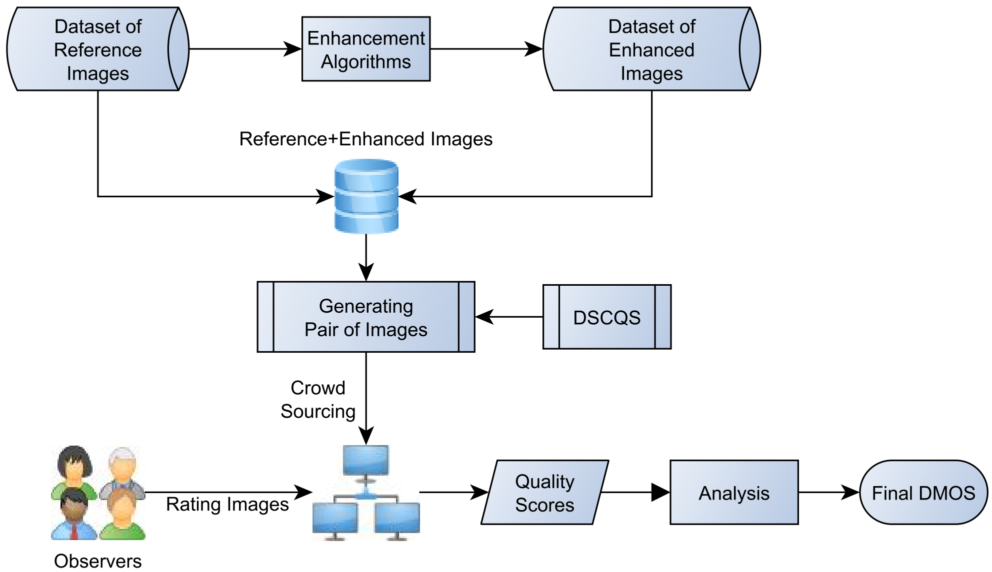

Figure 1: Block diagram of the strategy used to create the database and run the crowd-sourcing experiment.

Table 1: Enhancement algorithms and their corresponding Hypothetical Reference Circuits (HRC) in the experiment.

HRC	Enhancement Algorithm	Abbreviation
HRC01	Color Enhancement	CE
HRC02	Contrast Enhancement	CrE
HRC03	Brightness Enhancement	BE
HRC04	Sharpness Enhancement	SE
HRC05	Unsharp Masking	UM
HRC06	Gaussian Blur	GB
HRC07	Histogram Equalization	HE
HRC08	Dynamic Histogram Equalization	DHE
HRC09	Exposure Fusion Framework	EFF
HRC10	Average Histogram Equalization	AHE
HRC11	Hue Saturation Histogram Equalization	HSHE
HRC12	Joint Enhancement and Denoising Method via Sequential Decomposition	JEDMSD

rithms to be used as the different test conditions of our experiment. In this paper, we refer to these test conditions as Hypothetical Reference Circuits (HRC). We chose a total of 12 enhancement algorithms: Color Enhancement (CE), Contrast Enhancement (CrE), Sharpness Enhancement (SE), Brightness Enhancement (BE), Unsharp Masking (UM), Gaussian Blur (GB), Histogram Equalization (HE), Dynamic Histogram Equalization (DHE) [1], Exposure Fusion Framework (EFF) [22], Average Histogram Equalization (AHE), Hue Saturation Histogram Equalization (HSHE), and Joint Enhancement and Denoising Method via Sequential Decomposition (JEDMSD) [17]. Table 1 shows a list of the HRCs of the database and the corresponding enhancement algorithms. We used the traditional versions of these algorithms, which were implemented using the *Pillow package* [11] in Python [13] and Matlab [12]. To generate the test images, we process each of the 35 SRCs using all HRCs, generating a total of 420 enhanced images. In Figure 2, the last three rows correspond to images processed with HRC02, HRC03, and HRC04, which were generated from the SRC images shown in the top row.

Given the number of test images (420), we manually divided the stimuli set into 3 sub-sets. Each sub-set contains 140 images, with different combinations (non-factorial) of at least 10 HRCs and 35 SRCs. To avoid a presentation bias, we generated 4 versions of these 3 sub-sets, which had different choices HRC-SRC combinations. In total, there were 12 groups of different test images. This way, each participant did not rate all possible combinations of HRCs and SRCs, but only a third of them, which made the experiment less tiring. Naturally, this division demands more participants.

Experimental Methodology

To guarantee the validity, reliability, and reproducibility of subjective quality assessment methods, over the years several recommendations for experimental methodologies have been drafted. Examples of popular recommendations are Rec. ITU-R BT.500 [6] and Rec. ITU-T P.910. [7]. It is worth mentioning that a good number of these experimental methodologies have been derived from classic psychometric practices [9]. The experimental methodologies have different specifications, which include the stimuli presentation, the type of subjective scale, and the scoring procedure. Each methodology type has advantages and disadvantages and it is difficult to cover all factors to provide a specific set of recommendations. Pinson *et al.* [15] have detailed several aspects that should be taken into consideration when performing an experiment, like the environment conditions, the number of participants and stimuli, and the scoring procedure.

Figure 3 depicts the most popular types of experimental methodologies used for quality assessment. The experimental methodologies can be divided into rating or ranking based methods [15]. In rating based methods, participants assign a score to

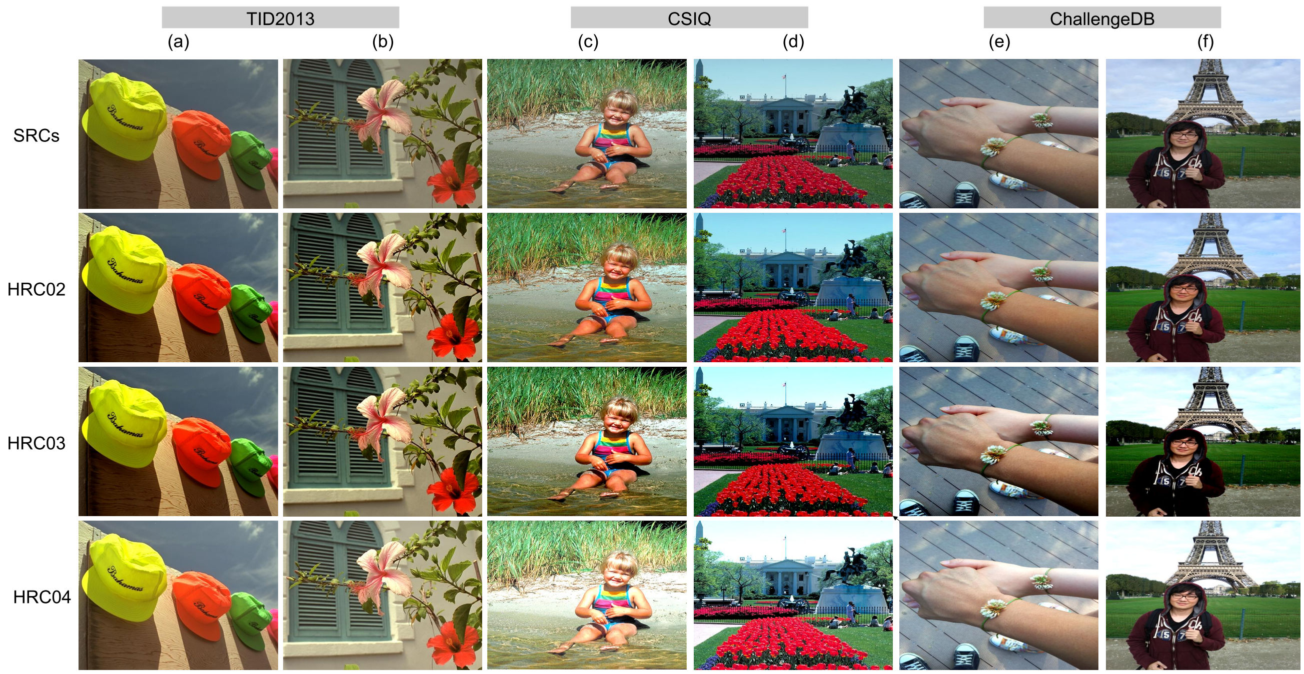

Figure 2: Sample source (SRC) images used in our database, processed with different enhancement algorithms (HRCs - see Table 1). SRC images were taken from the (a-b) TID2013, (c-d) CSIQ, and (e-f) ChallengeDB databases.

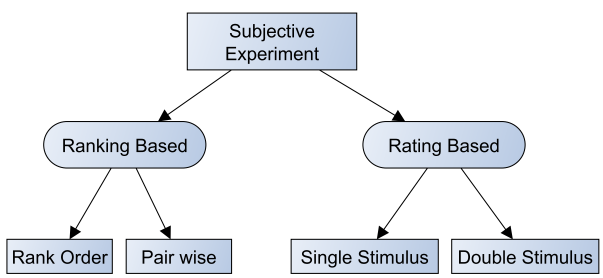

Figure 3: Most popular types of subjective experimental methodologies.

each stimulus presented to them, using either a numerical scale or a category scale. In ranking based methods, participants are asked to either rank the images in terms of their quality or to compare each pair of images of the set (a pairwise comparison - PC) [15]. PC methodologies are often considered when the differences between stimuli are small and, therefore, participants might have a hard time differentiating them. Unfortunately, since PC requires that the participants compare all possible combinations of image pairs, it is a very time consuming methodology. Given the large number of test images in our experiment, we used a rating-based experimental methodology.

In terms of stimuli presentation, the methodology can be single stimulus (SS), double stimulus (DS), or multi-stimulus (MS). In SS methodologies, participants rate the quality of just one stimuli (the test), without having a reference. In DS and MS methodologies, participants rate the quality of two or more stimuli, which are presented simultaneously or closely spaced in time. According to the rating scale used, DS methodologies are classified as Double Stimulus Continuous Quality Scale (DSCQS) [6] or Double Stimulus Impairment Scale (DSIS) [8]. In DSIS (also referred as Degradation Category Rating - DCR), participants rate both displayed stimuli using a discrete 5-point impairment scale: Imperceptible (5), perceptible but not annoying (4), slightly annoying (3), annoying (2), very annoying (1). In DSCQS, participants rate the quality of both the reference image and the test image using a 5-point quality scale: Excellent (5), Good (4), Fair (3), Poor (2), and Bad (1). In our subjective experiment, we have chosen to use the DSCQS methodology. DSCQS has been shown to be more accurate than SS, specially for stimuli with smaller quality differences, which is the case for enhanced images.

Given our database size, we decided to use a crowd-sourcing experimental methodology. The experiment was available online and the invitations for participation were made by email and by social media (Facebook and Twitter). We chose to use crowd-sourcing because it allows experimenters to collect data from a many participants with different backgrounds. There are currently several tools for running crowd-sourcing experiments, including Crowdflower, Crowdsource, Microtask, Taskrabbit, Amazon Mechanical Turk (Mtruk), and Microworkers. But, most of these tools are paid solutions, which require experimenters to (also) pay the participants [4]. For this reason, we have developed a web-based online crowd-sourcing experimental framework, which can be used on any browser that supports JavaScript. Each time a new participant starts the experiment, the system assigns him/her to one of the 12 groups of the database, prioritizing the groups with less participants to maintain a good distribution of participants per group. The framework tracks both the operating system and the device used by the participant.

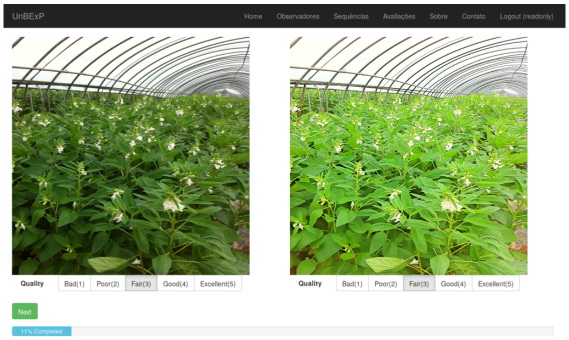

Figure 4: Screenshot of a trial of the crowd-sourcing experimental interface, displaying a SRC image and its version processed with a specific HRC.

Each experimental session is divided into four stages. In the first stage, called *the registration*, the participant fills out a form with personal information. In the second stage, called *the training*, the participant watches a set of sample images and their corresponding enhanced versions. The goal is that the participant familiarizes himself/herself with the quality range of the images in the database. In the third stage, called *the practice*, the participant tries out the scoring procedure by performing a small number of trials, identical to the ones in the experimental session. Finally, in the *main experimental session*, the participant rates the quality of a SRC image and its enhanced version, which was processed with a specific HRC. Figure 4 depicts a screenshot of the interface, showing this scoring procedure for each trial. In the interface, the 5-point quality scales are shown below each image. The positions (left or right) of the SRC and its enhanced version are randomized for each trial. The database created in this work is available for download at the GPDS site[1].

Crowd-Sourcing Experimental Results

A total of 108 participants took part in the crowd-sourcing experiment, with 78% of the participants being male and 22% female. The participant's age range varied from 17 to 63 years. To analyze the gathered data, we computed the mean observer scores (MOS) and the Difference MOS (DMOS). MOS is calculated by averaging the scores given by all participants for each HRC and each test image. DMOS is computed by taking the average of the differences between the scores given to a test image and the score given to its corresponding SRC. In other words, we average the values of the differences between the scores of the two images, which are shown jointly to participants (see Figure 4) in each experiment trial.

Figure 5 presents the DMOS for each one of the HRCs and SRCs. Each graph in this figure corresponds to the DMOS of a single HRC across the different SRCs (x-axis). Notice that, from this figure, it is hard to identify any patterns from the graphs. To take a closer look at the results and minimize the effect of image content, we compute the average of both DMOS and MOS values for each HRC across all SRCs. Figures 6 and 7 depict the average MOS and DMOS values, respectively, for each individual HRC, along with their confidence intervals. For DMOS, negative values indicate that the enhancement algorithm (on average) reduced the perceived quality of the SRC image, while positive values indicate that the algorithm (on average) improved the quality of the SRC image. Notice that HRC01, HRC04 and HRC05 produced positive average DMOS values, while the HRC06-HRC12 produced negative average DMOS values. For brightness (HRC02) and contrast enhancement algorithms (HRC03), the results were inconclusive. The Sharpness enhancement algorithm (HRC04) produced the highest DMOS values, followed by the Unsharp Masking (HRC05) and the Color Enhancement (HRC01) algorithms.

To verify if the participants perceived differences on the algorithm's image quality when comparing two HRCs, we executed a paired-samples t-test. This test checks if there is a statistically significant difference between the average DMOS considering pairs of HRCs. In each pair-wise comparison we considered only the cases where the participant scores both HRCs. The most relevant results of this test are shown in Table 2. Notice that the average DMOS of six HRC pairs are not statistically significant ($p > 0.05$): HRC01-HRC05, HRC03-HRC10, HRC07-HRC08, HRC07-HRC09, HRC08-HRC09, HRC11-HRC12. This means that participants (on average) did not see a difference in quality between these pairs of HRCs. For all other HRC pair combinations, the test found statistically differences between the average DMOS, which means that participants (on average) can distinguish the image quality produced by the remaining combinations.

Table 2: Paired Samples t-test pairs for which the DMOS differences were not statistically significant.

Pair	HRC	N	Mean	Std.dev	t	p
HRC01-HRC05	HRC01	222	0.4640	1.0746	-1.562	0.120
	HRC05	222	0.5946	0.9643		
HRC03-HRC10	HRC03	293	-0.1640	1.2165	0.838	0.403
	HRC10	293	-0.2389	1.3564		
HRC07-HRC08	HRC07	578	-0.9498	1.4431	0.795	0.427
	HRC08	578	-0.9896	1.3932		
HRC07-HRC09	HRC07	588	-0.9388	1.4568	0.836	0.404
	HRC09	588	-0.9796	1.4296		
HRC08-HRC09	HRC08	590	-1.0186	1.3856	-0.071	0.944
	HRC09	590	-1.0153	1.4195		
HRC11-HRC12	HRC11	624	-0.5048	1.4255	1.601	0.110
	HRC12	624	-0.6154	1.4532		

To verify if the content (SRC) affects the perception of quality, we executed a one-way ANOVA to determine if there is a statistically significant difference between the average DMOS, considering SRC as an additional factor. For all groups (HRCs), ANOVA returned a $p < 0.05$ value, meaning there is a least one pair of SRCs, considering each HRC, that has a difference in average DMOS that is statistically significant. To identify these pairwise comparisons, we performed a Tukey-Kramer post-hoc test, which allows different group sizes, but gives the same result the Tukey post hoc test would give if the group sizes were equal. The test shows that 1,151 combinations (out of 7,141 possible combinations) have differences in average DMOS that are statistically significant. This means that, in 16% of the cases, the participants found that an enhancement algorithm led to a visible difference for a specific pair of SRCs. We see this behavior by observing the DMOS values in Figure 5, which change according to the SCRs.

Another output parameter of the Tukey-Kramer test is the number of homogeneous sets. Table 3 depicts these results, where SCRs are classified as having one or more sets, depending on their distance to the mean. In a scenario where there is no statistically

[1] http://www.ene.unb.br/mylene/databases.html

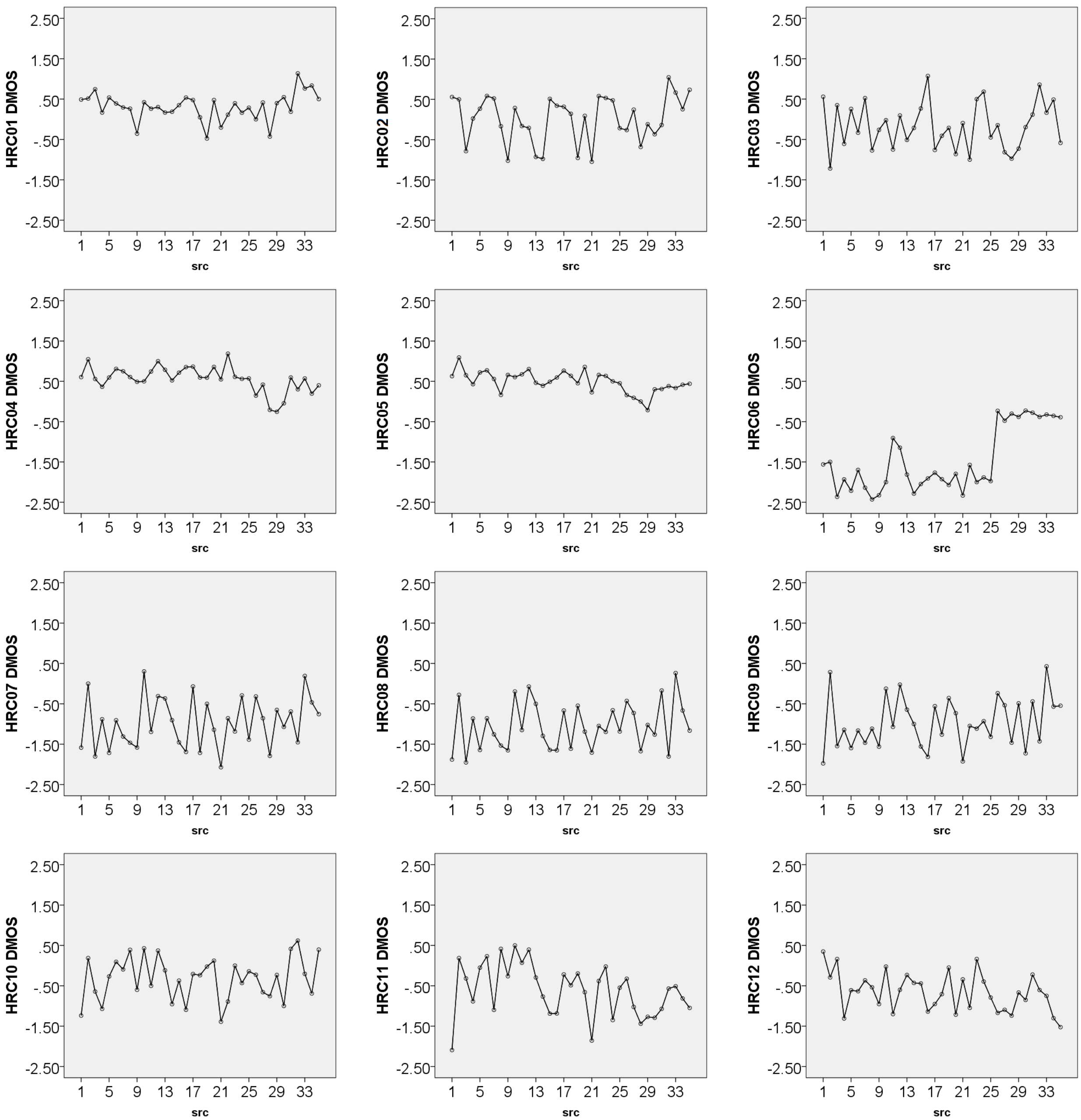

Figure 5: Average DMOS values versus the SRC image, for each HRC.

significant difference between the means, each HRC would have only one homogeneous set containing all SRCs. In other words, for this HRC, the results would be content independent. If the number of homogeneous sets are small and the sets are disjointed, we can extract the features of these groups that influence the quality differences. Notice that HRC05 has less homogeneous sets (4) than the others HRCs. HRC06 is the only HRC that has disjointed sets. But, since most SCRs have more than one set, it is not easy to identify a feature set that is able to distinguish the means. Further analysis is needed to understand how content (SRC) affects thee perceived quality.

Results from the Laboratory Experiment

Since the reliability of crowd-sourcing methodologies remains questionable [5], we have also performed an experiment in a controlled laboratory environment and performed a comparison of the results in the two experiments. Eighteen participants took part in the experiment, with 72% being male and 28% being female. The range of the participant's age varied from 18 to 70 years. We used the same experimental methodology and interface platform used in the crowd-sourcing experiment. The experimental set-up followed the ITU-R BT.500 recommendation [6]. Table 4 depicts the technical specifications used in the experiment. An experimental session lasted, on average, 40 minutes.

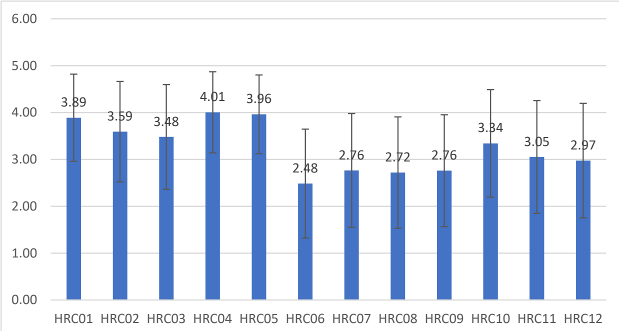

Figure 6: Mean Observer Score (MOS) computed across all SRCs for each HRC (see Table 1).

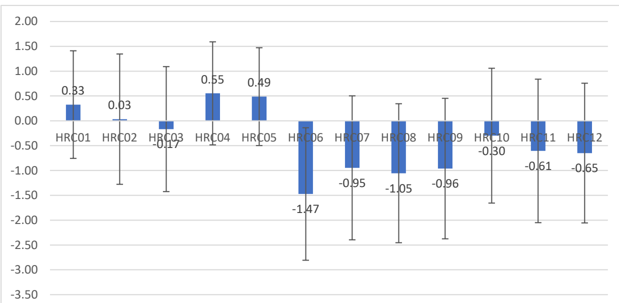

Figure 7: Difference Mean Observer Score (DMOS) computed across all SRCs for each HRC (see Table 1).

Table 3: Number of Homogeneous Sets found by the Tukey-Kramer post-hoc test.

HRC	Number of Sets
HRC01	6
HRC02	8
HRC03	13
HRC04	7
HRC05	4
HRC06	6
HRC07	11
HRC08	10
HRC09	11
HRC10	12
HRC11	12
HRC12	6

A one-way ANOVA was used to compare the DMOS values of these two groups of experimental results: the onsite (laboratory) and online (crowd-sourcing) experiments. Table 5 shows these ANOVA results. The last column of the table indicates if there is a statistically significant difference between the two groups. More specifically, if $p < 0.05$ the differences are statistically significant and, therefore, the participants in the two experiments rated the images enhanced with a particular HRC differently. We can see that for HRC07, HRC09, and HRC12 participants in the two experiments rated the images differently, while for the other 9 HRCs the differences are not statistically significant. So, for most of the HRCs, there is no statistically significant difference between the DMOS given by participants in both experiments.

Table 4: Technical specifications of the equipment used for the Laboratory Experiment.

Item	Specification
Monitor	BENQ XL2420z 1920x1080 144hz
Distance of the observer	60 CM
GPU	QUADRO K4000
Brightness	100%
Sharpness	50%

Table 5: ANOVA results comparing online and onsite groups.

HRC	Group	N	Mean	Std.dev	F	p
HRC01	OnSite	193	3.891	0.840	0.002	0.961
	Online	1,255	3.888	.9430		
HRC02	OnSite	196	3.515	.9844	1.344	0.247
	Online	1,250	3.611	1.090		
HRC03	OnSite	197	3.467	1.003	0.051	0.821
	Online	1,250	3.486	1.138		
HRC04	OnSite	197	4.015	0.854	0.060	0.806
	Online	1,248	4.031	.849		
HRC05	OnSite	197	4.015	.785	0.182	0.669
	Online	1,249	3.988	0.839		
HRC06	OnSite	201	2.408	1.146	1.525	0.217
	Online	1,246	2.518	1.172		
HRC07	OnSite	203	2.586	1.060	5.072	0.024
	Online	1,244	2.795	1.2495		
HRC08	OnSite	202	2.629	1.109	1.560	0.212
	Online	1,246	2.742	1.204		
HRC09	OnSite	203	2.542	1.044	8.176	0.004
	Online	1,249	2.801	1.219		
HRC10	OnSite	201	3.269	1.076	1.125	0.289
	Online	1,247	3.362	1.166		
HRC11	OnSite	196	2.913	1.131	2.990	0.084
	Online	1,246	3.074	1.221		
HRC12	OnSite	194	2.778	1.100	5.890	0.015
	Online	1,252	3.008	1.245		

Conclusions

The main contribution of this work is to create an image enhancement quality database, which is large and diverse. The database contains 35 SRC images and 12 HRCs (different enhancement algorithms), yielding a total of 455 test images. A crowd-sourcing (online) psychophysical experiment was performed to obtain quality scores for these images. We also conducted a controlled laboratory experiment (onsite) and compared its results with the crowd-sourcing (online) experiment. This database can be used to train image quality metrics that can detect both increases and decreases in the perceived quality.

In consideration to the algorithms used in this work, sharpness (HRC04 and HRC05) and color enhancement (HRC01) improved the quality of the SRC images. For brightness (HRC02) and contrast enhancement algorithms (HRC03), the results were inconclusive, i.e. they may increase or decrease the image quality depending on the SRC image. The remaining algorithms apparently reduced the perceived image quality. Furthermore, two (HRC07-HRC08) out of the four (HRC07-HRC08-HRC10-HRC11) histogram equalization techniques produced enhanced images with the same quality as the SRC images.

When comparing the results of the onsite (laboratory) and online (crowd-sourcing) experiments, we found that for most of the HRCs there is no statistically significant difference between the DMOS given by participants in both experiments. Further studies are needed to identify hidden factors that may be influencing the perceived quality of enhanced images.

Acknowledgments

This work has been supported by the Federal District Research Support Foundation (FAP-DF), Brazil, by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and by the University of Brasilia (UnB), Brasilia, Brazil.

References

[1] ABDULLAH-AL-WADUD, M., KABIR, M. H., AKBER DEWAN, M. A., AND CHAE, O. A dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics 53*, 2 (May 2007), 593–600.

[2] CHANDLER, D. M. Seven challenges in image quality assessment: past, present, and future research. *ISRN Signal Processing 2013* (2013).

[3] GHADIYARAM, D., AND BOVIK, A. C. Massive online crowd sourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing 25*, 1 (Jan 2016), 372–387.

[4] HOSSFELD, T., HIRTH, M., KORSHUNOV, P., HANHART, P., GARDLO, B., KEIMEL, C., AND TIMMERER, C. Survey of web-based crowdsourcing frameworks for subjective quality assessment. In *2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP)* (Sep. 2014), pp. 1–6.

[5] HOSSFELD, T., KEIMEL, C., HIRTH, M., GARDLO, B., HABIGT, J., DIEPOLD, K., AND TRAN-GIA, P. Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing. *IEEE Transactions on Multimedia 16*, 2 (Feb 2014), 541–558.

[6] ITU-R. Recommendation 500-11: Methodology for the Subjective Assessment of the Quality of Television Pictures. *Int. Telecomm. Union, Geneva* (2002).

[7] ITU-T. Recommendation P.910: Subjective Video Quality Assessment Methods for Multimedia Applications. *Int. Telecomm. Union, Geneva* (2008).

[8] ITU-T. Recommendation P.913: Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment. *Int. Telecomm. Union, Geneva* (2016).

[9] KEELAN, B. *Handbook of image quality: characterization and prediction*. CRC Press, 2002.

[10] LAHOULOU, A., BOURIDANE, A., VIENNET, E., AND HADDADI, M. Full-reference image quality metrics performance evaluation over image quality databases. *Arabian Journal for Science and Engineering 38*, 9 (2013), 2327–2356.

[11] LUNDH, F. Pillow: The friendly pil fork, dec 2019.

[12] MATHWORKS. Image filtering and enhancement, dec 2019.

[13] MIKE. Enhancing photos with python, oct 2017.

[14] MITTAL, A., MOORTHY, A. K., BOVIK, A. C., CHEN, C. W., CHATZIMISIOS, P., DAGIUKLAS, T., AND ATZORI, L. No-reference approaches to image and video quality assessment. *Multimedia Quality of Experience (QoE): Current Status and Future Requirements 99* (2015).

[15] PINSON, M. H., JANOWSKI, L., AND PAPIR, Z. Video quality assessment: subjective testing of entertainment scenes. *IEEE Signal Processing Magazine 32*, 1 (2014), 101–114.

[16] PONOMARENKO, N., JIN, L., IEREMEIEV, O., LUKIN, V., EGIAZARIAN, K., ASTOLA, J., VOZEL, B., CHEHDI, K., CARLI, M., BATTISTI, F., AND KUO, C.-C. J. Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication 30* (2015), 57 – 77.

[17] REN, X., LI, M., CHENG, W.-H., AND LIU, J. Joint enhancement and denoising method via sequential decomposition.

[18] SEGHIR, Z. A., AND HACHOUF, F. Full-reference image quality assessment measure based on color distortion. In *Computer Science and Its Applications* (Cham, 2015), A. Amine, L. Bellatreche, Z. Elberrichi, E. J. Neuhold, and R. Wrembel, Eds., Springer International Publishing, pp. 66–77.

[19] VU, C., PHAN, T., BANGA, P., AND CHANDLER, D. On the quality assessment of enhanced images: A database, analysis, and strategies for augmenting existing methods. pp. 181–184.

[20] WANG, Z., AND BOVIK, A. C. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine 26*, 1 (2009), 98–117.

[21] WINKLER, S. Video quality measurement standards—current status and trends. In *2009 7th International Conference on Information, Communications and Signal Processing (ICICS)* (2009), IEEE, pp. 1–5.

[22] YING, Z., LI, G., REN, Y., WANG, R., AND WANG, W. A new image contrast enhancement algorithm using exposure fusion framework. pp. 36–46.

Author Biography

Muhammad Irshad *received his B.Sc degree in Computer Science form Bahauddin Zakariya University, Multan, Pakistan in 2009 and M.S. degree from Air University Islamabad, Pakistan, in 2015. Currently, he is a Ph.D. candidate in the Electronic Engineering and System Automation program of the University of Brasília, Brasília, Brazil. His research interests includes image and video quality assessment, multimedia contents and machine learning*

Alessandro R. Silva *received his B.Sc degree in Computer Science from Federal University of Goias (UFG), Brazil, in 2000, his M.Sc degree in Computer science from Federal University of São Carlos (UFSCAR), Brazil, in 2003. Currently, he is a Ph.D student on University of Brasilia (UnB), working on immersive applications and media. He is also an Associate professor at Federal institute of Goiás-Anápolis. His current interests include stereo and multi-view video quality metrics.*

Sana Alamgeer *received her B.Sc degree in Computer Science form Bahauddin Zakariya University, Multan, Pakistan in 2009 and M.S. degree from Air University Islamabad, Pakistan, in 2015. Currently, she is a Ph.D. candidate in the electrical engineering and system automation program at University of Brasília, Brazil. She is a researcher of the GPDS and her current interests include video quality metrics, visual attention, Light Field Image Processing and Deep Learning.*

Mylene C. Q. Farias *received her B.Sc. degree in electrical engineering from Federal University of Pernambuco (UFPE), Brazil, in 1995 and her M.Sc. degree in electrical engineering from the State University of Campinas (UNICAMP), Brazil, in 1998. She received her Ph.D. in electrical and computer engineering from the University of California Santa Barbara (UCSB), USA, in 2004 for work in no-reference video* *quality metrics. Dr. Farias has worked as a research engineer at CPqD (Brazil) in video quality assessment and validation of video quality metrics. She has also worked as an intern for Philips Research Laboratories (The Netherlands) in video quality assessment of sharpness algorithms and for Intel Corporation (Phoenix, USA) developing no-reference video quality metrics. Currently, she is an Associate professor in the Department of Electrical Engineering at the University of Brasilia (UnB). Her current interests include video quality metrics, video processing, multimedia signal processing, watermarking, and visual attention. Dr. Farias is a member of IEEE, the IEEE Signal Processing Society, ACM, and SPIE*

Appendix

Table 6: List of SRC images of the experiment, which were taken from the Challenge database [3], TID2013 [16] and CSIQ [18] databases.

SRC	Database Name	Name in Databases
SRC01	Challenge database	10
SRC02	Challenge database	17
SRC03	Challenge database	129
SRC04	Challenge database	50
SRC05	Challenge database	113
SRC06	Challenge database	138
SRC07	Challenge database	147
SRC08	Challenge database	152
SRC09	Challenge database	167
SRC10	Challenge database	173
SRC11	Challenge database	186
SRC12	Challenge database	255
SRC13	Challenge database	261
SRC14	Challenge database	271
SRC15	Challenge database	338
SRC16	Challenge database	344
SRC17	Challenge database	414
SRC18	Challenge database	442
SRC19	Challenge database	444
SRC20	Challenge database	452
SRC21	Challenge database	455
SRC22	Challenge database	500
SRC23	Challenge database	525
SRC24	Challenge database	527
SRC25	Challenge database	820
SRC26	CSIQ database	1600
SRC27	CSIQ database	boston
SRC28	CSIQ database	child swimming
SRC29	CSIQ database	trolley
SRC30	CSIQ database	woman
SRC31	TID2013 database	103
SRC32	TID2013 database	104
SRC33	TID2013 database	107
SRC34	TID2013 database	111
SRC35	TID2013 database	123