Image Quality Assessment of Displayed and Printed Smartphone Images

Gaurav Sheth, Katherine Carpenter and Susan Farnand, Munsell Color Science Laboratory, Rochester Institute of Technology, Rochester, NY, USA

Abstract

Smartphones have become ingrained in our daily activities, driving Smartphone cameras to become better with every generation. As more and more images are being taken by smartphones it has become increasingly important to assess the quality of the images taken by different phones. The Cell Phone Image Quality (CPIQ) Group created the IEEE P1858 CPIQ Standard. To subjectively validate the group's metric, psychophysical tests were performed; each tested observers' preferences for a wide range of images. While many smartphone images are only viewed electronically, many images also get transformed into printed images, especially photobooks, as digital printing gets better and cheaper compared to traditional printing processes. The main goals of this research were to evaluate the image quality of smartphone images, both electronically displayed and for several printers, and to compare print quality to displayed quality. The subjective results indicated that the perceived quality of images is well-correlated with the objective results of the IEEE P1858 CPIQ Standard. It was also found that smartphones have a bigger impact on the image quality compared to the digital printers.

Introduction

Smartphones are becoming the most widely used cameras today. Currently, there are no regulations or guidelines for smartphone cameras. The manufacturers have proprietary approaches to image quality, leading to a wide range of quality in the images taken by phones with different manufacturers. The idea behind developing the IEEE P1858 CPIQ standard [1] was to allow for a determination of image quality of the phones based on objective performance of image characteristics. Several metrics were combined to generate one single metric that is intended to relate to perceptual quality. A series of experiments were conducted to test the image quality of the pictures taken by a range of smartphones. The same images were then printed on different digital printers to study the quality of the images when transitioned from display to print. This study employed several psychophysical methods including two-alternative forced choice, quality ruler assessment and indirect scaling. [2]

Digital presses have become a go-to for printing photos taken from smartphones, as they produce very high quality reproduction of the digital images and keep the cost of printing down. The question that arises now is how smartphone camera quality translates to digital printing. Do the press parameters affect the perception of the images? The simplest way to answer these questions is performing a psychophysical study of printed images taken with various smartphones and printed on different digital printers. For comparison of printed samples, rank ordering is chosen over paired comparison, as the number of pairs required to be compared would be large and the experiment would become long and cumbersome. As the image quality is high and it is not easy to distinguish between the images, there is bound to be confusion among the observers. If there was minimum confusion, a direct scaling method would have been a better approach; since confusion was predominantly present, an indirect approach was chosen. Once the observers were shown the images their response was used to calculate the scale values. This information was also useful for indicating the reliability of the choices made by the observers.

Methodology

The study involved the use of nine phones viz. iPhone 4, iPhone 5S, iPhone 6S Plus, LG G2, Nexus 6P, Nokia 1020, Galaxy S7 Edge, HTC One M8, and Sony Z5. Four digital presses were used to print the images used in the study, with modifications to the print settings on one of the presses. The presses involved were Xerox IGEN5, Xerox Versant 2100, Xerox Color Press 1000i [3] and Shutterfly [4]. An image enhancement tool was also used to modify the images and these images were considered as a separate press. The device cameras were evaluated in a lab setting before they arrived at RIT and were used in the experiment.

The first step in the study was to generate the image set. To capture images with similar fields of view, phones with larger pixel heights were moved farther from the target. A look-up table (LUT) was generated of the approximate distances required between the target and the phone. The pixel height of an object in the field of view was measured in each image; the images where the pixel height of the object was consistent between phones were then chosen. The settings under which the captures were made were carefully controlled. The phones were mounted on a tripod in order to keep them stable and the flash was turned off. Although tripods are rarely used with smartphones cameras, they were employed to remove the photographer as a possible factor in this study. The level of illumination at each target was measured with a Minolta CL200 lux meter and recorded, Table 1.

Table1: Illuminance level for each image scene.

	Scene	Illuminance (lux)
1	Empty Restaurant	257.8
2	Flowers – Blue LED	731.7
3	Flowers – Outdoors	>99,999
4	George Eastman House	14450
5	Handicapped Sigh	>99,999
6	House – Night	15.9
7	Person in Garden	>99,999
8	Portrait - Fluorescent	1150
9	Portrait – Low Light	15.4
10	Portrait - Outdoors	>99,999
11	Portrait - LED	140
12	Sign – Night	16.1

The scenes used for the experiment were chosen in such a way as to include a variety of feasible scenarios in which a person might use their smartphone to take a picture. The scenes included buildings, flowers indoors and outdoors, a fruit scene, the George Eastman house, a handicapped parking sign, a house at night, and people under low light, fluorescent lighting and outdoors. Figure 1 shows the images that were chosen for the study. These images were also chosen in such a way as to roughly correspond to an existing image set that is used in Quality Ruler experiments [6].

There was some variability in facial expressions between images of the models taken with the different phones. People are more likely to rate an image of a person smiling more positively than an image of a person frowning. Because of this, participants in the psychophysical experiments were given specific instructions to ignore the facial expressions of the models and to judge the image only on overall quality.

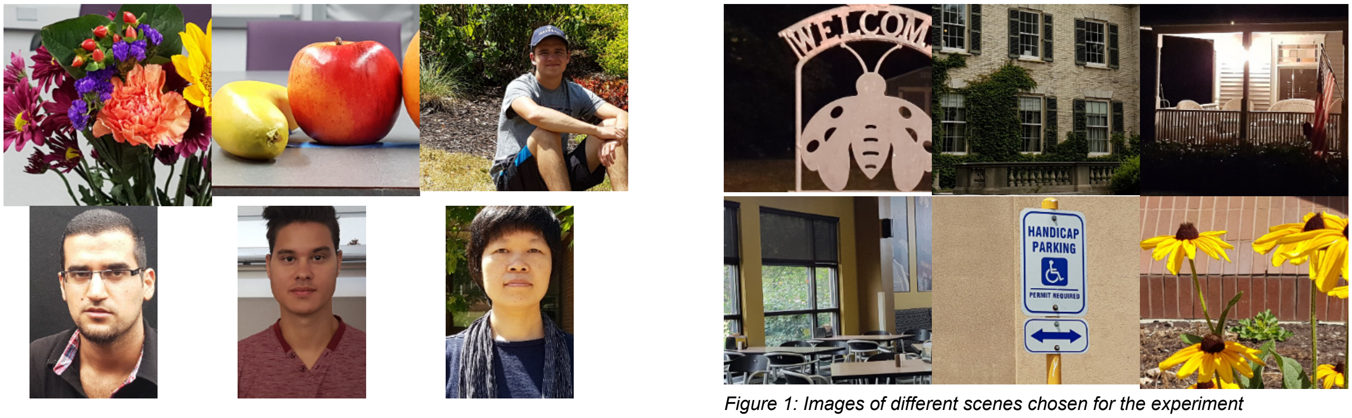


Figure 1: Images of different scenes chosen for the experiment

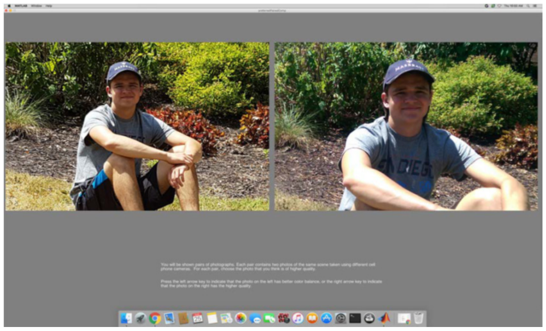
Experimental Setup and Participant Screening

The images were cropped in Adobe Photoshop to 1253 x 834 pixels for the landscape-oriented images and 834 x 1253 pixels for the portrait-oriented images. The images were then placed into a MATLAB graphical user interface (GUI). An HP ZR30w display was used for the experiment. Viewing conditions specified in the Quality Ruler experiments were used [6]. The wall behind the monitor, painted gray, was illuminated and the background of the GUI was set to a similar gray to make the participant's field of vision as uniform as possible. Gray paper was placed over the desk on which the display was placed for the same reason. The participants were seated 85 cm from the display. This distance was used as it what is required by the Quality Ruler methodology to make the sharpness scale of the anchor images perceptually uniform. Because the quality ruler test depends on visual acuity, for this test, the participants were required to use a chin rest to keep their heads stable and at a fixed distance from the display. All participants had normal color vision, as tested with an Ishihara plate test, and normal or corrected to normal visual acuity.

For the printing experiment, the presses that would produce the best image quality, like that needed for photobooks, were selected. Shutterfly is a commercial press; the Shutterfly layout was used to resize the images to be the same aspect ratio that is available on their website [4]. For the other presses (IGEN, Color press 1000 & Versant 2100), Adobe InDesign was used to make the raw files of the different images with same aspect ratio as Shutterfly. The page size used was 8.5" x 11" in landscape mode. Each page contained four images with the size of 4.15 in X 5.4 in each. The paper used for all presses was 200 gsm coated stock. Images from phone 7 of the display experiment were not used in the print experiment, making phones 8 and 9 in the display experiment phones 7 and 8 in the print experiment. Images were put in randomized order and each page had four images of the same scene, so a two-page spread contained all eight images of each scene. The background was chosen to be neutral gray across all prints.

Perceptual image Quality evaluation on an electronic display: Two-Alternative Forced Choice Experiment

The method of paired comparison used here is described in Engeldrum, 2000 [5]. For this experiment, each image was compared to all other images from the same set. The participants were given consistent instructions. They were told to ignore the facial expressions in the images and to only focus on the quality of the image. For the paired comparison experiment, the participants were instructed to choose the image that they felt was of superior quality. Quality was specified to be the overall result of the color balance, sharpness, noise, noise or graininess, and uniformity.



You will be shown pairs of photographs. Each pair contains two photos of the same scene taken using different cell phone cameras. For each pair, choose the photo that you think is of higher quality.

Press the left arrow key to indicate that the photo on the left has better color balance, or the right arrow key to indicate that the photo on the right has the higher quality.

Figure 2: GUI for Paired Comparison

Perceptual image Quality evaluation on an electronic display: Quality Ruler Experiment

The method of quality ruler assessment was developed in previous research [6] [7]. The participants placed their heads on a mounted chin rest. They were presented with two images and a slider bar. The test images were paired with images from the Quality Ruler set. The anchor images, displayed on the left side of the screen, were from the Quality Ruler image set. The test images were displayed on the right side of the screen. The slider bar adjusted the sharpness of the anchor image. With this, the participants were instructed to use the slider bar to adjust the anchor image to be of equal quality to the test image. The scenes were run one at a time, but the order of the images within each set was randomized.

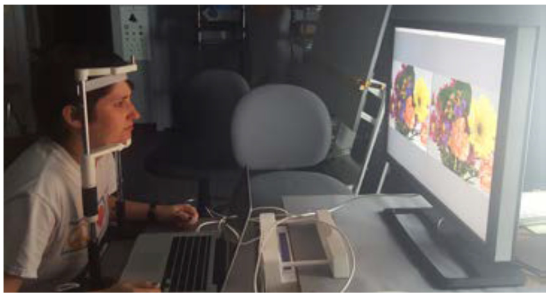


Figure 3: Experimental set up for the paired comparison

Perceptual image Quality evaluation in print: Rank Order Experiment

The images were presented in a light booth under a D65 light source. Twenty-four observers participated in the experiment in the Visual Perception lab in the Color Science Hall at RIT. (However, because the experiment proved to be rather lengthy and tiring for the observers, and because it was deemed that evaluating the effects of image enhancement was not of critical importance for this study, only 11 observers participated in that part of the experiment.) The observers ranged in age from 23 to 64, included 10 males and 14 females. 10 of the participants were Color Scientists and 6 people were from imaging backgrounds; the rest did not have imaging experience. Answer sheets were provided to the observers that were copies of the images being viewed but at 1% transparency of the originals, created in Adobe Photoshop©. This was done so that the answer sheet replicated the images being assessed. The observers were told to write their preference for each image on the answer sheet, going from 1 to 8 with 1 being the most preferred and 8 being the least preferred.

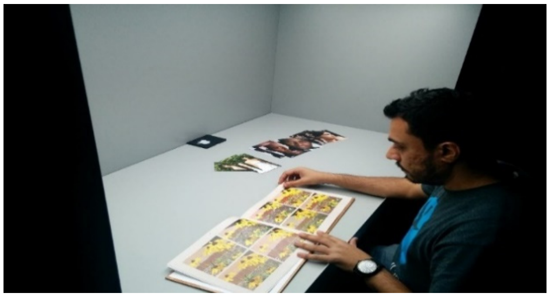


Figure 4: Experimental setup for Rank Order experiment

Results and Discussion

Display

The paired comparison results were, in general, expected: the newer, higher-end smartphones tended to perform better than older or cheaper models. However, it was also seen that the performance of each phone varied depending on the contents and lighting of the scene.

The quality ruler experiment found much the same results, indicating that the results from both experiments were consistent and valid. The mean correlation coefficient between the paired comparison and quality ruler experimental results for all 10 scenes was 0.893. Figure 5 shows the correlation coefficients for the Quality Ruler relative to the Paired Comparison results by scene. For 7 of the 10 scenes, the correlation coefficient is greater than 0.9. The results for the three remaining scenes (Outdoor Flowers, Person in Garden, and Portrait-LED) are also highly correlated, when Phone 8 was eliminated. This phone camera produced images that tended to differ in color from the images from other devices, which may have been impacted relative performance when they appeared as a single test images in the ruler experiment as opposed to being compared to other renditions of the same scene. For example, the brick in the outdoor flowers image was less reddish (a* value of about 9 versus 22 for the highest rated image as calculated in a four brick area of the scene in Matlab®) and, therefore, less chromatic than in the other renditions. When comparing the images directly, this may have been an important factor in its being rated the worst of the images. For the Portrait and Garden Person scenes, the skin tones were more reddish, which may be been appealing when seen one image at a time, but overdone when viewed in comparison to other images. If color is causing the shift in how the images from Phone 8 are being rated, this may suggest that the Quality Ruler approach may be more representative of third person evaluation, where less information regarding the original scene is available as compared to the paired comparison approach.

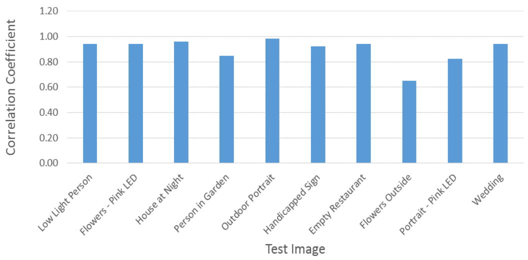


Figure 5: The correlation *coefficients between the paired comparison and quality ruler results for each scene*

With the knowledge that the results of both subjective psychophysical experiments are well-correlated with each other, attention then turned to the correlation between the objective and subjective results. The objective and subjective data were used to determine quality loss (QL) values in just noticeable difference (JND) units [1]. The objective QL values for each scene came from seven individual metrics: spatial frequency response, lateral chromatic displacement, chroma level, color uniformity, local geometric distortion, visual noise, and texture blur [7]. The Quality Ruler results are shown in Figure 6 relative to the JNDs generated from the probability values in the paired comparison test. These results show a high level of agreement between the two experiments for most of the phones, except Phone 8.

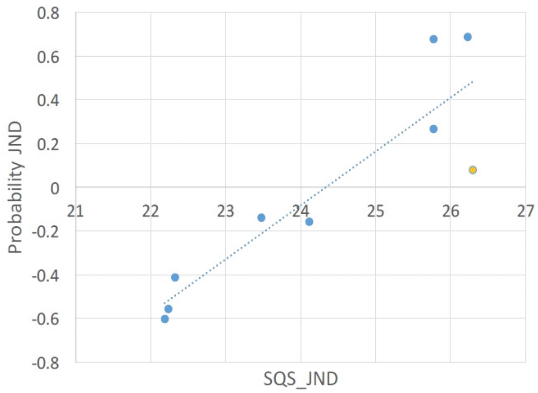


Figure 6: Paired comparison results relative to the Quality Ruler results averages over all scenes. Phone 8 is shown in orange

The subjective results were then compared to objective results as determined using the CPIQ metrics. As part of this, the ten scenes from the quality ruler experiment were split into three lighting categories: night/dim light, indoor, and outdoor scenes. Once QL values were obtained for the three lighting conditions, the subjective QL values were plotted versus the objective QL values, Figure 7.

The data is relatively well correlated, with a correlation coefficient of 0.83. This suggests that the objective metric is a reasonable estimation of actual perceived quality. In general, there is rough agreement between the QL values from the objective metrics and the subjective preference data. It is important to note that both outliers in Figure 7 are night scenes indicating that the objective metrics are overestimating the image quality loss. It will be of interest to investigate the reason behind it.

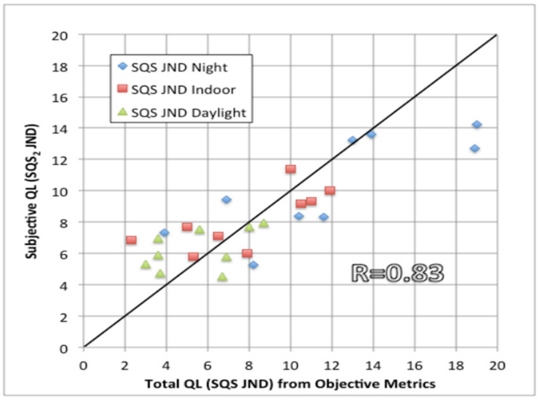


Figure 7: A plot of the correlation between objective and subjective results for each lighting category and all phones. [1]

Print

Printed samples were analyzed using Thurstonian Methodology described in [5]. The z-scores were adjusted to be all positive by adding a value of 1.5 for ease of viewing.

The bar graph in Figure 8 shows average preferences across all the presses. Note that the specific phones are not identified because this experimentation was evaluating the efficacy of the objective metrics relative to subjective results, not performance of a specific smartphone camera. The interval scale values were rescaled to remove negative values. Images from Ph1 were preferred across all presses, whereas images from Ph3 garnered the least preference. The performance for individual presses for each image set matched the overall performance; Ph1 was the most preferred and Ph3 was the least preferred. Phones 2, 5, 6 & 7 have similar performance across the board.

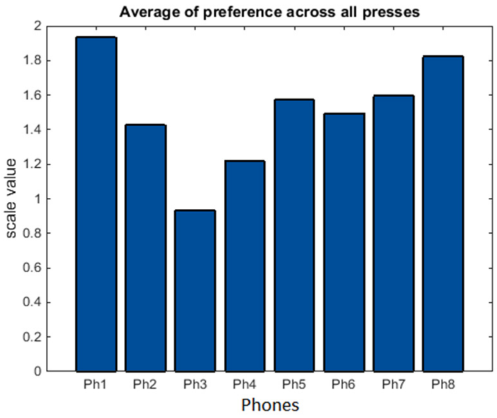


Figure 8: Preference plot of the images across all presses

Figure 9 shows the results by image with error bars calculated using the standard deviation and confidence interval for press 1. These plots correlate the data seen in the bar plot, with ph1 being the most preferred and ph3 being least preferred. The error bars describe the variability between observers. The error bars are scaled by a factor of 1.68 due to simulations following the Montag method but using rank order instead of all possible pairs (as used in paired comparison) [9].

The confidence interval of 95% defines a range in which the population estimate can be found 95% of the time if the experiment is repeated and results are recalculated. The observed standard deviation (empirical data fit) was estimated for the print data by using the following equation:

$$\sigma = 1.68(b_1(n-b_2)^{b^3}\ (N-b_2)^{b^5}\ )\ [9]$$

$$b_1 = 1.76, b_2 = 3.08, b_3\text{=-0.613}, b_4 = 2.55, b_5\text{= -0.491}$$

N is number of observer, n is number of stimuli
Confidence interval by following:

$$95\%\ CI = 1.96\ \sigma\ [10]$$

The confidence interval was calculated using the Montag method; for press 1, this value is 0.0356. The errors bars for CI are smaller and consistent and the error bars for the standard deviation are larger. Standard deviation shows the variation among observers for each image set and thus the variance for each image set can be calculated. The variance for press 1 is 0.008, implying that the observers did not vary much over their preference.

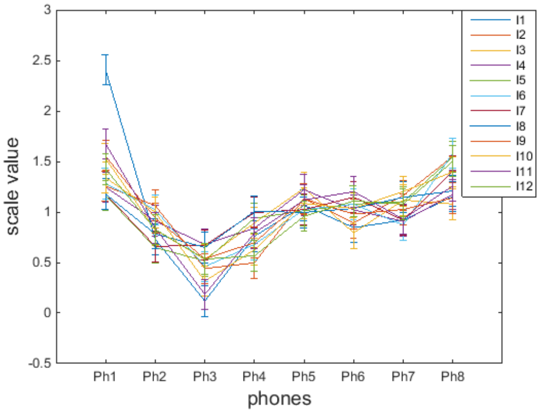
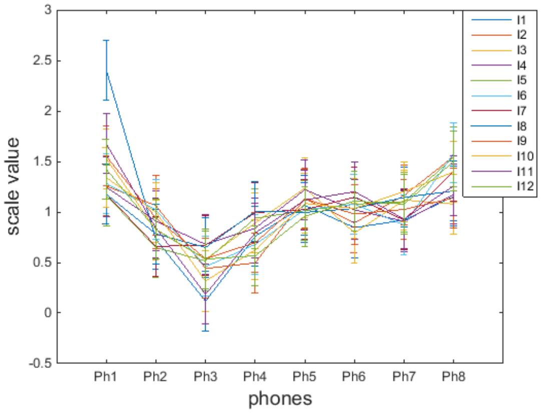


Figure 9: Errors bars with Standard Deviation



Errors bars with Confidence Interval

Preference of each phone across all presses shows its the performance. Figure 10 shows the preference plot of average performance of phones on different presses. It is observed that preference ratings are similar irrespective of the press used. This means the image quality from the phones translates well to printed image quality. Phones have bigger impact on the image quality compared to presses as the preference for each phone is similar irrespective of the presses on which they were printed. Ph1 and ph8 have high preference scale, whereas ph3 has the lowest preference across the board.

Image enhancement was used on one of the presses as a parameter for printing. All the images were treated in the Raster Image Processing (RIP) to improve contrast, sharpness, color tone, and to reduce noise and uneven tones before printing. Figure 11 shows the preferences for the press with and without image enhancement. Some of the enhanced images were highly preferred but others were affected negatively by the enhancement, making their reception poor. Images from Ph4 were not received well after enhancement whereas images from Ph1 got better preference. One of the contributing factors in this may be that only 11 observers performed the experiment with the Enhanced image set whereas 24 observers performed the default image set experiment. Relatively little change is observed due to enhancement in other phones.

The study involved three portrait pictures and nine landscape pictures. In comparing the average landscape scale values of landscape pictures to the average portrait pictures, it can be seen that the orientation of the pictures does not impact the preference scores. The bar plots of both the modes shown in Figure 12 validate the average bar plot that was observed for the overall experiment shown in Figure 8. In both cases, ph1 and Ph8 are still the most preferred and ph3 and ph4 are still the least preferred.

The CPIQ group studied the performance of smartphones images by scene light level. Low light levels make the images grainy and noisy. Newer smartphones try to compensate by applying noise reduction processing to improve the image quality. Ph1, Ph6, Ph8 generally captured the low light level scenes better than other phones. Phones 3 and 4 suffer under low light thus their preference rating is lower for low light images, though all these changes are minimal. The bar plot in Figure 13 shows the comparison between averages of low light levels compared to images with high illumination.

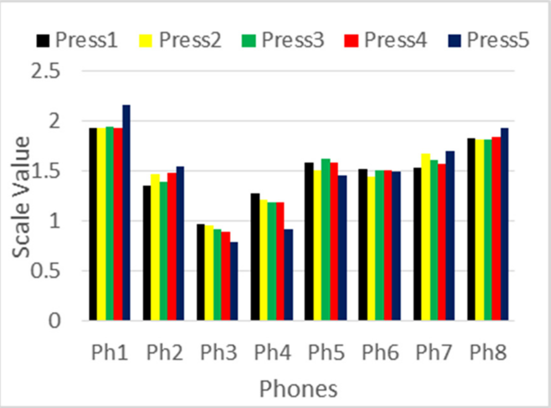


Figure 10: Average performance of each phone across each press.

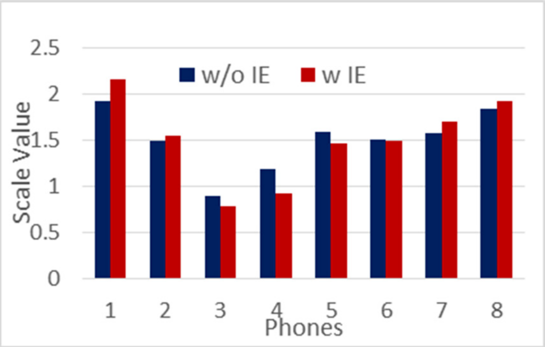


Figure 11: Preference change due to image enhancement

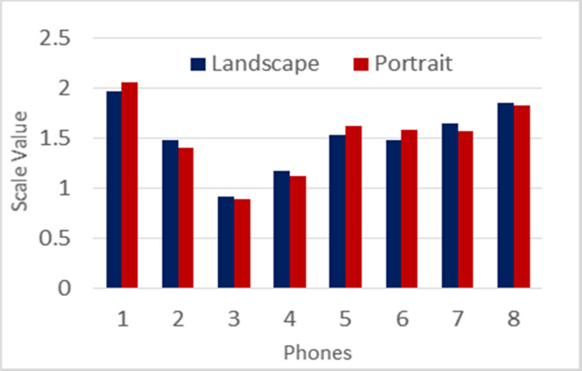


Figure 12: Average of landscape images against portrait images

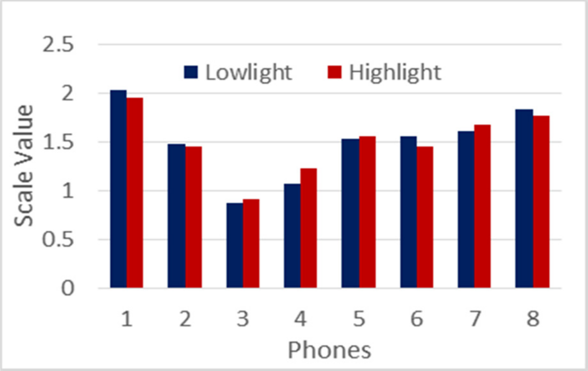


Figure 13: Average of Low light versus Highly illuminated images

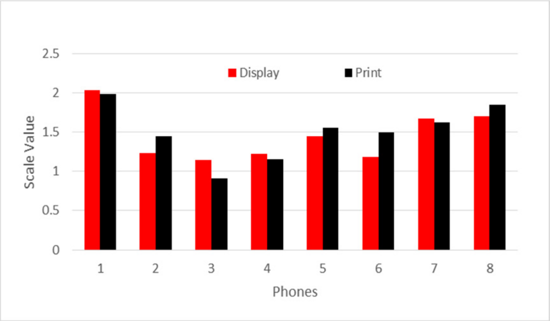


Figure 14: Average ratings of displayed and printed images

Display Relative to Print

The transition to printed images from displayed images retains the image quality, Figure 14. There is a high correlation between displayed and printed samples with a correlation coefficient of 0.86. Mostly displayed samples have very similar preferences to printed samples but as it is an indirect scale the preference for display is not necessarily higher than print. The important thing to note is that the most preferred phone images in display remained the most preferred in print and vice versa. Ph3 had the least amount of noise reduction, which translated to grainy printed samples, thus lowering their preference. Also, ph3 had lower resolution as compared to other phones, which decreases sharpness the displayed images making them undesirable.

Conclusion

Smartphones have been become a part of daily lives. The increasing number of images captured by smartphones generates the necessity to measure the image quality of smartphones. Digital presses have become a go-to for printing photos taken from smartphones as they produce very high quality reproduction of the images and keep the cost of printing down. Digital presses having a high perceived image quality are best suited for variable printing that goes hand in hand with the small number of prints of images that individual smartphones generate.

A subjective validation of an objective image quality metric, IEEE P1858, created by the CPIQ Standard group, was undertaken by the Munsell Color Science Laboratory as part of a group assessment in conjunction with six other labs. Subjective data on perceived image quality was gathered through a paired comparison experiment and a quality ruler assessment. The subjective data was then compared with the objective quality loss data. The comparison found the objective data and subjective data to be relatively well-correlated. This suggests that the CPIQ

Standard can be used to judge the perceived image quality of a phone based on objective measurements. The CPIQ group plans additional work on these metrics to further improve their efficacy in predicting perceived image quality.

The data shows that ph1 and ph8 have the highest preference across all digital presses ph3 and ph4 have the least preference. Ph5, ph6 and ph7 have similar preferences which are towards the high preference scale. The image quality of captured images translates with almost one to one correspondence to the printed samples with a correlation coefficient on 0.86. Image enhancement viz. noise reduction, tone correction etc. was used as a parameter for one of the presses. Low light images were more preferred after enhancement but other images lost their natural look decreasing their performance. The smartphones had a higher impact on image quality than the different digital presses. This drives the image quality observed in the printed samples as well. Thus smartphones affect image quality more than the presses themselves.

Acknowledgements

The authors wish to thank Xerox Corporation and IEEE for supporting this research and as well as all of the observers that participated in the experiment. Additionally, many thanks to Michael Murdoch for his guidance on adapting the Montag method for rank ordered results.

References

Jin, E., Phillips, J., Farnand, S., Belska ,M., Tran, V., Chang ,E., Wang, Y. and Tseng, B. “Towards the Development of the IEEE P1858 CPIQ Standard” – A validation study, Electronic Imaging, Image Quality and System Performance XIV, pp. 88-94(7).

G. Gescheider, “Psychophysics: the fundamentals” (3rd ed.), 1997

Xerox Corporation, "Digital Printers," [Online]. Available: https://www.xerox.com/en-us/digital-printing/technology/color-production-printer.

Shutterfly [Online]. Available: https://www.shutterfly.com/

P. Engeldrum, “Psychometric Scaling: A Toolkit for Imaging Systems”. Massachusetts: Imcotek Press, 2000.

Jin, E. W., Keelan, B. W., Chen, J., Phillips, J. B., and Chen, Y., “Softcopy quality ruler method: Implementation and validation,” Proc. SPIE 7242, 724206, 2009.

Jin, E. W., & Keelan, B. W., Slider-adjusted softcopy ruler for calibrated image quality assessment. Journal of Electronic Imaging, 19(1), 011009-011009, 2010.

L. Thurstone, “A law of comparative judgment”, 1927.

E. Montag, "Empirical formula for creating error bars for the method of paired comparison," Journal of Electronic Imaging, 2006.

G. P. Thomas Brown, "An enquiry into the method of paired comparison: reliability, scaling, and Thurstone's Law of Comparative Judgment," 2009.