©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.13.IQSP-202

A methodology for perceptual image quality assessment of smartphone cameras

Susan Farnand[1], Young Jang[2], Chuck Han[2], and Hau Hwang[2];
[1]Rochester Institute of Technology, Program of Color Science, Rochester, NY
[2]Qualcomm, San Diego, CA

Abstract

Having a methodology for assessing smartphone camera image quality is advantageous for both those who design and develop the cameras as well as those who use them. Camera engineers need to quickly and reliably assess the impact of the system decisions they make. Smartphone customers who are armed with a quantitative understanding of the image quality can include this information to make informed decisions between products. This research project was undertaken to develop a procedure for evaluating pictorial image quality for smartphone camera captures. Experiments were conducted to evaluate tone quality, color quality, and sharpness and noise using images captured with 20 cameras that were released primarily in the period between 2012 to late 2014. A variety of scenes were captured with each device. In each test, observers rated the test images for overall quality and then for a specific image quality characteristic using an anchored scaling experimental protocol.

The results indicated high correlations between the individual characteristics and overall quality. It was also determined that high correlations could be achieved between the visual results and objective measurements for sharpness and noise. Both analyses indicated that a two-step process in which devices are first sorted into categories of high and low quality followed by a second sort to further refine device quality may be required to successfully predict the visual results.

Introduction

It would be difficult today to visit almost any public space or event without seeing an array of smartphone cameras pointed at the scenery, the visitors, and the picture takers themselves. The process of picture taking has become so simple that anyone who can press a button can take a picture. Smartphone camera systems employ a range of automated systems including auto-focus, auto-exposure, auto white balance, and tone compression for high dynamic range scenes that aid smartphone users in capturing acceptable images. This degree of automation has resulted in smartphone cameras that range significantly in the quality of the images that are produced. Having a methodology for assessing image quality is essential for camera engineers who need to understand the impact of the system decisions they make – to understand the trade-offs they make between cost, speed, and sensor size and image quality. Such a methodology is also useful for smartphone customers who would like to have a quantitative understanding of the image quality of different smartphone products so that they can make informed decisions based on cost, functionality, and image quality. To develop such a methodology, the VICTOR (Visually Integrative Camera Test and Open Report) project was undertaken. This research project comprised a series of experiments aimed at developing a procedure for evaluating pictorial image quality, particularly for smartphone camera captures.

Evaluating pictorial image quality is a complex process that ISO Standards organizations have been attempting to tackle for many years. (For example, Farnand et al., 2006; Phillips and Christoffel, 2010) To make the problem more tractable in this study, image quality characteristics were first assessed individually, both with respect to the given characteristic and to overall perceived image quality. Sharpness, noise (uniformity of image areas that are intended to be uniform), color, and tone reproduction were investigated, each using an array of pictorial scenes.

Experimental Methodology

Three experiments were conducted: one for evaluating tone quality, one for color quality, and one for sharpness and noise. In each segment, observers were asked to first rate the test images for overall quality. Then they were shown the images a second time in random order and asked to rate the images for the specific image quality characteristic. In the third segment, the test images were shown to the observers three times. In addition to overall quality, they were also asked to rate noise and sharpness with half of the observers rating noise images first while the other half rating sharpness first. Twenty different observers from the RIT campus environment participated in each segment of the experiment. The gender, ethnicity, approximate age, and area of study of all observers were recorded. All observers who participated in the experiment had normal color vision and normal or corrected to normal visual acuity.

Twenty cameras were tested. These were devices that were generally released in the 2012 to Fall of 2014 timeframe, although there was one device from 2010. A variety of scenes were captured with each device for each of the experimental segments, Figures 1-3. In the segment in which tone was evaluated, Figure 1, scenes with high dynamic range, both brightly and dimly lit were included. Most of the scenes were captured outdoors although one indoor scene was included. The scenes in the color quality evaluation segment, shown in Figure 2, included humans, sky, wood, and food. Four of the scenes were captured indoor and three outdoor. Most scenes contained highly chromatic colors, although one scene did not. Four scenes were used in rating image noise, Figure 3. These included a scene with a large area of sky, two indoor, low-light scenes with and without flash, and a night scene. Five scenes were used in the assessment of sharpness (Figure 3), three that were also used in the noise evaluation (tower against the sky and the two indoor scenes) along with another, more brightly

DOI: 10.2352/ISSN.2470-1173.2016.13.IQSP-202

lit indoor scene and an image of the carved façade of a stone building.

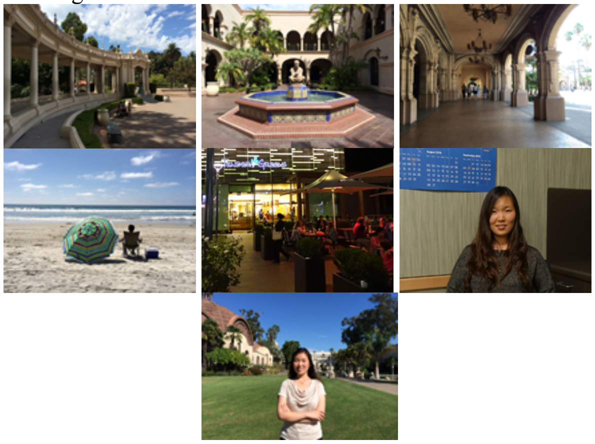

Figure 1 *Scenes used in the Tone Quality experiment, from top left – Colonnade, Courtyard, Piazza, Beach, Night, Low Light, and Memory Colors*

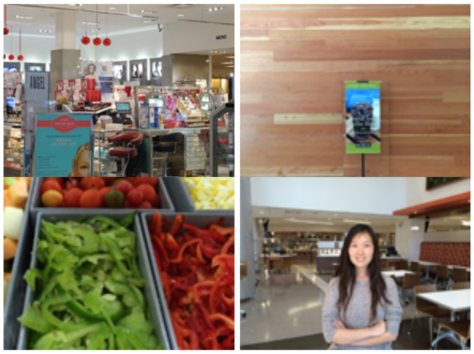

Figure 2 *Scenes used in the Color experiment, from left – Mall, Wood, Peppers, and Mixed Light. These were used in addition to the Courtyard, Beach, and Memory Colors scenes (Figure 1)*

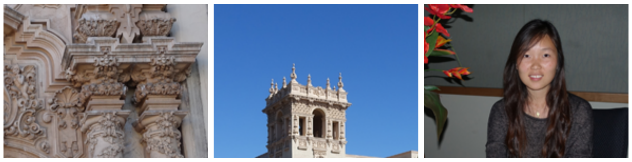

Figure 3 *Additional scenes used in the Sharpness and Noise experiments, from left – Sculpture, Tower, and Flash. The scenes used in the Sharpness experiment were Mall, Sculpture, Tower, Flash, and Low Light. Those used in the Noise experiment were Tower, Flash, Low Light, and Night.*

An anchored scaling experimental protocol was used (Engeldrum, 2000). This approach was chosen due to the potentially large differences in quality expected between some of the stimuli being evaluated. (Although a paired or triplet comparison approach is easier for the observers, if two stimuli are reliably different, no estimation of the magnitude of that visual difference is provided. Other choice would be Quality Ruler approach, which is quite powerful but requires most complex preparation to produce reliable results.) The observers were asked to scale the test stimuli relative to two reference images. The experimental set up is shown in Figure 4. A higher quality image was placed on the right of the display and a lower quality image on the left. The lower quality and higher quality anchor images were arbitrarily assigned the values of 30 and 75, respectively. The observers were approximately 18" from the HP_ZR30w display with their eyes aligned with the center of the display. They are instructed to type in their assigned values. These values appeared on the top right of the display. Each of the observers were presented with four training images and requested to rate these images as they would in the actual experiment. When the observers expressed comfort with the experimental process, the room lights were extinguished, and the testing was initiated.

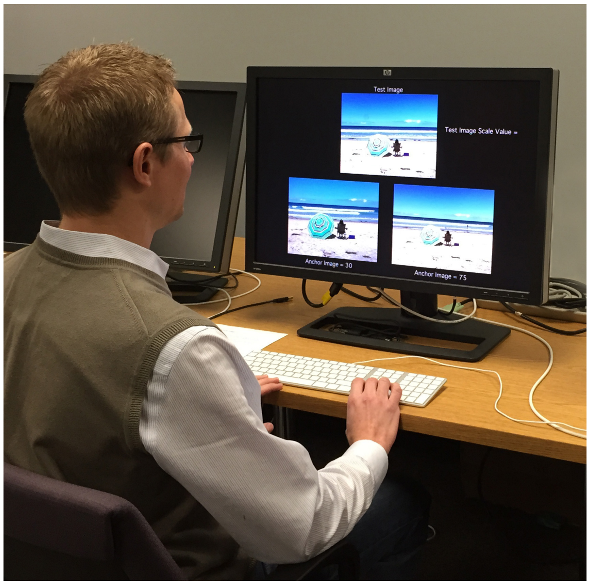

Figure 4 *Experimental setup. The room lights were turned off for the duration of the experiment.*

Results and Discussion

The experimental results included assessment of the overall ratings of image quality relative to the ratings for the individual characteristics, Tables I & II. These results showed relatively high correlation. There were a few exceptions. The correlation between tone and overall quality for the Beach Umbrella scene was lower than other scenes in the tone assessment segment. The correlation for Color and Overall Quality segment, however, was relatively high. The color quality of the sky and sand was more critical to the assessment of overall image quality than the tone quality, for this scene, which lacked substantial areas of shadows and highlights. Also, for the three scenes that were included in both tests, observers generally found that a test image with high tone quality had high color quality, though there was less agreement for the 'Beach' scene than the 'Courtyard' and 'Memory Color' scenes.

The correlation in the Color Quality test for the Peppers scene was much lower than other scenes. If three images, which were all poorly focused, are removed from the analysis, the correlation coefficient increases to 0.96. This demonstrates the impact of sharpness on, and the complex nature of, perceived image quality.

The correlations between Overall Quality and both Sharpness and Noise Quality ratings were generally very high, Table II. The highest correlation for any of the individual characteristics was for sharpness. The correlation for the Noise assessment for the Tower scene is much lower than the rest. This indicates that, for this scene, sharpness was more important for driving the perception of

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.13.IQSP-202

overall quality than noise. Also, there was no correlation (Correlation Coefficient = 0.03) between the overall quality rating of the Mall scene in the Color experiment relative to its cropped counterpart in the sharpness and noise experiment.

Table I: The correlation coefficients for the Tone quality and Color quality ratings relative to the overall quality ratings along with those of the tone versus color quality ratings for the Courtyard, Beach, and Memory Colors scenes

Scene	Tone v Overall	Color Quality Scene	Color v Overall	Tone v Color
Colonnade	0.98	Mall	0.94	
Courtyard	0.92	Courtyard	0.98	0.89
Piazza	0.91	Wood	0.98	
Beach	**0.82**	Beach	0.93	0.85
Night	0.95	Peppers	0.80 (0.96)	
Low light	0.91	Mixed light	0.90	
Memory Colors	0.95	Memory Colors	0.90	0.90

Table II: The correlation coefficients for the Sharpness quality and Noise quality ratings relative to the overall quality ratings along with those of the Sharpness versus Noise quality ratings for the Tower, Flash, and Low Light scenes

Scene	Sharp v Overall	Noise v Overall	Sharp v Noise
Mall	0.99		
Sculpture	0.97		
Tower	0.96	0.82	0.75
Flash	0.97	0.93	0.92
Low Light	0.97	0.90	0.91
Night		0.90	

Along with the subjective testing, objective measurements were made using the TE42 test target captured under several lighting conditions. (Altmann, 2015) The objective evaluation metrics included Visual Noise measured using OECF charts, Delta L noise, Resolution vMTF from Siemens star charts, Texture vMTF from 'Dead Leaves' charts, and Edge vMTF from slanted edge targets. The subjective ratings were evaluated with respect to these objective measures.

It was found that the most accurate predictions of perceived sharpness quality were made using a weighted combination of the Resolution vMTF and Texture vMTF measurements values made under D65 20 lux lighting conditions, Figure 5. For apparent noise quality, Visual Noise (with additional weight to the dL component) measurements made under D65 700 lux and D65 100 lux provided the closest predictions of the visual results, Figure 6. Note that there was one device with a much lower noise rating (.82) than was warranted by the subjective rating (40). The images captured by this device under low light conditions were generally over-exposed and had a substantial amount of noise reduction applied. While the level of noise apparent in these images was generally low, observers may have been unable to overlook the other obvious artifacts when making their assessments.

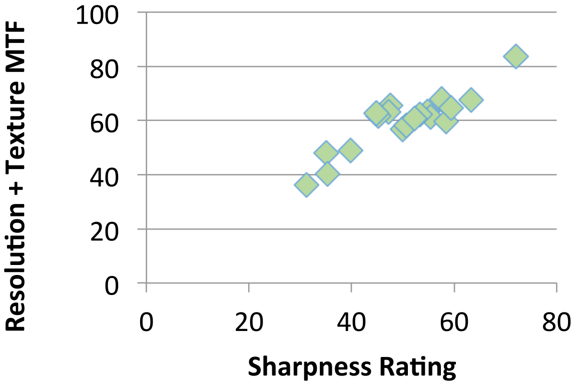

Figure 5 *Objective versus subjective ratings for sharpness. The objective measure is a weighted combination of the Resolution and Texture MTFs captured under a D65 20 lux light source.*

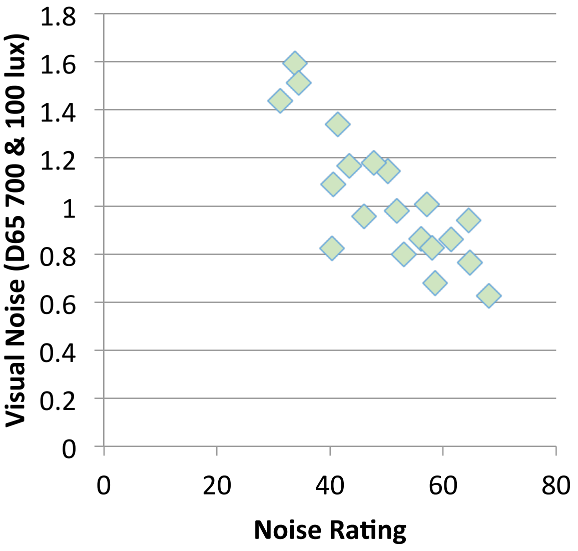

Figure 6 *Objective versus subjective ratings for image noise. The objective measure is a weighted combination of the Visual noise values each captured under both D65 700 lux and D65 100 lux light sources.*

A verification experiment was conducted to determine if results consistent with the initial experiment would be achieved with a smaller number of observers and with different devices. The experiment included twelve devices, eight that had been used in

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.13.IQSP-202

the initial experiment and four untested devices. The same three segments were conducted, each with four observers. Since new devices were added to the experiment, the scenes were reshot. Six scenes were used for each segment of the experiment.

The results indicated similar relationships between perceived overall quality and quality for the individual image quality characteristics, Tables III & IV and Figure 7. The correlations were highest for sharpness relative to overall quality. Tone, color and noise all had similar correlation levels with overall quality.

Table III: The correlation coefficients for the Tone quality and Color quality ratings relative to the overall quality ratings in the Verification experiment

Tone Quality Scene	Tone v Overall	Color Quality Scene	Color v Overall
Colonnade	0.70	Courtyard	0.91
Courtyard	0.94	Wood	0.91
Piazza	0.79	Beach	0.87
Night	0.91	Peppers	0.93
Low light	0.88	Mixed light	0.80
Memory Colors	0.79	Memory Colors	0.70

Table IV: The correlation coefficients for the Sharpness quality and Noise quality ratings relative to the overall quality ratings in the Verification experiment

Scene	Sharp v Overall	Noise v Overall	Sharp v Noise
Mall	0.91		
Sculpture	0.92		
Tower	0.96	0.81	0.73
Flash	0.97	0.91	0.82
Low Light	0.88	0.89	0.89
Night		0.88	

The relationships between the subjective and objective results were similar to those determined in the initial experiment. Predictions of sharpness made using D65 20 lux values again provided the highest correlation with the visual results, Figure 8. In this experiment, it was also noted that the predictions made with the D65 20 lux values accurately sorted the devices into low, medium, and high quality. With this pre-sorting, the D65 20 lux and D65 700 lux Edge MTF measurements accurately predicted relative performance of the three higher-end devices. Note that there were only three devices in this experiment and only one high-end device in 20-device experiment, so the overall body of data is extremely small. This approach of pre-sorting by sharpness may be useful in an overall image quality assessment methodology since results in the initial experiment indicated that sharpness was a stronger driver of quality than noise and color. (Keelan suggests using a Minkowski metric, which weights the characteristic that is furthest from optimal the most. Keelan, 2002.) In their study on Full Reference image quality assessment Larson and Chandler [Larson and Chandler, 2010] suggested that observers use different viewing strategies when assessing lower and higher quality images. They found that, for higher quality images, observers search images for distortions or artifacts, while for images having obvious distortions or artifacts, observers assess quality more globally. Given this finding of diverse strategies of image quality evaluation, constructing a methodology for predicting perceived overall image quality may benefit from a pre-sort into lower and higher quality devices to reflect this shift in assessment strategies.

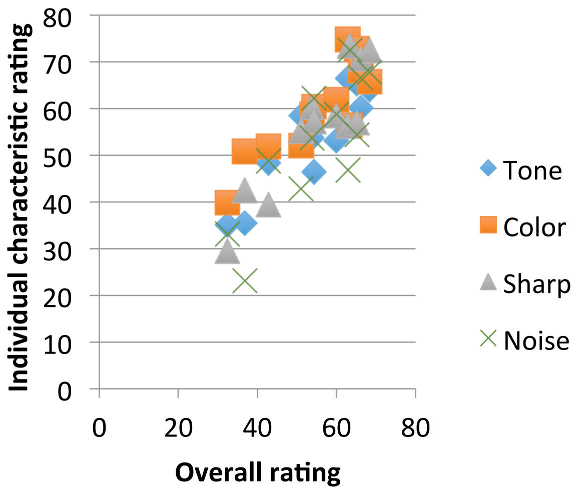

Figure 7 *Subjective ratings for individual image quality characteristics relative to the overall quality ratings for all scenes captured using each of twelve phone camera devices.*

For the predictions of perceived noise quality, Figure 9, again Visual Noise, with extra weighting to the dL component, made at high and D65 100 lux capture conditions performed well. In this instance, however, measurements were made at D65 2000 lux. These values slightly out-performed the predictions made with D65 700 lux values, as were used in the initial experiment. One device (the same one as in the original experiment) yielded a lower subjective rating for noise than the objective measurement would suggest, likely due to smoothing artifacts resulting from excessive 'noise cleaning'.

Initial evaluation of the color ratings relative to objective data was undertaken. Results were not as conclusive. The subjective ratings were compared to color, lightness, hue and chroma differences for the objective target capture. Correlations were poor. Relatively high correlations were achieved if the subjective ratings were compared to dominant colors in the scenes: for the green of the Peppers scene or the paneling color for the Wood scene, for example. Further work is underway.

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.13.IQSP-202

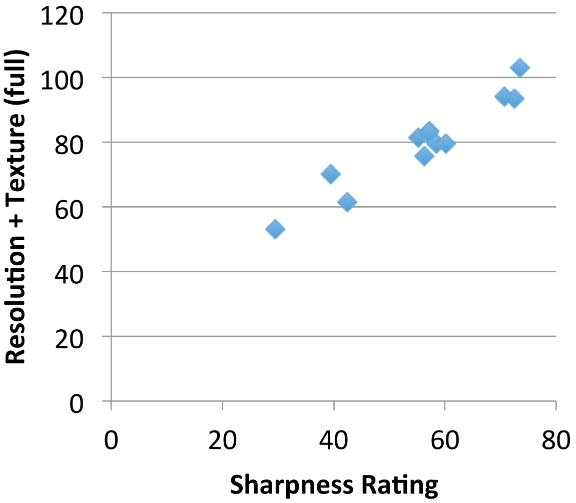

Figure 8 *Objective versus subjective ratings for sharpness. The objective measure is a weighted combination of the Resolution and Texture MTFs captured under a D65 20 lux light source.*

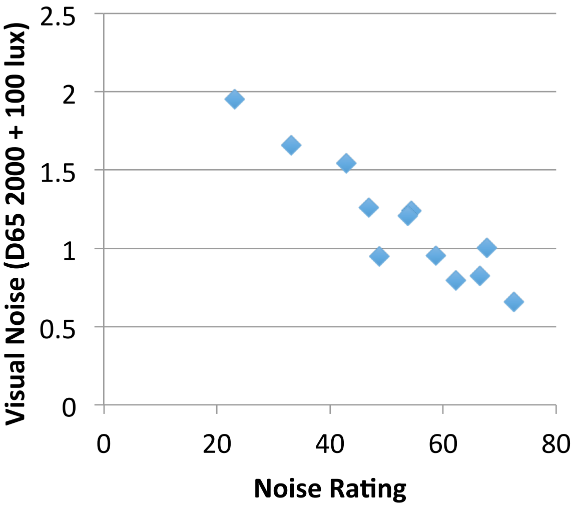

Figure 9 *Objective versus subjective ratings for image noise. The objective measure is a weighted combination of the Visual and dL noise values each captured under both D65 2000 and 100 lux light sources.*

Conclusions and Future Work

The long-term goal for this effort is to develop a methodology that brings the individual image characteristics together into an overall metric of the perceptual quality attainable by capture systems. As a first step toward this goal, subjective data was gathered for the individual characteristics of tone, color, sharpness, and noise were evaluated in perceptual experiments. The results of these tests are reported here. The correlations between the individual characteristics and overall quality were high, if the images were of generally high quality – in focus and correctly exposed. Correlations between the visual results and objective measurements were also investigated for sharpness and noise. It was determined that, for noise, the Visual Noise metric successfully predicted the visual results. For sharpness, a two-step process using Resolution vMTF and Texture (full) to sort the devices into categories of high, medium and low quality, followed by a second sort using Edge MTF for higher quality devices and Texture MTF for mid-to-lower quality devices successfully predicted the visual results.

Initial evaluation of the color results relative to objective data was not as conclusive. The subjective ratings were compared to color, lightness, hue and chroma differences for the objective target capture. Correlations were poor. Higher correlations could be achieved if the subjective ratings were compared to dominant colors in the scenes. Further work is being conducted toward developing a consistent approach for objectively assessing color.

References

Artmann, U, “Image quality assessment using the dead leaves target: experience with the latest approach and further investigations”, in proceedings of IS&T Electronic Imaging Symposium, Vol. 9396, San Francisco, CA, 2015.

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

Farnand, S. P., Dalal, E. N., and Ng, Y. S., *Recent progress in the development of ISO 19751*, in proceedings of SPIE/IS&T Electronic Imaging Symposium, Vol. 6059, San Jose, California, 2006.

Keelan, B. W. *Handbook of Image Quality: Characterization and Prediction*, Marcel Decker, NY, pp. 160-162, 2002.

Larson, E. C., Chandler, D. M., “Most apparent distortion: a dual strategy for full-reference image quality assessment”, J. of Elec. Imaging, 19(1), pp. 1-21, 2010.

Phillips, J.B. and Christoffel, D., “Validating a texture metric for camera phone images using a texture-based softcopy attribute ruler” in proceedings of SPIE/IS&T Electronic Imaging Symposium, Vol. 7529, San Francisco, CA, 2010.

Author Biography

Susan Farnand received her BS in engineering from Cornell University, her Masters in Imaging Science and her PhD in Color Science from the Rochester Institute of Technology. After beginning her career at Eastman Kodak, she moved to RIT where she currently works as a Visiting Professor in the Program of Color Science. Her research interests include human vision and perception and color science. She is the Publications Vice President of IS&T and serves as an Associate Editor for the Journal of Imaging Science and Technology, and has served as co-chair of the IQSP conference at EI.