https://doi.org/10.2352/ISSN.2169-2629.2018.26.116

Color Quality and Memory Color Assessment

Anku, Susan Farnand, Munsell Color Science Laboratory, Rochester Institute of Technology, Rochester, NY, USA

Abstract

With the rise in high quality displays and cameras following the mainstream adoption of smartphones, the color quality of images is becoming an essential aspect of engaging and attracting consumers. A color quality assessment (CQA) would provide insights into what users perceive and could be put to use when engineering cameras and displays. Twenty cameras were used to capture pictures of common objects like grass, sky, wood, and sand. CQA for common objects was performed using a rank order perceptual testing method, where the observers were asked to rank images of the same object captured by 20 different cameras according to their order of CQA. We confirm the results of the Rank CQA, by performing an Anchored Scaling experiment where the lower anchor is the least ranked and the higher anchor is the highest ranked image from the rank order CQA. The results of this test were generally consistent with the results of the CQA. We also evaluated memory colors using Method of Adjustment. Observers were asked to recreate their memory color for the common objects used in the CQA by adjusting a uniform color patch in CIELAB space. The results for the CQA shows that the preferred camera varies across the images of common objects. The results for the memory color experiment vary across the observers as they can be influenced by geographical locations, cultural backgrounds and other such factors. The results for both the experiments were then combined to compare how the memory color of observers differs from the actual color of the object images.

Introduction

Image quality has always been an important factor for various consumer electronics like televisions, cameras and smartphones. The widespread adoption of smartphones has raised the bar for image quality over the past few years. Color plays an important role in conveying information about an image. Beyond this, it engages the users and plays a big role in how they "like" an image. The rise of popular applications like Instagram and Snapchat, that use filters and other processing techniques to change the color of images, reflects that.

This paper talks about the color image quality assessment (CQA) for smartphone and digital cameras. The assessment is done on images of common objects such as grass, sand, wood, and skin. These images were captured on 20 different smartphones. Color quality of these images is assessed on the basis of the results of perceptual testing experiments.

Perceptual testing is a powerful tool that helps us to measure the perception of observers and record their preferences. Paired Comparison, a technique used in perceptual testing, is the standard way of comparing images, where observers are asked to pick out their preference between two images. Rank Ordering is where observers have to rank the images according to their most preferred to least preferred color image quality. We opt to use Rank Ordering for comparing the color quality of these cameras, because the number of comparisons required for Paired Comparison would have been prohibitively large (Seth et al., 2017). We performed an Anchored Scaling perceptual test to validate the results of the Rank Order Color Quality Assessment, where the lowest ranked image from the rank order color quality experiment served as the low anchor and the highest ranked image as the high anchor.

We also conduct another perceptual experiment to better understand the memory color of the same observers for these common objects. Memory Color is one a person recalls for common natural objects, such as the ones we included in our experiment. Humans tend to have a very short-term memory for the color of objects, and are not very accurate at recreation. However, the color of certain objects that are observed by a person very frequently, like the sky, grass, or skin tend to be recreated with relative consistency and may provide information regarding preferred reproduction of these objects. These recreations will obviously vary between subjects as they can be influenced by geographical locations, cultural backgrounds and other such factors (Fernandez et al., 2005).

Methodology

This study was performed in two parts. The first part focuses on color quality assessment and the second, on memory color assessment. For color quality assessment, we used images captured by Qualcomm using 20 devices retailing in 2014 (Farnand et al., 2017) which were a mix of smartphones and digital cameras. The scenes used for the pictures were taken in various common scenarios. These scenes include beaches, parks, restaurants, statues and food. Lighting conditions and framing were kept constant across all pictures. The set of images is shown in Figure 1.

For color quality assessment, we cropped a 200*200-pixel patch from the original images. We chose familiar objects from the pictures for cropping, like grass, sky, face, beach sand, wood, arm, vegetables, brick and foliage as shown in Figure 1. There was minor variability in the position of the textures among the cropped pictures due to different resolutions of the devices. The observers were given instructions to ignore these differences and to judge the color quality of the overall image. For our memory color experiment, we showed them uniform color patches for the familiar objects used in color quality assessment experiment.

Experimental Setup

The experiment was performed in the Perception laboratory at the Munsell Color Science Lab with a setup shown in Figure 2. A laptop was connected to the *Eizo CG248* color display, which was used to show the content to the observers for our perceptual testing experiments. The display was calibrated using the *Photo Research Spectrometer PR655* (Day et al., 2004).

The wall behind the monitor was painted gray (Munsell N5) and was illuminated by a metal halide lamp whose luminance was equivalent to that of the display. This light source and display were switched on 30 minutes before the experiment started to make the illumination uniform. Other lights in the laboratory and the surrounding rooms were kept switched off. The table on which the display was kept was covered with a similar gray sheet of paper to make the observers field of view as uniform as possible. The observers were seated 85cm from the display. These conditions were consistent with CPIQ viewing conditions (Jin et al. 2009). The stimuli were presented on the display using MATLAB. A Graphical User Interface (GUI) was used to present the stimuli and it enabled users to pick their preference among them. The background of the GUI was set to a similar gray as the wall and the table sheet. The setup was similar for all experiments with the exception that the GUI background for the Anchor Scaling experiment was black.

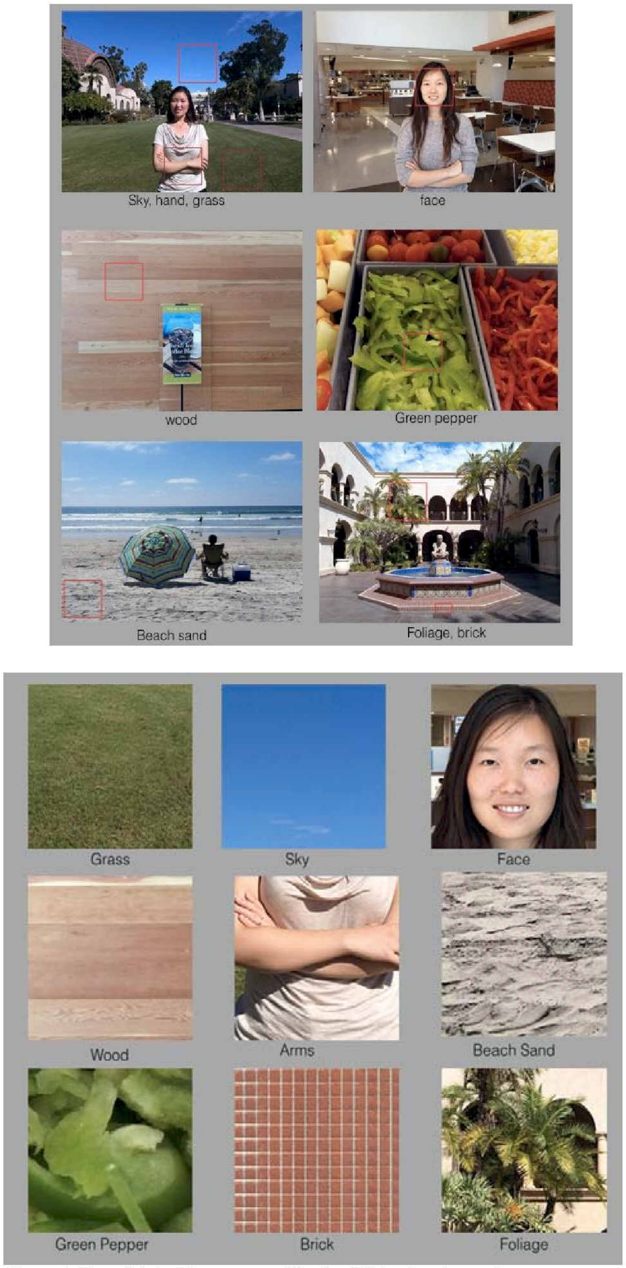


Figure 1: Top: Original images used for the CQA, showing various common objects in typical scenes with different types of lighting. These images are cropped to create texture patches used for the CQA testing. Bottom: Texture patches of familiar objects cropped from original images.

Observers

All the observers were tested for color vision using the Ishihara plate test. There were 26 total observers for the color quality assessment and memory color assessment among which 6 were female and 20 were male. All the observers had normal color vision and normal or corrected to normal visual acuity. Fourteen observers had some color or imaging science background whereas the other twelve did not. For the Anchor Scaling experiment, 30 people participated, 7 females and 23 males. Among the 30, 13 observers did not perform the color quality and memory color assessment experiment, whereas the remaining 17 did all 3 experiments. Seventeen of the observers had color or imaging science background, the rest did not.

Procedure: Color Quality Assessment

In this experiment, we study the preference for color quality for cropped images. The GUI displays 20 images at once as shown in Figure 3. The images were arranged in a 4*5 grid. The order of the images was randomized to avoid bias in the data. The observers were asked to rank the images by color quality by sorting the images in the GUI. They were given a set of instructions before the experiment to judge the images according to their preference for color quality. The observers adapted to the room lighting while receiving instructions, which took about 5 minutes. The average time taken for this experiment was 40 minutes.

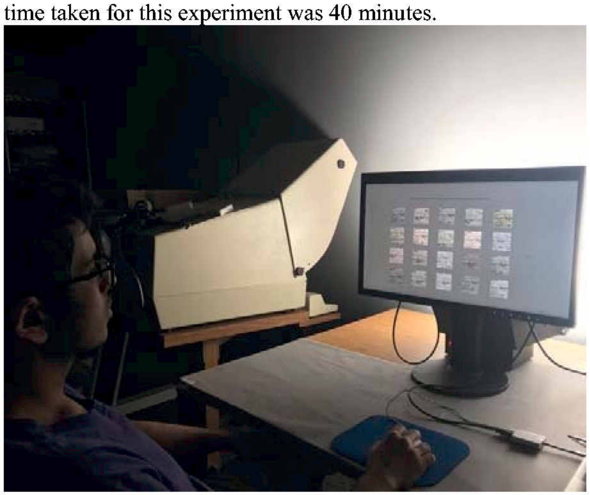


Figure 2: Experimental setup consists of a display showing a GUI consisting of 20 patches for the color quality assessment or 2 anchor images and a test image for the Anchor Scaling. The room lights were switched off and a back light was used.

Procedure: Anchor Scaling Color Quality Assessment

This experiment was conducted to verify the results of the Rank Order Color Quality Assessment experiment, in which observers reported the task as being difficult to perform. In this experiment, the GUI displays two anchor images - a low-quality image, as determined in the Rank Order experiment, on the lower left-hand side and a high-quality image on the lower right-hand side. The test images were displayed between the anchor images. The anchors were arbitrarily assigned scores of 30 for the lower quality image and 75 for the higher quality image. The observers were asked to assign a score to the test image relative to the scores of the anchor images. If the color quality was between the anchor images then the score assigned would be between 30 and 75. If the observer thought that the test image was of greater quality than the high-quality anchor image, then a score greater than 75 was assigned. Similarly, if the test image was of a lower quality than the low-quality anchor image, then a score below 30 was assigned. The minimum score that can be assigned was 0 and the maximum was 999. Observers were asked to ignore the sharpness, blurriness, and other image-quality characteristics. The focus was just on the color quality. Previous testing, including work with the QA ruler, indicates observers can successfully separate color quality for other image characteristics (Farnand et al., 2016; Keelan & Urabe, 2003).

Procedure: Memory Color Assessment

In this experiment, we study the color memory of the same objects as used in Experiment 1. The GUI showed uniform patches of the color of these objects as shown in Figure 4. The initial uniform patch color was randomized for each of the objects. The randomization was done within a close range of the object's hue, and various lightness levels. The observers were asked to adjust them according to their color memory of these objects using three sliders. Slider 1 was lightness scale, slider 2 was CIELAB a* and slider 3 was CIELAB b*. The non-color science observers were given a brief introduction and demo about CIELAB color space and how the 3 sliders worked. We chose CIELAB because it's a perceptually uniform space.

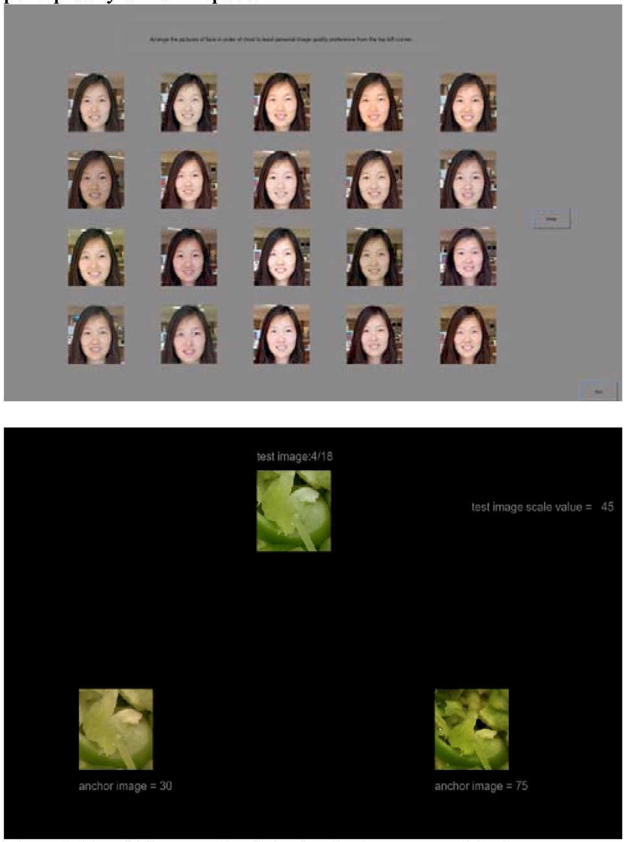


Figure 3: Top: GUI created for Color Quality Assessment. The images are randomly arranged on the grid, to avoid bias. Observers rank images from best to worst. Bottom: GUI created for Anchor Scaling Color Quality Assessment. The lower left and lower right are the lower quality anchor image and higher quality anchor images respectively. These anchor values were assigned randomly as 30 and 75 respectively. Observers were asked to assign a value for the test image relative to the anchor value.

Each color patch was accompanied with a brief description of the object it represented, like "grass", "beach sand", and "sky". No additional information was provided about the scenes in which the object was captured, since we are concerned more about the memory color of the participant and not their ability to recall specific object colors. This procedure is common across several studies involving memory color (Bartleson, 1960).

Results and Discussion

Color quality assessment:

The expected results for the color quality assessment would be that, the camera of a higher quality would be preferred over the others. However, we see that the results vary depending upon the content of the image.

Figure 5 shows the overall performance of the cameras for all 9 image sets. The x-axis is the z-score scale value, calculated using Thurstonian Analysis (Gescheider, 2013), and the y-axis values represent the cameras. The red, green and cyan bars represent the first, second and third best choices respectively. The top choices are the images for camera numbers 14, 12 and 13, respectively, which are not statistically different from each other. Devices 4, 3, and 2 are also not reliably different. As we observed earlier, there is variance in preference across the scenes. That is, certain cameras were preferred for some scenes while not being preferred for others. This is reflected in the scale values for the overall result, which shows that the values do not differ much for the most preferred devices. This is also true for the least preferred ones.

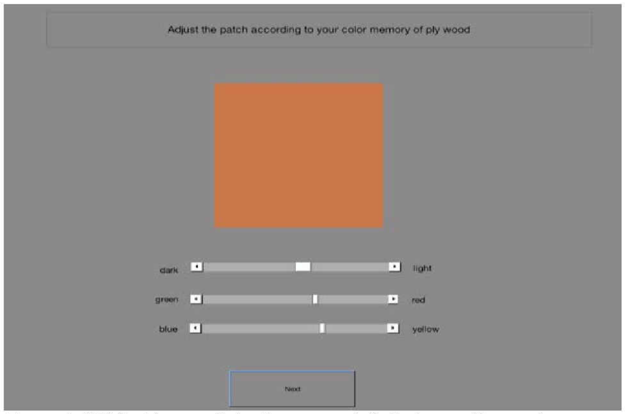

Adjust the patch according to your color memory of ply wood

Figure 4: GUI for Memory Color Assessment. It displays uniform color patches of the common objects used in CQA. The observers are asked to adjust them according to their color memory of these objects using three sliders. Slider 1 is the lightness scale, slider 2 is a of CIELAB and slider 3 is b* of CIELAB.*

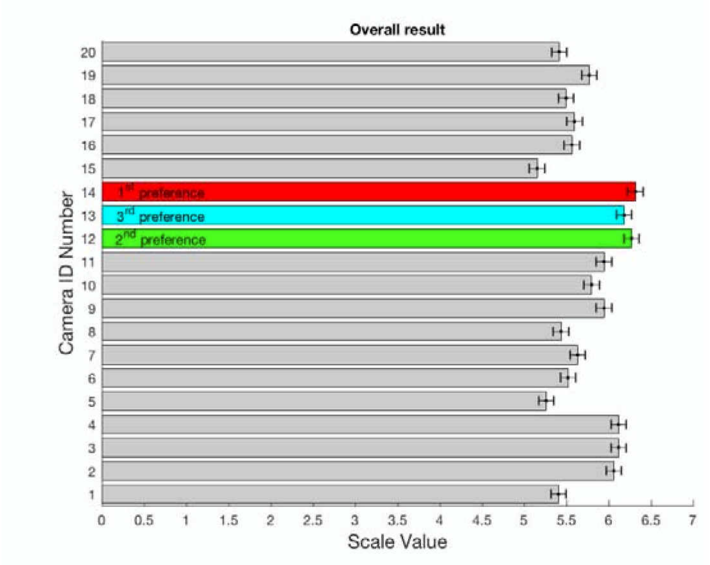


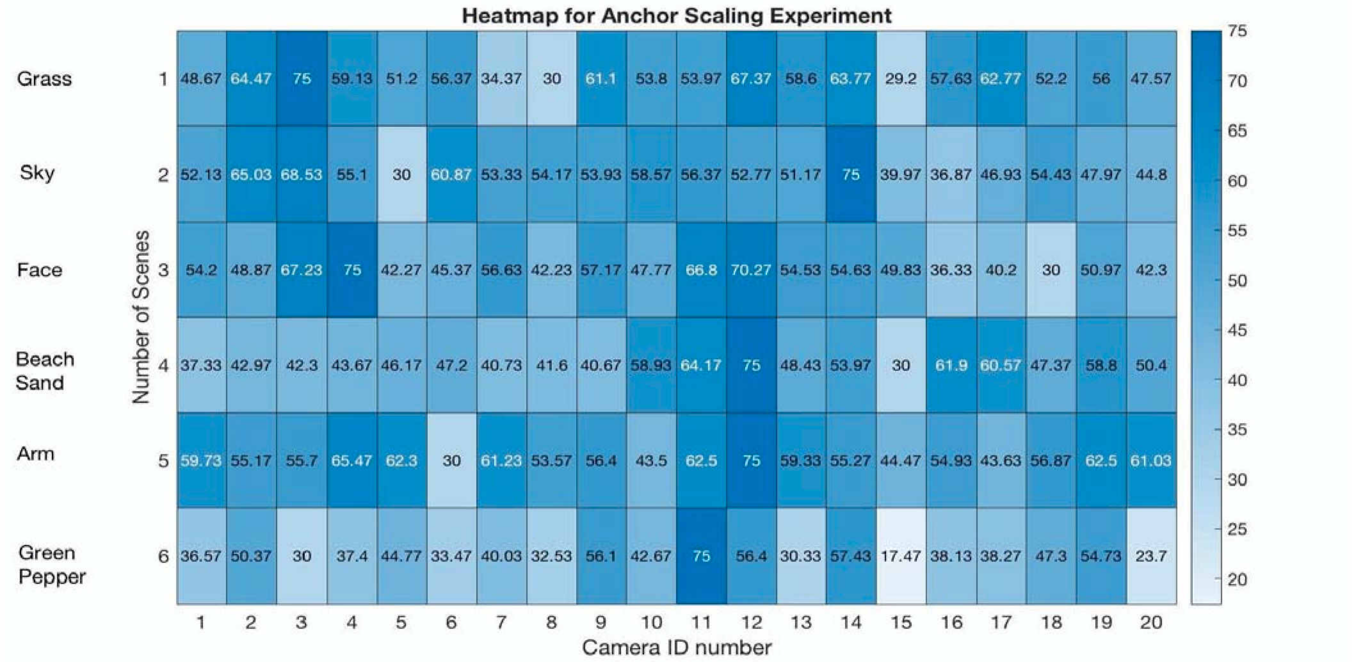
Figure 5: Overall ranking for the entire image set. The bars represent scale values, and have error bars at the end to indicate standard deviation. The red, green and cyan bars represent the first, second and third top choices respectively. There is no statistically significant top choice since color quality preference varies between different images.

Anchor Scaling Color Quality assessment:

Figure 6 shows a heat map plot for the anchor scaling results. The x-axis represents the camera number from 1 to 20 and the y-axis represents the number of scenes used for the experiment. The scenes numbered from 1 to 6 are grass, sky, face, beach sand, arm and pepper, respectively. On the right-hand side of the map we observe a color bar with score values on it. The score values start from 75 with the darkest blue down to 20 with lightest blue. The values inside the map are the average score of all of the observers for each scene and for each camera. Darker hues represent high scores and vice-versa for lighter hues.

The top choice for anchor scaling is for camera 12 with a mean score (for all the scenes) of 66.133. The second and third top choices are for cameras 11 and 14 respectively with mean scores of 63.133 and 60.00. The score value for the top choice (camera 12) varies significantly across the 6 scenes. For instance, it has two scores of 75 for scenes 4 and 5, whereas for scenes 2 and 6 it has scores of 52 and 56. The lower scores for scenes 2 and 6 could probably be due to the variance in the preference of saturation. Camera 14, the third highest-scored camera has a more consistent mean score across all scenes. The lowest scored is for camera 15 which has a mean score of 35.15 across all scenes. The second and third lowest scored are for Cameras 8 and 20 with mean scores of 42.35 and 44.96 respectively. Camera 15 had the lowest score for scenes 6 and 1, which are green pepper and grass, respectively, and its highest scores for face and hand, though these scores are still low relative to other cameras. For Camera 8, the scores are relatively consistent. Its lowest scores are also for the pepper and grass scenes. Camera 20 has a very low score for the green pepper scene but for other scenes it had score values between 40 to 60.

From Figure 6, we observe that, relative to the high-anchor image, scores for the green pepper scene are low for all cameras. Whereas the arm scene performed well for all cameras, with all images scaled high relative to the low anchor image. This could imply that some cameras do most things well, while some may have difficulty with specific subject matter. It is important to rate a device's color quality based on a wide variety of content. (Smet, 2011).



		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Grass	1	48.67	64.47	75	59.13	51.2	56.37	34.37	30	61.1	53.8	53.97	67.37	58.6	63.77	29.2	57.63	62.77	52.2	56	47.57
Sky	2	52.13	65.03	68.53	55.1	30	60.87	53.33	54.17	53.93	58.57	56.37	52.77	51.17	75	39.97	36.87	46.93	54.43	47.97	44.8
Face	3	54.2	48.87	67.23	75	42.27	45.37	56.63	42.23	57.17	47.77	66.8	70.27	54.53	54.63	49.83	36.33	40.2	30	50.97	42.3
Beach Sand	4	37.33	42.97	42.3	43.67	46.17	47.2	40.73	41.6	40.67	58.93	64.17	75	48.43	53.97	30	61.9	60.57	47.37	58.8	50.4
Arm	5	59.73	55.17	55.7	65.47	62.3	30	61.23	53.57	56.4	43.5	62.5	75	59.33	55.27	44.47	54.93	43.63	56.87	62.5	61.03
Green Pepper	6	36.57	50.37	30	37.4	44.77	33.47	40.03	32.53	56.1	42.67	75	56.4	30.33	57.43	17.47	38.13	38.27	47.3	54.73	23.7

Figure 6: *The heat map for the anchor scaling results. The x-axis is the camera number from 1 to 20 and the y-axis is the number of image sets (scenes) used for the experiment. The number of image sets used from 1 to 6 used were for grass, sky face, beach sand, hand and pepper respectively. On the right side of the map we can see a color bar with score values on it.*

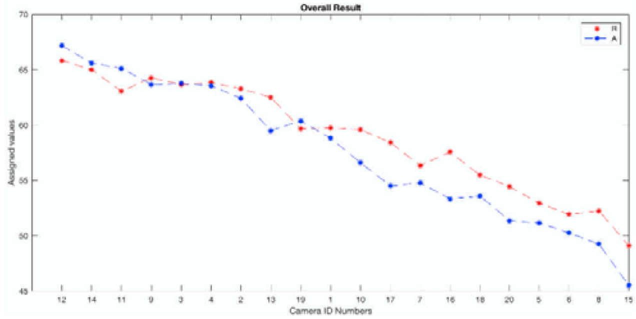


Figure 7: Rank Order vs. Anchor Scaling color quality assessment. The x-axis is the Camera ID and the y-axis is the assigned score value. The red stars represent the rank order and the blue stars the anchor scaling results.

When we discard the green pepper scene from our evaluation, the highest ranked cameras (mean scores) change to 12, 3, 11, 14. Similarly, the bottom camera choices change to 15, 8, 5, 6. Camera 3 which was not among the top 3 choices for anchor scaling became top 2 choice after discarding the green pepper scene. Camera 5 and 6 became among the bottom choices whereas camera 20 is no longer in the bottom 4 choices, after discarding the green pepper image.

Figure 7 shows the overall result for the comparison of Rank Order vs. Anchor Scaling. The x-axis is the camera ID number and the y-axis is the assigned values (score value). The overall result is the average of all six scenes used in the testing. We can observe that the results of the methods are consistent.

Memory Color assessment:

For the memory color experiment, we showed the observers uniform color patches for the familiar objects used in the color quality assessment. Recreation of memory color for the objects would be expected to vary as the observers are influenced by factors such as geographical and cultural background (Fernandez, 2005). We were interested to see the consistency of the memory color of the observers and how preferred memory color relates to the colors in the top ranked images.

The results in Figure 8 show that the grass memory color has some linearity between the a* and b* channels as well as lightness and chroma. The trend seems to be that people recall grass as having a saturated shade of green. Some points are scattered from the common result point which could be attributed to how varied the color of grass can be geographically.

When the observers were asked what they pictured when they were adjusting the grass color patch according to their memory color, most of them said that it was the appearance of the turf on a soccer field, some people recollected the color of their lawn, and a few people imagined grass on a golf course. It is difficult to set an average on the color of grass, as it can be wet, dry, and can have different hues in various geographical locations. For example, in hot and dry regions, the color of grass would not be as saturated as in cooler places. Other factors that play an important role in influencing memory color, are the time of day, season and weather. This experiment was conducted in Rochester, NY, USA in October.

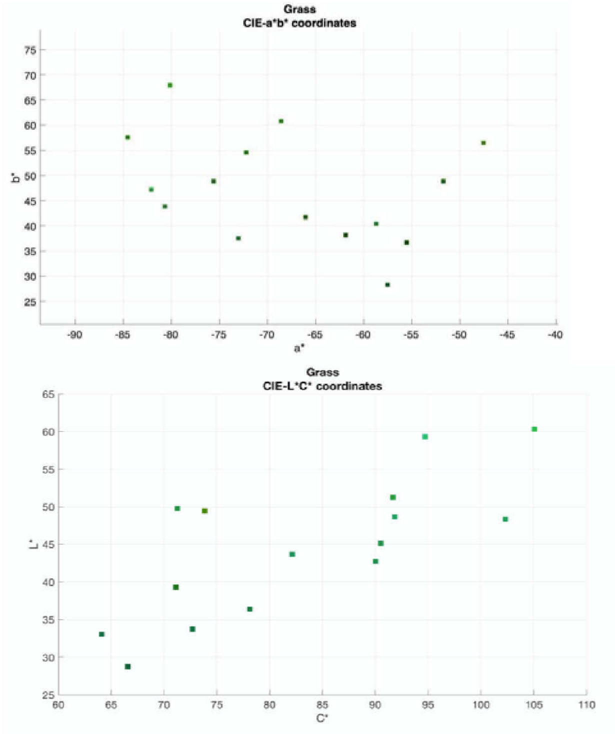


Figure 8: Top: Memory Color of Grass. Here, the x-axis is the a channel and y-axis is the b* channel. Bottom: Memory Color of Grass. Here, the x-axis is the C* channel and y-axis is the L* channel.*

Memory Color assessment vs Color Quality assessment:

We compared the results for both the experiments. The pixel color values for the images in the dataset were averaged in CIELAB space and compared with the results of the memory color assessment. This was done to see how close the memory color of the objects is to the actual image and also especially to the higher or lower ranked images. We would expect some distinction between memory and object colors, because when we think of an object's color we generally tend to not consider the texture and other characteristics that influence its appearance.

Figure 9 shows the results for the comparison between memory color and CQA for sky. The plot on the left of Figure 8 has a* values on the x-axis and b* values on the y-axis, and the right plot in Figure 7 has b* on the x-axis and L* on the y-axis.

We observe how the CQA results, represented by the circular points, are clustered on the plots whereas the memory color results, represented by the square data points, are scattered. The a* values for most of the CQA results vary between -10 and 10, whereas for the memory color assessment they vary between -40 and 5. Hence, there is considerable variance in the memory color results, which implies that people do not have consistent recollection of the color of the sky. But we can see that observer's memory of sky is always cyan or blue, and generally more chromatic than the original image. We can see from Figure 7 on the left that the lower ranked images have positive a* values, which may be perceived as purplish.

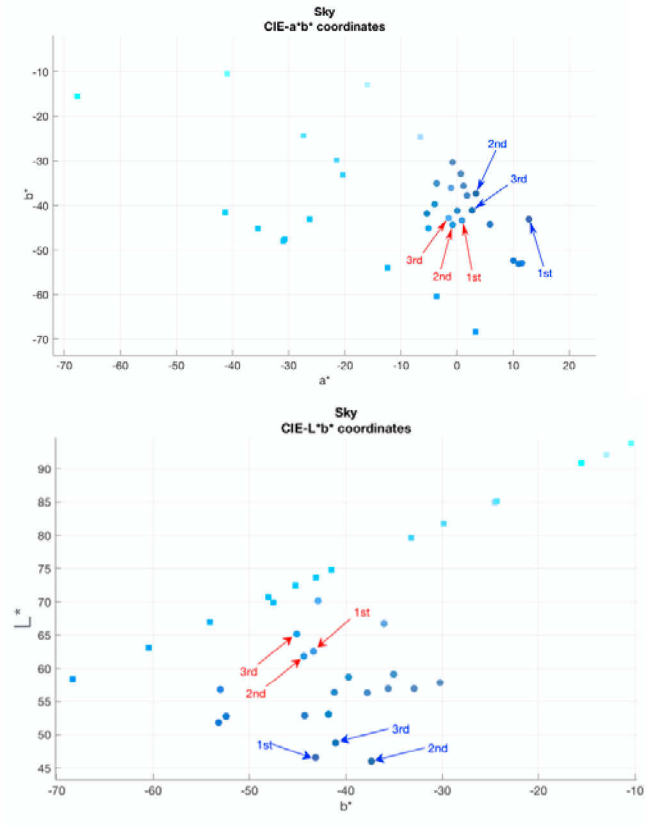


Figure 9: Comparison of Color Quality and Memory Color Assessment for sky. The square points are the results of the memory color assessment and the circular points are the results of CQA. The red arrows point to the top 3 preferences and the blue arrows to the bottom 3 preferences. Top: The x-axis is the a channel and y-axis is the b* channel. Bottom: The x-axis is the b* channel and y-axis is the L* channel.*

If we look at the lightness scale, on the right of Figure 8, there is also a big spread of the memory color results. For CQA, the CIELAB values are very close for the most preferred ones, which have the among highest L* values. The least preferred renditions, in contrast, have the lowest L* values. These results suggest that people prefer the sky to be of relatively high lightness. The memory color results also reflect this in that their L* values are almost all higher than the image values. Further, we see that the lightness is as high as it can be for a given b* value, which results from the adjustment being at the gamut limit of the display. This information could prove useful for determining how sky should be rendered in a preferred color reproduction.

An explanation for the variation in memory color could be the geographical location, season, time of day, or other reasons. Big and busy cities in countries like China and India are very polluted, and this influences how the sky can appear. It is not entirely uncommon for the sky to have a faded hue because of pollutants. We can see that the results for memory color are more saturated and lighter than the color of the sky in the images. We tend to remember the color of an object to be more saturated (Bartleson, 1960). When asked about what they pictured for the color of the sky, many observers spoke about how the sky looked in the morning versus the afternoon, or after rains.

Conclusion

Color quality assessment is important in understanding what colors are preferred by people and hence is important for phone, camera industry to improve their cameras accordingly for their customers.

We conducted a rank order perceptual experiment where the observers were asked to rank familiar object images from 20 different cameras according to their preference in color image quality. The results show that the camera preference varies with the image content. This indicates that a variety of scene content is necessary when evaluating device color quality.

We verified the results of the Rank Color Quality Assessment by performing an anchored scaling experiment where the lower anchor was the lowest ranked and the higher anchor the highest ranked image from the rank order experiment. The results between the tests were consistent.

We conducted another experiment on memory color where we asked the observers to recreate the color of familiar objects, same as the ones used in color quality assessment by using the method of adjustment in CIELAB space. The results show that observers tend to choose more saturated colors for familiar objects. The results from color quality assessment and memory color experiments were compared by evaluating the average of images from 20 cameras for eight familiar objects relative to the memory color results. This showed that the image color varies from memory color. This can be explained by the fact that humans do not tend to average the factors involved in object appearance like sand under the grass, dust particles in beach sand, melanin in skin, and clouds in the sky. Also, the lighting conditions, weather, surrounding are not constant for everyone in their memory. The geographical location, personal interest, cultural background of observers influences the memory color results for familiar objects. However, the memory color results for sky show that observers were remarkably consistent in selecting the highest lightness possible for a given b* value. The memory color of wood, while not as consistent as for sky, displayed a consistent hue angle value regardless of lightness.

Future Work

In the memory color evaluation reported here, we used uniform patches. However, as noted, true memory colors are not uniform patches and involve non-uniformities such as blades of grass and freckles on skin. To evaluate the effect of non-uniformity, we will repeat this experiment using texture patches of familiar objects rather than uniform patches for adjustment and determine how the results differ. We will use highly and poorly rated patches from the CQA experiment as starting points in this experiment. We may also explore the effect of nationality in this follow-up experiment.

Among the most important memory colors for color quality assessment are skin tones. We have collected some female face data which we will use as training and test sets for deep learning techniques, which will be employed to evaluate tone and color quality, particularly for skin color and white balance.

Acknowledgments

Authors would like to thank Qualcomm for supporting this study and thank all the participants who graciously volunteered to participate in the study. We would also like to thank Dave Wyble, who helped in so many ways.

References

Bartleson, C. (1960). Memory Colors of Familiar Objects. Journal of the Optical Society of America, 50(Number 1).

Day, E. A., Taplin, L., & Berns, R. S. (2004). Colorimetric characterization of a computer-controlled liquid crystal display. Color Research & Application, 29(5), 365-373.

Engeldrum, P. (2000). Pychometric Scaling: A Toolkit for Imaging Systems. Massachusetts: Imcotek Press.

Fairchild, M. D. (2013). Color appearance models. John Wiley & Sons.

Farnand, S., Jang, Y., Choi, L. K., & Han, C. (2017). A methodology for perceptual image quality assessment of smartphone cameras–color quality. Electronic Imaging, 2017(12), 95-99.

Fernandez, S. R., Fairchild, M. D., & Braun, K. (2005). Analysis of observer and cultural variability while generating "preferred" color reproductions of pictorial images. Journal of Imaging Science and Technology, 49(1), 96-104.

Gescheider, G. (2013). Pychophysics: The Fundamentals. Psychology Press.

Hunt RWG, P. M. (2011). Measuring colour. NJ: Wiley.

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.

Keelan, B. W., & Urabe, H. (2003, December). ISO 20462: a psychophysical image quality measurement standard. In Image Quality and System Performance (Vol. 5294, pp. 181-190). International Society for Optics and Photonics.

Sheth, G., Carpenter, K., & Farnand, S. (2017, September). Image Quality Assessment of Displayed and Printed Smartphone Images. In Color and Imaging Conference (Vol. 2017, No. 25, pp. 13-19). Society for Imaging Science and Technology.

Smet, K., Ryckaert, W. R., Pointer, M. R., Deconinck, G., & Hanselaer, P. (2011). Colour appearance rating of familiar real objects. Color Research & Application, 36(3), 192-200.

Thurstone, L. (1994). A law of comparative judgment. Psychological review, 101(34.4), 266-27