Evaluating Camera Performance in Face-Present Scenes with Diverse Skin Tones

Megan Borek, Alexander Schwartz, Amelia Spooner; Imatest LLC; Boulder, CO, USA

Abstract

Consumer cameras are indispensable tools for communication, content creation, and remote work, but image and video quality can be affected by various factors such as lighting, hardware, scene content, face detection, and automatic image processing algorithms. This paper investigates how web and phone camera systems perform in face-present scenes containing diverse skin tones, and how performance can be objectively measured using standard procedures and analyses. We closely examine image quality factors (IQFs) commonly impacted by scene content, emphasizing automatic white balance (AWB), automatic exposure (AE), and color reproduction according to Valued Camera Experience (VCX) standard procedures. Video tests are conducted for scenes containing standard compliant mannequin heads, and across a novel set of AI-generated faces with 10 additional skin tones based on the Monk Skin Tone Scale. Findings indicate that color shifts, exposure errors, and reduced overall image fidelity are unfortunately common for scenes containing darker skin tones, revealing a major short-coming in modern-day automatic image processing algorithms, highlighting the need for testing across a more diverse range of skin tones when developing automatic processing pipelines and the standards that test them.

Introduction

The VCX Standard

The Valued Camera Experience (VCX) WebCam 2023 specification defines test procedures and metrics for a wide range of video quality concerns, including contrast, dynamic range, exposure, spatial frequency response, color accuracy, and white balance [1, 2]. We assess the performance of consumer web and smartphone cameras under various lighting conditions and scenes defined in Version 1.0 of the standard. Though the specification is intended for the testing of web cameras, including integrated laptop, stand-alone, and conference cameras, the same procedures and metrics can be applied to the testing of smartphone cameras. The VCX association has also developed a separate PhoneCam testing specification [1], but it is not considered in this study in order to test both web and phone cameras under identical conditions.

Lab Setup

The VCX WebCam specification proposes a comprehensive chart design to measure a variety of IQFs. A prototype chart designed by Imatest according to standard specifications is shown in two scenes in Fig. 1. The chart contains the Calibrite Classic ColorChecker, Siemens star, spilled coins/dead leaves texture, and slanted edge targets. Of particular interest is the use of two different mannequin heads—one with a dark skin tone (Richard) and a second with a light skin tone (Alexis), which are used to compare the behavior of auto white balance (AWB) and auto exposure (AE) algorithms under various conditions. For this study, all scenes are captured with the chart and face mounted in front of a neutral gray backdrop. Videos are captured according to VCX procedures, and frames are extracted from a point in the video after convergence and averaged for analysis. Frames are analyzed using Imatest software and supporting scripts to obtain objective metrics quantifying auto exposure, white balance, and color accuracy performance.

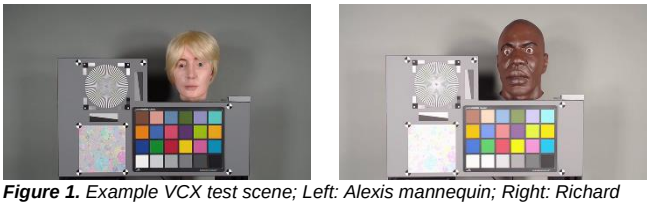

Figure 1. *Example VCX test scene; Left: Alexis mannequin; Right: Richard mannequin; Both scenes captured under identical illumination conditions.*

Initial Results and Analysis Methods

Initial results are obtained using three consumer webcams with a range of price points across the three scenes shown in Fig. 2—a control scene containing only the chart, a scene containing the Richard mannequin, and a scene containing the Alexis mannequin. The analyzed images are frames extracted from videos of the scenes captured under identical lighting conditions, with the default, out-of-box camera settings. All frames depicted in Fig. 2 are captured under 6200K correlated color temperature (CCT) illumination ("cool white", per VCX definition) at a brightness of 250 lux at the chart surface. It is important to note that Webcams 2 and 3 utilize facial recognition to adjust scene exposure, which becomes particularly evident for scenes containing darker skin tones. Only a small subset of VCX required metrics are calculated and discussed in this paper to emphasize the performance of AE, AWB, and color accuracy across various skin tones.

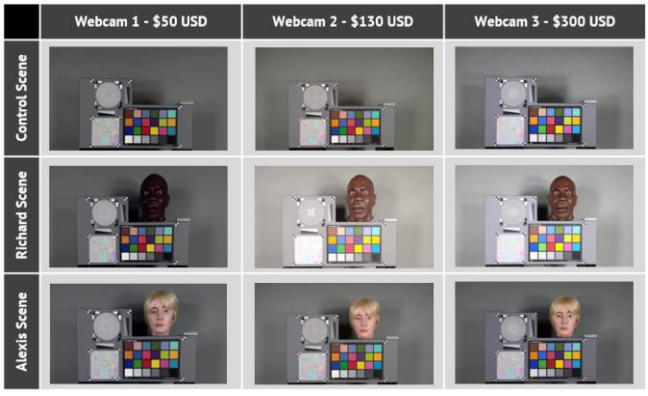

Figure 2. *Three webcams of varied price points were tested across three scenes, with and without mannequins. All scenes were captured under 6200K CCT illumination at 250 lux.*

Figure 3. *Simulated ColorChecker patches from each scene compare the target patch color against the corresponding average measured patch value. Bolded borders indicate saturation in one or more color channels.*

Color Error - ΔE_{2000}

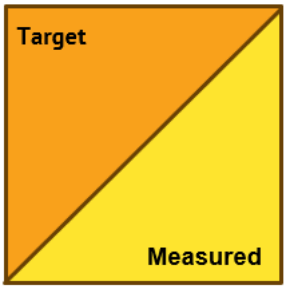

Figure 4. *Example simulated ColorChecker split patch comparing target and averaged measured patch colors*

Assuming the camera under test has global performance (e.g., there is not local tone mapping), the ColorChecker in each of the captured scenes gives a basic idea of the color reproduction of scene content surrounding a face. Simulated ColorChecker patches for each scene are shown in Fig. 3. Each tile represents a split view of each patch on the ColorChecker, comparing the target color to the average measured value, as enlarged in Fig. 4. This is most meaningful when viewed on a calibrated display, but still of interest are the saturated patches, indicated by bolded red borders in Fig. 3. The two cameras with facial recognition—Webcams 2 and 3—show significant oversaturation of several patches in scenes containing the Richard mannequin.

Calculating the average ΔE_{2000} (CIEDE2000) error across all 24 patches of the ColorChecker in each scene reveals that there is higher error in scenes containing the Richard mannequin captured with the face detection cameras. (See Fig. 5(a)). For example, the color error for Richard is 20 with Webcam 2, but only 3.5 for Alexis.

Exposure – Lightness (L)*

Though the consistency of the ΔE_{2000} errors across all three scenes captured by Webcam 1 is desirable, when faces are taken into consideration, there is a conflict between this objective metric and a subjective analysis of a scene. Fig. 6 shows crops of the mannequin faces captured under identical illumination conditions using the three webcam devices. The Richard mannequin, when captured by Webcam 1, is severely underexposed in comparison to captures by Webcams 2 and 3. The Alexis mannequin appears adequately exposed across all devices.

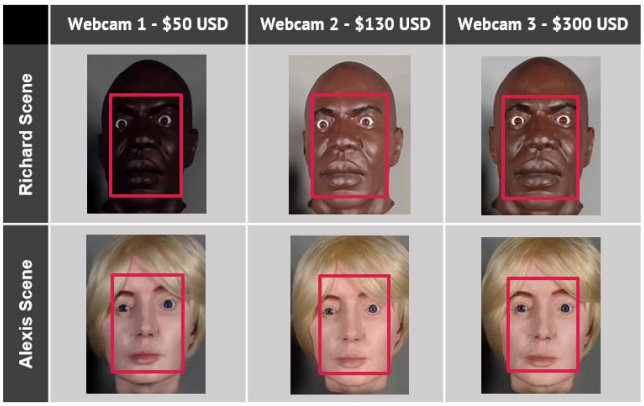

Figure 6. *Close-up of mannequin faces captured under identical illumination by three devices. Boxes indicate the region used to calculate the average L* and C* values.*

Webcams 2 and 3 make use of face detection to automatically adjust exposure to prioritize a face in the scene, which is arguably more desirable for webcam applications, where the primary content in a frame is a face. A major drawback of using global exposure in face-present scenes is that for darker skin tones, the camera either adjusts the exposure based on the scene as a whole, risking underexposure of the face, or it can adjust the exposure based on a detected face, which risks overexposure of other scene content. With the inclusion of facial recognition in modern image processing pipelines, there is now a major flaw in using flat color patches to measure skin tone reproduction for some devices. As shown by the oversaturated patches in the ColorChecker in Fig. 3, it is no longer

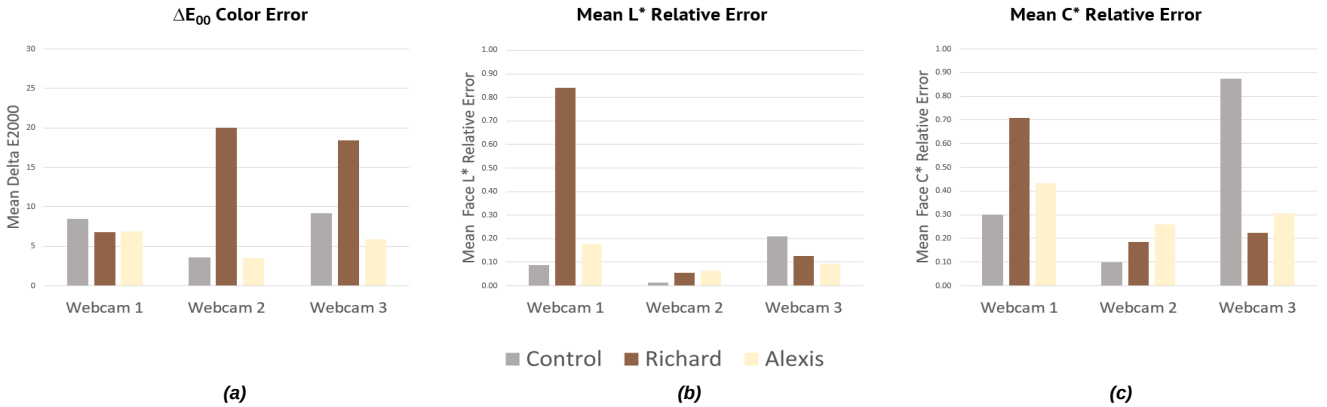

Figure 5. *Errors for three primary metrics calculated for each webcam across each of the three scenes, including mean ΔE_{2000} (a) of the ColorChecker, mean L* error (b) of the face ROI, and mean C* error (c) of the face ROI.*

accurate to judge color reproduction from the ColorChecker alone when a detectable face is present in the scene.

The average lightness, or L* value of a region of interest (ROI) containing the face, as depicted by the bounding boxes in Fig. 6, is one indication of whether a face is properly exposed in a scene. Exposure accuracy of the surrounding scene is based on the lightness of patch #21 (neutral 6.5) of the ColorChecker. In this paper, we use the average L* of patch #21 to examine exposure accuracy in control scenes, and face ROIs to examine exposure for face-present scenes. In a full VCX dataset, both are analyzed for all scenes. VCX provides a target average L* range for each of the mannequins. We use the central value of these ranges to calculate the relative error in the measured average L* value of each face ROI. The relative error for ColorChecker patch #21 is calculated using the corresponding reference L* value of 66.766. L* errors are plotted in Fig. 5(b), where we see expectedly higher errors in the Richard scene captured by Webcam 1, and fair performance across all scenes by Webcams 2 and 3.

White Balance – Chroma (C)*

White balance is analyzed similarly to exposure, but by looking at the average C* value across the face (or patch #21 of the ColorChecker) after convergence. This value indicates how saturated colors appear in the captured scene. Again, we calculate the relative error based on central VCX and ColorChecker target values and plot the results in Fig. 5(c). Webcam 2 shows the least deviation from target C* values across all scenes. Interestingly, Webcam 3 performs better in scenes containing a face than in the control scene. Looking at all of these metrics in Fig. 5, we can see that no single metric, scene, or skin tone can provide a complete understanding of system performance or image/video quality. Though this may not be surprising, it does highlight that image and video processing algorithms can and do perform very differently depending on whether there are faces in the scene, and what the characteristics of those faces are.

Testing Diverse Skin Tones

The noticeable differences in performance based on the primary skin tone in the scene point to a limited understanding of camera performance by testing only two skin tones. Beyond VCX, the industry is looking closely at ways to better understand the impact of a broader range of skin tones, and we can apply these solutions to image and video testing.

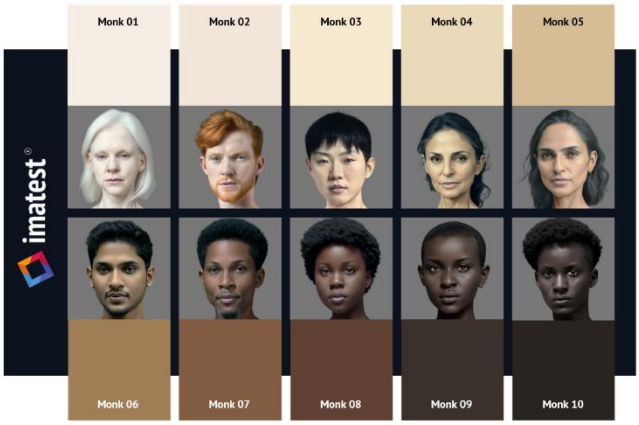

Figure 7. *Ten simulated faces with a range of skin tones based on the Monk Skin Tone Scale are created using generative AI and printed for data capture*

Monk Skin Tone Scale

The Monk Skin Tone (MST) Scale was developed by Dr. Ellis Monk at Harvard and is comprised of 10 different tones [3]. The scale is currently being used by Google Research and is designed to represent a broader range of geographic communities [4] than the commonly used Fitzpatrick scale [5], which is skewed towards lighter skin tones due to its dermatological background.

AI-Generated Diverse Human Faces

To test how automatic image processing algorithms perform across a wider range of skin tones without access to a diverse group of real people, we used artificial intelligence (AI) to simulate humans instead. We worked with Generated Photos [6] to create 10 AI-generated faces with skin tones based on Google's Monk Skin Tone Scale. This AI tool offers 16 default skin tone options, 120+ ethnicities, and custom AI prompt input. Initial parameters were chosen to achieve target skin tones (rather than focusing on specific ethnicities), followed by manual adjustments in Adobe Photoshop to match MST values more accurately and create a uniform 18% gray background. The resulting faces are depicted in Fig. 7. Each

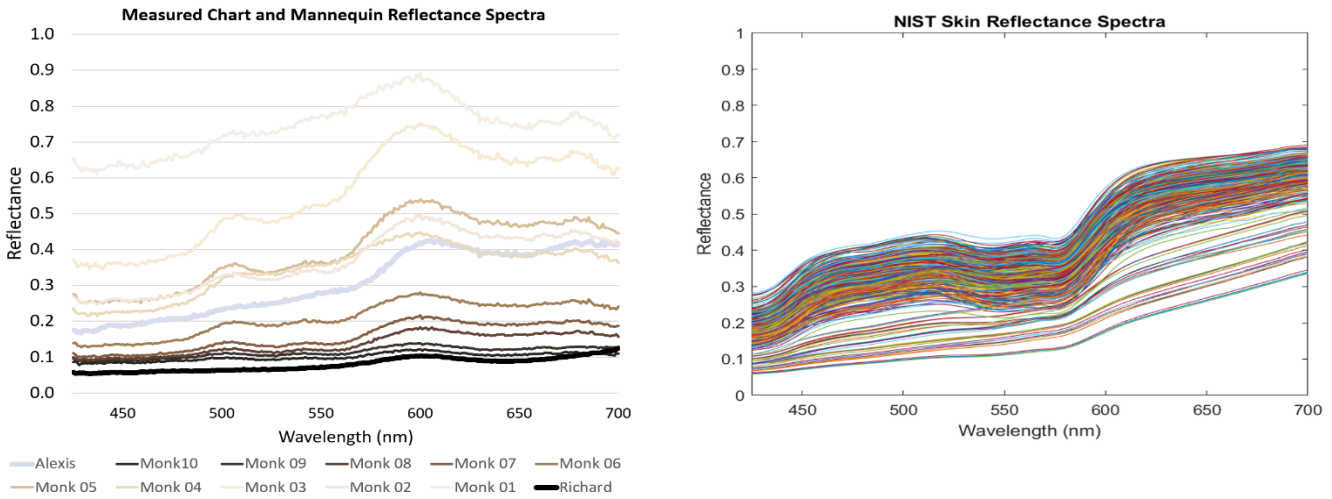

Figure 8. *Measured reflectance spectra of simulated skin from mannequin heads and printed face charts (left) compared to 100 sample reflectance spectra of actual human skin (right) measured by the National Institute of Standards Technology (NIST) [7].*

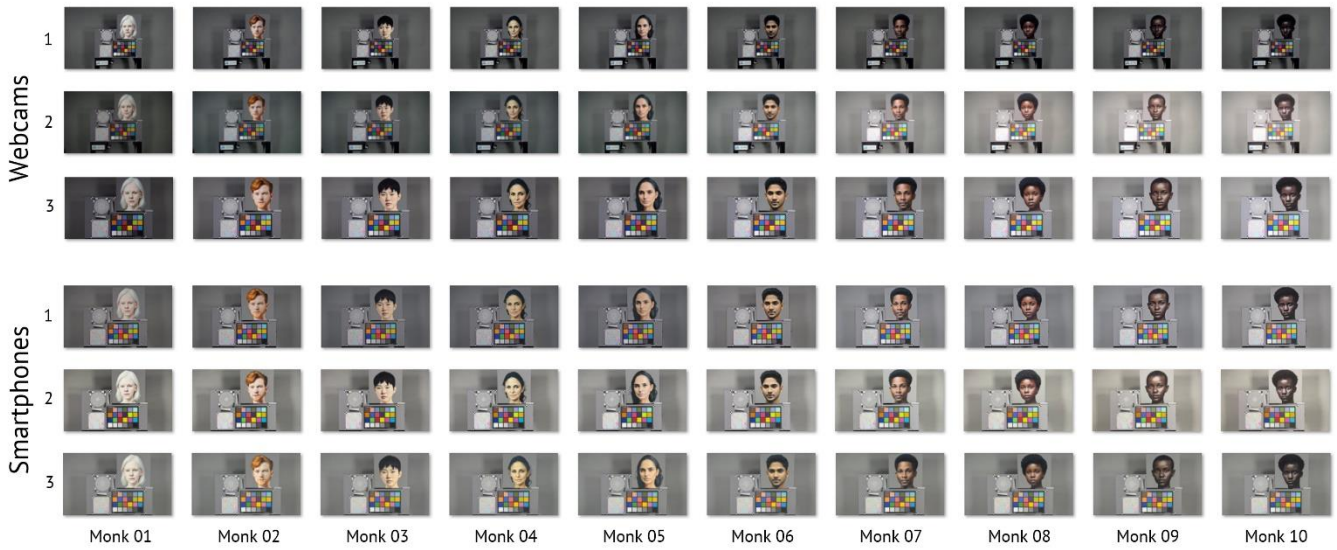

Figure 9. *Frames extracted from video clips captured by three webcams and three smartphones under 6200K 250 lux illumination. Note the gradual increase in overall brightness as the skin tone in the scene becomes darker, despite being captured under identical lighting conditions.*

face is printed at life-size scale using color-accurate printing methods and mounted individually.

Spectral reflectance of simulated versus real skin

A primary concern when using mannequins or printed targets in place of humans is the loss of the spectral nuances of actual skin—simulated or printed skin will behave differently from real skin, which has unique absorption, reflection, and sub-skin scatter properties that are dependent on the amount of melanin present. Reflectance spectra of the mannequins and each of the 10 printed targets are measured using a spectrophotometer and are plotted in Fig. 8. These are compared with 100 sample reflectance spectra of actual human skin (right, in Fig. 8) measured by the National Institute of Standards Technology (NIST) [7]. As expected, the reflectance spectra of the simulated skin—both mannequin and print—do not directly match the spectra of actual skin. However, we do see comparable trends across the spectral properties of both simulated and real skin, including similar peaks in the range from 500-600 nm. More research is required to better understand the impact of the differences between simulated and real skin, which is beyond the scope of this paper.

Results across devices, targets, and lighting conditions

Each of the 10 face charts is captured in a scene with the VCX chart by the same three webcams. The face is mounted on the same plane as the chart at a position comparable to the mannequins. Additionally, three common smartphones (which use face detection) are also tested for each skin tone under the same conditions using VCX WebCam procedures. This process is repeated across several lighting conditions. Frames extracted from videos captured by all six devices at 250 lux and 6200 CCT are shown in Fig. 9. Notice the visible increase in overall scene brightness as the skin tone of the face in the scene darkens for many of the devices, as seen

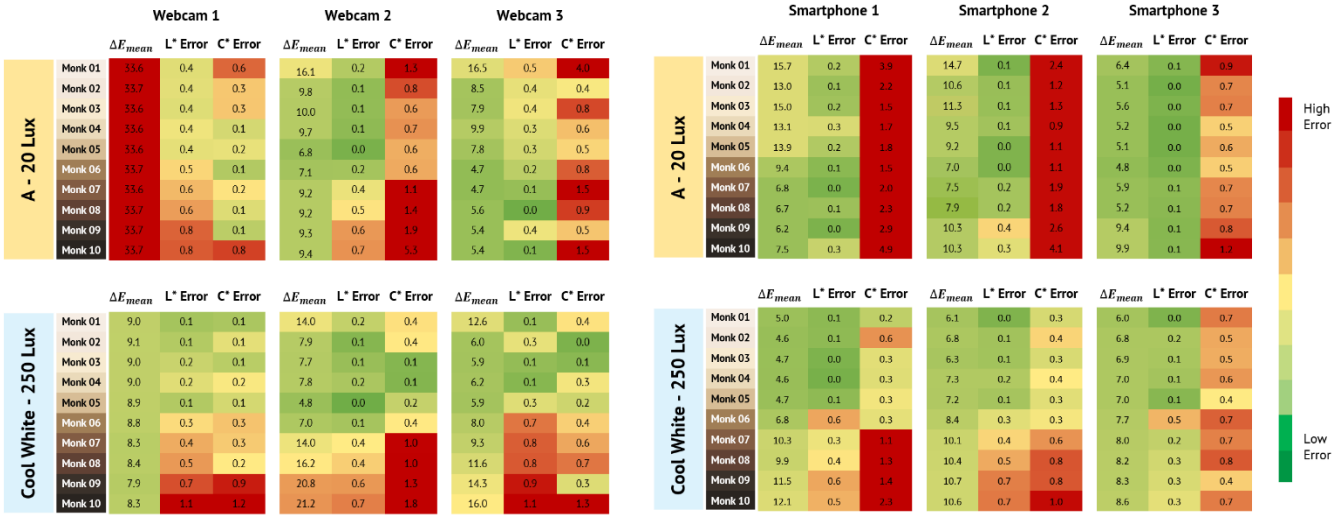

Figure 10. *Three metrics of interest—average ΔE2000, L* error, and C* error—calculated across devices, skin tones, and select lighting conditions. Values are color-coded according to deviation from target values.*

progressing from left to right. The only two devices that do not illustrate this phenomenon are Webcam 1 (which does not adjust exposure based on a detected face) and Smartphone 3.

The same three metrics— average ΔE_{2000}, L* error, and C* error—are calculated for each skin tone scene across devices and lighting conditions. Results for all six tested devices at two select VCX-specified illumination conditions are summarized in Fig. 10. While these three metrics are by no means a comprehensive analysis for these scenes, they offer a high-level understanding of system performance across a greater range of skin tones than is required by most testing procedures.

We tend to expect higher C* errors in modern non-scientific use cameras, especially in smartphones, due to human preference for more saturated colors, so these results are not particularly surprising. Perhaps more telling is the relationship between the ΔE_{2000} and L* errors across the skin tones, plotted in Fig. 11. These, in a simplified way, compare how good the chart (or surrounding scene content) appears versus how good the face appears.

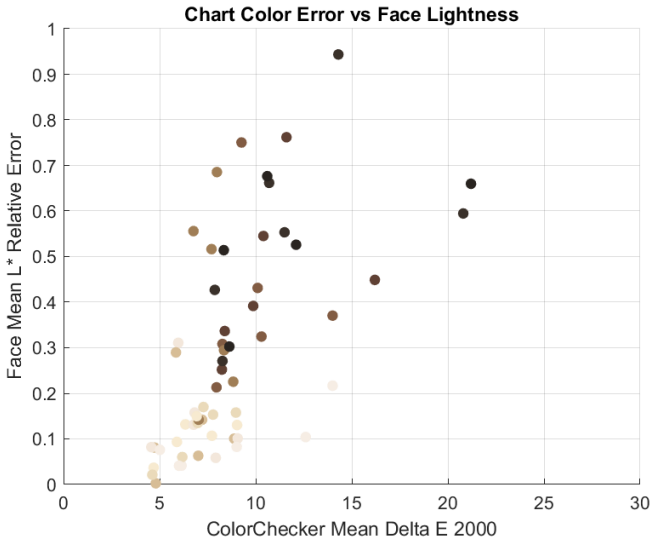

Figure 11. *Scatter plot of average chart ΔE2000 error versus face L* error for each of the 10 skin tones tested across all six devices.*

Fig. 11 clearly shows that both ΔE_{2000} and L* errors tend to be larger for scenes containing darker skin tones. It also shows how the handling of darker skin tones varies more drastically within and across devices than for scenes containing lighter skin tones, whose errors fall within a noticeably tighter range.

Conclusion

This paper seeks to illustrate the importance of testing a diverse range of skin tones for camera systems designed for human and face-present scenes. Industry efforts such as VCX and groups within the International Organization for Standardization (ISO) seek to better understand the appropriate range of skin tones that are tested in standards, but there is still much to be done to expand skin tone testing and performance requirements to be more inclusive of a wider range of human diversity, eliminating the trade-off between high-quality depictions of scene content or high-quality depictions of faces.

Future Work

The presented findings merely scratch the surface of understanding how cameras perform in face-present scenes. There is additional work to be done in the future that would supplement this work:

- Significant amounts of data were collected as part of this study, only a small portion of which is reflected in this paper. More than 100 videos were captured across 15+ illumination conditions, from which dozens of metrics can be derived.
- Additional scenes containing multiple faces and skin tones should be captured to evaluate camera performance in more complex and diverse scenes.
- Further investigate the effects of using simulated skin versus actual human subjects.
- Consider and test the effects of dimensionality for 3D mannequins or human targets in comparison to 2D printed charts.
- Investigate other color charts that have more focus on human skin tones.

References

[1] The VCX Website, VCX-Forum, https://vcx-forum.org/ (accessed Jan. 31, 2024).

[2] "VCX-WebCam Standard 2023 Rev. 1.0," VCX-Forum, https://vcx-forum.org/content/press/VCX_WebCam_2023_v1.pdf (accessed Jan. 31, 2024).

[3] E. Monk, The Monk Skin Tone Scale (MST), May 2023. doi:10.31235/osf.io/pdf4c.

[4] "Developing the Monk Skin Tone Scale," Skin Tone Research @ Google, https://skintone.google/the-scale (accessed Jan. 31, 2024).

[5] W. H. Ward, "Fitzpatrick classification of skin types I through VI," Cutaneous Melanoma: Etiology and Therapy [Internet]., https://www.ncbi.nlm.nih.gov/books/NBK481857/table/chapter6.t1/ (accessed Feb. 1, 2024).

[6] "AI Human Generator," AI Human Generator – Generate and Modify People Online, https://generated.photos/human-generator (accessed Jan. 31, 2024).

[7] "Reference data set of human skin reflectance," National Institute of Standards Technology, https://catalog.data.gov/dataset/reference-data-set-of-human-skin-reflectance-e18fa#sec-dates (accessed Feb. 1, 2024).

Author Biography

Megan Borek received her BS in Imaging Science from the Rochester Institute of Technology (2022), where her studies focused on remote sensing, image processing and computer vision, and image sensor analysis. She currently works as an Imaging Scientist at Imatest LLC in Boulder, CO, where her work is focused on imaging standards, software development, and test lab services. She is an active contributor on the VCX WebCam committee, among other standards bodies such as IEEE and ISO.