Robustness and Generalization in Smartphone-Based Colorimetry: A Comparative Analysis of Deterministic and Stochastic Correction Paradigms

Zhi Jing [*] **Enrico Vesentini** [*] **Tài Thái** [*] **Jonas Thumbs** [*] **Baisu Zhou** [*]

Abstract

While colorimetry devices can accurately measure colors of objects, they are too expensive for daily usage such as detecting color of a scratched surface for repainting. Color picking from images captured by a smartphone is subject to distortion due to ambient light. We statistically analyze a deterministic and a data-driven color correction method that improves the accuracy of color detection using smartphone images, showing that simple, post-hoc correction can greatly improve the quality of color measurements done by smartphones.

1. Introduction

Measuring color accurately is a common problem in daily life, e.g., when a surface needs to be repainted in the same color. Devices that can carry out this task, the so-called *colorimeters*, are available commercially and usually cost several hundred dollars. There are smartphone apps that claim to offer the same service for free by using the built-in camera. A challenge in this approach is how to deal with the effect of ambient light. Simply picking the color from a photo is not ideal because hue, saturation, and brightness depend heavily on the lighting conditions. One way to address this issue is to capture the color of a white reference object (e.g., a white piece of paper) at the same time and use this to correct for the ambient light.

In this study, we want to find a sufficiently accurate approach to correct for ambient light and analyze the usability of measurements from such an app for color detection. While smartphone cameras store colors in RGB format, the RGB color space is not suitable for our analysis, because the distance between two points in the RGB space does not align with the perceived difference between two colors. In our analysis, we use the CIELAB (or $L^*a^*b^*$) color space, designed by Commission Internationale de l'Éclairage (CIE)[1] for colorimetric applications. Every color in the RGB space can be mapped non-linearly to a point (L^*, a^*, b^*) in the $L^*a^*b^*$ space, where the L^* coordinate represents lightness and the a^*, b^* coordinates control the hue. The Euclidean distance on the $L^*a^*b^*$ space reflects the perceived difference between colors.

To evaluate the performance of the correction methods, we define an accuracy threshold. The smallest wavelength difference that can be perceived as a chromatic difference under constant luminance conditions ("just noticeable difference") is approximately 1 nm, which corresponds to about an Euclidean distance of approximately 1.27 in the $L^*a^*b^*$ color space (Oleari, 2015). The authors further mentioned that a few multiples of the just noticeable difference can still be considered small. Accordingly, we define the Euclidean error 6 (approximately five times the just noticeable difference) as the *small color difference threshold* in our study.

We have developed our own color measurement app with built-in correction using white reference. As the transformation from RGB to $L^*a^*b^*$ is intended for measurements under CIE's D65 lighting condition (Oleari, 2015), we collected data using our app under a lighting condition similar to D65 (Section 2.1). We further compared the app's correction-by-scaling method (Section 2.2) with a modeling approach (Section 2.3). While the model provides insights into the color correction mechanism, the scaling method proves to be more robust (Section 3).

Mention that our app can do this, but we do not do this in data collection?

2. Data and Methods

2.1. Data Collection

We used a color reference sheet designed for camera calibration as the object from which we collect data. The reference sheet consists of 24 color cards and provides accurate ground-truth RGB values for each color. We used two phones, Google Pixel 8 Pro and Samsung Galaxy S21

[*]Equal contribution . Correspondence to: BZ <baisu.zhou@student.uni-tuebingen.de>.

Project report for the "Data Literacy" course at the University of Tübingen, Winter 2025/26 (Module ML4201). Style template based on the ICML style files 2025. Copyright 2025 by the author(s).

[1]In Enligh: International Commission on Illumination. Website: https://cie.co.at/.

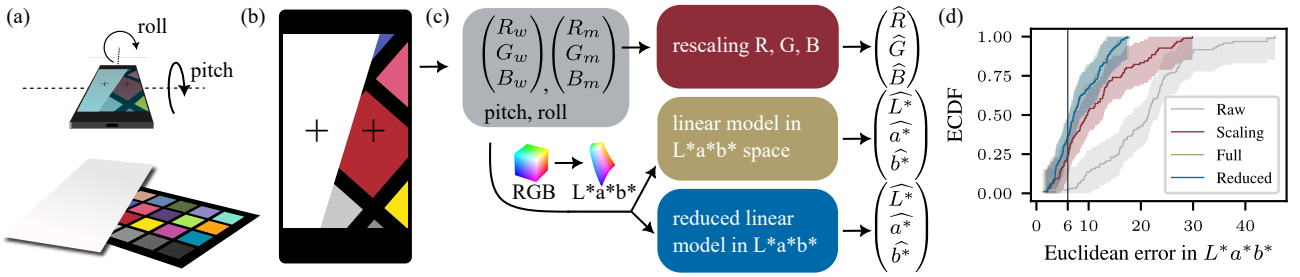

Figure 1. Workflow of data collection and color correction. (a) Color measurements taken by smartphone with white paper reference. (b) Illustration of our color detection app. During data collection, we covered all color cards but the one under measurement rather than exposing them as in the illustrative figure. (c) Color correction methods. (d) ECDFs of Euclidean distances from the raw and corrected colors to the ground truths. The shaded areas are 95% Dvoretzky–Kiefer–Wolfowitz confidence bands (Dvoretzky et al., 1956; Wasserman, 2004). The dashed vertical line indicates the small color difference threshold.

Ultra, for data collection. The data were collected indoors on a desk in front of the window under natural daylight on November 19, 2025 in Tübingen, Germany. We assume our lighting condition approximates the D65 standard lighting condition, so that the transformation from RGB to $L^*a^*b^*$ is valid. The data collection procedure is as follows.

We placed a white piece of paper (called *white reference* in the following) with a cut-out square hole over the array of colors to ensure similar conditions for all measurements. Only one color card was visible to the camera at any time. We took 10 photos for each color along with the white reference and stored the pixel RGB values. In addition, the app stores the pitch and roll angle (see Figure 1) from which every measurement was made. To make the procedure of taking the photos as similar as possible for the 24 colors, all measurements in one run were taken by one person, and images were taken with a similar distribution of angles and distances over the 24 colors. We tried not to vary the yaw angle during dataset collection, since the app only detects pitch and roll. The dataset we use in subsequent analysis consists of 480 measurements taken in two runs.

2.2. Correction by Scaling

Our app is equipped with a simple, deterministic correction algorithm. Let $(R_\mathrm{m}, G_\mathrm{m}, B_\mathrm{m})$ be the measured RGB values of a color card. We assume that the measured color $(R_\mathrm{w}, G_\mathrm{w}, B_\mathrm{w})$ of the white reference is the brightest color one can measure under the ambient light. To recover the full RGB space $[0, 255]^3$ from the measurable color space $[0, R_\mathrm{w}] \times [0, G_\mathrm{w}] \times [0, B_\mathrm{w}]$, we apply a channel-wise rescaling. The corrected RGB values are given by

$$\begin{pmatrix} \widehat{R} \\ \widehat{G} \\ \widehat{B} \end{pmatrix} = \begin{pmatrix} \frac{255}{R_\mathrm{w}} & 0 & 0 \\ 0 & \frac{255}{G_\mathrm{w}} & 0 \\ 0 & 0 & \frac{255}{B_\mathrm{w}} \end{pmatrix} \begin{pmatrix} R_\mathrm{m} \\ G_\mathrm{m} \\ B_\mathrm{m} \end{pmatrix}.$$

This method was suggested by de Greef et al. (2014) for monitoring newborn jaundice using smartphones. We consider it the simplest yet meaningful method for color correction using white reference.

2.3. Correction by Model

As a more complex approach to color correction, we fit a regression model on our measured data. To optimize the model toward reducing perceived color difference between detected and true colors, we transform the measured RGB values into the $L^*a^*b^*$ space before modeling. We consider a multi-target linear model defined by

$$\begin{aligned} \begin{pmatrix} \widehat{L^*} \\ \widehat{a^*} \\ \widehat{b^*} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} + \begin{pmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{pmatrix} \begin{pmatrix} L^*_\mathrm{m} \\ a^*_\mathrm{m} \\ b^*_\mathrm{m} \end{pmatrix} \\ + \begin{pmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{pmatrix} \begin{pmatrix} L^*_\mathrm{w} \\ a^*_\mathrm{w} \\ b^*_\mathrm{w} \end{pmatrix} \\ + \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \\ D_{31} & D_{32} \end{pmatrix} \begin{pmatrix} \text{pitch} \\ \text{roll} \end{pmatrix}, \end{aligned} \tag{1}$$

where $(L^*_\mathrm{m}, a^*_\mathrm{m}, b^*_\mathrm{m})$ and $(L^*_\mathrm{w}, a^*_\mathrm{w}, b^*_\mathrm{w})$ are the measured colors of a color card and the white reference, respectively, and $a_i, B_{ij}, C_{ij}, D_{ij}$ are parameters.

The model is fitted by minimizing the empirical risk associated with the loss function

$$\ell(y, \hat{y}) = \left(L^* - \widehat{L^*}\right)^2 + \left(a^* - \widehat{a^*}\right)^2 + \left(b^* - \widehat{b^*}\right)^2,$$

where $y = (L^*, a^*, b^*)$ denotes the ground truth color and $\hat{y} = (\widehat{L^*}, \widehat{a^*}, \widehat{b^*})$ denotes the prediction given by the model. For simplicity, we refer to $\sqrt{\ell(y, \hat{y})}$, the Euclidean distance between y and $\hat{y}$, as the *Euclidean error* of a correction.

Our assumption is that the white reference can sufficiently characterize the ambient light, so that additional information of pitch and roll contributes little to color correction. We validate this assumption by comparing the full model (1) with a reduced model with $D = 0$.

3. Results

3.1. Effect of Correction

We partition the measurements into an 80/20 train-test split, training both the full and reduced models on the former. To evaluate the efficacy of the correction step, we apply scaling correction directly to the held-out test set, allowing for a rigorous comparison against the model-based predictions. Figure 1(d) shows the empirical cumulative distribution function (ECDF) of the measurement-wise Euclidean error in the $L^*a^*b^*$ space on the test set. Although the corrected colors remain visually different from the ground truths, they reduce the gap by a large margin in comparison to raw measurements. The median error of raw measurements is 21.46. After correction by scaling, the median error reduces to 9.71; After correction by the full and reduced models, the median errors drop to 7.19 and 7.17, respectively. Without correction, only 2% of measured colors fall below the small color difference threshold (Euclidean error ≤ 6). Using correction by scaling, the small color difference rate is increased to 26%. The full and reduced models yield a small color difference rate of 31% and 35%, respectively.

Table 1 shows the pitch and roll coefficient estimates of the full model fitted on the training set along with their 95% bootstrap confidence intervals. We observe that both angles have non-significant effects on color correction. This aligns with our observations on the test error distributions. The error ECDFs of the two model variants shown in Figure 1(d) are almost identical, which indicates that adding information about pitch and roll does not change the behavior of the model, as we expected. The statistics reported in the previous paragraph also show no improvement of the full model over the reduced one. Thus, we focus on the reduced model in subsequent analysis.

le results?

While the error ECDF of the models lie above that of the scaling method, their confidence bands overlap. At the higher end of error distributions, we see an improvement of the models over simple scaling. The maximum error incurred by either model is lower than the scaling method.

To gain some insights into how the correction methods reduce the Euclidean error, we visualize the error per channel in Figure 2. One can observe a shift of error distribution in the L^* coordinates toward zero resulted from the correction methods. The error distributions in the a^* and b^* coordinates are all centered at approximately zero, but the correction methods leads to a reduction in variance. The error distribution of the model is more concentrated around zero than that of the scaling method. These observations show that the correction methods are capable of removing bias toward darker colors, but cannot eliminate uncertainty in hue.

Table 1. Pitch and roll coefficient estimates on training set with 95% percentile bootstrap confidence intervals (Davison & Hinkley, 1997; Efron & Tibshirani, 1994) obtained from 1000 bootstrap trials. To construct the confidence intervals, we resample training instances with replacement, which corresponds to "resampling cases" in Davison & Hinkley (1997) and "bootstrapping pairs" in Efron & Tibshirani (1994).

	pitch		roll	
$\widehat{L^*}$	-0.07	$[-0.11, -0.03]$	0.00	$[-0.02, 0.03]$
$\widehat{a^*}$	0.02	$[-0.02, 0.06]$	-0.01	$[-0.04, 0.02]$
$\widehat{b^*}$	0.05	$[-0.01, 0.11]$	-0.00	$[-0.04, 0.04]$

3.2. Generalizing to Unseen Colors

Our dataset covers only 24 colors, but the models ought to generalize to arbitrary colors. We simulate the scenario of generalization to unseen colors by employing a cross-validation strategy. For each $k = 1, \ldots, 20$, we randomly reserve k colors for validation and use the remaining $24 - k$ colors for training. The results are shown by Figure 3. The average-case performance of the model, measured by the average Euclidean error, remains comparable with the scaling method even if we omit up to 15 colors from the training set. When leaving out up to 5 colors, the frequency of the model achieving small color difference is, on average, higher than the scaling method, but there is large fluctuation across choices of colors to leave out.

To identify colors with high impact on generalization, we perform leave-one-out cross validation, training the model on all but the j-th color for $j = 1, \ldots, 24$ and testing it on the single color left out. The results of leave-one-out cross validation (Figure 4) reveal an uneven distribution of generalization performance across colors. We find that color $j = 11$, which possesses the largest green (G) channel intensity in the dataset, is the primary driver of failure of generalization.

4. Discussion & Conclusion

Our study provides empirical evidence that rescaling RGB channels according to a white reference can improve the quality of color measurements made by smartphones. Under usual daylight, a multi-target linear model can further avoid large deviations from the true color, although the median improvement over the scaling method is marginal. We confirmed that the pitch and roll angle of the camera have little

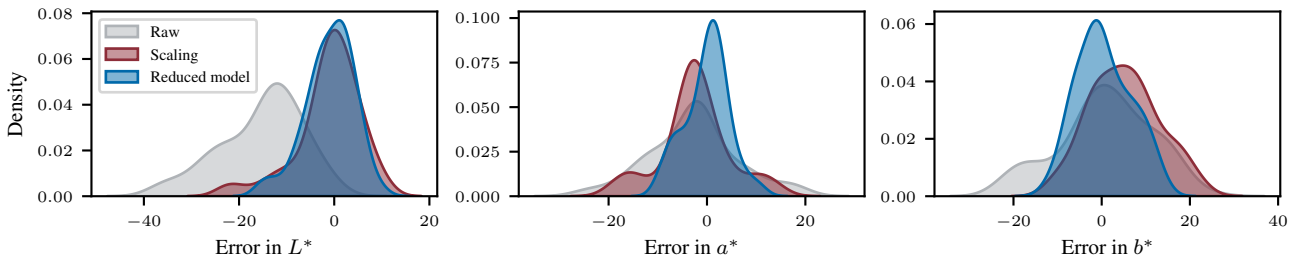

Figure 2. Kernel density plots for channel-wise difference from ground truth.

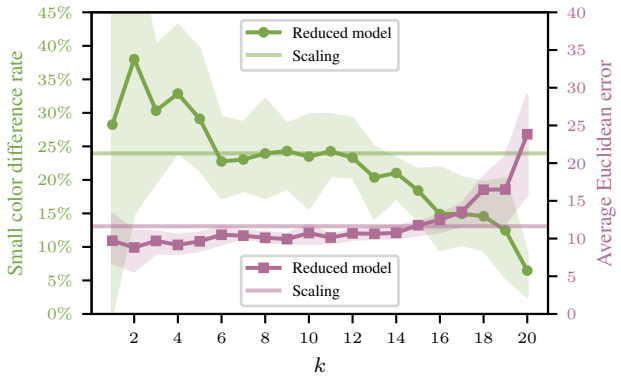

Figure 3. Validation performance of the reduced model on k left-out colors. The shaded bands show the upper and lower quartiles over 20 trials per k. The horizontal lines represent the average performance of the scaling method.

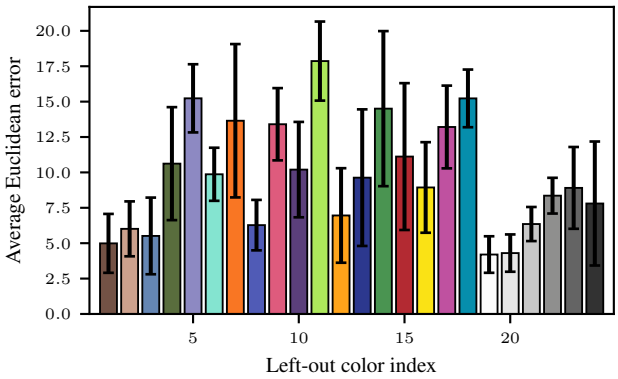

Figure 4. Validation average Euclidean error (mean $\pm$ one standard deviation across 20 measurements) of the reduced model on single left-out color. The color of each bar is the color left out.

influence on color correction using white reference, and identified that green-toned colors play an important role in the generalization of our models.

We acknowledge the limitations of our study and generally of color detection using smartphones. Firstly, raw camera measurements are affected by sensor noise and illumination variability. According to Gueli et al. (2019), for a better color adjustment, both black and white references should be considered. Black reference is obtained by taking a picture in a black box, such that to capture the camera noise; white reference is taken as before. The black–white correction can be implemented as an affine transformation consisting of a translation that maps the black reference to the origin of the $L^*a^*b^*$ space, and a rotation that aligns the white reference with the L^* axis. Secondly, given that green-toned colors appear to have high influence in the model's generalization performance, we propose normalizing the spectral components of the light according to their relative perceptual influence. One possible implementation would be to weight each RGB coefficient by its corresponding influence prior to the conversion to $L^*a^*b^*$. After the full processing pipeline, the inverse weighting could be applied to the corrected colors in order to recover values in the original color space, while preserving the corrections introduced by the normalization and enabling a meaningful comparison with the reference.

Contribution Statement

All authors participated in data collection. Zhi Jing assisted in exploratory analysis, implemented the models, and conducted preliminary experiments. Enrico Vesentini identified the objects that theoretically satisfy the properties required for the purposes of this study, and suggested improvements to the workflow. Tài Thái engineered the core codebase for exploratory analysis, performed bootstrapping for robustness analysis, designed the cross validation experiments, architected and improved the models. Jonas Thumbs developed the app for data collection, performed exploratory analysis, and designed the graphical abstract. Baisu Zhou performed exploratory analysis, engineered the project infrastructure, participated in the analysis of the methods, and finalized the figures and the report.

References

Davison, A. C. and Hinkley, D. V. *Bootstrap Methods and their Application*. Cambridge University Press, 1997.

de Greef, L., Goel, M., Seo, M. J., Larson, E. C., Stout, J. W., Taylor, J. A., and Patel, S. N. Bilicam: using mobile phones to monitor newborn jaundice. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 331–342. Association for Computing Machinery, 2014. ISBN 9781450329682.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956.

Efron, B. and Tibshirani, R. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994.

Gueli, A. M., Pasquale, S., Politi, G., and Stella, G. The role of scale adjustment in color change evaluation. *Instruments*, 3(3), 2019.

Oleari, C. *Standard Colorimetry: Definitions, Algorithms and Software*. John Wiley & Sons Inc, 2015.

Wasserman, L. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer, 2004.