Adapting Color Difference for Design

Danielle Albers Szafir[1,2], Maureen Stone[2], and Michael Gleicher[1]
[1] Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI; [2] Tableau Research, Seattle, WA

Abstract

CIELAB is commonly used in design as it provides a simple method for approximating color difference. However, these approximations model color perception under laboratory conditions, with correctly calibrated displays and carefully constrained viewing environments that are not reflective of complexity of viewing conditions encountered in the real world. In this paper, we present a data-driven engineering model for parametric color difference that extends CIELAB to be more broadly applicable to real-world conditions. Our model can be tuned to a desired range of viewers and conditions using a simple modeling procedure, while minimally increasing the complexity of the model. We demonstrate our approach empirically by modeling color differences for the web by leveraging crowdsourced participants.

Introduction

Color difference models are often used in design applications to predict how noticeably two colors will differ. These models serve several purposes, such as determining sets of colors that are subtly different (Fig. 1); however, they model perception under laboratory conditions, with correctly calibrated displays and constrained viewing environments. Given the rapid proliferation of visual content on the web and the increasing mobility of digital devices, visual media is becoming increasingly diverse, making factors that influence color difference perception, such as lighting conditions and display properties, highly variable in everyday viewing. Existing color difference models, while powerful descriptors of human vision, do not consider this variability, limiting their utility in design.

CIELAB is commonly used in design scenarios as it offers a color difference formulation based on Euclidean distance (ΔE^*_{ab}). This metric is not as accurate as other appearance models, such as CIECAM02 [1], but its simplicity makes it practical for design. In this work, we present an approach to adapt CIELAB to model color difference perception for real-world viewing populations that preserves its simplicity. As the range of viewing factors in these populations is too complex to model each independently, we capture these factors in aggregate by sampling difference perception across target viewers and use these samples to derive scaling factors for CIELAB.

The resulting model parameterizes CIELAB with respect to a given population and desired level of noticeable difference. Providing a parametric model tuned empirically to the target population exchanges the ability to make exacting claims about perceptual mechanism to instead create an *engineering* model that captures color difference in practice. This engineering model has several desirable properties for designers. It is *parametric* as it can be tuned to reflect a desired range of viewers and conditions. It is *data-driven* as it derives these parameters from observations under the target viewing conditions, yet *practical* as this data can be collected quickly using a simple task and requires only small modifications to common design metrics. Additionally, the model explicitly considers the *probabilistic* nature of the data, providing designers simple controls for defining how "noticeable" their desired color difference will be. We use this approach to model and validate discriminability on the web, using crowdsourced participants from Amazon's Mechanical Turk—a web-based platform that has gained popularity for quantitative experiments involving real-world participants. The resulting model provides an empirically validated metric for just-noticeable color difference for web content that aligns with common designer intuitions.

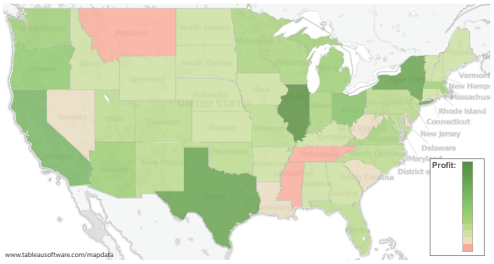


Figure 1. *Visualizations often use color sets with large numbers of subtly distinct colors for encoding quantitative data. We develop an engineering model that offers insight into the discriminability of different candidate color sets for their target audience to help guide design.*

Contributions: We introduce an engineering model for color difference that adapts CIELab to diverse viewing conditions using a relatively small amount of data. We present a procedure for deriving model parameters from limited experiments and apply the model to quantify discriminability for web applications.

Related Work

CIELAB was designed such that one unit of Euclidean distance equals one just-noticeable color difference (JND). However, prior work suggests that this may be an overly optimistic estimate. Several experiments have quantified just-noticeable color differences for CIELAB, such as the empirical benchmark from Mahy et al. we use in this study ($\Delta E^*_{ab} = 2.3$) [2]. These studies also demonstrate that CIELAB is not fully perceptually uniform even under ideal conditions [3]. Revised models of color difference have been developed for CIELAB, such as ΔE^*_{94} and CIEDE2000 [4], that account for these nonuniformities using hue and chroma (see [5] for a survey). However, these models tend be more mathematically complicated and less intuitive than the Euclidean ΔE^*_{ab} metric, improving accuracy at the expense of simplicity. Because of this trade-off, designers commonly use ΔE^*_{ab} in practice [3].

Existing CIELAB distance metrics quantify difference under laboratory conditions: lighting, display parameters (e.g. gamma and peak outputs), viewer position, and surround are all controlled. However, these factors substantially impact color difference perception [6, 7, 8]. Some efforts have attempted to account for variation caused by individual viewing factors [9, 10, 11], but do not consider interactions between factors. More general models exist for specific contexts [12, 13], but it is unclear how well these models generalize beyond their target applications. Our goal is to provide a color difference model that can be readily tuned to different environments and offers designers control over discriminability within those environments. We do this using a data-driven model sampled under the target conditions (e.g. mobile devices or clinical settings). With this approach, we do not need to consider each factor independently, but rather can account for expected variation factors in a more manageable way.

Crowdsourcing has greatly increased the feasibility of collecting data for diverse sets of viewers. Crowdsourced color studies can generate the same general insights as equivalent laboratory experiments [14, 15]. By conducting data collection via the web, crowdsourcing offers easy access to a large, low-cost participant pool under natural conditions, such as at home or at work. For this study, we used Amazon's Mechanical Turk to explore our modeling approach. Mechanical Turk has been shown to produce reliable results for quantitative evaluations of design techniques involving diverse participant pools [16]. We use our modeling technique to explore color difference for the web, taking advantage of the diversity of users on Mechanical Turk and compare these results to known color difference benchmarks.

A Parametric Color Difference Model

Our color modeling methodology builds on the CIELAB color difference model. CIELAB provides an effective approximation of color perception to create a space that is relatively perceptually uniform, yet sufficiently practical to use. The color space was designed such that the following assumptions hold [3]:

- **A1:** The axes are perceptually orthogonal, so they may be treated independently.
- **A2:** Euclidean distance (ΔE^*_{ab}) is an effective metric for perceived color difference.
- **A3:** The axes are perceptually uniform: differences at the higher end of the scale and lower end of the scale are the same.
- **A4:** The axes are scaled such that one unit along any axis corresponds to one just-noticeable difference.

While prior work shows that these assumptions do not always hold, they are still frequently assumed in design as they exchange a small amount of perceptual accuracy for a degree of practicality desirable for many design applications. This trade-off is often worthwhile for all but **A4**, which is generally addressed using non-empirical intuitions.

Our model aims to empirically adapt CIELAB such that **A4** holds for the designer's desired definition of "just noticable." Accepting the first three assumptions allows us to do so using a simple extension to the CIELAB model. We model discriminability along each axis independently on the basis of **A1**, which has the added benefit of empirically correcting any imbalance between lightness and chroma. **A1** and **A3** collectively allow us to determine a single scaling factor for each axis, denoted as $ND_L(p)$, $ND_a(p)$, and $ND_b(p)$, such that a difference of one unit along the scaled axis is noticeable for $p\%$ of the target viewing population. We derive these scaling factors using simple tasks to quickly measure discriminability across color differences for the target population and model this discriminability linearly. We use these scaling factors to normalize each color axis of CIELAB according to the designer's desired discriminability threshold using only one multiplication for each axis.

(a) Color Matching Stimulus (b) Forced Choice Stimulus

Figure 2. *Free-response color matching tasks provide insight into discriminability, but are of limited utility for probabilistic modeling. We use a forced choice microtask to measure discriminability as a function of color difference.*

The resulting adapted color model meets our desired goals: it is parametric, as the scaling factors can adapt to the viewing conditions; it is data-driven, as these parameters can be determined empirically from observations of the color difference; it is practical, as data collection can be done quickly and easily and model computations require only three multiplications beyond the standard CIELAB computation; and it is probabilistic, as we can dynamically define our desired noticeable differences and adjust the model accordingly. We confirm our approach by modeling color difference perception for the web. The subsequent sections discuss three studies that construct and validate this example model. The first describes a color matching pilot that provides insight into our modeling assumptions. The second illustrates our data collection and model construction methods. The third validates this model on 161 web viewers.

Insight from a Color Matching Task

Color difference is commonly measured using a free-response color matching task, where participants manually adjust a stimulus color to match a given reference color [17, 7]. The adjusted response colors provide a distribution of colors that appear to match the stimulus color. This procedure, akin to Maxwell's color matching experiments [3], has been used to measure JND for cross-media applications and provides substantial insight into color difference across color space.

We conducted a crowdsourced color matching experiment on Mechanical Turk to verify our modeling assumptions. Participants saw a 2° colored reference square centered on a 500 pixel-wide white background with a slider that adjusted the color of a second stimulus square (Fig. 2(a)) and were instructed to drag the slider until the adjusted color matched the reference color as closely as possible. Tested colors were sampled from the Swedish Natural Color System primaries [18] and varied at equal intervals along each axis of CIELAB within the gamut defined by $\gamma = 2.2$ and a D65 whitepoint [19], resulting in 24 distinct colors per axis. Our modeling procedure uses a constant gamma and whitepoint as, in practice, designers cannot feasibly adjust content to such

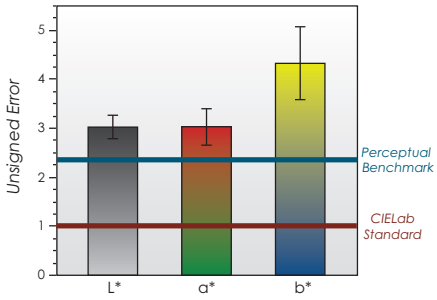


Figure 3. *Mean error for the color matching task. Web viewing discriminability thresholds exceed existing benchmarks and may vary between axes.*

display-dependent conditions. By holding these factors constant and measuring difference perception across multiple displays, we parameterize color difference for design using only information immediately available to designers. To simplify the color matching task for Turk workers who may be unfamiliar with CIELAB, participants were shown only one slider, corresponding to one axis of CIELAB (L^*, a^*, or b^*), as in [20, 7]. The slider displayed the full color range within the gamut along the tested axis through the reference color. Axis was a between-participants factor.

We recruited 48 participants (16 per axis, 33 female, 15 male) from age 18 to 61 ($\mu = 34.32, \sigma = 13.22$) with normal or corrected-to-normal vision and no color vision deficiencies. Participants were screened for color-vision deficiencies using five digital renderings of Ishihara plates [21] and asked self-report their approximate distance from the monitor, which was used in conjunction with DPI to compute the size of the 2° stimulus square. They completed a simplified tutorial to test task understanding and then asked to complete the color matching task for each of the 24 reference colors described above in a random order. Participants were given unlimited time for each response.

We analyzed the difference between response and reference colors (*error*) using a two-way ANCOVA (reference color and tested axis) with display distance and question order as covariates to account for interparticipant variation. Our results confirm that color difference perception is reduced on the web: per-axis mean errors ($\mu_L = 3.21, \mu_a = 3.03, \mu_b = 4.33$, Fig. 3) were significantly larger than both the theoretical JND ($\Delta E^*_{ab} = 1.0$) and our empirical benchmark ($\Delta E^*_{ab} = 2.3$), suggesting that existing metrics underestimate color difference for the web. Error varied significantly between axes ($F(2,997) = 11.2693$, $p < .0001$), but not within axes ($F_L(1,871) = 1.6072$, $p_L = .2052$; $F_a(1,862) = 1.8942$, $p_a = .1691$; $F_b(1,755) = 0.1875$, $p_b = .6651$). These findings support **A3**, but suggest that each axis should be modeled independently.

However, these results do not provide probabilistic insight into color discriminability— the adjusted colors identify a window of indiscriminable colors around a given reference, but do not capture how likely it is that these colors will appear distinct from the reference for different viewers. Without controlled insight into the likelihood that colors will appear different, designers cannot tune the model to their desired application settings. In the next section, we propose a task model which addresses this limitation in order to quantify the likelihood distribution of color differences for a target population and allows us to collect this data quickly and efficiently.

Constructing the Engineering Model

While the slider task provides insight into color difference perception, it suffers from a number of limitations. One limitation unique to crowdsourcing is that the sliders essentially provide continuous responses. Participants seek to complete a large number of tasks as quickly as possible to maximize their overall reward. For continuous response tasks, participants can optimize their time by providing answers that are "close enough" rather than taking the time to respond as accurately as possible.

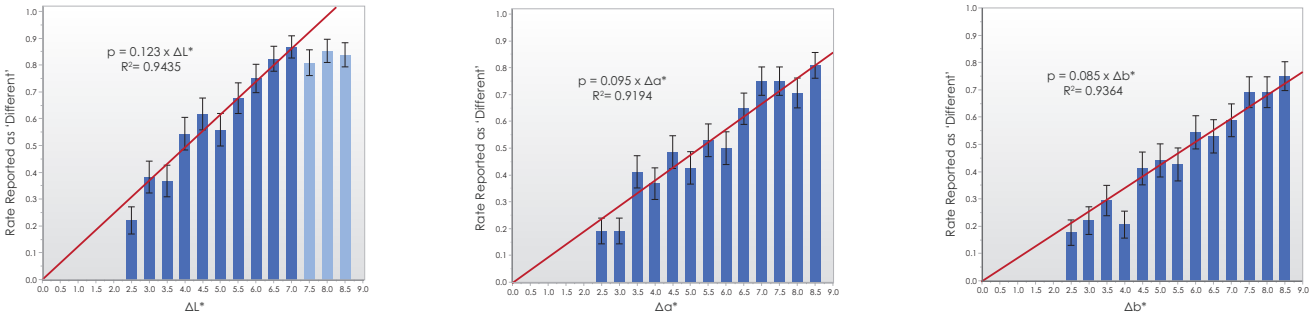
Effective crowdsourced studies use a *microtask* model, providing simple tasks that require roughly as much time to answer accurately as to answer "close enough" [16]. We designed a data collection microtask to measure how frequently colors appear to be different at specific color differences (*discriminability rate*) and use this measure to parameterize our color difference model. The task is a binary forced choice comparison of two colored squares based on the method of constant stimuli [22] (Fig. 2(b)). The squares differed in color by a controlled amount along one axis. Participants were asked whether the squares appeared to be the same color or different colors. We can quantify discriminability as a probabilistic function of color difference for our sample population by measuring how frequently the squares appeared to be different colors at different levels of color difference. We can also obtain a large amount of discriminability data efficiently: median response time for was 5.8 seconds per color pair.

Sampling Method

We can use the above microtask model to estimate per-axis scaling parameters ($ND_L(p)$, $ND_a(p)$, and $ND_b(p)$) representing the color differences along each axis perceived by p% of the target population. We computed these parameters for our web viewing model through a crowdsourced experiment involving 75 participants (37 female, 38 male) age 19 to 56 ($\mu = 31.05, \sigma = 9.75$) with normal or corrected-to-normal vision and no color vision deficiencies. Participants were asked to directly compare two 2° squares placed at opposite ends of a 8° white plane (Fig. 2(b)). One square was colored using a reference color randomly selected from a set of 316 colors from $L^* = 10$ to $L^* = 90$ evenly sampled from the CIELAB color space and within the color gamut defined by a standard PC gamma and whitepoint ($\gamma = 2.2$ and D65 whitepoint)[19]. The color of the second square differed from the reference color by a controlled amount along exactly one color axis (between 2.5 and 8.5 ΔE^*_{ab} sampled at 0.5 ΔE^*_{ab} increments).

Forced choice tasks are vulnerable to gamed responses: participants could provide random answers to complete the study quickly. To help mitigate such gaming and also to unbias the stimulus set, we included 20 stimuli with identically colored squares and two with obviously different colors. Three participants answered less than 65% of the same-color questions or one extreme difference correctly and were excluded from our analyses.

Participants were first screened for color vision deficiencies using digital renderings of Ishihara plates [21] and self-reported their distance from the display. They then completed three tutorial questions to ensure their understanding of our definition of "same" and "different" colors—two colors that varied in hue, two that varied in luminance, and two identical colors—and could not proceed until each was answered correctly. Participants were then shown a sequence of 61 stimuli (39 modeling stimuli and 22 validation stimuli) in a random order, with each reference color ap-



(a) (b) (c)

Figure 4. *An illustration of our modeling approach. A linear model (red) is fitted to the rate of 'different' responses across measured differences and forced through zero to account for sampling. Only color differences where discriminability changes with distance are modeled (dark blue).*

pearing twice and each color axis × color difference once. A two-second grey screen separated subsequent stimuli to minimize adaptation effects. Participants had unlimited time to respond.

We analyzed discriminability rates using a three-way ANCOVA (reference color, tested axis, and magnitude of difference) with display distance and question order as covariates to account for interparticipant variation. The magnitude of color difference significantly affected discriminability ($F_L(1,2846) = 169.0197$, $p_L < .0001$; $F_a(1,2846) = 163.0631$, $p_a < .0001$; $F_b(1,2846) = 148.5278$, $p_b < .0001$). Discriminability also varied significantly between tested axes ($F(2,2846) = 3.1380$, $p = .0244$). Reference color L^* significantly influenced discriminability ($F(1,2846) = 17.3941$, $p < .0001$), but the effect was small—the lightest colors were 0.3% more discriminable than the darkest. We found no significant effects of reference a^* or b^*.

Parameterizing CIELAB

To derive the parameters of our engineering model, we create linear models of this sampled discriminability data. These models express the sampled discriminability rates as a linear function of color difference for each axis of CIELab on the basis of **A1** and **A3** (Fig. 4). While identical colors should always have a discriminability rate of zero, sampling error can introduce noise that skews these linear models. To correct for such discrepancies, we construct the models with intercepts forced through zero and further account for sampling errors by treating interparticipant variability as a random factor. Likewise, these models are only fit to data below the upper bound of discriminability (e.g where tested differences are not immediately perceivable; dark blue in Fig. 4(a)), a point referred to as the *knee* [23].

The resulting models have the form $p = V_x d$ where x is the color axis, p is the desired discriminability rate, V_x is the slope of the model, and d is the color difference in ΔE^*_{ab}.We derive the parameters $ND_x(p)$ of the engineering model using the function

$$ND_x(p) = p/V_x \tag{1}$$

Assuming **A3** holds, we can divide color difference along each axis by $ND_x(p)$ to renormalize CIELAB such that $p\%$ of people modeled under our target conditions will detect color differences at $\Delta E_p = 1$. Given two colors (L^*_1, a^*_1, b^*_1) and (L^*_2, a^*_2, b^*_2), ΔE^*_{ab} can be adapted to the viewing population as:

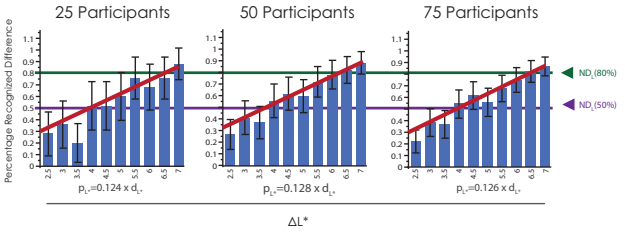


Figure 5. *Models can be generated using a relatively few samples. As the number of samples increases, the confidence in model increases, but the parameters estimated by the model remain roughly constant.*

$$\Delta E_p = \sqrt{(\frac{L^*_1 - L^*_2}{ND_L(p)})^2 + (\frac{a^*_1 - a^*_2}{ND_a(p)})^2 + (\frac{b^*_1 - b^*_2}{ND_b(p)})^2} \tag{2}$$

For our crowdsourced data, a traditional $p = 50\%$ JND maps to $ND_L(50) = 4.06\Delta E^*_{ab}$, $ND_a(50) = 5.26\Delta E^*_{ab}$ or $ND_b(50) = 5.88\ \Delta E^*_{ab}$ on the web (Fig. 4), which roughly aligns with popular designer intuitions. These color differences are notably larger than the CIE standard (1.0) and laboratory benchmark (2.3). Also unlike these benchmarks, these parameters vary for each axis.

We constructed models for 25, 50, and 75 participants (Fig. 5). These models yielded nearly identical parameter values, although the larger sample sizes provided greater statistical confidence in the parameters. This points to the practicality and reproducibility of our model: data from relatively few participants sufficiently characterize the model, and this characterization remained consistent across groups.

Validating the Adapted Model

We wanted to confirm that the engineering model derived in the previous section generalizes from the smaller tuning population to the larger target population and that it makes effective predictions when the color changes are not axis aligned. Our model predicts that if two colors are $\Delta E_p = 1.0$ different, then the viewing population will perceive them as different $p\%$ of the time. Colors with smaller ΔE_p will be perceived as different less frequently, and larger ΔE_p will be seen as different more frequently. We validated these predictions empirically using a second, larger group from the target population and a broader range of colors and

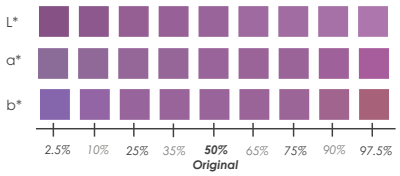


Figure 6. *We use the error distributions from the color matching experiment to generate per-axis color differences for validating our model.*

color differences, including cross-axis differences. We collected data from 182 crowdsourced participants (106 female, 76 male) ages 18 to 64 ($\mu = 30.60, \sigma = 9.78$) with normal vision and no known CVD to evaluate the model.21 participants were excluded for poor performance on the validation questions.

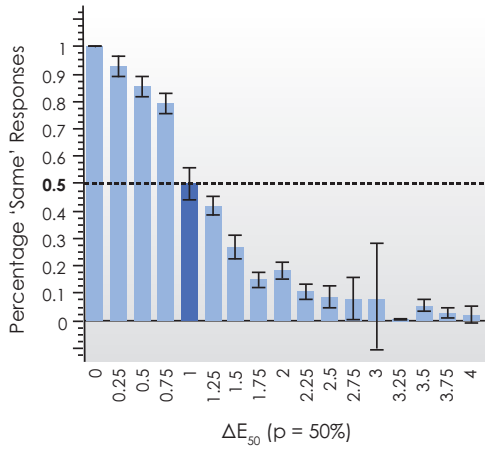
Our validation procedure modified the parameter sampling task to include a broader range of colors and color differences. Reference colors were sampled uniformly from CIELAB, but more densely than in the parameter sampling task. Color differences were drawn from the error distributions of the color matching experiment to create nine color difference levels per axis (Fig. 6). For each stimuli, one difference level was applied to each axis of the reference color to emphasize variation across multiple axes. Difference level combinations were drawn randomly for each participant and counterbalanced between participants. The procedure otherwise matched the parameter sampling task.

Our findings validated the model predictions (Fig. 7). Tuning the model to $p = 50\%$ discriminability predicts that participants can distinguish colors with a difference of $\Delta E_{50} = 1 \pm 0.125$ as different roughly 50% of the time. Our validation data confirmed this prediction: colors at this difference were differentiable in 49.81% of trials. These predictions are considerably better than the CIELAB specification ($\Delta E_{ab} = 1.0$) or our laboratory benchmark [2] ($\Delta E_{ab} = 2.3$), which were perceived as different for 7% and 13% of samples respectively. Further, colors less than $\Delta E_{50} = 1$ apart were consistently less discriminable, and more distant colors were perceived as more discriminable.

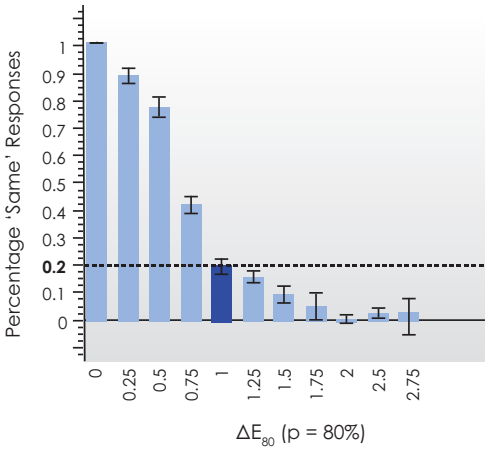
Model predictions were robust across discriminability levels. For example, tuning the model to 80% discriminability yields the parameters $ND_L(80\%) = 6.5$, $ND_a(80\%) = 8.42$, and $ND_b(80\%) = 9.41$). Applying this to our validation data, our population identified colors at $\Delta E_{80} = 1.0$ to be different in 80.62% of trials, confirming the model predictions. Across models at all discriminability levels, predictions were accurate to within 7% on average, and within 3.5% for models with $p \geq 50\%$.

Discussion: Limitations and Applications

The core feature of our modeling approach is that it is empirical: we can tune the model parameters by sampling the target viewing population. A myriad of factors can influence color difference perception, ranging from displays to viewing environments to the viewers themselves. Rather than trying to analyze each potential factor, we instead capture their effects in aggregate empirically. This allows our model to adapt to specific settings if necessary: we can model senior citizens using tablets in dimly lit cafes or students viewing projectors in classrooms by simply sampling these populations to construct specific model instances. The fact that we capture anisotropy in color difference perception helps the models provide good predictive performance.



(a) Validation results for $\Delta E_{50\%}$



(b) Validation results for $\Delta E_{50\%}$

Figure 7. *Plotting the percentage of perceived matches against ΔE_p tuned to (a) $p = 50\%$ and (b) $p = 80\%$ differentiability for the crowdsourced model shows that the model effectively predicts noticeable difference.*

The data-driven nature of our model is both a strength and limitation. Sampling quickly captures the specific conditions of a population; however, it offers no insight into how well an adapted model transfers between target populations. The efficiency of the microtask modeling approach helps alleviate this concern: large amounts of data can be collected quickly from a lay population. An additional limitation of this sampling method is that it does not characterize specific local viewing factors, such as gamma, ambient lighting, peak color output, and whitepoint. While this is a significant limitation from a colorimetry standpoint, it is a strength from a design standpoint as designers using color difference for cross-media applications do not necessarily have access to these variables when creating visual content; assuming constant parameters mirrors what designers do in practice.

Aspects of stimulus presentation, such as stimulus size and background color, may effect the results. We are exploring these issues [24], and, in practice, still expect our model to achieve reliable results. Also, we have only assessed a small number of applications to date. However, the fact that our approach works well on the challenging case of the web makes us optimistic that

it will be effective in other scenarios. We hope to explore new scenarios, such as mobile devices, in future work.

Color difference models are useful in a wide range of design applications including marketing, graphic design, digital art, watermarking [25], and visualization [26]. As displays become increasingly mobile, designers must consider a broader range of conditions and devices when designing for such applications. Metrics for color difference in design need to consider how to generalize laboratory models to fit these real-world design requirements. Our parametric color difference model attempts to capture real-world perceptions for specific populations using a relatively small amount of data. Our model also helps normalize color difference *between* color axes in practice, for example, to balance lightness and chroma for color map construction in visualization.

Conclusion

In this work, we present an engineering model of color difference for applications in design. This model attempts to account for variation in viewing condition by reparameterizing CIELab using data sampled directly from a target viewing population. This approach allows us to account for the broad variety of factors encountered in modern design scenarios by creating a model that is parametric, data-driven, probabilistic, and practical.

Acknowledgments

This work was funded in part by NSF awards IIS-1162037 and CMMI-0941013. We thank Vidya Setlur and Justin Talbot for their technical input.

References

[1] N. Moroney, M.D. Fairchild, R.W.G. Hunt, C. Li, M. R. Luo, and T. Newman. The CIECAM02 color appearance model. In *10th IS&t/SID Color Imag. Conf.*, number 1, pages 23–27, 2002.

[2] M. Mahy, L. Van Eycken, and A. Oosterlinck. Evaluation of uniform color spaces developed after the adoption of CIELAB and CIELUV. *Color Res. Appl.*, 19(2):105–121, 1994.

[3] M.D. Fairchild. *Color appearance models*. J. Wiley, 2005.

[4] M.R. Luo, G. Cui, and B. Rigg. The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Res. Appl.*, 26(5):340–350, 2001.

[5] A.R. Robertson. Historical development of cie recommended color difference equations. *Color Res. Appl.*, 15(3):167–170, 2007.

[6] B. Oicherman, M.R. Luo, B. Rigg, and A.R. Robertson. Effect of observer metamerism on colour matching of display and surface colours. *Color Res. Appl.*, 33(5):346–359, 2008.

[7] A. Sarkar, L. Blondé, P. Le Callet, F. Autrusseau, P. Morvan, J. Stauder, et al. A color matching experiment using two displays: design considerations and pilot test results. In *Proc. CGIV*, 2010.

[8] M. Stokes, M.D. Fairchild, and R.S. Berns. Precision requirements for digital color reproduction. *ACM T. Graphic.*, 11(4):406–422, 1992.

[9] K. Devlin, A. Chalmers, and E. Reinhard. Visual calibration and correction for ambient illumination. *ACM TAP*, 3(4):429–452, 2006.

[10] M.D. Fairchild and R.S. Berns. Image color-appearance specification through extension of CIELAB. *Color Res. Appl.*, 18(3):178–190, 1993.

[11] M.C. Stone. Color balancing experimental projection displays. In *9th IS&t/SID Color Imag. Conf.*, volume 7, 2001.

[12] L.D. Silverstein and R.M. Merrifield. Color selection and verification testing for airborne color crt displays. In *AACD Symp.*, pages 39–81, 1982.

[13] S.M. Pizer and FH Chan. Evaluation of the number of discernible levels produced by a display. In *Inf. Process. Med. Imaging*, pages 561–580. Editions INSERM, Paris, 1980.

[14] J. Heer and M. Stone. Color naming models for color selection, image editing and palette design. In *Proc. CHI*, pages 1007–1016. ACM, 2012.

[15] S. Zuffi, C. Brambilla, G.B. Beretta, and P. Scala. Understanding the readability of colored text by crowd-sourcing on the web. Technical report, J. of Color Res. Appl., 2009.

[16] M. Buhrmester, T. Kwang, and S.D. Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.*, 6(1):3–5, 2011.

[17] BH Crawford. Colour matching and adaptation. *Vision Research*, 5(1):71–78, 1965.

[18] A. Hård and L. Sivik. NCS Natural Color System: a Swedish standard for color notation. *Color Res. Appl.*, 6(3):129–138, 2007.

[19] M. Stokes, M. Anderson, S. Chandrasekar, and R. Motta. A standard default color space for the internet-sRGB. *Microsoft and Hewlett-Packard Joint Report*, 1996.

[20] R.L. Alfvin and M.D. Fairchild. Observer variability in metameric color matches using color reproduction media. *Color Res. Appl.*, 22(3):174–188, 1997.

[21] H. Legrand, G. Rand, and C. Rittler. Tests for the detection and anlysis of color-blindness I. The Ishihara test: An evaluation. *JOSA*, 35:268, 1945.

[22] E.G. Boring. Urban's tables and the method of constant stimuli. *Am. J. Psychol.*, 28(2):280–293, 1917.

[23] R.C. Carter and L. D. Silverstein. Size matters: Improved color-difference estimation for small visual targets. *J. Soc. Inf. Display*, 18(1):17, 2010.

[24] M.C. Stone, D. Albers Szafir, and V. Setlur. An engineering model for color difference as a function of size. In *22nd IS&t/SID Color Imag. Conf. (to appear)*.

[25] C.I. Podilchuk and W. Zeng. Image-adaptive watermarking using visual models. *IEEE J. Sel. Areas Commun.*, 16(4):525–539, 1998.

[26] S. Silva, B. Sousa Santos, and J. Madeira. Using color in visualization: A survey. *Comp. & Graph.*, 35(2):320–333, 2011.

Author Biographies

Danielle Albers Szafir received her BS in Computer Science from the University of Washington (2009) and is currently completing her PhD in Computer Sciences at the University of Wisconsin–Madison under Prof. Michael Gleicher. Her work focuses on understanding scalability and interpretability in visualization design.

Maureen Stone is a Research Scientist at Tableau Software, where her work focuses on enhancing the effectiveness of information visualization by blending principles from perception and design. Prior to joining Tableau, she worked first at Xerox PARC, then independently as StoneSoup Consulting. She has a BS and MS in Computer Engineering from the University of Illinois, and a MS in Computer Science from Caltech.

Michael Gleicher is a Professor in the Department of Computer Sciences at the University of Wisconsin–Madison. His research interests span the range of visual computing, including visualization and character animation. Prior to joining the university, Prof. Gleicher was a researcher at Autodesk and Apple Computer. He earned his Ph.D. in Computer Science from Carnegie Mellon University and a BSE from Duke University. Prof. Gleicher is an ACM Distinguished Scientist.