Color Compatibility From Large Datasets

Peter O'Donovan
University of Toronto

Aseem Agarwala
Adobe Systems, Inc.

Aaron Hertzmann
University of Toronto

Abstract

This paper studies color compatibility theories using large datasets, and develops new tools for choosing colors. There are three parts to this work. First, using on-line datasets, we test new and existing theories of human color preferences. For example, we test whether certain hues or hue templates may be preferred by viewers. Second, we learn quantitative models that score the quality of a set of five colors, called a *color theme*. Such models can be used to rate the quality of a new color theme. Third, we demonstrate simple prototypes that apply a learned model to tasks in color design, including improving existing themes and extracting themes from images.

Links: DL PDF WEB DATA CODE

1 Introduction

Graphic design relies on effective use of color, and choosing colors is a difficult but crucial task for both amateur and professional designers. Designers often look for inspiration from many sources, such as art, photography, and color palette books. Color choice is guided largely by intuition and qualitative rules, such as theories of complementary colors and warm versus cool colors. It is generally believed that certain color combinations are harmonious and pleasing, while others are not. In the past two centuries, many theories of color compatibility have been proposed to describe and explain these phenomena, but there has been little large-scale testing.

On-line communities provide new ways for graphic designers to create and share color designs. Two websites, Adobe Kuler and COLOURLovers, allow users to create *color themes*, i.e., ordered combinations of 1-5 colors, though the vast majority have 5-colors. Each theme has a name, but is otherwise free of context. Users may rate, comment on, and modify previously-created themes. Over two million themes have been created on these sites, by tens of thousands of users. The datasets produced by these websites provide an opportunity for quantitative study of color theories and development of new color compatibility models.

This paper employs on-line datasets to study color compatibility, with three main goals. First, we test new and existing theories of color compatibility. For example, we test to what extent certain hues or hue templates may be preferred by viewers. Second, we learn quantitative models to rate the quality of a color theme. Third, we demonstrate simple prototypes that apply these learned models to tasks in color design, including improving existing themes and extracting themes from images. Together, these prototypes illustrate how the development of effective color compatibility models could be useful for various tasks in graphic design and computer graphics.

Our studies are based on three datasets, each of which comprises a collection of color themes and their ratings. We derived two datasets from Kuler and COLOURLovers, and created the third using Amazon Mechanical Turk ("MTurk"). These datasets exhibit different advantages and disadvantages. For example, Kuler users have more exposure to color theory than MTurk workers, while MTurk data is collected in a more controlled fashion. However, taste in color can vary widely, and users in these datasets have varying goals, backgrounds, and viewing environments; not surprisingly, there is substantial variation. Nonetheless, analysis of the data reveals many regularities and patterns.

We first analyze these datasets to understand which colors people use, and how colors are combined. Our main observations are as follows. User-created themes are far from random; themes form clusters or manifolds in the space of 5 colors, and themes farther from this manifold tend to be rated worse. People also have strong preferences for particular colors. The data reveals a preference for warm hues and cyans in color themes, which is distinct from preferences for purples and blues with single colors. Hue templates, the most popular models of color compatibility, are tested in several ways, and no evidence is found that they predict compatible colors. We examine the number of distinct hues people prefer in a theme, and find users generally prefer themes which are neither too simple (i.e., monochromatic), nor too complex (more than 2-3 different hues). Further MTurk experiments indicate that theme names usually do not affect the rating, though evocative names can have an impact.

We offer a new color compatibility model for predicting ratings, and examine which features of color themes are most important. Our model is distinct from previous work in that it uses a large number of features in many color spaces. The model is learned by linear regression with an L1-norm, thereby selecting the most relevant features for predicting the aesthetic rating. In particular, lightness features are important; dark themes are poorly rated and gradients from light-to-dark or vice-versa are preferred. Choosing popular adjacent color pairs is important, and theme colors should not be too similar to each other.

Aside from their scientific value, effective compatibility models would be useful for numerous tasks in graphic design and computer graphics, where selecting colors is often challenging. To that end, we demonstrate simple prototype applications, such as improving an existing color theme, extracting a compatible theme from an image, and suggesting colors given some existing colors. Pilot user studies on MTurk show that users prefer our results over simple baselines. Our learned predictors with source code, datasets (aggregate ratings only), and supplementary material are available on-line.

2 Background

Color has intrigued philosophers since the ancient Greeks [Gage 1999]. Modern color theory began with Newton, who developed a color wheel with hues arranged according to wavelength. Color wheels allow color relationships to be represented geometrically. Goethe [1810] arranged the color wheel according to physiological vision phenomena such as after-images; he proposed that compatible contrasting colors are opposite on the color wheel.

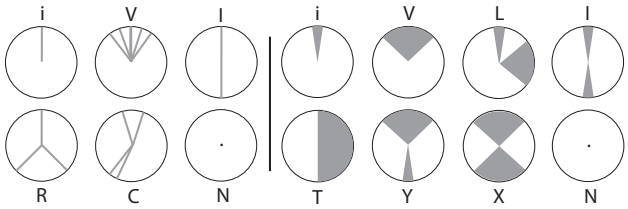

Figure 1: ***Hue templates implemented in Kuler (left), and those proposed by Matsuda [1995] (right).*** *Kuler implements several color selection rules (similar to Matsuda's i, V, I), as well as others: t(R)iad, (C)ompound. Each theme is described by a color wheel, with gray areas for the hues used by that theme. COLOURLovers implements the i, V, I, R, Y, X templates (see Supplemental Material §4). Matsuda uses sectors over the hue wheel, whereas Kuler and COLOURLovers use fixed angle distances, similar to classical theory (see Appendix B for discussion).*

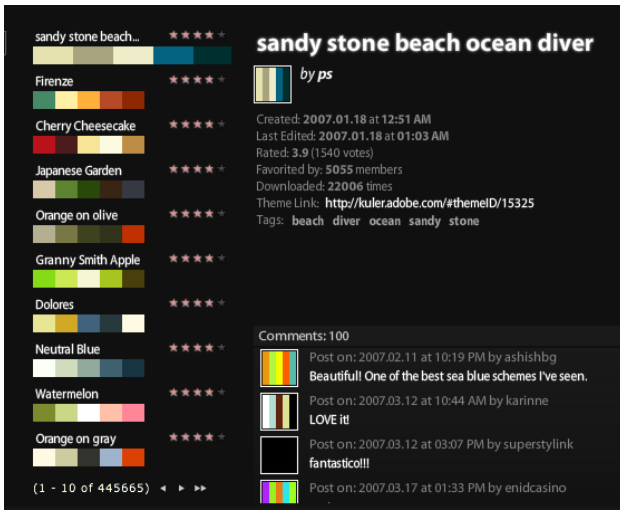

Figure 2: ***All-time most popular color themes from Kuler,*** *as of January 7, 2011. Each color theme consists of five colors. The interface shown here also provides some statistics about the ratings, as well as commentary from other users.*

Hue Templates. One of the most popular theories of color compatibility is the notion of *hue templates*, which generalizes Goethe's theory by describing compatible colors as fixed rotations about the color wheel. Hue templates are taught in many texts on art and design [Itten 1973; Krause 2007]. Itten [1973] proposed that sets of 2, 3, 4, and 6 hues equidistant on the color wheel were harmonious. Templates are considered equally harmonious, and rotationally invariant along the color wheel. However, designers often treat templates as starting points, rather than strict rules [Meier et al. 2004]. One weakness of hue templates is they are defined independently of the underlying hue wheel. Kuler uses a BYR color wheel (the "artists' color wheel"), whereas COLOURLovers uses an RGB color wheel, which suggest different colors using the same template rules. Fig. 1 shows the templates implemented in Kuler and COLOURLovers, which cover the most popular hue templates.

Matsuda's color harmony model [1995] has been used in several computer vision and graphics projects [Cohen-Or et al. 2006; Li and Chen 2009; Tokumaru et al. 2002]. Matsuda derived a set of 8 hue templates (Fig. 1(right)) and 10 tone templates from fashion questionnaires given to female students in Japan over a nine-year period, and from color themes provided by fashion companies. To our knowledge, these templates have never been rigorously evaluated. We do not evaluate tone templates, focusing on hue templates due to their wider usage.

Color Harmony. Numerous theories have been proposed for color harmony beyond just hue [Chevreul 1839; Munsel 1921; Ostwald 1932; Nemcsics 2003]. While the underlying color spaces typically vary, these theories often have similar rules. Many suggest colors are harmonious if one dimension of the space (such as saturation or value) contrasts while the others remain fixed, or that colors along lines in the color space are harmonious. For example, the Munsell system suggests that color themes with fixed hue and value but varying saturation are harmonious. The Ostwald system suggests that colors are harmonious with equal white or black content. These sets of colors form lines in that color space.

In recent decades, psychologists have begun controlled studies of color compatibility and preferences [Granger 1952; Ou et al. 2004; Ou and Luo 2006; Szabó et al. 2010; Matsuda 1995; Palmer and Schloss 2010; Schloss and Palmer 2010; Neumann et al. 2005]. While this work is often contradictory [Schloss and Palmer 2010], a few trends emerge: colors harmonize if they have the same hue, equal or similar color saturation, and contrasting lightness values. The data comes from tightly-controlled laboratory experiments, which forces a small number of participants (usually less than 100), a small range of colors (usually less than 100), and a small number of combinations (usually 1-3). An exception is the Coloroid system [Neumann et al. 2005], derived from several large-scale experiments on sets of up to three hues, though often with a small number of stimuli (e.g., only 108 color combinations for 3 hues). By contrast, our approach uses vast datasets of 5-color themes from thousands of participants from across the globe with a very broad range of colors and viewing conditions. Another significant difference is that we explore compatibility in user-generated color combinations.

There is little consensus among these different models. In our work, we use an exploratory learning procedure to identify relevant features from a large set in several color spaces. In addition, we also evaluate hue templates, the most widely-used color harmony model.

Learning Aesthetics and Applications. Rated training data is increasingly being used to learn computational models of aesthetics, including for photographs [Datta et al. 2006], Impressionist paintings [Li and Chen 2009], and videos [Moorthy et al. 2010]. Color features are often important to these methods. However, these results cannot necessarily generalize to other contexts. We study color themes without context, which permits more generic analysis, though may not apply as well to a specific context. Csurka et al. [2010] learn associations between 5-color themes and keywords, and use themes for image recoloring, but do not study preferences.

A few applications of color theory and themes have begun to emerge in computer graphics and vision. Color harmonization [Cohen-Or et al. 2006] optimizes the histogram of hues in an image to lie within the closest of Matsuda's hue templates. Images may also be harmonized with respect to a theme from other sources such as flags. Wang et al. [2010] modify the colors of an image to match a user-selected color theme. Lalonde and Efros [2007] use color compatibility to evaluate image realism for realistic recoloring and compositing. We focus on evaluating and improving color themes, and our methods could provide inputs for such algorithms.

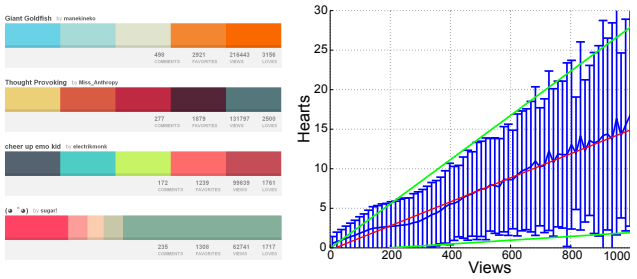

Figure 3: ***Top: All-time most-hearted themes from COLOURLovers****, as of January 7, 2011. COLOURLovers themes may also include proportions, which we do not use in this paper.* ***Bottom: views vs. hearts for COLOURLovers dataset.*** *The red line is the fit of the histogram means $\bar{h}(v)$, the green is the fit of the standard deviations $\sigma(v)$. Ratings are estimated as $r(h, v) = (h - \bar{h}(v))/\sigma(v) + 3$. For example, at 800 views, a theme requires more than 10 hearts to receive a rating over 3.*

3 Datasets

Our work employs three datasets which we first describe at a high level. See Appendix A for detailed statistics of the datasets and how they were created. The *Kuler dataset* comprises 104,426 5-color themes created by visitors to the Kuler website (`kuler.adobe.com`). Each theme can be rated on a discrete scale of 1 to 5 (Figure 2). The dataset includes 327,381 ratings from 22,376 users. Except where noted, in all of our experiments we omit themes with less than 2 ratings, leaving 46,137 themes with 266,239 ratings.

The COLOURLovers site (`www.colourlovers.com`) includes over one million 2-5 color themes created by users. While themes are not rated directly, users may "heart" any theme. The total number of hearts h and user views v is given for each theme. We define a rating function $r(h, v)$ that computes a score based on h and v (see Fig. 3 and Appendix A). Our *COLOURLovers dataset* includes 383,938 five-color themes downloaded from the COLOURLovers website. Ratings for 178,086 themes with over 100 views were estimated.

In order to obtain data under more controlled conditions, we created a third dataset using Amazon Mechanical Turk (*MTurk dataset*) with 40 user ratings per theme. We selected 10,743 Kuler themes covering a range of poorly-rated to highly-rated themes, and with at least three user ratings each.

These datasets each have advantages and limitations. Kuler and COLOURLovers are used by highly-motivated designers, including professionals with an interest and aptitude for color. However, no demographic information is available to control for the different tastes, backgrounds, or goals of users. Hence, these datasets may mix together very different aesthetics and design goals. Kuler and COLOURLovers have interface biases, namely, specific affordances for creating themes with one of several standard hue templates. Both sites encourage users to name their themes.

The data also includes numerous rating biases. Themes may be rated non-uniformly—with some exceptions, most themes in Kuler have very few ratings, and our ratings for COLOURLovers are inferred (see above). Both sites promote highly-rated themes, so popular themes are more likely to get more ratings, the so-called "rich-get-richer" effect [Easley and Kleinberg 2010]. Also, a user's opinion of a theme can affect whether or not they rate it. For example, a few users only rate themes with 4 or 5 star ratings. When missing ratings are not "missing-at-random," learned estimators can perform poorly on random test data [Marlin and Zemel 2007].

By contrast, MTurk users are assigned random themes to rate. MTurk also allows us to ensure that we get sufficient numbers of ratings for a wide range of themes; we can avoid community biases and naming biases. MTurk workers are much less likely to be professionals with interest or experience in working with color. MTurk experiments are less controlled than the in-person experiments common in the psychological literature, but they can be run on a far larger scale. Heer and Bostock [2010] have demonstrated the viability of MTurk for graphical perception experiments by comparing to classic results from the literature.

All our datasets may have variation due to differences in users' monitors, viewing conditions, and color blindness. Color calibration in on-line experiments is a challenging problem. However, most graphics applications are aimed at uncalibrated viewing, so finding colors which are compatible on average over many viewing conditions is important for graphic designers. Lastly, our conclusions are restricted to the color gamut of conventional monitors.

4 Model-Free Data Analysis

In this section, we consider general questions of color compatibility, independent of specific learning algorithms. First, the data density of user-created themes is measured to evaluate whether themes form a manifold or clusters in the space of 5 colors. Preferences for individual colors and colors in combination are examined, as well as the popularity of adjacent hue pairs. Hue templates are evaluated in detail, specifically looking at the rotational invariance assumption, as well as the prevalence of templates not implemented in the user interfaces. Based on our experiments with hue templates, we define hue entropy, a theme complexity metric which roughly corresponds to how many distinct hues are present in a theme. Lastly, the impact of theme names on user ratings is evaluated.

4.1 Distribution of Themes

To what extent do user-created themes lie on a manifold or in clusters, as opposed to being uniformly distributed throughout the space of possible themes? We can explore the existence of this manifold by measuring the distance from a single theme to a separate set of user-created themes. Fig. 4 (left) shows a histogram of distances from 10,000 Kuler themes to a disjoint set of 27,000 Kuler themes. For comparison, the distribution of distances from 10,000 themes uniformly sampled in each of RGB and HSV space is shown as well. A similar histogram for COLOURLovers to Kuler themes is also shown. The distance was calculated by the sum of CIELab distances for each color in the theme. The mean distance to the nearest 10 themes indicates the distance to the manifold. The shape of these plots indicate the degree of clustering: Kuler and COLOURLovers demonstrate similar clustering, whereas the random themes are considerably more spread-out. Kuler and COLOURLovers themes therefore form a manifold in the space of 5 colors.

Given the non-uniform distribution of themes, does proximity to other themes help indicate the rating of a new theme? Fig. 4 (right) plots mean rating for a test theme against the distance to the 10 Nearest Neighbors. A downward trend appears in all three datasets (especially in Kuler and MTurk), indicating that unusual themes are more likely to have lower scores. Hence, a new theme that is similar to existing themes is more likely to receive a higher rating.

4.2 Preferred Colors and Color Pairs

We next examine overall preferences for colors, both for single colors and in combination. Previous work on color preferences, based on in-person surveys of small groups of individuals (see included

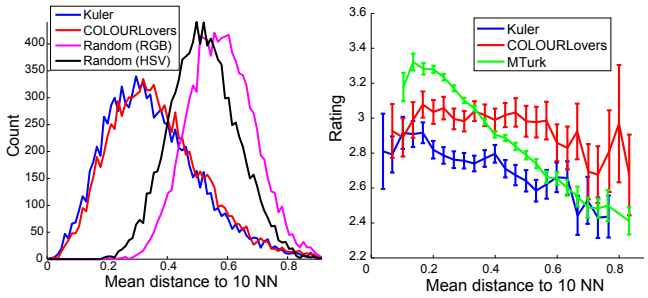

Figure 4: ***Data density vs. rating. Left:*** *Theme distance to Kuler training set. This suggests that the Kuler and COLOURLovers datasets are similar, and quite different from data randomly sampled in RGB or HSV space.* ***Right:*** *Distance to Kuler training set vs. mean rating. Themes similar to the other data are more likely be highly rated. Error bars show 2 std error. These results suggests that color themes occupy a manifold of higher aesthetic scores.*

references in [Ou et al. 2004]) find that most people generally prefer blues and purples and dislike yellows. For comparison, we performed a study in which 100 MTurk participants were asked to rate 126 colors, each shown on a uniform grey background. Colors were densely chosen in HSV space and participants asked to rate colors between 1-5. The color preferences we measured match the previous findings (Fig. 5, MTurk-Colors), which helps validate the use of MTurk for color preference studies. Both plots were first convolved with a Gaussian filter of width 5 and $\sigma = 1$.

Fig. 5 indicates the hue histogram for the Kuler and COLOURLovers datasets, and indicate greater density for warm hues (red, orange, yellow) and blues. A large spike appears at pure red, perhaps because red lies at the ends of the hue sliders in both interfaces. Fig. 5 shows the average rating assigned to themes containing that hue. These ratings mix together the contribution of each color to the rating of the theme; unmixing the contributions of the colors yields similar results (see Supplemental Material §1). We find that single-color preferences do not match preferences for color combinations, a result that echoes previous findings from in-person studies [Ou et al. 2004]. The preferences of Kuler and COLOURLovers users appear more similar to each other than to the MTurk users, which might reflect that the former are more likely to be professional designers. We conjecture that designers may be rating themes on their usefulness, whereas MTurk users are asked to rate solely based on visual appeal.

Fig. 6 shows the pairwise co-occurrence of adjacent hues in themes. The overall joint distribution of hues in a theme is quite similar to the pairwise distribution (see Supplemental Material §3). Warm hues around yellow and red have strong adjacency, and yellow and cyan are often paired with many other hues. Green and purple are relatively unpopular hues, and are more commonly paired with similar hues.

4.3 Hue Templates

Perhaps the most prominent theory of color compatibility is the notion of *hue templates*: fixed sets of rotations around the color wheel which produce compatible colors (see Figure 1 for examples). Here we investigate whether these templates describe the themes that users create, and whether the use of a template predicts better ratings. Previous research for 2 and 3 color combinations did not find complementary and triadic hue templates to be harmonious [Ou and Luo 2006; Szabó et al. 2010] but these studies were limited to 17 and 9 participants, respectively. To our knowledge, ours is the first

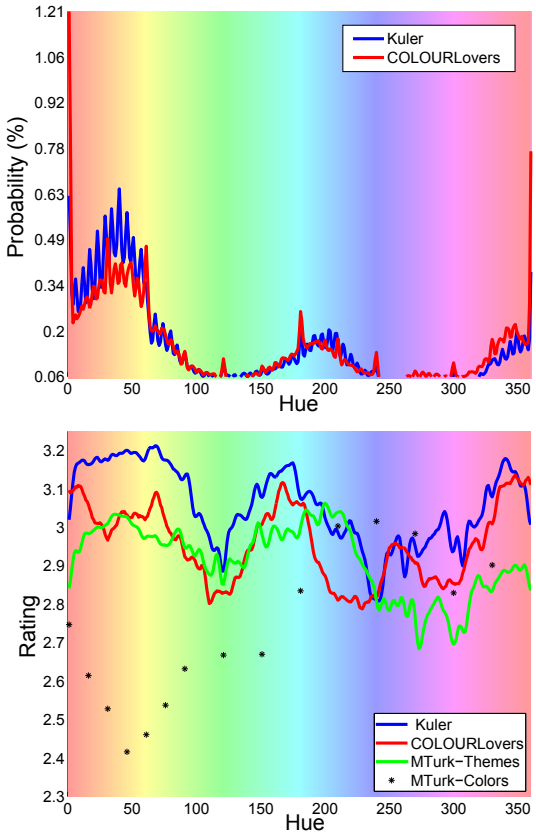

Figure 5: ***Color preferences. Top:*** *Histogram of hues for Kuler and COLOURLovers datasets.* ***Bottom:*** *Mean rating of themes containing each hue, and individual color ratings from MTurk. While the 3 datasets vary, they display similarities in relative hue preference, particularly compared with individual color ratings. Both datasets have a spike at pure red ($hue = 0$) caused by an interface bias.*

study exploring compatibility of combinations of up to five colors, or with user-created combinations.

The use of color templates can be clearly seen in the Kuler pairwise color histogram (Fig. 6(bottom)): diagonal lines in the histogram correspond to fixed rotations about the color wheel. This may be explained as a result of interface bias: though both Kuler and COLOURLovers provide tools for creating themes with templates, they are harder to find and use in COLOURLovers. This suggests that designers do not create themes that match templates unless encouraged to do so by the interface. These histograms also strongly suggest that, contrary to belief, preferences are *not* rotationally-invariant about the color wheel: green's complement is purple, yet these plots suggest users prefer to pair green with blue or yellow instead. On the other hand, orange often pairs with cyan, its complement on the hue wheel.

We investigate hue templates in more detail by assigning each theme to the closest template (see Appendix for details). Figure 7 gives a histogram of distances from themes to the templates implemented in Kuler, as well as to Matsuda's templates. A commonly-used template should appear as a spike at zero distance. We see that themes implemented in the Kuler interface do appear often, whereas none of Matsuda's additional themes do. This indicates that designers are not gravitating to Matsuda's templates of their own accord. The histogram also shows that monochromatic (i), analogous (V), and complementary (I) templates are the most popular. These are the most elementary and basic templates. COLOURLovers shows even less use of templates, as can be seen in Figure 6. Plots of tem-

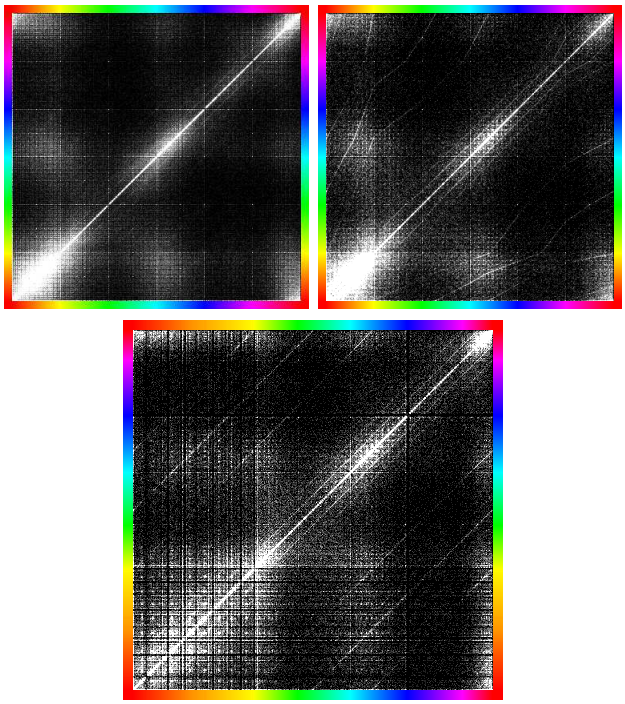

Figure 6: ***Pairwise histogram for adjacent hues****. High brightness indicates higher probability.* ***Top left:*** *COLOURLovers dataset.* ***Top right:*** *Kuler dataset.* ***Bottom:*** *Kuler dataset remapped to the BYR color wheel used in the interface. Diagonal lines correspond to hue templates and show a lack of rotationally invariance in the dark bands around purple and green. Lack of smoothness in the data is apparent in the BYR histogram, also visible in Fig. 5(top).*

plate distances for the COLOURLovers dataset indicate the presence of only the (i, V, I) templates (Supplementary material §4).

Does the distance to a hue template help predict ratings? Fig. 8 plots the rating of themes as a function of distance to their nearest template. Distance to the template does not appear to be closely related to the rating. However, in Kuler and COLOURLovers, themes which are close to templates have slightly lower scores. We hypothesize these templates are created by more inexperienced users, or the community penalizes themes close to the interface defaults. More importantly, the MTurk results have no interface biases, and also show little evidence that template distance affects rating. We also plot the relation between rating and template distance for each individual template (see Supplemental Material §4), and none show a positive relationship between rating and adherence to a template.

We also computed an assignment of the themes to each of the templates (see Supplemental Material §4). These results indicate a preference for themes near simpler templates like i, V, I compared to more complex templates like R, X. However, these results are dependent on the distance threshold used.

In short, although users appear to use templates built into the Kuler interface, we find little evidence that people gravitate to templates naturally, or that matching a template produces higher scores.

4.4 Hue Entropy

As noted in the previous section, simple themes—typically consisting of one hue, two hues, or a blend—tend to be more popular

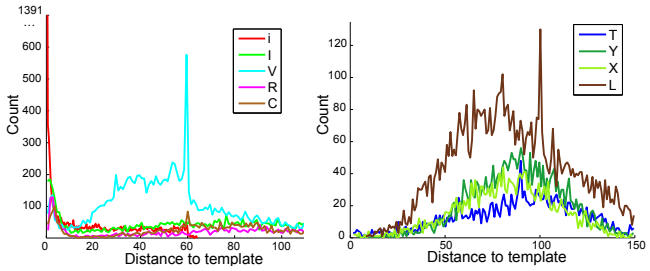

Figure 7: ***Template distances*** *for Kuler-implemented templates, and for the rest of Matsuda's templates. Note the spike around 0 for the implemented templates not present for the other templates. The spike at 60 degrees for the V template is caused by monochromatic themes with a single accent color between 30-60 degrees.*

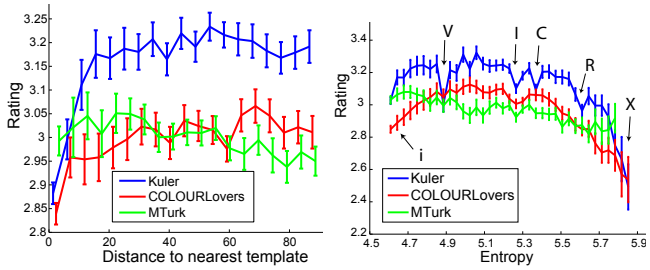

Figure 8: ***Left: Nearest template distance vs. theme rating.*** *Themes that closely match templates tend to score lower than those that do not. Beyond this, increasing the distance to a template does not significantly affect the rating.* ***Right: Hue entropy vs. theme ratings.*** *The entropy values for equally-spaced hues are, from 1 hue to 5 hues: (4.62, 5.29, 5.65, 5.81, 5.87) where the first 4 correspond to the i, I, R, X templates. The analogous template V has an entropy value of 4.89, the C template is 5.33. Monochromatic or complex themes (i.e., more than 2-3 hues) tend to rate worse. Themes too close to an interface template result in a significant ratings drop in the Kuler dataset (see Sec. 4.4).*

than complex ones. Here we propose *hue entropy* $H(\mathbf{t})$ as a single measure of the simplicity of a theme $\mathbf{t}$. Let $\theta_1, ..., \theta_5$ be the hues in a theme, represented as angles. We convert these hues into a probability distribution as a mixture of von Mises distributions: $p(\theta) \propto \sum_{i=1}^{5} \exp\left(\kappa \cos(\theta - \theta_i)\right)$. The hue entropy is then the entropy of this distribution, computed numerically. $\kappa = 2\pi$ was selected by fitting a mapping from entropy to rating using cross-validation. Hue entropy is lowest when all values of θ are identical ($H = 4.62$), and highest when they are uniformly spread about the circle ($H = 5.87$). Colors spread over a narrower range will have lower entropy than a wide range (e.g., red-to-magenta has lower entropy than red-to-green).

Figure 8 (right) shows the relation of hue entropy to theme rating in the three datasets. The data shows a clear trend: for the Kuler and COLOURLovers ratings, there is a preference for entropies roughly in the range 4.7–5.4. This corresponds to themes with about 2-3 hues. MTurk ratings are more uniform, with no penalty for monochromatic themes, though scores trend downward for themes with more colors. This highlights a difference between the datasets: Kuler and COLOURLovers users are likely evaluating the usefulness of themes for design tasks, and monochromatic themes are rarely useful for design. By contrast, MTurk users are rating purely visual appeal. The figure also shows clearly that using templates provided by the interface correlates with lower ratings.

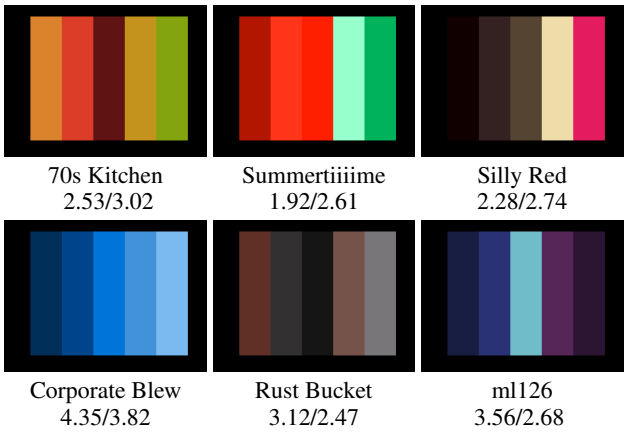

Figure 9: ***Effect of theme names.*** *For each theme, the name, and the average MTurk rating without and with name, respectively, are shown. In most cases (95% of our test themes), names have no statistically significant impact on rating. However, in a few cases, names improve scores (top row), or worsen them (bottom row).*

4.5 Theme Names

All themes on Kuler and COLOURLovers are given a name. Names provide some context to themes, describing what they are meant to evoke or how they might be used. Palmer and Schloss [2010] provide evidence that color preferences are affected by real-world associations, and color names might affect these associations. To what extent does the name affect the rating of a theme?

Using MTurk, we asked 40 participants to rate 216 themes where each participant was shown all themes either with or without their original names from Kuler. Of the 216 themes, only 10 showed a significant difference (using the Mann-Whitney U-test) in scores with the name versus without. Hence, names affect rating only in a minority of cases, though as Fig. 9 illustrates, evocative names can have a significant impact. The mean of the absolute difference for named vs. unnamed ratings was 0.20, median 0.16, and max 0.88 (the name "ml126" decreased the rating from 3.56 to 2.68).

5 Learning Color Compatibility

We now describe methods for learning to rate color themes. These learned models allow us to perform additional analysis of the data, and to create new color-selection applications.

The input data comprises pairs $(\mathbf{t}_i, r_i)$, where $\mathbf{t}_i \in R^{15}$ represents theme i, and $r_i \in [1...5]$ is the mean user rating for this theme. Our goal is to predict the mean rating r_{new} for a new theme $\mathbf{t}_{new}$. For performance, we only use themes with at least 2 user ratings each. We learn separate models for all datasets. For the COLOURLovers dataset, we restrict our regression tests to 60,000 randomly-selected themes.

Feature vectors. As input to the learning algorithm we define a *feature vector* $\mathbf{y}$ that can be computed from any input theme $\mathbf{t}$. However, due to the exploratory nature of this work, we do not know in advance what the best features are. The feature vector comprises a large set that might be useful; finding relevant features is valuable for understanding color compatibility in general.

The feature vector has 334 dimensions and is constructed from four color spaces: RGB, CIELab, HSV, and CHSV[1]. We create the following features in each space: the five colors themselves, colors sorted by lightness, differences between adjacent colors, sorted color differences, mean, standard deviation, median, max, min, and max minus min across a single channel. Differences in hue are computed with wraparound. Note that some features are redundant. For example, lightness/value is represented in many spaces. Since many themes lie along lines or planes in color space, we also include plane-fitting features in RGB, CIELab, and CHSV. A 2D plane is fit to the 3D color coordinates using PCA and the plane normal, eigenvalues, and sum-of-squared error used. Hue entropy is also included (Section 4.4). All features are normalized to the range 0...1.

Lastly, we use the color histograms from the Kuler training set to produce scores for individual colors and pairs of colors. Let p_c be the percentage of colors in the training themes with hue c (Figure 5), p^a_{bc} be the percentage of adjacent colors b and c (Figure 6), and p^j_{bc} be the percentage of colors b and c in the same theme. A list of hues from saturated colors and light colors (both are determined by thresholding) is first created from a theme. The probability of each hue is taken, and the mean, standard deviation, min, and max computed. Features are also computed with log probabilities. The same features are then computed for the pairwise probabilities p^a_{bc} and p^j_{bc}. When there are no saturated or light colors, the features are set to 0.

Regression. In regression, we learn a continuous mapping from feature vector $\mathbf{y}$ to rating r. We test the following regression algorithms: LASSO [Tibshirani 1996], robust linear regression using Iteratively-Reweighted Least-Squares, Support Vector Machine regression (RBF kernel), and K-Nearest Neighbors [2]. As a baseline, we use a fixed regressor that outputs the mean rating of all themes $\bar{r}$. The SVM RBF kernel width was $\gamma = 4$, and the margin $C = 0.5$. To reduce overfitting for SVM and KNN, we apply them only to the top N features selected by the LASSO regressor. All parameters were set with cross-validation, which selected $N = 40$.

We find that the LASSO algorithm generally gave best results, and also performs automatic feature selection. Sample results are shown in Figure 10, and Table 1 shows the regression results using a $0.6/0.4$ train/test split. The LASSO regressor is a linear function of the features: $r(\mathbf{t}) = \mathbf{w}^T\mathbf{y}(\mathbf{t}) + b$, learned with L_1 regularization:

$$\min_{\mathbf{w},b} \sum_i (\mathbf{w}^T\mathbf{y}_i + b - r_i)^2 + \lambda||\mathbf{w}||_1 \quad (3)$$

The optimal parameters $\mathbf{w}$ and b are computed via a convex optimization [Friedman et al. 2010] (glmnet package), with $\lambda = 0.00016$ selected by cross-validation. We use the LASSO model for all further regression tests. On the MTurk dataset, LASSO gives a 32% decrease in MAE over the baseline, and a 53% decrease in MSE. Table 1 shows several examples and we plot the full test set in Figure 11. In Supplemental Material §6, we show the effect of increasing the minimum number of ratings for the Kuler dataset.

Classification. We also learn classifiers to distinguish "good" themes from "bad". Classification provides another way to test the

[1] A space where hue θ and saturation s are remapped to Cartesian coordinates: $d_1 = s\cos(\theta)$ and $d_2 = s\sin(\theta)$.

[2] Given an input feature vector $\mathbf{y}$, the K-NN regressor computes the k training feature vectors $\mathbf{y}_j$ most similar to $\mathbf{y}$, and returns a weighted linear combination of their mean ratings r_j, taking into account the distances between feature vectors, the number of ratings N_j, and the rating variance v_j:

$$r_{knn}(\mathbf{y}) = \sum_j w_j(\mathbf{t}) r_j / \sum_j w_j(\mathbf{t}) \quad (1)$$

$$w_j(\mathbf{y}) = \frac{\exp(-||\mathbf{y}-\mathbf{y}_j j||\sigma_{rbf})}{1+\exp(-N_j\sigma_{cnt})} \frac{\sigma_{var}}{1+v_j} \quad (2)$$

where K=50, σ_{rbf}=1, σ_{cnt}=0.05, σ_{var}=0.5, set by cross-validation.

Set	Fixed	LSO	IRLS	SVM	KNN	Class.
KL MAE	0.572	0.521	0.523	0.531	0.533	64.2%
CL MAE	0.703	0.664	0.654	0.650	0.674	60.1%
MT MAE	0.267	0.179	0.179	0.182	0.205	77.4%
KL MSE	0.525	0.448	0.449	0.466	0.470	64.2 %
CL MSE	0.763	0.688	0.695	0.725	0.708	60.1 %
MT MSE	0.115	0.052	0.053	0.053	0.068	77.4%

Table 1: ***Regression and classification results****, including mean average and squared error (MAE/MSE). For classification, the LASSO regression output is compared to the mean theme rating to distinguish "good" themes from "bad."*

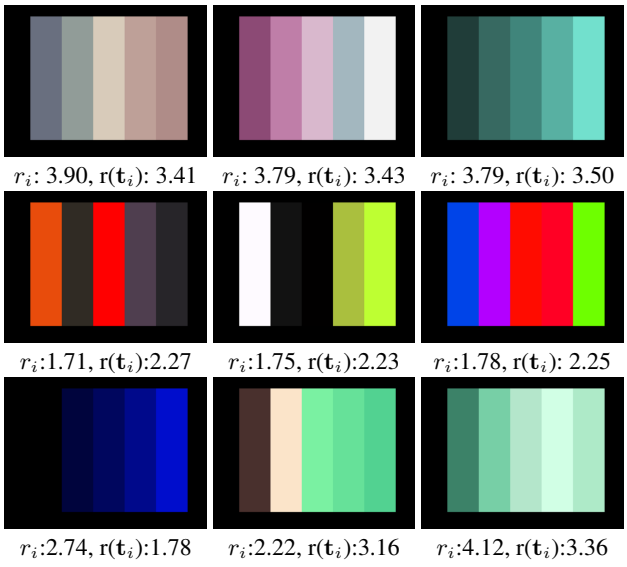

Figure 10: ***Some example themes****, including well-rated (top row), poorly-rated (middle row), and themes with high regression error (bottom row). User rating r_i and LASSO regression r($\mathbf{t}_i$) results shown for each theme, with LASSO from MTurk training set. All themes and scores are from the MTurk testing set.*

predictability of the data; we do not use classification in any of our applications. A training point is marked as "good" if its mean rating is above the mean rating of all themes, and "bad" otherwise. Table 1 shows classification results using the LASSO regressor. We also experimented with other classifiers (Logistic Regression, Gentle AdaBoost, SVM) but found results only improved slightly ($< 1\%$).

6 Model-Based Data Analysis

We now perform further analysis and experiments on the datasets using regression. Specifically, we investigate which features are most predictive of the rating, and test the value of learning different models for different subsets of users.

6.1 Important Features

Because LASSO uses a linear predictor, inspecting the feature weights $\mathbf{w}$ of a learned model gives a sense of which features are most predictive of rating for that dataset. All features are normalized to the range 0...1, so weights are directly comparable. However, it is important to remember that the LASSO ratings depend nonlinearly on the input theme since the features are non-linear. Therefore, examining individual weights gives only a partial picture

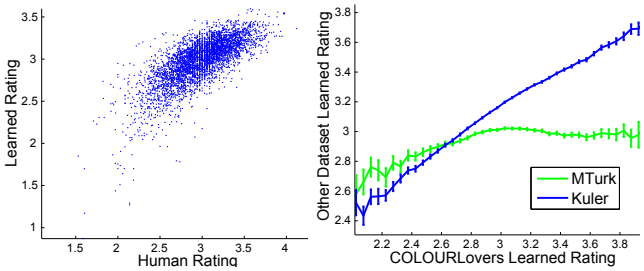

Figure 11: ***Left: regression results for MTurk testing set.*** *Predicted ratings compared to human ratings for all themes.* ***Right: evaluating COLOURLovers estimated ratings.*** *Predicted ratings from each model are compared with a COLOURLovers test set. The estimated rating model matches the Kuler model closely.*

of the predictor's behavior. The weights are listed in Supplemental Material §5.

Lightness features are among the most predictive of good themes. MTurk's most important feature is Mean Lightness, indicating a preference for overall bright themes. However, in Kuler, the Max Lightness feature is important, indicating that users like darker colors, as long as there is at least one bright color. The feature measuring difference between min and max lightness is highly weighted, further indicating that a spread of lightness is important. These features support previous research showing that lightness contrast is important for harmony [Szabó et al. 2010; Schloss and Palmer 2010]. However, all models heavily penalize a high standard deviation in lightness, suggesting that pairing several bright and dark colors is a poor choice (e.g., Figure 10 (center)). Combined, these two features promote high contrast with low standard deviation, that is, a gradient; indeed, many highly-rated themes in MTurk are simple gradients from light to dark, e.g., Figure 10 (upper-right).

All three datasets have a positive weight on the mean of hue probability p^j_{bc}, indicating that pairwise relationships between colors are important, as are choosing warm hues and cyans (which have higher probability). A significant negative feature for all datasets is the min of the pairwise hue probabilities, p^j_{bc}; a large min value indicates that all the colors in a theme are the same or similar, since, as can be seen in Fig 6, p_{bc} is largest when b and c are the same. Hence, a set of good colors should be reasonably popular, but not too similar.

Some important differences arise between models. The most positive feature in the Kuler model is the standard deviation in CIELab's b dimension (roughly, blue to yellow). This feature favors a bimodal set of similar colors instead of a transition of different colors. A high weight on the max minus min in this dimension keeps the colors from becoming too saturated. A similar preference for blues and yellows can be found in the COLOURLovers dataset in a CHSV standard deviation. The lack of this feature in MTurk may reflect some of the differences in color preference (Figure 5).

Hue entropy is an important feature for all models. For MTurk and COLOURLovers, it is negative, indicating a penalty for having too many distinct colors. In Kuler, entropy has a positive weight. Since the relationship between entropy and rating is roughly parabolic for Kuler, this suggests that the penalty for having too few colors (e.g., monochromatic), outweighs the penalty for having too many colors.

6.2 Ratings Across Datasets

In order to validate the COLOURLovers' estimated ratings $r(v, h)$, in Fig. 11 (right) we plot the predicted ratings using all three

Set	KL Select	CL Select	MT Select	Fixed
KL MAE	0.552	0.546	0.545	0.572
CL MAE	0.692	0.686	0.686	0.703
MT MAE	0.207	0.217	0.201	0.268

Table 2: ***Regression MAE using only top-20 LASSO features selected from another dataset.*** *The similar performance suggests significant overlap in the features relevant for each dataset.*

LASSO models on a single dataset (50,000 COLOURLovers themes not used for training). The strong positive trend between COLOURLovers and Kuler validates the estimated ratings. Between the COLOURLovers and MTurk ratings, a positive trend appears at lower scores, flattening out for higher scores. This indicates a higher discrepancy between COLOURLovers and MTurk users than between COLOURLovers and Kuler users. A similar comparison showed an intermediate trend between MTurk and Kuler.

To further explore the overlap between dataset features, we choose the top 20 features (according to absolute value) from each dataset and trained the other datasets on only those features. Table 2 shows that performance is similar for most models regardless of which dataset was used to select the features, indicating a significant overlap in features relevant for each dataset.

6.3 Collaborative Filtering

Color preferences often vary across individuals [Palmer and Schloss 2010], suggesting that collaborative filtering might be used to improve individualized predictions. We test the presence of variations in style as follows. We first selected the 597 Kuler users that had rated at least 50 themes. We say that two users *agree* on a theme if both users rate that theme higher than their own respective mean ratings, or if both users gave ratings less than their means. The average agreement across all these users and themes is only 52%, only a slight improvement over chance. However, it is possible to cluster users into groups and only compare their intra-cluster agreement. A simple clustering algorithm was used where each user was repeatedly compared against the agreement with all users in each cluster and re-assigned to the cluster with most agreement. For 10 clusters, within-cluster agreement jumps to 81%.

We learned a Mixture-of-Experts (MoE) [Jacobs et al. 1991] model to predict color ratings by collaborative filtering. For each cluster, we first trained a LASSO regressor. Given a new user with some current ratings, the MoE first identifies which cluster's regressor best fits the new user's ratings. That regressor then predicts all missing scores for that one user. A set of test users with 25 to 50 ratings were chosen. Given 1 rating we choose the best cluster based on a score from a single sample theme. The resulting mean average error on the test set is 1.30. For 5 ratings, MAE=1.07, for 10 ratings MAE=1.03, for 20 ratings MAE=1.00. By contrast, if a single cluster is used for all expert users, the MAE is 1.30, indicating a 30% improvement over using the same model for all users.

6.4 Clustering By Demographics

Another question is how different demographic groups compare. The demographics of the MTurk workers were 504 Indians, 351 Americans, and 146 from other countries. There were 593 female and 410 male participants. 548 participants were age 15-30, 300 were age 31-45, and 152 were age 46+. A LASSO model was first trained on each group. To account for improved scores for groups with more participants, each theme was given the same number of ratings from each group. Women's scores were generally more varied than men, though prediction was similar (MAE: 0.26 vs 0.25). Indians rated more consistently than Americans and had lower prediction error (MAE: 0.54 vs 0.58). Prediction error also tended to decrease with age (MAE: 0.50/0.49/0.48). These results suggest that demographics groups do rate differently. However, a further test comparing models trained on demographic groups to a model trained on the full dataset revealed little improvement in prediction. This suggests that personal preferences outweigh demographic preferences, so pooling between groups helps performance overall.

7 Applications

Choosing colors is a challenging problem in many design scenarios. We now demonstrate several simple prototype applications that illustrate how our model may be useful in design and computer graphics. In each case, we make use of a LASSO regressor that, given a color theme $\mathbf{t}$, outputs a predicted rating $r(\mathbf{t})$. Except where noted, we use a regressor trained on the MTurk data, as we find this gives the best results when testing applications on MTurk users.

7.1 Theme Optimization

First, we use our learned model of color compatibility to improve existing themes. Given an input theme $\mathbf{t}_{\text{in}}$, we seek a similar theme $\mathbf{t}$ with the highest rating. The ratings given by our model depend on the order of the colors. Hence, the simplest way to improve the rating is to search for the best permutation of the colors: $\mathbf{t} = \arg\max_{r \in \mathcal{P}(\mathbf{t})} r(\mathbf{t})$, where $\mathcal{P}(\mathbf{t})$ denotes the 120 permutations of a theme $\mathbf{t}$.

More generally, we can search for a theme that maximizes the score while staying within a given distance $d_{\min}$ of the input theme:

$$\begin{aligned} &\max_{\mathbf{t}} \max_{\mathbf{t}' \in \mathcal{P}(\mathbf{t})} r(\mathbf{t}') \\ &\text{subject to } \min_{\mathbf{t}'' \in \mathcal{P}(\mathbf{t})} ||\mathbf{t}'' - \mathbf{t}_{\text{in}}||_2 \leq d_{\min} \end{aligned} \tag{4}$$

with the L_2 distance computed in CIELab space. After running this optimization, the optimally-ordered $\mathbf{t}$ is returned. Though not technically necessary, permutations are included in both the objective and the constraints in order to reduce local minima. We optimize this function with Covariance Matrix Adaptation (CMA) [Hansen 2006], with $d_{min} = 35$ and constraints enforced by assigning very large penalties to themes that violate the constraints. CMA is run for 50 iterations with a sample size $N = 30$, taking approximately 5 minutes per theme. Optimal permutations are found by brute force enumeration. d_{min} can also be varied for each color to enforce that certain colors be fixed (e.g., corporate colors) or allow more significant changes. Sample optimized themes are shown in Fig. 12.

Evaluation. MTurk users were shown the original theme and an optimized version and asked to select their preferred theme in an A-B test. Users could also select "neither" if they had no preference. Three sets of themes were used: the 50 worst-rated themes from the MTurk dataset, the 50 best-rated, and 100 random. We used 40 comparisons per task, and duplicates were added to identify and remove inconsistent users. After removal, the median number of participants per theme was 46. Two optimization tests were performed. In the first, themes were optimized only by color re-ordering (only themes where the order changed were evaluated), and a second test where colors were also optimized. Fig. 13 shows results indicating that only the worst-rated themes could be improved through color re-ordering. By allowing colors to be optimized as well, the new themes were preferred on average for all groups. An additional outcome of this experiment is to show that ordering affects user ratings.

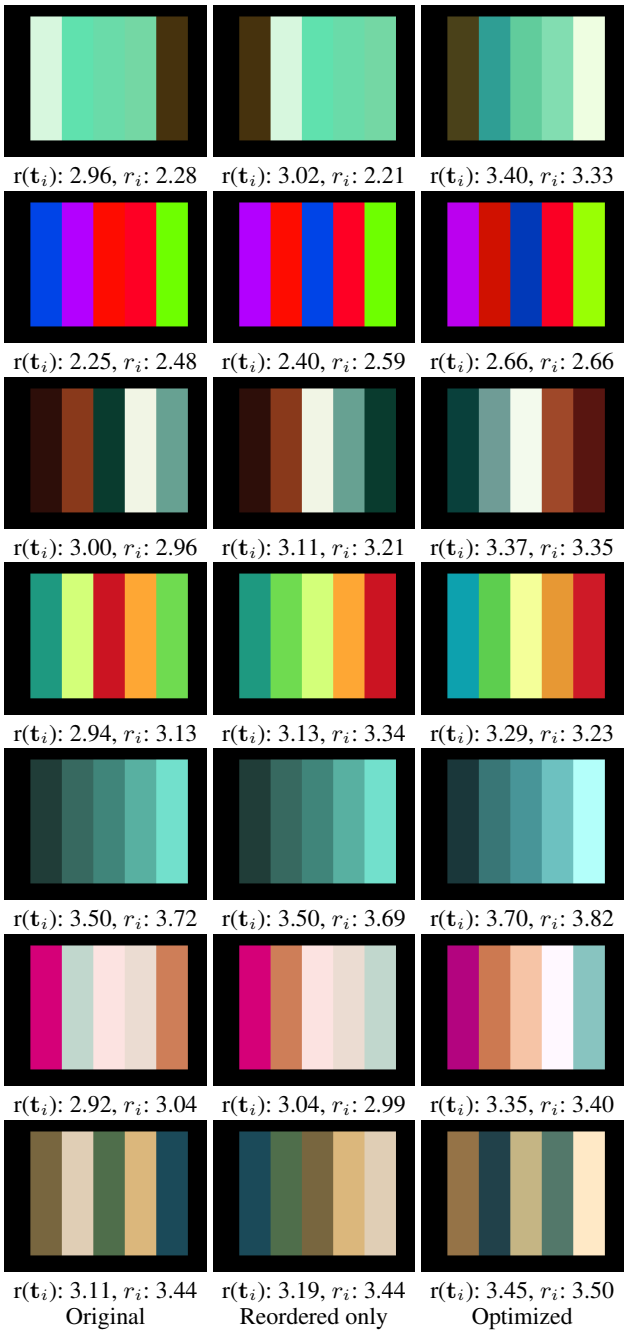

Figure 12: ***Theme Optimization.*** *Left column: original themes. Middle column: optimized order. Right column: optimized color and order. Regression rating $r(\mathbf{t}_i)$ and mean ratings r_i from a follow-up MTurk test listed below. Note in the fourth row a case where optimizing colors gives a lower rating than just re-ordering.*

7.2 Theme Extraction

Designers often look to photographs for inspiration for selecting colors, and both Kuler and COLOURLovers provide tools for creating themes from images. Here we extract a color theme from an image I. We extract themes with an objective function that attempts to represent or suggest an image while also being highly rated:

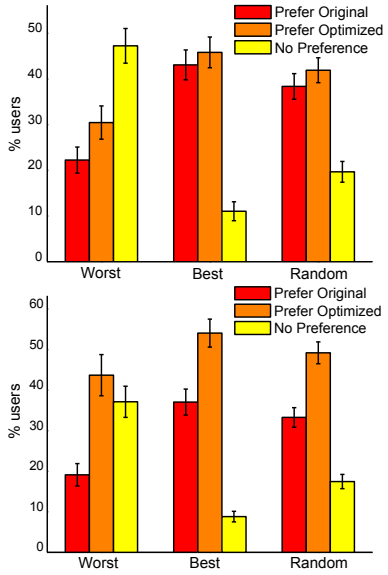

Figure 13: ***Top: theme optimization using only color re-ordering.*** *Preferences for worst rated, best rated, and random themes in MTurk dataset (2 std error).* ***Bottom: optimizing both order and color.*** *Re-ordering only improves the worst-rated themes, whereas optimizing colors improves all themes.*

$$\max_{\mathbf{t}} \quad \alpha r(\mathbf{t}) - \frac{1}{N}\sum_{i} \min_{1\le k\le 5} \left(\max ||\mathbf{c}_i - \mathbf{t}_k||_2, \sigma\right) - \frac{\beta}{M}\max_{k} \sum_{j\in\mathcal{N}(\mathbf{t}_k)} \max\left(||\mathbf{c}_j - \mathbf{t}_k||_2, \sigma\right) \tag{5}$$

where $\mathbf{c}_i$ is a pixel color, $\mathbf{t}_k$ a theme color, and N is the number of pixels. The first term measures the quality of the extracted theme. The second term penalizes dissimilarity between each image pixel $\mathbf{c}_i$ and the most similar color $\mathbf{t}_k$ in the theme. Optimizing this term alone would be equivalent to K-means clustering with a modified distance function. The third term penalizes dissimilarity between theme colors $\mathbf{t}_k$ and the M most similar image pixels $\mathcal{N}(\mathbf{t})$, to prevent theme colors from drifting from the image. We use $M = N/20$, $\beta = 0.025$, and $\sigma = 5$. We use the DIRECT algorithm for optimization [Jones et al. 1993], since it performs a deterministic global search without requiring a good initialization. Fig. 14 shows several examples including the original image, the extracted theme without the rating term ($\alpha = 0$), and with $\alpha = 3$. For $\alpha = 0$, colors were sorted by value.

Evaluation. MTurk users were shown the original image and the extracted themes with and without the compatibility model, and asked to select their preferred theme (or neither if there was no preference). Tasks were structured as in previous tests, with a median of 35 participants comparing each image. Figure 17 (left) shows that themes with the compatibility model are preferred for theme extraction than themes without the model ($p < 0.01$).

7.3 Color Suggestion

A common problem is to choose additional colors for a design, given some that have already been selected. Here, we consider a version of this problem in which we begin with a specific design taken from COLOURLovers (Figure 15), and a specific theme created by COLOURLovers users for this design. Four colors

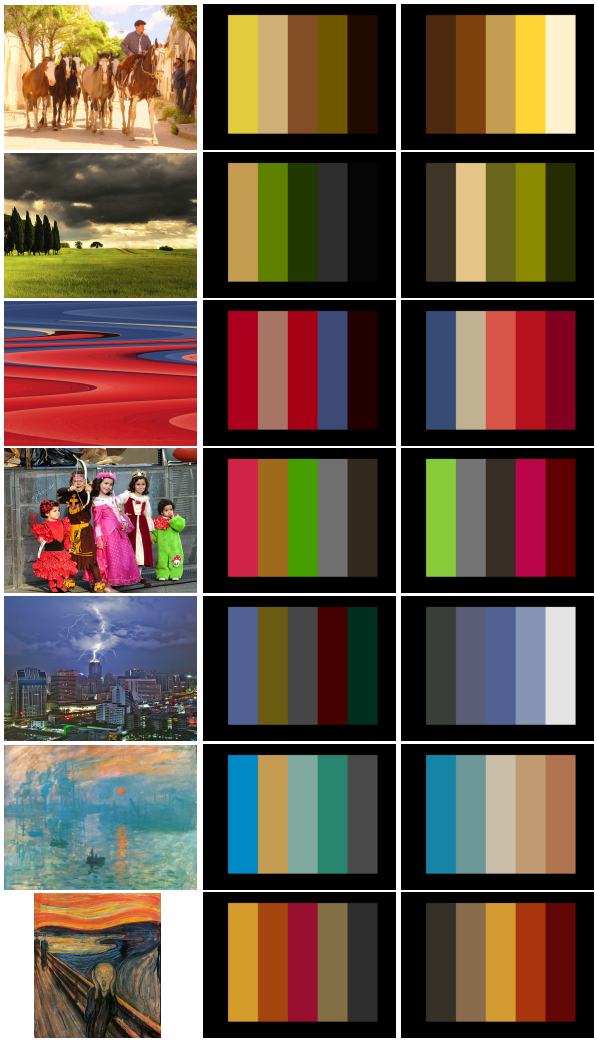

Figure 14: ***Theme Extraction. Left:*** *the original images.* ***Middle:*** *the extracted theme without compatibility model.* ***Right:*** *extracted theme with compatibility model. Creative-Commons photographs courtesy of Flickr users bombeador (Eduardo Amorim), marzinians (Dimitri Boisdet), szacharias (Stephen Zacharias), epsos, mikebehnken (Mike Behnken) respectively.*

$\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4$ from this theme are fixed. We suggest candidates for the fifth color $\mathbf{c}$ so that the suggestions are compatible with the inputs colors. However, because the new color is assigned to a region, we also want the suggestions to contrast with neighboring regions.

To pick the first suggestion $\mathbf{c}^{(1)}$, we optimize:

$$\begin{aligned} &\max_{\mathbf{c}^{(1)}} \quad \max_{\mathbf{t} \in P([\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4, \mathbf{c}^{(1)}])} r(\mathbf{t}) \\ &\text{subject to} \quad ||\mathbf{c}^{(1)} - \mathbf{c}_i||_2 \geq d_i \qquad i \in \{1, 2, 3, 4\} \end{aligned} \tag{6}$$

The constraint enforces that the new color contrasts with each previous color $\mathbf{c}_i$ by d_i, which is computed as CIELab distance. d_i is set by the user to enforce scene-dependent constraints (e.g., a flower petal should contrast more with the background then a flower stem). Optimization is performed by brute-force search in color space.

We produce a sequence of suggestions $\mathbf{c}^{(j)}$ recursively. Specifically,

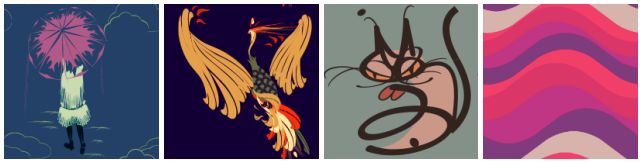

Figure 15: ***Examples of graphic designs used to evaluate our model.*** *Graphic designs include color themes with an alpha matte for each color. First 3 designs courtesy of COLOURLovers users mcmp, Yv-chan, and ycc2106.*

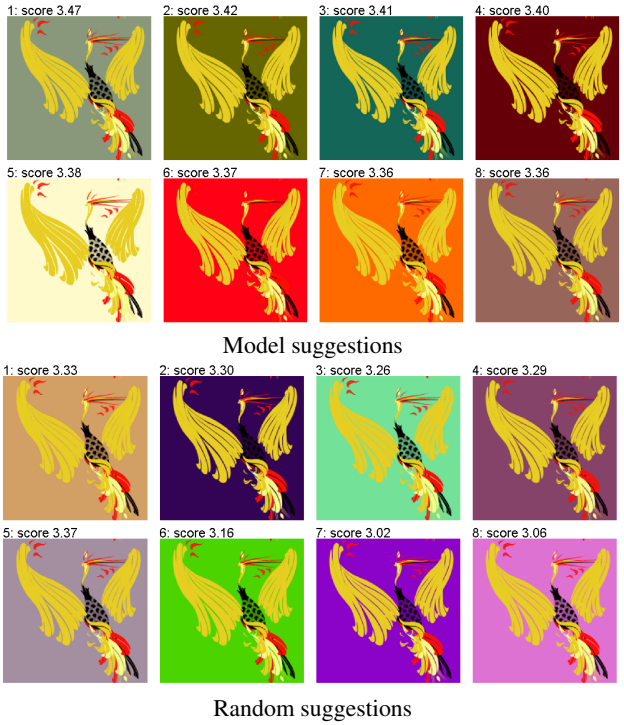

Figure 16: ***Color suggestions generated from an input design****. The same set of constraints was used (see text) for both sampling methods. Color compatibility scores are listed for both sets, though not used for random sampling.*

when optimizing the j-th suggestion, we perform the same optimization as above, but add a constraint that the next suggestion be dissimilar from all previous:

$$||\mathbf{c}^{(j)} - \mathbf{c}^{(k)}||_2 \geq d \qquad \text{for all } k < j \tag{7}$$

Fig. 16 shows an example of our model's suggestions, as well as random sampling of colors satisfying the same constraints. The Kuler model was used as it lacks MTurk's bias for brighter colors.

Evaluation. MTurk users were shown sets of color suggestions and asked to select the best and worst image from each set. Users were shown 24 sets, with each set consisting of 4 suggestions from our model, and 4 from random sampling, with all 8 randomly shuffled. 6 duplicate sets were included, in order to detect inconsistent users. A median of 64 people were used for each comparison. The initial 10 sets were ignored to allow user training. Figure 17 (right) shows that our model is both better at choosing the "best" color as well as avoiding the "worst" colors. For the "best" image, our model suggestions were chosen 56% of the time (Student t-test $p < 0.01$). For the worst, our model was chosen 40% of the time ($p < 0.01$). While both results are statistically significant, the standard deviation of the

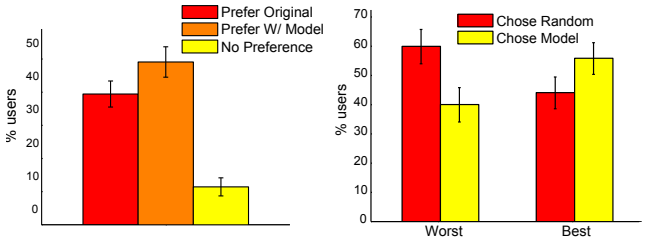

Figure 17: ***Left: theme extraction results.*** *User preferences with and without compatibility model (2 std err).* ***Right: color suggestion results.*** *Mean user preferences for worst and best background suggested randomly or using our model.*

tests is quite high for both (0.128 and 0.138 respectively). This may be attributed to the importance of other factors in each design context, which could be considered more carefully in future work.

To further evaluate the color suggestion model, we performed an experiment optimizing single colors from a theme. We measured the distance of the model's suggested color to the original color and compared to the theme's human rating. As expected, the model's suggestions are closer to highly-rated original colors than to poorly-rated original colors. See Supplemental Material §7 for details.

8 Discussion

In this paper we describe the first large-scale on-line studies of color preference, the first studies based on five-color themes, and the first studies of user-generated color combinations. A number of observations emerge from this work. Designers creating themes do not uniformly sample the space of all possible themes. We find no support for the notion of hue templates as guides to aesthetics, except for the simplest, most basic themes. Instead, a number of simpler rules emerge, including a preference for a small range of colors, typically 2-3. The simplicity of a theme can be quantified by the hue entropy. The ordering of colors affects ratings. Certain color pairings are preferred, as is significant lightness variation; preferably a gradient. Segmentation of users by preference can lead to more accurate predictions. We also confirm previous findings that certain colors are preferred over others, but that preferences change in combination, which helps validate MTurk for studies of aesthetics.

Along with our observations on color preferences, we describe a learned model that can predict a rating for new color themes, and several demonstrations of tools that use this predictive model. Though our prototypes are only initial explorations, we hope that further refinement will yield tools that can help non-color experts navigate the sometimes daunting task of choosing colors. We also hope that our findings will inform future studies into color, and provide clues as to which aspects of color are or are not important.

Finally, while color plays a significant role in most forms of visual design, it is not the only important factor. Our work provides an example of how an evidence-based approach to aesthetics can aid understanding and development of new algorithms. A deeper understanding of the factors that influence human aesthetic preferences can have substantial impact on how we design tools in computer graphics, and we believe that large-scale, on-line studies are a promising approach to increasing our knowledge of aesthetics.

Acknowledgements

Thanks to Koji Yatani for translating Matsuda [1995]. We thank Maneesh Agrawala, Dan Goldman, and Sylvain Paris for helpful discussions. This work was supported in part by Adobe, NSERC, and CIFAR.

References

CHEVREUL, M. 1839. *The Principles of Harmony and Contrast of Colors and Their Application to the Arts*.

COHEN-OR, D., SORKINE, O., GAL, R., LEYVAND, T., AND XU, Y.-Q. 2006. Color Harmonization. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH) 25*, 3, 624–630.

CSURKA, G., SKAFF, S., MARCHESOTTI, L., AND SAUNDERS, C. 2010. Learning Moods and Emotions from Color Combinations. In *Proc. ICVGIP*.

DATTA, R., JOSHI, D., LI, J., AND WANG, J. Z. 2006. Studying Aaesthetics in Photographic Images Using a Computational Approach. In *Proc. ECCV*, 7–13.

EASLEY, D., AND KLEINBERG, J. 2010. *Networks, Crowds, and Markets*. Cambridge University Press.

FRIEDMAN, J. H., HASTIE, T., AND TIBSHIRANI, R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software 33*, 1 (2), 1–22.

GAGE, J. 1999. *Color and Culture: Practice and Meaning from Antiquity to Abstraction*. University of California Press.

GOETHE, J. 1810. *Theory of Colors*.

GRANGER, G. W. 1952. Objectivity of Color Preferences. *Nature 170*, 4332, 778–780.

HANSEN, N. 2006. The CMA evolution strategy: a comparing review. In *Towards a new evolutionary computation*. 75–102.

HEER, J., AND BOSTOCK, M. 2010. Crowdsourcing Graphical Perception. In *Proc. CHI*, 203–212.

ITTEN, J. 1973. *The Art of Color*.

JACOBS, R. A., JORDAN, M. I., NOWLAN, S. J., AND HINTON, G. E. 1991. Adaptive Mixtures of Local Experts. *Neural Computation 3*, 79–87.

JONES, D. R., PERTTUNEN, C. D., AND STUCKMAN, B. E. 1993. Lipschitzian Optimization Without the Lipschitz Constant. *J. Opt. Theory. Appl. 79*, 3 (Oct.).

KRAUSE, J. 2007. *Color Index 2*.

LALONDE, J.-F., AND EFROS, A. A. 2007. Using Color Compatibility for Assessing Image Realism. In *Proc. ICCV*.

LI, C., AND CHEN, T. 2009. Aesthetic Visual Quality Assessment of Paintings. *J. Sel. Topics in Signal Processing 3*, 2, 236–252.

MARLIN, B. M., AND ZEMEL, R. S. 2007. Collaborative Filtering and the Missing at Random Assumption. In *Proc. UAI*.

MATSUDA, Y. 1995. *Color Design*. Asakura Shoten.

MEIER, B. J., SPALTER, A. M., AND KARELITZ, D. B. 2004. Interactive Color Palette Tools. *IEEE Computer Graphics and Applications 24*, 3, 64–72.

MOORTHY, A. K., OBRADOR, P., AND OLIVER, N. 2010. Towards Computational Models of Visual Aesthetic Appeal of Consumer Videos. In *Proc. ECCV*.

MUNSEL, A. 1921. *A Grammar of Color*.

NEMCSICS, A. 2003. Coloroid Colour System. In *Hungarian Electronic Journal of Sciences*.

NEUMANN, L., NEMCSICS, A., AND NEUMANN, A. 2005. Computational Color Harmony Based on Coloroid System. In *CAGVI*.

OSTWALD, W. 1932. *Color Science, Vol 1*.

OU, L.-C., AND LUO, M. R. 2006. A Color Harmony Model for Two-Color Combinations. *Col. Res. Appl 31*, 5, 191–204.

OU, L.-C., LUO, M. R., WOODCOCK, A., AND WRIGHT, A. 2004. A Study of Colour Emotion and Colour Preference. Part III: Colour Preference Modeling. *Col. Res. Appl 29*, 5, 381–389.

PALMER, S. E., AND SCHLOSS, K. B. 2010. An ecological valence theory of human color preference. *Proc. Nat. Acad. Science 107*, 19, 8877–82.

SCHLOSS, K. B., AND PALMER, S. E. 2010. Aesthetics of color combinations. In *Human Vision and Elec. Imaging XV*, vol. 7527.

SZABÓ, F., BODROGI, P., AND SCHANDA, J. 2010. Experimental Modeling of Colour Harmony. *Col. Res. Appl 35*, 1, 34–39.

TIBSHIRANI, R. 1996. Regression Shrinkage and Selection Via the Lasso. *Royal. Statist. Soc B 58*, 1, 267–288.

TOKUMARU, M., MURANAKA, N., AND IMANISHI, S. 2002. Color Design Support System Considering Color Harmony. In *Proc. FUZZ-IEEE*, 378–383.

WANG, B., YU, Y., WONG, T.-T., CHEN, C., AND XU, Y.-Q. 2010. Data-Driven Image Color Theme Enhancement. *ACM Transactions on Graphics (Proc. SIG. Asia) 29*, 6.

Appendix A: Dataset Details

For Kuler themes, the mean rating is $\bar{r} = 3.14$ with a variance $\sigma^2 = 0.52$. The mean variance $\overline{\sigma^2}$, indicating the typical disagreement of users on themes, is 1.31. These statistics are defined as follows. For theme $\mathbf{t}_i$ with N user ratings $r_{i,j}$, the theme mean rating is $r_i = \frac{1}{N}\sum_j r_{i,j}$. The mean rating over M themes is $\bar{r} = \frac{1}{M}\sum_i r_i$ with variance $\sigma^2 = \frac{1}{M}\sum_i (r_i - \bar{r})^2$. The mean variance $\overline{\sigma^2} = \frac{1}{M}\sum_i \frac{1}{N}\sum_j (r_{i,j} - r_i)^2$. The mean number of ratings per theme is 5.77, median 3 and max 1440. The distribution of ratings is (17%,11%,25%,24%,24%) for 1-5 stars.

For COLOURLovers themes, we define a numerical rating for each theme as: $r(h, v) = (h - \bar{h}(v))/\sigma(v) + 3$, where $\bar{h}(v)$ and $\sigma(v)$ are linear functions corresponding to the mean and standard deviation of hearts for a given number of views. Specifically, $\bar{h}(v) = av + b$. The parameters a and b are fit as follows. We discretize the number of views v into bin i, and define m_i be the mean hearts per theme with v_i views. The parameters are set by minimizing L_1 error: $\sum_i |(av_i + b) - m_i|$, yielding values of $a = 0.0152$ and $b = -0.263$. Similarly, $\sigma(v)$ is fit to the variances, giving $a = 0.0128$ and $b = 0.218$. See Fig. 3 for a plot. The mean of the estimated ratings $r(h, v)$ over all themes is $\bar{r} = 2.98$ and variance $\sigma^2 = 0.75$.

MTurk tasks required each participant to rate 30 themes on a discrete scale of 1 to 5. Themes were shown on black to match the Kuler interface. Theme names were not shown. Each task was worth \$0.02. See the Supplementary Material for example tasks. Each theme was rated by 40 participants, with 1,301 participants total. Workers were required to indicate their gender, age, and the country they have lived longest in. Participants were evaluated for consistency by including duplicate themes, with 263 workers removed. Inconsistency was measured with the standard deviation for each pair of duplicate ratings (σ_d). The average of all duplicates ($\bar{\sigma_d}$) was found, and if $\bar{\sigma_d} > 0.7$, the user was removed. If the standard deviation of all ratings was less than 0.6, the user was removed for using too few rating values.The mean number of ratings per theme is 39.3, median 35, and max 80. The mean rating is $\bar{r} = 2.98$ with a low variance of $\sigma^2 = 0.11$ since theme ratings are aggregated over many user ratings. The mean variance $\overline{\sigma^2} = 1.14$. Since MTurk themes are sampled from Kuler, we can also directly compare the Kuler statistics for those 10,743 themes. Of that Kuler subset, the mean rating $\bar{r} = 3.19$, variance $\sigma^2 = 0.40$, and mean variance $\overline{\sigma^2} = 1.64$. This indicates that for the same themes, MTurk users generally give lower scores with less disagreement. The distribution of MTurk ratings is (10%,22%,35%,24%,9%) for 1-5 star ratings. Note in Kuler 24% of ratings were 5 stars with only 9% in MTurk.

Appendix B: Theme to template distance

We compute the distance between a theme and a template as follows. For a theme, we first sort by hue and then compute the hue differences for saturated and light colors, with unsaturated or dark colors given a hue difference of 0. We use a threshold of 15 for both saturation and lightness.

For example, a theme with only 2 exactly complementary hues has a hue difference vector of (180,180,0,0,0). For a theme with 3 triadic hues, the difference vector would be (120,120,120,0,0). These vectors are then compared to exemplars for each template. The vectors above are the exemplars for the I and R templates respectively. The distance between a theme and template is the sum of absolute differences. Note that for symmetric hue templates (all except C and L), the hue difference vectors and exemplars can be sorted. C and L require multiple exemplars.

For a concrete example of the theme to hue template distance, consider the middle-left theme of Fig. 12, defined by the 5 HSV values (8,80,18), (16,80,54), (164,85,23), (75,7,96), (164,36,63). We remove the 4th color since it is lower than the saturation threshold and sort the remaining hues: (164,164,16,8). The hue differences (with wrap-around for the final color) are then (0,148,8,204) , which are sorted to give (204,148,8,0) and padded with zeros for the removed color. Intuitively, the two large numbers indicate two widely separated hue clusters. As expected, the sum of absolute hue difference from the complementary template is 64, which is much smaller than the triad template distance of 224.

Templates in [Matsuda 1995] are defined as sectors over the hue wheel instead of fixed angular differences. For our exemplars, we use the centers of the template sectors, or equally spaced hues in the sectors, to calculate the distance. Directly using hue sectors has several problems. For example, simpler templates are contained within larger ones; a monochromatic theme would have a distance of 0 to all templates. It is also difficult with sectors to evaluate slight differences from the geometric ideal of classical templates.