A Probabilistic Model of the Categorical Association Between Colors

Jason Chuang[†], Maureen Stone[‡], Pat Hanrahan[†]; Stanford University[†] (USA) and StoneSoup Consulting[‡] (USA)

Abstract

In this paper we describe a non-parametric probabilistic model that can be used to encode relationships in color naming datasets. This model can be used with datasets with any number of color terms and expressions, as well as terms from multiple languages. Because the model is based on probability theory, we can use classic statistics to compute features of interest to color scientists. In particular, we show that the uniqueness of a color name (color saliency) can be captured using the entropy of the probability distribution. We demonstrate this approach by applying this model to two different datasets: the multi-lingual World Color Survey (WCS), and a database collected via the web by Dolores Labs. We demonstrate how saliency clusters similarly named colors for both datasets, and compare our WCS results to those of Kay and his colleagues. We compare the two datasets to each other by converting them to a common colorspace (IPT).

Introduction

There has been growing interest in how to use color naming data to improve color models. Better color name databases[7, 10, 11, 12, 14, 2] and online naming studies[18, 8] have stimulated recent work. Color naming databases and associated models have been been useful in color transfer[5], gamut mapping[19, 20], and methods for specifying or selecting colors in an image[15, 16, 17].

In this paper, we examine the issue of how to represent and quantify the association between colors induced by names. Current methods that incorporate naming data represent the category associated with a color using either a single name[5, 6], a vector[19], or by a set of fuzzy logic memberships[1, 2, 17].

We present a probabilistic framework for working with colors. We define the categorical association of a color c as a conditional probability $P(C|c)$ over colors C in the color space $\mathscr{C}$. For a color c, the probability $P(C|c)$ represents how likely other colors in the space $\mathscr{C}$ are assigned the same linguistic label as c. Our choice of using a probability over *colors* in our framework is motivated by the following criteria not met by current approaches.

Our model satisfies three design goals. (1) Our approach can incorporate categorical effects from any number of color words, expressions involving multiple words, and different languages. (2) Our framework is based on a non-parametric model which can capture the differences in color name distributions such as "yellow" having a narrow focus and "green" having a wide distribution[21]. (3) Embedding our representation in a probabilistic framework enables us to apply a wide array of statistical and probabilistic tools to further analyze and study the effect of categories on colors.

We implement our model on two datasets. We extract color naming data from six languages in the World Color Survey which contains naming information at 330 colors on the surface of the Munsell solid[7]. We also investigate online naming data collected by DoloresLabs which contains names given to 10,000 randomly sampled colors in the RGB cube[8]. Our framework can incorporate cross-linguistic data and combine contributions from color words with similar meanings. We introduce the concept of salient colors based on the statistical notion of entropy. Salient colors from our approach show good correspondence basic color terms identified by Berlin and Kay[3]. Our approach also reveals two regions that are consistently named in the sRGB cube *not* corresponding to typical basic color terms. We compare qualitatively the differences in salient name regions between the World Color Survey and the DoloresLabs datasets.

Motivations and Related Work

The goal of this paper is to present a computational framework for modeling color categories derived from experimental data. Our framework is motivated by three issues that are at best partially addressed in the current literature.

1. We would like a framework that can include all possible words for describing a color and not be limited to a predefined list of terms.
2. We would like a non-parametric model capable of capturing the details in categorical association but still be robust to noise in the naming dataset.
3. We would like a framework that can support a rich set of computational and mathematical operations, so that more in-depth studies of categorical effects can be built on the framework. In particular, our approach is grounded in probability theory.

The first issue addresses how to account for the many potential expressions for describing a color. In 1969, Berlin and Kay defined color words as *basic color terms* if their meanings cannot be derived from other words, and proposed that there are a total of eleven basic color terms. Basic color terms were shown to be universal across languages. While some languages such as English contain all eleven terms, others may have developed only a subset of the words[3]. Subsequent studies confirmed that basic color terms are words with the highest consensus between speakers[4], but found twelve basic color terms in Russian contradicting the limit on the number of terms[22]. Kay and McDaniel hypothesized that as languages evolve, some individuals may consider additional words such as aqua/turquoise (green and blue), chartreuse/lime (yellow and green), and maroon/burgundy (red and black) as basic color terms[9].

Many existing methods assume eleven or a fixed number of color categories and cannot process the full set of responses from recent surveys such as the HP Labs Multilingual Naming Experiment[18] and the DoloresLabs Naming Dataset[8], which

have hundreds of color words. Chang et al.'s category-preserving color transfer algorithm defines eleven convex regions in the color space corresponding to the basic color terms[5]. Motomura's categorical color mapping algorithm maps foci of the eight chromatic basic color terms between the source and target gamuts[19]. Moroney's system for translating colors to names operates on the n most frequently used color words. We want a framework where all words are included and contributions from words with similar cognitive concepts such as "maroon" and "burgundy" are combined based on their similarity.

Secondly, color names exhibit different naming distributions. Colors such as "red" and "yellow" are known to have a narrow and well-defined center while colors such as "green" and "blue" are known to be composed of a broad range of hue.[21]

We want our framework have the flexibility to capture the details in the distributions while being robust to noise in the data. Current approaches tend to model color categories as a volume in color space, using various parameterized models, or using non-parameterized approaches such as histograms. Partitioning the color space[12, 5] assume color names occupy discrete and non-overlapping regions in the color space. Motomura's gamut-mapping algorithm assumes that each basic term has an ellipsoid-shaped distribution and models the distributions using an 81-parameter covariance matrix[19]. Benavente models the color naming space using a set of 6-parameter Sigmoid-Gaussian distributions[1]. One advantage of parameterized models is that they are constructed from a small number of parameters which can be estimated accurately. In his adaptive lexical classification system, Moroney proposes an alternative implementation in which color names are represented as non-parametric histograms[16]. While histograms can capture any shape of distribution, Moroney reported noise in the data due to limited number of data points and suggests that smoothing operators or hedging be applied to post-process the histograms.[1]

Finally, we would like a framework capable of supporting a rich set of computational and mathematical tools. Instead of being merely a representation, the framework should allows us to perform further computation and analysis on how categories affect the way we associate colors. Treating the association between colors as a probability distribution positions our framework within the well-studied domain of probability theory.

Methodology

Colors and Color Words

A naming dataset consists of a list of responses in the form of "color"-"color word" pairs that record the words used to describe a color. A "color" refers to the stimuli shown to a respondent and varies between datasets from Munsell color chips viewed under controlled lighting to rectangles of colors displayed on uncalibrated monitors. Unconstrained surveys allow respondents to use any expression whereas constrained surveys ask respondents to choose from a predefined list of words. An unconstrained color expression could include, e.g., "granny smith apple green", "light robin's egg pastel blue", or "mix all the paint together". In practice, most expressions recorded in unconstrained surveys consistent of a single word or a simple set of words such as "blue" or "bluish green". We will use the term "color words" from this point on even though it could refer to any possible expressions for describing a color.

A naming dataset can be tabulated using a word count table where the list of all colors presented in the survey is displayed along the columns, and a list of all color words recorded is displayed along the rows. Each entry in the table indicates the number of times a corresponding color word is used to describe the corresponding color.

Depending on the nature of the naming dataset, the density of word count table may vary. The World Color Survey (WCS)[7] is cross-linguistic and unconstrained, and collects naming data on a set of 330 colors. The word count table for the WCS consisting of 2300 rows by 330 columns with 20% non-zero entries. In comparison, the DoloresLabs color name dataset[8] while also unconstrained uses 10000 randomly-sampled colors. A total of 1966 expressions were recorded creating a 1966-by-10000 table that contains non-zero values for only 0.05% of the entries.

We use $\mathscr{C}$ to denote all colors in the color space. For certain datasets such as the World Color Survey where naming data is collected on the surface of the Munsell solid, $\mathscr{C}$ is a two-dimensional surface in the color space. We use $\mathscr{W}$ to denote the set of color words in a dataset.

We define two random variables C and W in our framework. C is a random variable that takes on different colors $c \in \mathscr{C}$. The probability that C takes on the value c is $P(C = c)$. W takes on values over the set of color words $\mathscr{W}$.

We use T to refer to the word count table. $T(w,c)$ is the number of times the word w was used to describe the color c.

The first relationship of interest is the conditional probability $P(W|c)$. Given a color c, $P(W|c)$ refers to the frequency of color words W being used to describe c, and can be computed by selecting the column in the word count table corresponding to c, and normalizing the column. This produces a probability over the set of possible color words that sums up to 1 in proportion to their frequency of use.

$$P(w|c) = T(w,c) / \sum_{w} T(w,c) \tag{1}$$

The second relationship of interest in the conditional probability $P(C|W = w)$. Given a color word w, $P(C|w)$ describes the likelihood of colors C being referred to by the word w. This probability can be computed by selecting the row in the word count table corresponding to w, and normalizing the row.

$$P(c|w) = T(w,c) / \sum_{c} T(w,c) \tag{2}$$

Categorical Association of a Color

We now compute a conditional probability $P(C|c)$ for each color c that summarizes how color c is associated with all other colors C due to categorical effects. This distribution describes, given a color c, likelihood of colors C in the space $\mathscr{C}$ being given the same name as c. We compute this distribution as follows:

1. For each color word w, we compute the conditional probability $P(C|w)$. This distribution describes, *if* a color word w is used, the likely colors that w is referring to.
2. We compute the conditional probability $P(W|c) = (P(w_1|c), P(w_2|c), \cdots)$. The color c may be associated

[1] We should emphasize our application differs from Moroney's in that his work is on modeling the distribution of color names while our work is on modeling the association between colors due to naming effects.

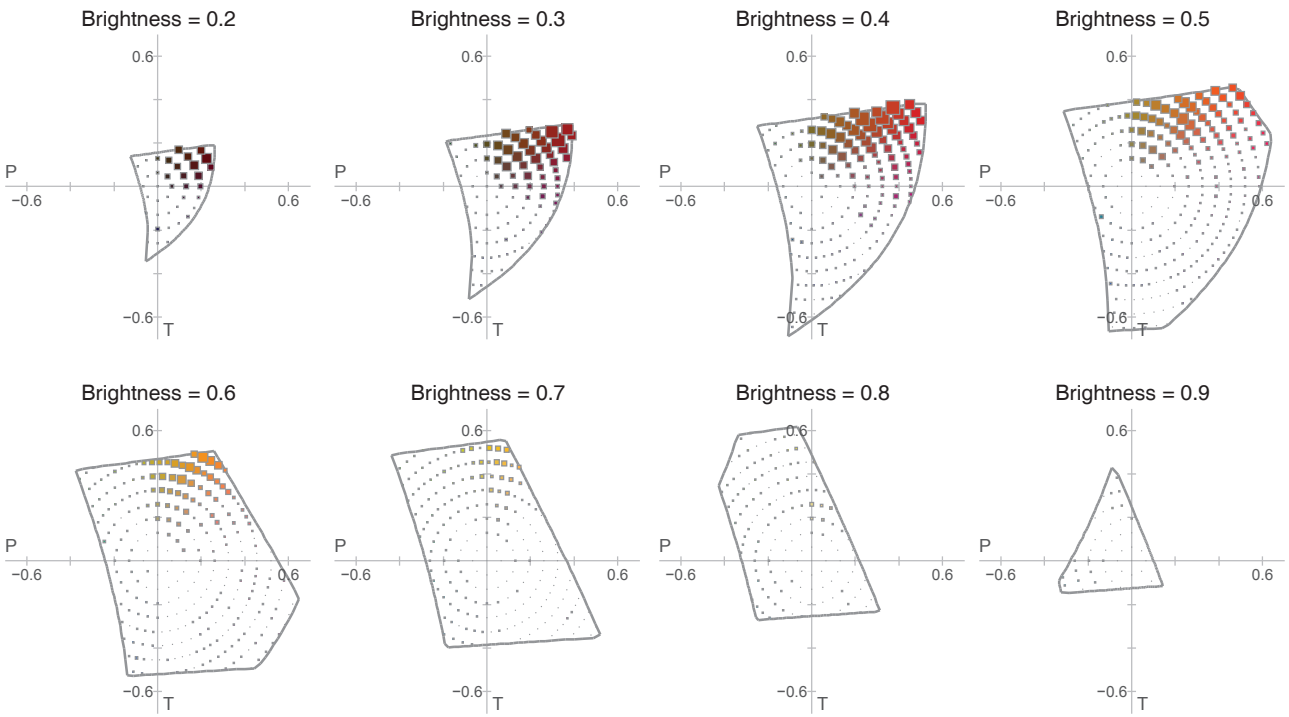

Figure 1. *Categorical association $P(C|c)$ for color ■ $c = sRGB(178,75,32)$. The graphs show planes through eight levels of brightness in IPT space. The size of the squares is proportional to the the likelihood of other colors given the same name as c. The solid line indicates the boundary of the sRGB gamut. This distribution over colors is used to computationally represent the category of color c.*

with multiple color words. This distribution describes the frequency that color words $w_1, w_2, \cdots$ are used on color c.

3. We then iterate over all color words. Suppose color c is referred to by a word w, we tally up how likely other colors are named w. We sum over the contribution in proportion to how frequently w is applied to c.

$$P(C|c) = P(C|w_1)P(w_1|c) + P(C|w_2)P(w_2|c) + \cdots \quad (3)$$

$$= \sum_w P(C|w)P(w|c) \quad (4)$$

The categorical association between colors is now completely expressed in terms of colors removing the linguistic labels from the representation. Figure 1 shows the categorical association for a color in the DoloresLabs naming dataset.

Color Saliency

We define the quantity "color saliency" as a way for determining colors that have strong name association versus colors that have ambiguous names. High saliency implies that a color c is strongly associated with a small set of colors; the category to which c belongs is well defined. Low saliency implies c is weakly associated with a large number of colors. This occurs when the names given to c is ambiguous; the respondents put color c in the same category with a wide variety of colors but not consistently.

Entropy measures the amount of randomness in a probability distribution. As the categories associated with a high-salient color is less random than low-salient color, we define color saliency for a color c as negative entropy of $P(C|c)$.

$$\text{Saliency}(c) = -H(P(C|c)) = \sum_{c' \in C} P(c'|c) \log P(c'|c) \quad (5)$$

Implementation

World Color Survey Dataset

The World Color Survey (WCS) collects naming information in 110 languages from 2616 respondents. Conducted in the respondents' native language, the survey records color names on 330 Munsell color chips (320 saturated colors on the surface of the Munsell solid plus 10 achromatic colors).[7] The WCS contains spoken languages only as the original intent of the survey is to compare color names from societies that developed independently from one another. English and other languages from industrialized societies are *not* included in the survey.

We select six languages[2] that have at least 11 basic color terms, so the results could be easily interpreted by people who speak English. The number of color words recorded in the six languages were 28, 31, 19, 25, 60, and 13 respectively. As the languages share no common vocabulary, there are a total of 176 distinct words. Each value in C corresponds to a color chip in the WCS stimulus array. W is a discrete variable of length 176. This gives us a 176-row by 330-column word count table.

DoloresLabs Dataset

The DoloresLabs naming dataset[8] is an online survey that collected English naming information from 10,000 respondents via Mechanical Turk using 10,000 randomly-sampled colors from the RGB cube. The colors are displayed as 60- by 40-pixel rectangles ten at a time against white background on each respondent's

[2]The languages selected are: Cakchiquel (from Guatemala), Camsa (from Colombia), Chavacano (from the Philippines), Kriol (from Australia), Mazahua (from Mexico), and Yakan (from the Philippines)

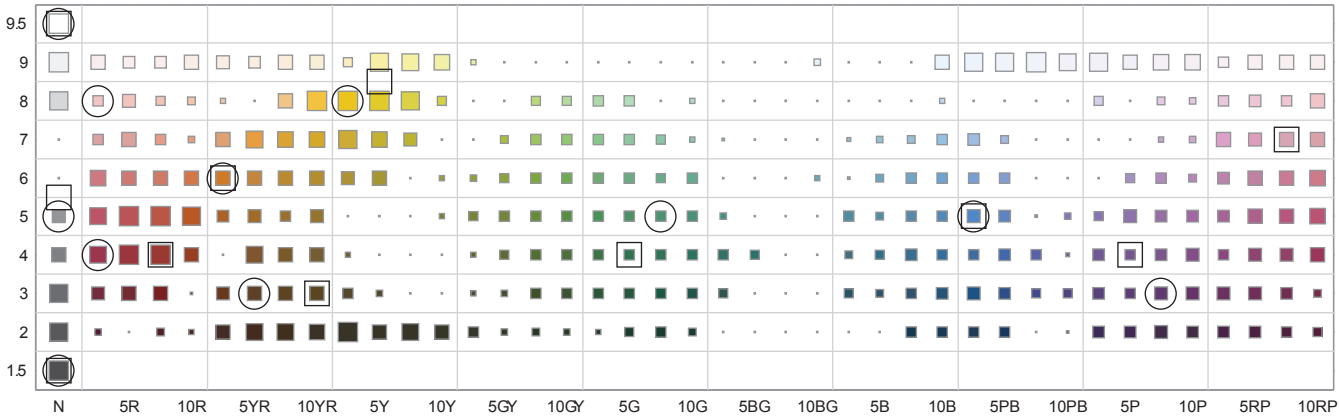

Figure 2. *Color saliency for colors on the surface of the Munsell solid. Hue is on the horizontal axis, and value is plotted on the vertical axis. The size of the colored square is proportional to color saliency. Circles indicates the foci of the basic color terms in the Munsell Space reported by Berlin and Kay[3]. Squares indicates the foci of the basic color terms reported by Sturges and Whitfield[24].*

own monitor.

The raw DoloresLabs naming dataset contains 1966 distinct ASCII strings. After correcting for spelling, leading/trailing white spaces, and hyphenation, there are 1740 distinct expressions. We break apart expressions with multiple words into single words, so that W is a variable over 302 individual words. To facilitate the comparison with the WCS dataset, we resample the colors in the DoloresLabs dataset so we have a cylindrical grid of points. We assume colors are in sRGB coordinate, and convert them to IPT colorspace. We then create a three-dimensional grid of cells. We divide brightness $[0.00, 1.00]$ into 10 equally-spaced intervals, divide chroma $[0.00, 0.78]$ into 12 equally-spaced intervals, and use a variable number of hue intervals depending on the chroma ranging from 1 (for chroma = 0), 8 (for chroma = 0.065), up to 81 (for chroma = 0.455). The grid contains 7469 vertices, 1234 of which lie inside the sRGB colorspace. We resample at these 1234 vertices of the grid. C is a variable over the 1234 resampled colors. The final word count table consists of 302 rows by 1234 columns.

Results

The goal of our saliency measure is to identify regions in the color space that are consistently named. Comparisons of naming regions across viewers, languages, media, or viewing conditions have traditionally been conducted by matching the foci (or centroids) or by comparing boundaries of specific color names. We examine the actions of computing saliency by matching colors to colors.

Figure 2 shows color saliency for the 330 colors on the World Color Survey stimulus array. Known locations of basic color foci reported by Berlin and Kay[3] and by Sturges and Whitfield[24] are marked as circles and squares respectively. We observe that regions of colors with high saliency appear to match the locations of the basic color term foci. Brown, green, and blue foci from both sources are all situated within their respective high-saliency regions. Red, orange, and yellow foci correspond to high-salient colors in their respective part of the color space. We also observe several areas with low color saliency separating green from all other colors, separating blue from all other colors, and marking the red-brown, red-black, and pink-yellow boundaries. This provides some initial evidence that saliency can be used to describe the shape of name regions and identify naming boundaries in a color space.

Figure 3 shows color saliency computed from the DoloresLabs dataset and displayed on planes corresponding to eight levels of brightness in IPT space. For illustration purposes, we create regular grid in a cylindrical IPT space with 8 brightness levels, 12 chroma levels, and up to 81 hue intervals. Naming data is linearly interpolated to the closest vertices. While we do not have experimental data on the locations of the basic color terms foci in sRGB space, salient regions appear qualitatively to correspond to most of the basic color terms. The is a lack of white salient region which we suspect is due to the fact that DoloresLab color patches are displayed against a white background. We observe that many of the salient regions are clustered in the gamut corners such as red, orange, green, blue, and pink. Interestingly, some of these regions do *not* appear to correspond to basic color terms; examples of such regions include cyan (at a brightness level of 0.9 corresponding at the bottom left corner of the gamut) and lavender (light purple at a brightness level of 0.7 situated half way between light blue and pink).

Figure 4 compares color saliency in the World Color Survey dataset with that of the DoloresLabs dataset. We convert both set of data to IPT coordinates. For the World Color Survey, we assume that the colors were viewed under D65 lighting for all speakers in all languages. For the DoloresLabs dataset, we assume that coordinates to be sRGB. We show the comparison on eight separate plots. In each plot, we display the color saliency in the WCS using a ring of Munsell colors of constant value. We also display DoloresLabs dataset using a constant-brightness plane in IPT space.

We observe a shift in hue angle for salient regions corresponding to green and blue between the WCS and DoloresLabs data set at Munsell values of 2 and 3. At Munsell value of 4, we also observe that the salient regions corresponding to red, brown, and purple are aligned between the two surveys. As mentioned earlier, we observed an additional salient region corresponding to lavender in the DoloresLabs dataset. The lack of a corresponding

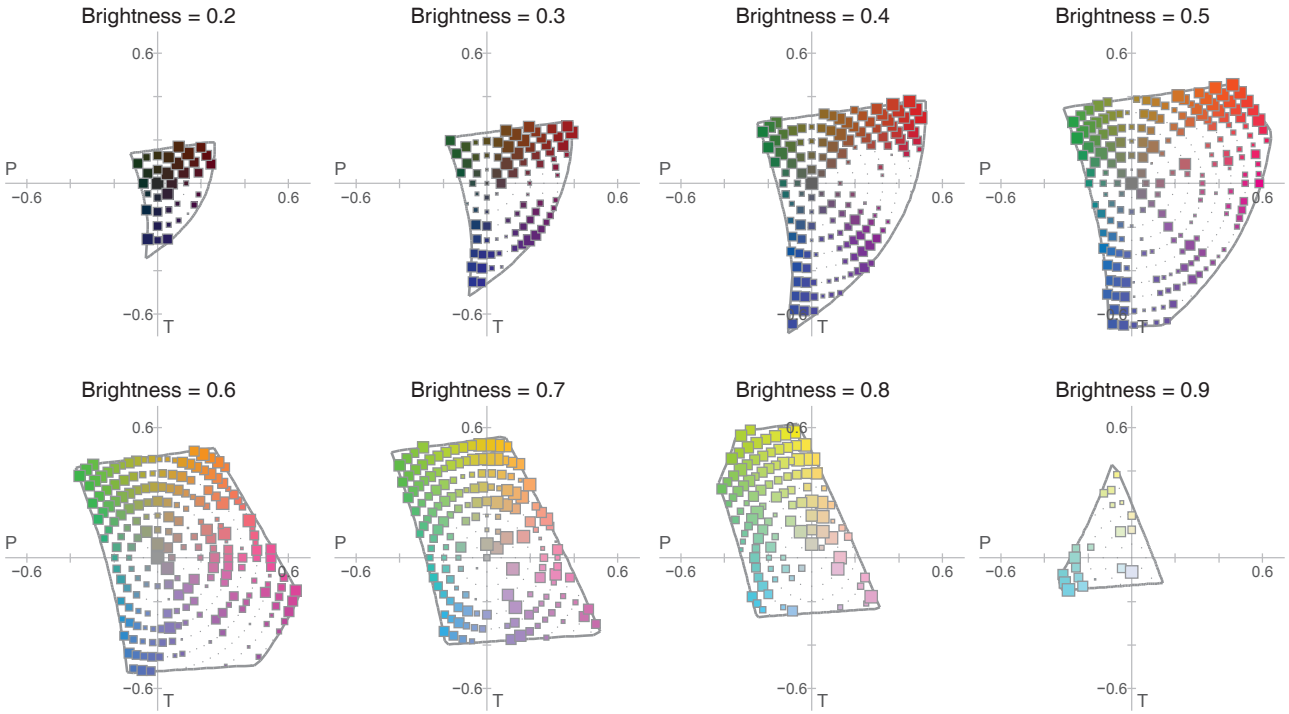

Figure 3. *Color saliency for planes corresponding to 8 brightness levels in IPT space. The size of the colored square is proportional to color saliency.*

salient region in the WCS dataset is demonstrated on the plot for Munsell value of 7. Similarly, there is a lack of an equivalent cyan salient region in the WCS dataset at a value of 8.

Discussion and Future Work

Our preliminary results showed initial support for our framework in three ways.

First, we were able to identify high-saliency regions in the color space. Salient colors from six languages in the World Color Survey match known foci locations of basic color terms. We observed that the sRGB colors cluster in the gamut corners, and include clusters not typically associated with basic names such as "lavender" and "cyan".

By creating a saliency plot for the two datasets in a common perceptual space, we were able to qualitatively compare the name regions on a display to those produced by reflective samples (Munsell chips). As expected, these are different due to both gamut limitations and appearance differences.

We showed that it is possible to build a non-parametric naming model for capturing detailed information about how colors categorically associated with one another without being negatively affected by noise in the data. Traditional naming dataset generally collects responses at a fixed number (usually in the hundreds) of colors which yields a large number of responses per color but lacks naming information between the colors. In contrast, the DoloresLabs naming dataset contains naming information for a large number of colors, but yields very few responses per color. Our model was able to interpret this sparse data to provide naming information at 1,234 points. In contrast, there are 330 colors in World Color Survey, 387 colors in Benavente's fuzzy English naming dataset, or the 216 colors in the HP Labs Multilingual Experiment.

We plan to further refine our interpretation of saliency and look into methods for quantitatively correlating the saliency regions across gamuts in a way that could create gamut mapping algorithms that preserve categorical names. We are currently looking into methods for computing categorical distance between pairs of colors, and hope such a measure can benefit further visualization and analysis of color naming datasets. We will also explore applications of our model in digital imaging, information visualization and computer graphics.

References

[1] R. Benavente, F. Tous, R. Baldrich, and M. Vanrell. Statistical modelling of a colour naming space. In European Conference on Colour Graphics, Imaging, and Vision, pages 406–411, 2002.

[2] R. Benavente, M. Vanrell, and R. Baldrich. A data set for fuzzy colour naming. Color Research and Appl., 31(1):48–56, February 2006.

[3] B. Berlin and P. Kay. Basic color terms: Their universality and evolution. University of California Press, 1969.

[4] R. M. Boynton and C. X. Olson. Salience of chromatic basic color terms confirmed by three measures. Vision Research, 30(9):1311–1317, 1990.

[5] Y. Chang, S. Saito, and M. Nakajima. A framework for transfer colors based on the basic color categories. In Computer Graphics International, pages 176–181, 2003.

[6] Y. Chang, K. Uchikawa, and S. Saito. Example-based color stylization based on categorical perception. In Applied Perception in graphics and visualization, pages 91–98, 2004.

[7] R. Cook, P. Kay, and T. Regier. The World Color Survey database: history and use. Elsevier, 2005.

[8] Dolores Lab. Color names experiment. http://blog.doloreslabs.com/2008/03/where-does-blue-end-and-red-begin/, 2008.

[9] P. Kay and C. K. McDaniel. The linguistic significance of the mean-

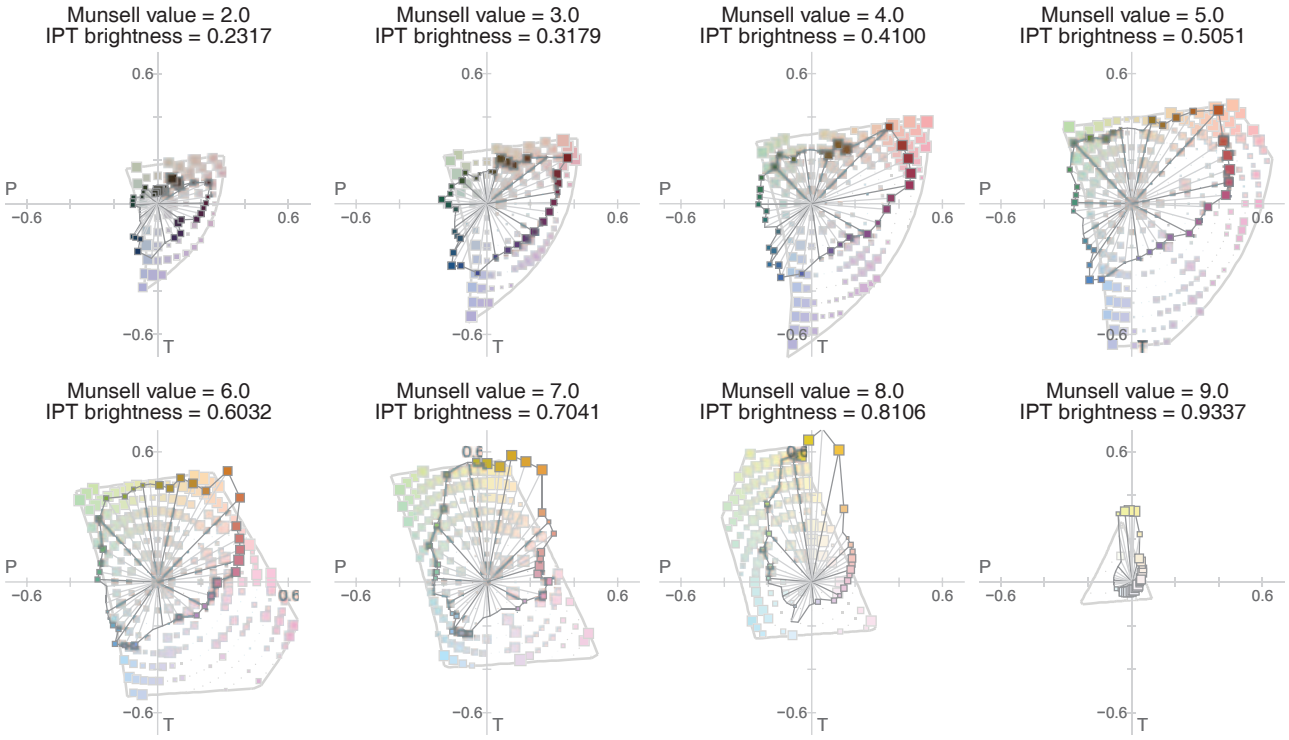

Figure 4. *Comparison of consistently named regions on the surface of the Munsell solid versus in the sRGB cube at eight Munsell values*

ings of basic color terms. Language, 54(3):610–646, September 1978.

[10] H. Lin, M. R. Luo, and L. W. MacDonald. A cross-cultural colour-naming study. Part I: Using an unconstrained method. Color Research and Application, 26(1):40–60, 2001.

[11] H. Lin, M. R. Luo, and L. W. MacDonald. A cross-cultural colour-naming study. Part II: Using a constrained method. Color Research and Application, 26(3):193–208, 2001.

[12] H. Lin, M. R. Luo, and L. W. MacDonald. A cross-cultural colour-naming study. Part III: A colour-naming model. Color Research and Application, 26(4):270–277, 2001.

[13] J. Lin. Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory, 37(1):145–151, January 1991.

[14] C. Moore, A. K. Romney, and T. Lien Hsia. Shared cognitive representations of perceptual and semantic structures of basic colors in chinese and english. 97(9):5007–5010, April 2000.

[15] N. Moroney. US Patent App. 11/285,850: Method, system, and computer software product for selecting elements of a digital image, 2005.

[16] N. Moroney. US Patent Application 11/259,597: Adapative lexical classification system, 2007.

[17] N. Moroney. US Patent Application 11/264,575: Lexical classification system with dynamic modifiers, 2007.

[18] N. Moroney. Multi-lingual color naming experiment. http://www.hpl.hp.com/personal/Nathan_Moroney/mlcn.html, 2008.

[19] H. Motomura. Categorial color mapping using color-categorical weighting method. Part I: Theory. J. of Imaging Science and Technology, 45(2):117–129, March/April 2001.

[20] H. Motomura. Categorial color mapping using color-categorical weighting method. Part II: Experiment. J. of Imaging Science and Technology, 45(2):130–140, March/April 2001.

[21] H. Motomura. Analysis of gamut mapping algorithms from the viewpoint of color name matching. J. of the Society for Information Display, 10(3):247–254, 2002.

[22] G. V. Paramei. Singing the russian blues: An argument for culturally basic color terms. Cross-Cultural Research, 39(1):10–34, Feb. 2005.

[23] B. Sayim, K. A. Jameson, N. Alvarado, and M. K. Szeszel. Semantic and perceptual representations of color: Evidence of a shared color-naming function. J. of Cognition and Culture, 5(3–4):427–486, 2006.

[24] J. Sturges and T. W. A. Whitfield. Locating basic colours in the munsell space. Color Research and Appl., 20(5):364–376, Oct. 1995.

[25] K. Yokoi and K. Uchikawa. Color category influences heterogeneous visual search for color. J. of Optical Societies of America A, 22(11):2309–2317, November 2005.

Author Biography

Jason Chuang received his B. Sc. in mathematics from the University of British Columbia (2002) and his M. Sc. in Scientific Computing and Computational Math from Stanford University (2005). He is currently a Ph. D. Candidate in Computer Science at Stanford University.

Maureen Stone (StoneSoup Consulting) is an independent consultant working in the areas of digital color and interactive information visualization. She is an adjunct professor at the Simon Fraser University School of Interactive Art and Technology, Editor-in-Chief of IEEE Computer Graphics & Applications, and is a member of ACM, IEEE and IS&T.

Pat Hanrahan is a professor of Computer Science and Electrical Engineering at Stanford University working in the areas of information visualization, visual analytics, graphics systems and architectures, and rendering algorithms.