What is the Color of Chocolate? – Extracting Color Values of Semantic Expressions

Albrecht Lindner[1]***, Nicolas Bonnier***[2]***, Sabine Süsstrunk***[1]
[1] ***Ecole Polytechnique Fédérale de Lausanne,*** [2] ***Océ – Print Logic Technology***

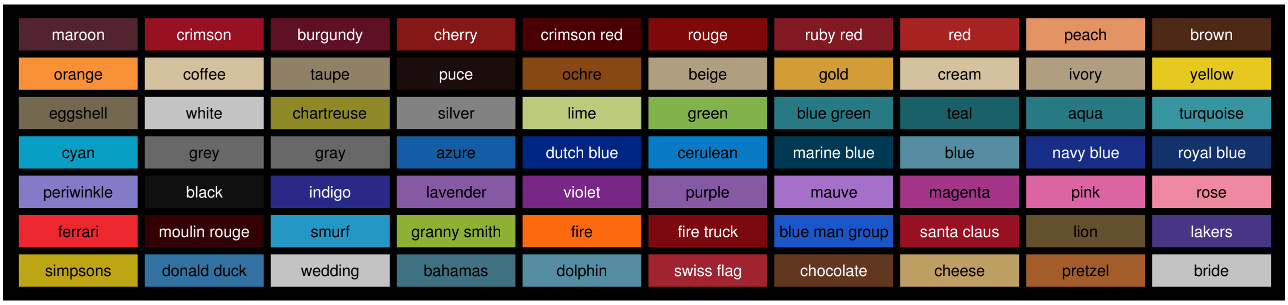


Figure 1: 70 semantic terms with their associated color patches. Our algorithm is able to determine the color values not only for an arbitrary color name (rows 1-5) but also for any other semantic expression (last 2 rows). It is fully automatic without any human intervention using only freely available images from Flickr.

Abstract

We present a statistical framework to automatically determine an associated color for a given arbitrary semantic expression. The expression can not only be a color name but any word or character string. In addition to the color value, we are also able to compute the result's significance, which determines how meaningful defining the color is for the expression. To demonstrate the framework's strength we apply it to two well known tasks: assessing memory colors and finding the color values for a given color name (color naming). We emphasize that we solve these tasks fully automatic without any psychophysical experiment or human intervention. Further, we outline the potential of our automatic framework and in particular the significance for the imaging community.

Introduction

For a computer to answer the question "What is the color of chocolate?" with the color triplet (R, G, B) = 98, 55, 32, it is necessary to bridge the semantic gap. The semantic gap comes from the fact that, on the one hand, *chocolate* is an expression that is meaningful to human beings. Almost everybody has seen and also tasted chocolate multiple times in his/her life. On the other hand, 98, 55, 32 is an abstract representation of color suited for a computer's bitwise representation of information. Bridging the semantic gap has been and still is a main challenge in computer science. In this article, we present a framework to bridge the semantic gap for semantic expressions (e.g. *chocolate*) and color values such as 98, 55, 32.

Relating a semantic expression with a numeric representation can be seen as a translation problem [2]. In order to learn the relation between the two "languages" one needs examples that have already been translated. We thus need a large corpus of data with numerical color representations that are linked to corresponding semantic expressions.

The just described properties of the necessary data corpus are perfectly matched by annotated images. The pixel values are the numeric representation of color and the annotated text/keywords are a semantic description of them. Learning from annotated images thus is a promising approach to train a computer to answer questions like the one in the title and similar ones.

There can be a large variety of semantic expressions linked to an image: keywords, title, filename, captions, comments and so forth. In this article we focus on image keywords since they are single entities. They do not need to be parsed for a grammar structure, such as a title or even associated sentences in the caption.

Image keywords themselves again cover a widespread semantic variety. They can represent color names (e.g. *red*, *turquoise*), sceneries (e.g. *sky*, *sunset*), objects (e.g. *chocolate*, *apple*), abstract terms (e.g. *love*, *war*) and more [11]. We do not try to distinguish between different categories, but rather use a unified framework for all keywords.

Our system uses annotated images from Flickr[1], an online image sharing community. This enables us to leverage the semantic input from millions of people. The publicly available API eases access to their database. It offers functions to query for a keyword and to download image metadata.

The framework uses well-known statistical methods to extract the necessary information from the abundance of data. This so called data-mining is robust against inevitable noise such as mis-typed keywords or wrong annotations.

We use two well established tasks in color science to demonstrate the strengths and capabilities of our framework. The first one is to determine memory colors. Memory colors are colors of familiar objects/scenes that strike human observers if not correctly reproduced. The three most cited colors are: green grass, blue sky and skin tone. We use our framework to determine the color values for these three categories and variations thereof as can be seen in Figure 2. The second task is color naming, where the goal is to find the color values associated to a color name and vice versa. We use color names and values from Moroney's online experiment [14] and compare our estimations against it. The top Figure 1 shows the color names along with the automatically estimated color values.

After demonstrating that our framework automatically solves these known tasks, we go a step further and present new possibilities this approach offers. The first is to automatically determine the associated color values of an arbitrary semantic expression. The second is to assess the association strength between a semantic expression and a color.

The great potential this approach offers is to show a way to bridge the semantic gap between color values and semantic ex-

[1] `http://www.flickr.com`

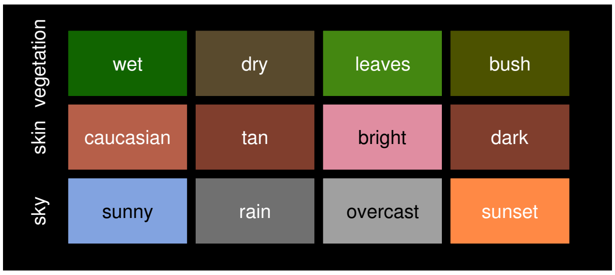


Figure 2: Example memory colors from our automatic algorithm. The three basic categories (vegetation, skin and sky) are further refined by the additional keywords indicated in the center of each patch.

pressions. The ability to perform this automatically offers new applications that have not been possible so far since acquiring results was expensive and labour-intensive. We believe this can help to improve applications such as image rendering, image tagging or image search.

Related Work

This work covers the areas of memory colors, color naming and image semantics. Memory colors and color naming are well established fields with many contributions from different researchers. Image semantics is comparatively more recent, but still, a lot of interesting research has already been done. In this section we aim to give an overview of the fields, highlighting major contributions and aspects related to our work.

At the beginning, memory colors were mostly discussed from a psychologist's point of view [8, 1]. Adams discusses in his article [1] the appearance of grass, snow, coal, gold, and blood under different illumination conditions. However, the lack of adequate color spaces limited research and applications in this field.

The invention of the Munsell Color System [16, 19] allowed to describe memory colors in a perceptual color space. In 1960, Bartleson defined ten different memory colors in the Munsell hue and chroma plane using 50 observers [3]. The categories had subtle nuances such as "green grass", "dry grass", "summer foliage", and "evergreen trees".

Memory colors have become important to assess different qualitative aspects in image reproduction. Yendrikhovskij et al. showed that a deviation from the memory color prototype is perceived as unnatural [29]. Taplin et al. demonstrated that if a color shift is unavoidable, observers agree on a preferred hue angle of the shift [24].

The active tuning of memory colors in image reproduction systems is a common application in industry. Park et al. proposed a method to adjust skin colors for a more preferred image rendering [20]. You and Chien presented a framework to enhance blue sky [30]. Other work focusses on segmenting memory colors in images [18, 10, 21]. The extracted maps can be used for further image processing.

In color naming, the well known study of Berlin and Kay [4] proposes that a language has, depending on their stage, two to eleven basic color terms. The simplest language distinguishes only black and white. As a language evolves new color terms are added in a strict chronological order: red, green, yellow, blue and so forth. Thus a language of a higher stage contains all color terms of the previous stages. Fully evolved languages all have the same eleven basic color terms. As this study suggests, color naming is a research subject in many fields such as linguistics, psychology, and ethnology.

Despite the importance of the different aspects of color naming, we focus on the acquisition of a numeric model for a given color expression. This is usually very labour intensive since the responses of many observers have to be gathered in order to achieve statistical significance. Recent publications used web-based approaches to crowd source the task to a large public [15, 17]. Moroney's experiment still continues online and the color names and their corresponding RGB values are accessible [14]. Another interesting study discusses the feasability of color adjustment by non-experts through the use of language [28].

Image semantics is a growing research field [22]. Online image sharing communities stimulate social tagging, which provides a rich resource for semantic research. Flickr makes its database accessible via a public API, where images can be downloaded together with their annotations and other metadata.

Uncontrolled tags from non-experts might be of lower quality, but the plentifulness of data and modern data mining techniques make their exploitation inexpensive and competitive. This is known as the "wisdom of crowds" and has become a more and more widespread approach for a variety of applications in the last years [23, 22]. In a recent study, van de Weijer et al. [25] use images from Google image search to learn a generative model for colors. The authors use a modified PLSA based model with a Dirichlet prior and enforced unidimenionality. The method performs well, but requires a retraining of the entire statistical model if a new color term is added. In our framework, it is possible to add a new color term without affecting previous estimations.

ImageNet [6] is a recently published database that aims at populating each synset in WordNet [13] with 1000 images on average. Deselaers and Ferrari computed image descriptors for all images in this database and made an interesting observation: the more two images are semantically related, the more their visual descriptors are similar [7].

In this work, we focus on one visual descriptor, color, and present a unified framework to determine numeric color models of semantic expressions. The input can be a color name (e.g. *red*), something related to a color (e.g. *tanned skin*) or any other expression.

Statistical Analysis

Our database contains image/annotation pairs $(I_i, A_i) \in I_{db}$. An image's annotation is a set of one or more keywords $A_i = \{w_1, w_2, \ldots\}$. The database can be split into two parts with a keyword w: $\mathbb{I}_w = \{I_i | w \in A_i\}$ and $\mathbb{I}_{\overline{w}} = \{I_i | w \notin A_i\}$ that contain all images annotated with keyword w and all remaining images, respectively. Each image is in exactly one of the two subsets, i.e. $\mathbb{I}_w \cap \mathbb{I}_{\overline{w}} = \emptyset$ and $\mathbb{I}_w \cup \mathbb{I}_{\overline{w}} = I_{db}$.

For each image I_i we compute a characteristic j, denoted C_i^j. Like the images, the characteristics can be split into two disjoint sets $\mathbb{C}_w^j = \{C_i^j | I_i \in \mathbb{I}_w\}$ and $\mathbb{C}_{\overline{w}}^j = \{C_i^j | I_i \in \mathbb{I}_{\overline{w}}\}$, containing all j characteristics of those image that are annotated with keyword w or not, respectively.

The goal is to assess whether a specific keyword w influences the characteristic j in the associated images. Therefore the two sets $\mathbb{C}_w^j$ are $\mathbb{C}_{\overline{w}}^j$ are compared against each other. This is done by a statistical test that assesses whether the values in the first set are significantly larger (or smaller) than the ones in the latter set.

Since the distribution of the characteristic is not known a priori, we have to use non parametric statistics. A well known test is the Mann-Whitney-Wilcoxon ranksum test [27, 12]. It assesses whether one of two sets tends to have larger values than the other.

For a characteristic j, let the computed values from the first set be $\mathbb{C}_w = \{C_1, C_2, C_4, \ldots\}$ and those from the second set $\mathbb{C}_{\overline{w}} =$

$\{C_3, C_5, C_6 \ldots\}$. We first sort the joint set $\mathbb{C}_w \cup \mathbb{C}_{\overline{w}}$ and give each element an associated integer rank. The ranksum is defined as the sum of the rank indices of the elements of $\mathbb{C}_w$. Wicoxon denoted this statistic with T.

In a concrete example we consider $\mathbb{C}_w = \{2.4, -1.0, 5.0, 4.1\}$ and $\mathbb{C}_{\overline{w}} = \{-0.4, 7.1, 2.9, 11.0, 3.0\}$. The sorted list of the joint set is: -1.0, -0.4, 2.4, 2.9, 4.1, 5.0, 7.1, 11.0 and $T = 1+3+5+6 = 15$. The expected mean and variance of T are [27, 12]:

$$\mu_T = \frac{n_w(n_w + n_{\overline{w}} + 1)}{2} \qquad (1a)$$

$$\sigma_T^2 = \frac{n_w n_{\overline{w}}(n_w + n_{\overline{w}} + 1)}{12} \qquad (1b)$$

where $n_w = |\mathbb{C}_w|$ and $n_{\overline{w}} = |\mathbb{C}_{\overline{w}}|$ are the cardinalities of either set, respectively.

The expected mean and variance can be used to compute the normalized z statistic, which is a widely used quantity in statistics [12]:

$$z = \frac{T - \mu_T}{\sigma_T}. \qquad (2)$$

For our application, the characteristics are bin values of color histograms in CIELAB color space. It is a simple perceptual color space that can be computed efficiently on a large database of images. Our assumption is that the original images from Flickr are encoded in sRGB [9]. We uniformly divide the CIELAB space into $15 \times 15 \times 15$ histogram bins in the ranges $0 \le L \le 100$, $-80 \le a \le 80$ and $-80 \le b \le 80$, respectively. Values outside the range on the a and b axis are clipped to the closest bin. In the above notation the j-th histogram bin of image I has the value C_I^j.

Example

Let us exemplarily consider the keyword *ferrari*. When comparing the two sets $\mathbb{C}^j_{ferrari}$ and $\mathbb{C}^j_{\overline{ferrari}}$ one finds that the first set has significantly larger values in the reddish bins. This is expressed by a high z value and can be easily detected as the z value distribution's maximum.

In this example the highest z value is $z^{j^*}_{ferrari} = 14.4$ where j^* is the bin with center $[L_{j^*} = 63.3,\ a_{j^*} = 53.3,\ b_{j^*} = 21.3]$ and describes a red as depicted in Figures 3 or 1.

Figure 3 shows the $z^j_{ferrari}$ values in a 3-dimensional heat map. The three orthogonal planes are defined by $L = L_{j^*}$, $a = a_{j^*}$ and $b = b_{j^*}$; thus, the intersection of the planes is the bin center j^* with maximum z value. For better orientation the bottom plane shows the bin centers' colors for $L = L_{j^*}$.

We show a similar plot for the keyword *color* in Figure 4. As the term suggests, there are many colors present and little neutral white, gray or black pixels. This is why there are negative z values all along the center axis and elevated z values everywhere else. There is no clearly located maximum value as opposed to the *ferrari* example in Figure 3. Also the z values are lower than in the previous example.

Discussion

The task we are trying to solve is to find the color value that corresponds to a semantic expression. This high-level task is drastically reduced in complexity by using the presented statistical framework. We do not use insecure pre-processing steps like trying to find salient regions, to label image regions, to classify images, to estimate the credibility of an annotated keyword and so forth. Neither do we use laborious manual work or crowd sourcing with online experiments.

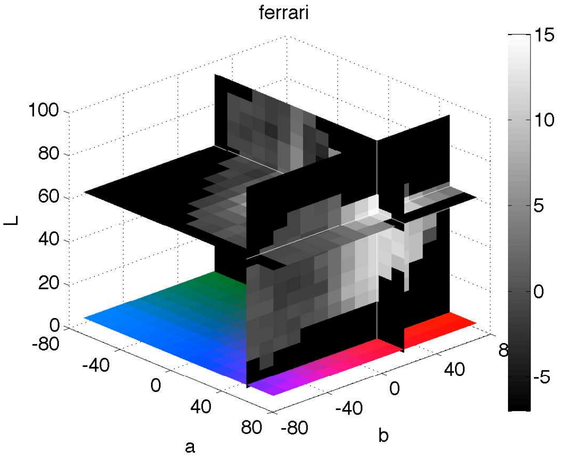


Figure 3: The $z^j_{ferrari}$ values in a 3-dimensional heat map. The maximum is $z^{j^*}_{ferrari} = 14.4$ and is at the crossing of the three orthogonal planes. The homogeneous dark areas along the plane borders are out of gamut values. For better orientation in the *ab*-plane we show on the plot's floor a plane indicating the colors for the bins with $L = L_{j^*}$.

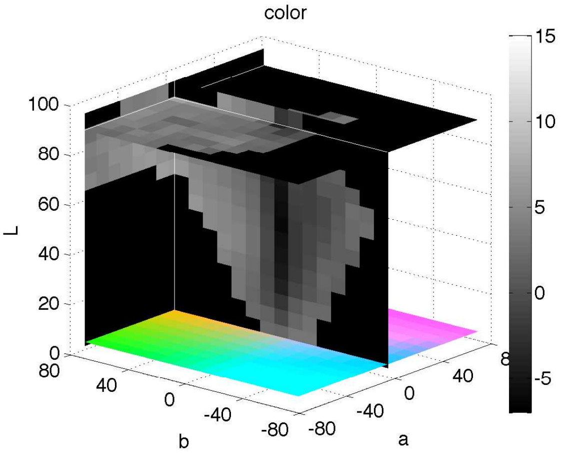


Figure 4: Same plot as in Fig. 3 but for z^j_{color} values. For better comparison the heatmap has the same scale as above. The only certainty is that there are less image pixels along the neutral gray axis as indicated by the negative z values. Since all colors are more or less equally present there is no clearly located maximum.

The simplistic nature of our approach allows us to process a lot of data on relatively little hardware. The quantity outweighs a possible input data's lack of quality.

The *ferrari* example shows an important property of our statistical framework. Even though the official color, *rosso corsa*, is well defined in controlled lab conditions, in the real world the cars are exposed to different lighting conditions and shadows. Thus the pixel values of red Ferraris are spread in color space. The z values inherently reflect this fact: the more frequent a particular tint of ferrari red in images, the higher the corresponding z value; less frequent values that occur under extreme lighting conditions have lower z values.

Picking the corresponding color for a semantic expression can be as easy as finding the bin with maximum z value. But also the other bins contain information about the semantic expression. Analyzing the whole z value distribution indicates how confined the corresponding region is. As can be seen in Figure 3, there is a strong peak in the $z_{ferrari}$ value's distribution around the red bin. In general, stronger peaks imply well defined colors whereas flatter peaks imply more color variation of a semantic concept.

The number of histogram bins has to be chosen with respect to two limitations. If one chooses too few bins, the sampling in color space is very coarse and estimating fine color nuances is not possible. If there are too many bins, the histogram begins to be sparse and is more affected by noise. We address this further in the third experiment of the following section.

The code for our experiments along with example images is available for download and re-use: `http://ivrg.epfl.ch/color/color_of_chocolate`.

Experiments

We present five different experiments to show the strength and potential of our approach. We demonstrate the efficiency of the statistical framework with the examples of memory colors and color naming. The framework solves these tasks fully automatically. We also provide experimental justification for our choice to set the one parameter the framework depends on. Then we outline how the framework can be further exploited to add semantic understanding of color to image applications.

1. Memory colors: We use the three basic memory colors, which are *vegetation*, *skin*, and *sky*. We then chose other additional keywords that modify the tint of the memory colors in a distinct way. We combined *vegetation* with *wet*, *dry*, *leaves*, *bush*, further *skin* with *caucasian*, *tan*, *bright*, and *dark*, and finally *sky* with *sunny*, *rain*, *overcast*, and *sunset*. We then downloaded 500 images for each combination.

The rows in Figure 2 show the output of our statistical analysis for the different combinations of memory colors. It is clearly visible how the shade of a memory color varies with the specific context; e.g. tanned skin is darker than caucasian skin. The variations of a memory color can be very extreme, such as for sky under different environmental conditions.

To give an intuition of the z value distribution and how it changes for different semantic expressions we show in more details the cases *sky+sunny* and *sky+sunset*. In order to show the z values on a plane we computed also color histograms on the *ab*-plane, discarding the luminance information.

Figure 5 shows the bin centers' colors in the *ab*-plane and the corresponding z value distribution as a heat map. One sees how the expression *sunset* causes the z values to rise in the orange and red regions of the histograms. For *sunny* the highest z values are as expected in the blue region.

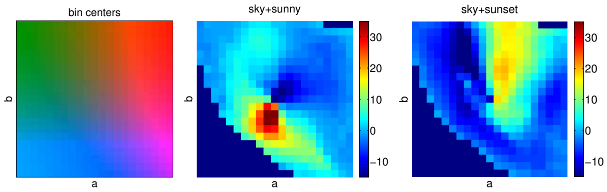
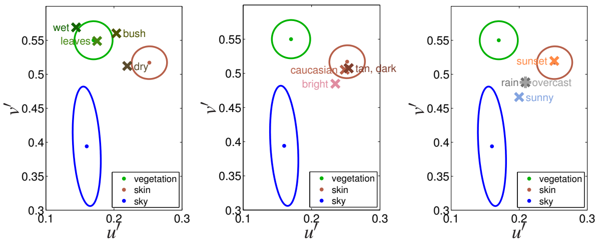


Figure 5: The map on the left shows the colors of the bin centers. In the middle and on the right are the z value distributions for *sky+sunny* and *sky+sunset*, respectively. The dark blue homogeneous areas are out of gamut values.

We compare our memory color values with Yendrikhovskij's values [29]. Figures 6(a) to 6(c) each show his ellipses for *vegetation*, *skin* and *sky* in the $u'v'$ plane, respectively. For clarity we do not show the whole z distribution for each keyword combination, but only the value with maximum z value. They are plotted as labeled cross-marks in the respective color.

Our values lie within or relatively close to the ellipsis for *vegetation* and *skin*. However, they differ significantly for *sky*. The reason is that sky drastically changes under the different weather conditions we used for this experiment. Even a *sunny sky* is brighter than what is usually considered as sky blue. The cross-mark is thus shifted outside the ellipse towards the neutral white point.



(a) vegetation (b) skin (c) sky

Figure 6: All three subfigures show the ellipses from Yendrikhovskij [29] for *vegetation*, *skin* and *sky*. Our results are visualized with crossmarks: (a) variations of *vegetation*, (b) variations of *skin*, (c) variations of *sky*. For clarity only the color estimates with maximum z values are shown, not the complete z distribution.

The distinction between different varieties of a memory color is significant. This is crucial for high quality image rendering since images with wrong memory colors appear unnatural [29]. There is not a single vegetation green in the world, but it visibly changes across landscapes and human observers expect to see it the way they know it. The same holds for skin tones and sky blues.

2. Color names: We use the 50 most common[2] color names from Nathan Moroney's color naming experiment [14]. Using Flickr's API we downloaded for each color name 200 images. The search query was simply the color name itself. The color patches in Figure 1 represent the bin with maximum z value for each semantic expression (i.e. color name).

Figure 7 shows the ΔE distances between Moroney's estimations and our color value with the highest z score. The distance distribution shows that the two estimations are relatively close to each other with a few outliers. It is worth to point out that due to the binning in CIELAB histograms we introduce an inherent quantization error. A color within a bin can have a distance to its center of up to $\Delta E = 8.2$, which is indicated by the dashed red line. The outliers with highest ΔE distances are: *puce* ($\Delta E = 55.2$), *royal blue* ($\Delta E = 45.3$) and *lime* ($\Delta E = 42.1$).

In the following we explain for these three color names why the estimations differ so much since it helps to get a better understanding of the framework's functioning:

PUCE: Our estimate is 27, 13, 10 while Moroney's values are 171, 134, 55. Even though this color name is rarely used and opinions about its correct tint diverge (a very complete online color database [5] reports 204, 136, 153), our estimate is clearly too dark.

The reason for this is, that the term *puce* has two other meanings: *Puce Moment*, a music group and *puce* as the french translation of *microchip*. It turns out that *puce* is more often used to refer to the band or the microchip than to the color. The images from the band's concerts have, like most live stage acts, a black background with the band members illuminated in the foreground. And microchips almost all have a dark body. Thus, black is over-proportionally present in images with keyword *puce* and thus our framework finds this association.

[2] status on October 20, 2011

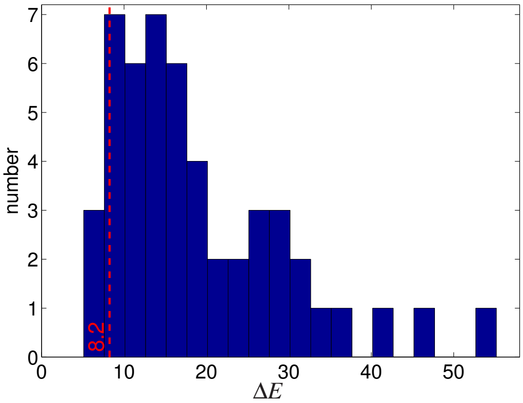


Figure 7: The ΔE distance in CIELAB color space when comparing the color values with maximum z value with the values from Moroney's database. The distance a color within a bin can deviate from its center is up to $\Delta E = 8.2$.

The true origin of the color name *puce* is different. It comes from the French word for flea, puce, possibly a reference to the 16-19th century source of the carmine dye colour that was extracted from Mexican scale insects (resembling fleas)[3].

We see that the reason for the large deviation is semantic ambiguity. This can be seen as a positive and a negative point. If the task was to find the exact values for the color *puce* it is better to do a color naming experiment since human observers do understand the semantic ambiguity. However, if the task is to find what the semantic expression means for the majority of images, our estimate is better.

ROYAL BLUE: There exist two kinds of royal blue: a traditional royal blue 0, 35, 102 [5] and a modern royal blue as defined by the Word Wide Web Consortium (W3C) 65,105,225 [26]. Society's perception must have changed from the darker version to the brighter version over time. Our and Moroney's estimates are 19, 49, 107 and 39, 41, 212, which are closer to the original and the modern version, respectively.

The reason why the statistical framework ranks the traditional royal blue first is due to the "Royal Blue Coach Services", an English coach operator from 1880 to 1986. Their coaches were varnished in traditional royal blue; which is obvious when considering the early founding year. The coaches seem to have a very active fan community that preserves them for nostalgic reasons. They also post many pictures online so that the analysis ranks this color first.

Again, our estimate is different from what one would expect at first sight but it is not wrong. The distance between the color *traditional royal blue* and our estimate of the semantic expression *royal blue* is $\Delta E = 12.6$. The distance between Moroney's estimate and the W3C's definition of *royal blue* is much higher: $\Delta E = 40.5$.

LIME: There is no straight forward explanation why the estimate 186, 204, 124 is not bright and saturated enough. Moroney's estimate for this color name is better: 106, 239, 59. The best explanation to give is that our estimation, which is based on 200 images per color name, is only correct with a certain probability.

For a given semantic expression (i.e. color name) our system computes a z value for all possible color values. So far we considered only the color value with the highest z value, but for a deeper insight also the other z values need to be analyzed.

In order to consider all z values we rank for a given color name the color estimates with decreasing z value. We computed for the best 1000 color estimates the ΔE distance to Moroney's value. The values for the first rank are thus the ones shown in the histogram in Figure 7 (the histogram shows the color estimate with highest z values, i.e. 1st rank). The following 999 distances are the deviations for the less significant colors.

The results for all the 50 color names are summarized in Figure 8. It shows the rank on the logarithmic horizontal axis and the ΔE distance on the vertical axis. The deviations continuously grow for increasing ranks and become more prone to noise. The graph illustrates that it is not possible to guarantee a specific error. But it shows that, from a probabilistic viewpoint, colors that are ranked first are better estimates.

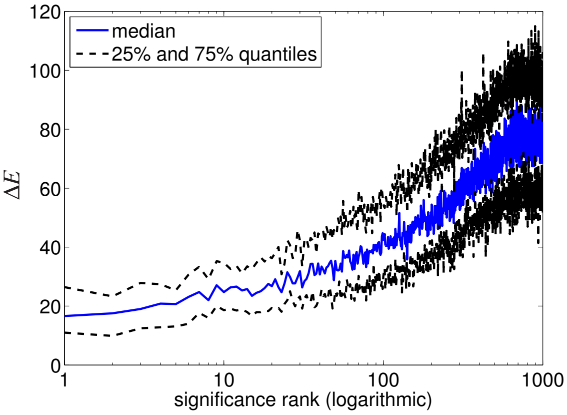


Figure 8: ΔE distances between Moroney's 50 color names and our estimations. We compare not only our best estimate (color value with the highest z value) but the first 1000 estimates (sorted by decreasing z values). This significance rank is plotted along the logarithmic horizontal axis. It is clearly visible that color estimates on the first ranks have smaller errors.

3. Dependency on the number of bins: The only parameter our framework depends on is the number of bins in the histogram. To show its effect on the results, we compute the ΔE distances between ours and Moroney's estimates for $2^3, 3^3, 4^3, \ldots, 32^3$ histograms bins. Figure 9 shows the median and 25% and 75% quantiles of the ΔE distance as a function of the number of bins. The additional red curve is the maximum quantization error, which is the distance between the bin center and bin corner. Please note that the horizontal axis is not linear, but cubic.

It is visible that the error is high for very small bin numbers and then decreases for higher number of bins. The plot also shows that the error stops improving for approximately 12^3 or more bins. Our choice of 15^3 bins is thus on the safe side, but not excessively high.

4. Arbitrary semantic expressions: In the next experiment we do not limit the semantic expressions to memory colors or color names. We downloaded for 20 semantic expressions 200 images each. The two bottom rows in Figure 1 illustrate the semantic expressions[4] and their associated colors. Even though one might want to argue about the correct tint of the one or the other example, they are all very good estimates. Precisions are in the range of human disagreement.

[3] We thank the anonymous reviewer for this knowledgeable contribution.

[4] *granny smith* is a green kind of apple and *lakers* is a basketball team from the United States with a violet and yellow outfit.

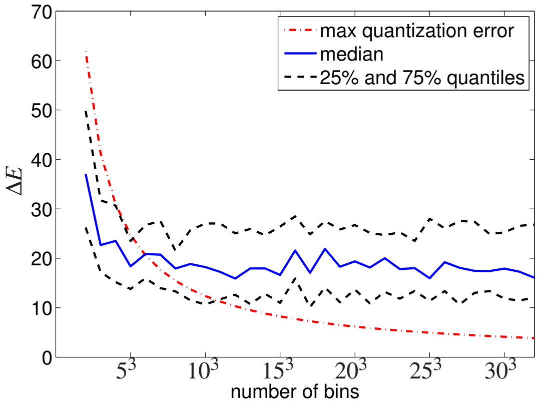


Figure 9: Median and 25% and 75% quantils of ΔE error between ours and Moroney's estimates as a function of the number of bins. Please note that the horizontal axis is not linear, but cubic. The red curve shows the maximum quantization error, the distance between a histogram bin's center and corner.

5. z **significance:** We finally show that, apart from assessing an associated color value, the z values can be used to estimate the association strength. The higher the z value the more an association is significant. Thus, semantic expressions that are very meaningful in terms of color have a higher z value.

We compare the maximum z values from the color names (experiment 2) and the arbitrary semantic expressions (experiment 4). Figure 10 shows the maximum z values of both sets in a histogram plot. The highest z values are solely from color names. This is not surprising since color names have a stronger link to colors by definition. The highest values stem from *red* (28.8), *yellow* (26.3) and *purple* (26.0). Among the arbitrary semantic expressions the highest z_{max} values are obtained for *granny smith* (14.9), *ferrari* (14.4) and *smurf* (13.6).

For the sake of completeness we downloaded also 200 images with keywords for which we expect low z_{max} values. The results are: *poster* (4.5), *painting* (4.0) and *boredom* (2.5). The reason why the z values are low is straight forward. None of these semantic expressions can be associated with a specific color, even though *poster* and *painting* might be colorful in general (see also example in Figure 4).

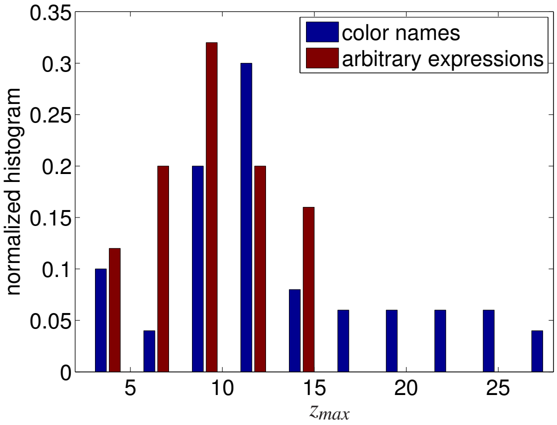


Figure 10: Histogram of the maximal z values for the color names (experiment 2) and the arbitrary semantic expressions (experiment 4). The color names have higher z_{max} values and are thus stronger associated with color.

Conclusions

In this article we presented a novel approach to link colors and semantics. Our algorithm is able to automatically give an answer to the question "What is the color of X". X can be a color name such as *red*, but also any other semantic expression such as *chocolate*, *vegetation* or *dolphin*.

As easy as this task is for human beings, as hard it is for computers. The challenge is to bridge the semantic gap. The semantic expression (e.g. *chocolate*) is meaningful for a human being, but only a character string for a computer. In order to find the correct answer 98, 55, 32, the computer has to retrace the semantic meaning.

We argue that the semantic gap can be seen as a translation problem from a semantic to a numeric language. In order to learn this translation process, a computer needs examples that are already translated.

A rich source of examples that contain both semantic expressions and numeric values are annotated images. When annotating images, human beings translate from the numeric pixel values to a semantic representation of the images. Our goal is develop an algorithm that learns from annotated images to invert this process.

The image acquisition is done on Flickr, an online image sharing community. The API allows to query the database, to download the corresponding images and their metadata. The abundance of data on Flickr makes our framework very versatile.

We compute color histograms in CIELAB color space on all images. We then compute for each keyword and each color bin a statistical test. It assess whether the color bin has the tendency to be more populated under the presence of the keyword. This results in a z score. The bin with the highest z value designates the corresponding keyword's color value.

This approach's potential is demonstrated on two common tasks of the color community: assessing memory colors and color naming. Usually these tasks are carried out with psychophysic experiments. However, our approach is fully automatic and gives reasonable results. We compare our results with state-of-the-art data. For the experiment on memory colors we use Yendrikhovskij's ellipses of memory colors and for the color naming experiment we compare against Moroney's online color naming experiment [15]. We show in both cases that our approach gives comparable results.

The same framework can be used to determine the associated color of an arbitrary semantic expression. We show examples such as *chocolate*, *smurf* or *pretzel* in Figure 1.

We finally show how the z value can be used to determine the association strength. Color names such as *red* and *yellow* have high z values and are thus significant. The arbitrary semantic expressions have lower z values than the color names. This is reasonable since we estimate the significance for colors. The highest z values among arbitrary expressions are deduced from *granny smith* (a green kind of apple) and *ferrari*.

We believe that our framework brings great value to the imaging community. It shows how laborious psychophysics can be replaced with statistical processing on annotated (i.e. semantically enhanced) images. The significance values can be exploited for applications such as image processing, image tagging and image search.

Future work should include word sens disambiguation methods. This can help to separate different meanings of a keyword such as *puce* (see experiment 2). Further, we want to demonstrate on different imaging applications that our framework adds a semantic understanding of color and thus improves quality.

References

[1] G. K. Adams. An Experimental Study of Memory Color and Related Phenomena. *The American Journal of Psychology*, 34(3):359–407, July 1923.

[2] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, February 2003.

[3] C. J. Bartleson. Memory Colors of Familiar Objects. *Journal of the Optical Society of America*, 50(1):73–77, January 1960.

[4] B. Berlin and P. Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, 1969.

[5] Color Database. http://www.perbang.dk, last checked oct. 2011.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: a large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, 2009.

[7] T. Deselaers and V. Ferrari. Visual and semantic similarity in imagenet. In *Conference on Computer Vision and Pattern Recognition*, pages 1777–1784, 2011.

[8] K. E. K. Hering. *Zur Lehre vom Lichtsinne*. G. A. Agoston, 1878.

[9] International Color Consortium. *sRGB, IEC 61966-2-1:1999*.

[10] M. Jaber, E. Saber, and F. Sahin. Extraction of Memory Colors Using Bayesian Networks. In *IEEE International Conference on System of Systems Engineering*, pages 1–6, May 2009.

[11] A. Lindner, N. Bonnier, M. Candemir, and S. Süsstrunk. Automatic grouping of semantic keywords to improve image rendering. In *Proceedings of Color and Imaging Conference*, pages 217–222, November 2010.

[12] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[13] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[14] N. Moroney. http://www.hpl.hp.com/personal/nathan_moroney/, last checked oct. 2011.

[15] N. Moroney. Unconstrained web-based color naming experiment. In *Color Imaging VIII: Processing, Hardcopy, and Applications*, volume 5008, pages 36–46. Annual IS&T/SPIE Symposium on Electronic Imaging, 2003.

[16] A. H. Munsell. *Munsell Book of Color*. The Munsell Color Company, 1929.

[17] D. Mylonas, L. MacDonald, and S. Wuerger. Towards an Online Color Naming Model. In *Proceedings of Color and Imaging Conference*, pages 140–144, 2010.

[18] F. Naccari, S. Battiato, A. R. Bruna, A. Capra, and A.Castorina. Natural scenes classification for color enhancement. *IEEE Transactions on Consumer Electronics*, 51(1):234 – 239, 2005.

[19] D. Nickerson. History of the Munsell Color System and Its Scientific Application. *Journal of the Optical Society of America*, 30(12):575–586, December 1940.

[20] D.-S. Park, Y. Kwak, H. Ok, and C.-Y. Kim. Preferred skin color reproduction on the display. *Journal of Electronic Imaging*, 15(4), October 2006.

[21] B. T. Ryu, J. Y. Yeom, C.-W. Kim, J.-Y. Ahn, D.-W. Kang, and H.-H. Shin. Extraction of Memory Colors for Preferred Color Correction in Digital TVs. In *Color Imaging XIV: Displaying, Processing, Hardcopy, and Applications*, volume 7241. Annual IS&T/SPIE Symposium on Electronic Imaging, 2009.

[22] N. Sawant, J. Li, and J. Z. Wang. Automatic image semantic interpretation using social action and tagging data. In *Multimedia Tools and Applications*, volume 51, pages 213–246, 2011.

[23] J. Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday; Anchor, May 2004.

[24] L. Taplin and G. Johnson. When Good Hues Go Bad. In *Conference on Color in Graphics, Imaging and Vision*, pages 348–352, April 2004.

[25] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus. Learning color names for real-world applications. *IEEE Transactions in Image Processing*, 18(7):1512 – 1523, 2009.

[26] W3C Recommendation, World Wide Web Consortium. *CSS Color Module Level 3*, 07 June 2011.

[27] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[28] G. Woolfe. Making color adjustment accessible to non-experts through the use of language. In *Proceedings of Color and Imaging Conference*, pages 3–7, 2007.

[29] S. Yendrikhovskij. *Color reproduction and the naturalness constraint*. PhD thesis, Thesis, Technische Universiteit Eindhoven, 1998.

[30] J.-Y. You and S.-I. Chien. Saturation Enhancement of Blue Sky for Increasing Preference of Scenery Images. *IEEE Transactions on Consumer Electronics*, 54(2):762–767, May 2008.

Author Biography

Albrecht Lindner is a PhD student at EPFL Lausanne, Switzerland. He is supervised by Professor Sabine Süsstrunk and is sponsored by Océ. He graduated from the University of Stuttgart (Germany) and Télécom ParisTech (France) within a German-French exchange program in 2008. His two Master degrees are in the domains of electrical engineering (Stuttgart) and signal and image processing (Paris). His Master thesis was directed by Professor Francis Schmitt and Professor Joachim Speidel, sponsored by Océ.