A Similarity Measure for Large Color Differences

Nathan Moroney, Ingeborg Tastl and Melanie Gottwals
Hewlett-Packard Laboratories, Palo Alto, CA

Abstract

Hundreds of large color differences, of magnitude 20 ΔE_{00}, were generated and used in a visual sorting experiment. The process of generating these color differences and two specific experiments are described in detail. The results show that small color difference metrics, such as ΔE_{00}, do not consistently model the visually sorted differences for large differences. A new similarity measure, based on a cosine similarity between categorical vectors of colors, is described and used to more consistently model large color differences. This similarity metric can be used to better characterize large color errors during reproduction, for image processing operations such as segmentation or as a feature for content retrieval. The new measure can also be applied to visual phenomena, such as categorical perception, in which within category color differences are perceived as smaller than across category differences.

Introduction

There has been decades of research on the topic of color difference metrics.[1-3] This work has primarily considered small color differences or just-noticeable differences (JND's). These distance metrics include CIELab 1976 ΔE^*_{ab}, ΔE_{94} and ΔE_{00} and are most applicable to quantifying whether or not two colors match. There has been some work in the area of large color differences but even for these publications[4-8], the emphasis is on understanding how well existing color differences metrics scale up to larger color differences. Indeed unlike a threshold or JND, there is perhaps even limited consensus on what constitutes a large color difference. Likewise existing color difference metrics are based on geometric distance computations and advanced color differences metrics differ mainly in the complexity of the weighting schemes applied to underlying geometric quantities. One area where larger, non-threshold color differences have been researched, is the phenomena of categorical perception[9-10]. This phenomenon is the increase in perceived color differences for color pairs that cross categorical boundaries relative to perceived differences for pairs within a single category. However, this has been a separate research activity from formulation and testing of color difference metrics.

One challenge for color difference research is the fact that while a given color can be perceived and described using a three dimensional representation, a color pair is six dimensional. This means that color sampling is a challenge and even simple, concise, systematic sampling results in rapid explosion in the number of color pairs. For example a 5 by 5 by 5 sampling in a given color space will result in thousands of possible different pairs of color differences. An additional challenge is the time and expense required for the collection of visual evaluations, which are typically based on the paired-comparison forced-choice technique. There are methods to sub-sample the full sampling of pairs but even with a substantially more efficient technique, this method also limits all visual evaluations to pair-wise assessments. Finally, the caveat that color difference equations, like ΔE_{00}, are to be used to quantify small color differences of less than 5 is widely known, but mostly ignored. How else is the maximum or 95th percentile error for a device characterization to be expressed?

Experiments

Given the challenges described in the introduction, we spent considerable time designing the specific experiment to investigate large color differences. The first area of effort was on the sampling of color pairs and the second area of effort was the specific experimental task. The sampling and task were used to conduct two experiments, although for this paper we focus mainly on the first of the experiments. The first experiment makes use of the World Wide Web to collect distributed data from online volunteers. A second experiment used full sampling and a single characterized display to collect data from a dozen participants in a controlled, laboratory setting.

For the generation of color differences a random walk with farthest point sampling was used. This algorithm consisted of the following four steps. First, random RGB values for a known additive display were generated using a uniform random number generator. From a specific point of this set a random walk was taken away from the point in RGB space. This walk was terminated once the distance between the first point and the second point was within some error term away from a specified ΔE_{00} value. For this paper, the sRGB color space was used, the error term was 0.0001 and the target ΔE_{00} value was 20. Thousands of candidate pairs were generated before the next processing step. Second, given a large number of possible color pairs, farthest point sampling[11] was applied to narrow down the pairs to several hundred pairs. The farthest point sampling was based on the average CIELAB coordinate of the two colors. The farthest point sampling based on the average CIELAB value effectively eliminates similar pairs of colors. Third, a nearest pair thresholding was applied to the minimum of the average color differences (between first-first/second-second points and first-second/second-first points of the pairs). A threshold of 15 was used and the result is a further reduction in similar pairs and elimination of mostly symmetric pairs of differences. Forth, the target number of pairs was selected, shuffled and assigned to 18 blocks of 9 color difference pairs. Additional hierarchical sampling or drawing of points from different blocks, was also applied but will not be described in this paper. The resulting set of color differences is shown in patch form on the left of Figure 1 and the corresponding a* versus b* plot on the right of Figure 1. Note that the end result of the steps described above is a collection of color difference vectors that fairly uniformly covers most of the gamut with minimal overlap of the end-points or the vectors. These vectors also lack any consistent orientation and vary randomly in their lightness, chroma and hue values. This sampling is purposely quite different from a regular, systematic sampling and can be tuned to

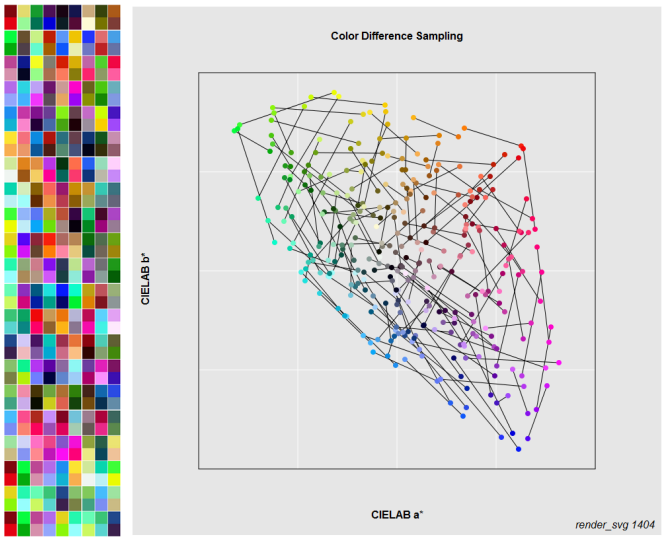

Figure 1. *The color differences pairs used in the experiment shown as color patches, on the left, and as vectors in an a* versus b* plot on the right. All of the plotted color differences are approximately 20 ΔE00 in magnitude. However, the differences vary from 20 to 80 1976 ΔE*ab.*

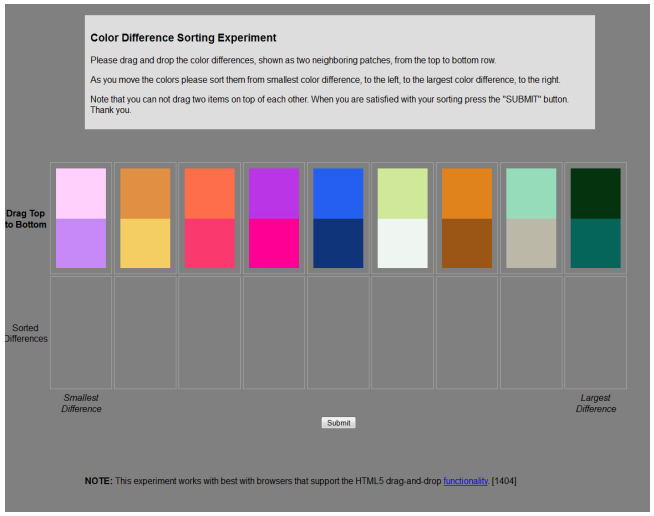

Figure 2. *Screen shot of the color difference sorting task used for the experiment.p*

larger and smaller sizes and to various magnitude color differences.

The experimental task consisted of an interactive sorting of color differences. Specifically, HTML5 drag-and-drop functionality was used with nine random starting locations and a forced sorting to nine ending locations. The color differences pairs were oriented horizontally and rendered to 90 by 180 pixels. The observers were instructed:

Please drag and drop the color differences, shown as two neighboring patches, from the top to bottom row. As you move the colors please sort them from smallest color difference, to the left, to the largest color difference, to the right. Note that you cannot drag two items on top of each other. When you are satisfied with your sorting press the "SUBMIT" button. Thank you.

A screen shot of an initial view of the experiment is shown in Figure 2. This figure shows the color differences pairs before sorting. The experiment also recorded the time taken to complete the sorting and the user sort sequence. Based on 285 participants the median time to complete the sorting of nine color difference pairs was 91 seconds. The experiment was developed as a web-based or online experiment and volunteers were recruited through various means.

The sorting task can be contrasted with the typical paired-comparison forced-choice task. Many participants reported that the task was challenging and one aspect of the effort was deciding which color difference attribute (e.g. lightness, chroma, hue, or combinations thereof) to use. At the same time, several participants reported that the task was interesting enough that they voluntarily did multiple trials because it was enjoyable. Finally, given the design the observers were able to simultaneously view the whole range of color difference pairs of a block at once as opposed to in isolation. Observers were not required to attempt to remember color differences as part of the task. The mathematic or modeling implications of sorting versus forced-choice binary decisions will be discussed in the discussion section.

In parallel a laboratory experiment was also conducted using a single known and characterized display. This experiment used a full sampling of all blocks of color differences, or 22 trials per session. Twelve color normal observers participated and an HP DreamColor Z27x Display in sRGB mode was used. The median time to complete the sorting for the laboratory experiment was 46 seconds. Some additional details about this version of the experiment will be mentioned in the discussion but for the following analysis section the results shown are for the web-based experiment.

Analysis

To analyze the experimental results, the data was aggregated and sorted by block. The raw data consisted of a random block identifier and a final observer sort sequence for the randomized color difference pairs. The pair numbers were lists as columns and the result is a matrix of ranks with each element consisting of indices 1 to 9, each occurring only once (no ties). This raw data was transformed into frequencies by taking the summed counts for each color pair for each column or rank order and dividing them by the total number of times the block was sorted. This resulted in a 9 by 9 matrix of frequencies for each block. Each row of this matrix corresponds to a color difference pair and each column is a rank order. Finally, each row of a given block was approximately sorted by the peak rank for the color difference pair. This was initially plotted as a stacked area plot but can be more easily interpreted when drawn as a column of area charts. Example renderings of this analysis are shown in Figures 4-6 where the results are shown for three different blocks of color differences. For each block the top area chart is the color pair that was most consistently sorted to have the smallest color difference. Similarly the bottom area chart is the color pair that was most consistently sorted to have the largest color difference. The intermediate charts vary between these extremes and all charts also show an

approximate rendering of the color difference pair to the right of the charts. The order of the color pairs can be considered as an aggregated ordering of the all the observers that performed the experiment for data of a specific block. The number varies from block to block but is typically around 15 for the data set that we have collected so far. The unique color pair ID is displayed on the y-axis. Thus it is quite easy to analyze the specifics of the pairs that are observed to be the most similar and the most different for the different blocs.

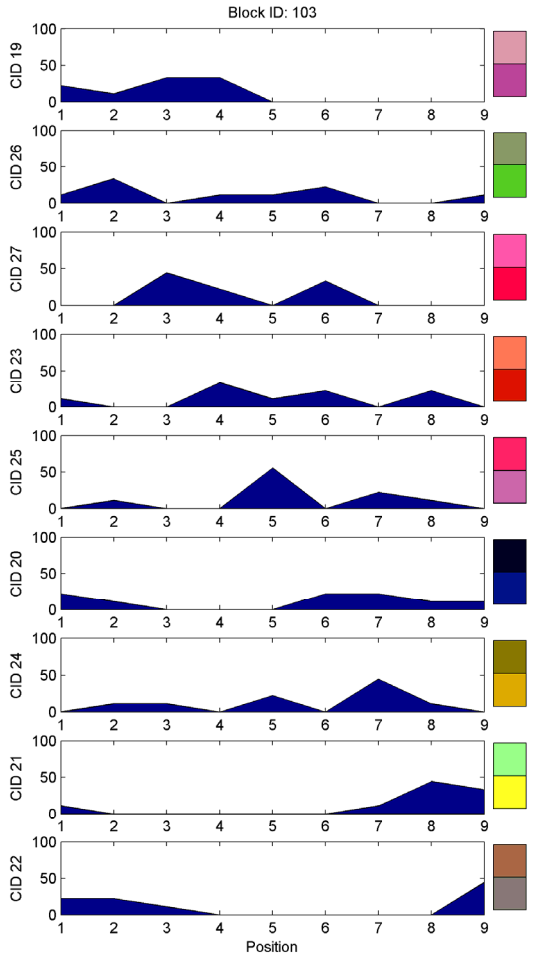

Figure 4. *Area charts per color difference pairs for three blocks of experimental data. In all sub-plots the x-axis is an ordinal scale of position or rank going from smallest, to the left, to the largest on the right. The y-axis is frequency that the color difference was placed in that position or rank across all participants. The color difference identifier is to the left of each chart while an approximate rendering of the color difference pairs is to the right.*

These results exhibit a number of noteworthy features. First, all charts show at least one peak and one or more positions of zero frequency, indicating that observers were able to differentiate between the pairs of color differences. If all of the color differences were equal then the results would be a roughly uniform distribution for all color differences. This is not the case for these blocks of data. Qualitatively the smallest and largest color differences have the most well defined peaks. A second noteworthy feature lies in the intermediate area charts. These distributions show multiple peaks. That is there are color difference pairs that in some cases were consistently sorted as having both smaller and larger color differences. This might be related to the choice of color attribute used for sorting by the observers. Additional data and analysis are required before making more specific hypotheses but these results show that these differences can be sorted by the participants and that the result of the sorting is complex.

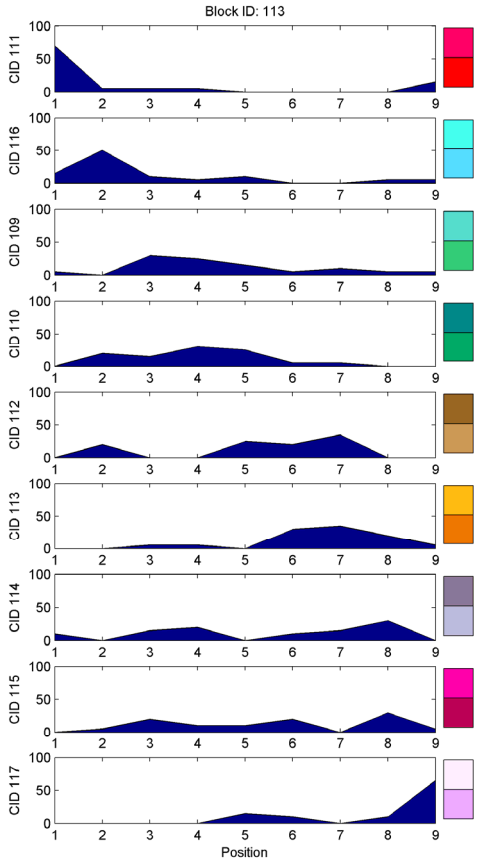

Figure 5. *Additional area charts for a second block of color differences. Format is identical to Figure 4.*

Finally, some explicit clarifications are required regarding these charts. The area charts are shown connected and the x-axis is continuous although strictly speaking this axis is a ranking and should be shown as discontinuous bars. However this representation should be taken as closer to a means to visualize the data and not an indication of the actual analysis. Likewise, the x-axis is actually a block specific ordinal scale. That is while an automatic processing tools allows all blocks of color differences to be rendered as a column of area charts, each block of data may

vary based on the range and distribution of the member color pairs. More complex analysis will be discussed briefly in the discussion section. The final feature to note is a consistent trend concerning color difference pairs sorted to be smallest and largest. For the smallest differences across the different blocks, the color pair consists of two relatively similar colors are members of a single color category, such as two shades of "red". For the largest differences, the color pair consisted of two colors from two different categories, such as an "pink" and a "purple" or a "light" color and a "dark" color.

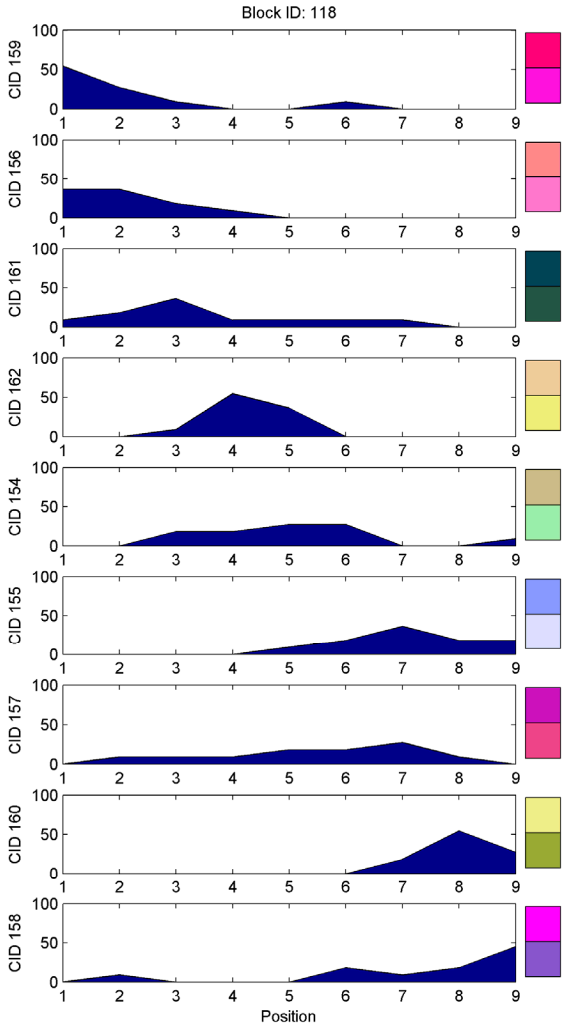

Figure 6. *Additional area charts for a third block of color differences. Format is identical to Figure 4.*

Color Similarity Measure

Given a consistent increase in the magnitude of the sorted color difference for color difference pairs crossing a categorical boundary it is reasonable to consider an alternative to weighted geometric difference schemes. We proposed a cosine similarity measure based on a categorical vector derived from the input color. This section describes how to compute this color similarity measure.

The first step is to select a vocabulary of color terms for processing. In this example we use red, green, yellow, blue, purple, pink, brown, orange, black, white and gray. This vocabulary is used to query a large scale color naming database, in this case the Munroe and Ellis database.[12] Given a subset of training data based upon only the selected vocabulary, the next step is a machine color naming step. There are multiple algorithms with many trade-offs but here we use the k-nearest neighbors algorithm.[13] For any given input color, this algorithm selects the k-nearest neighbors in the database to perform classification. For our implementation we used a k of 50 and the input color space was CIELAB values derived from assuming an sRGB display for the RGB values of the database. Note that unlike a conventional classification scheme where the k-nearest neighbors is used to classify input to a single classification, we use this algorithm to construct a categorical vector representation of the input color. This categorical vector has zero values if there were no instances of that color term being used to describe the input CIELAB value. Likewise the vector can have a maximum of k or the input was only described using that color term. The color similarity measure is finally computed as the cosine similarity of the categorical vectors for two colors. Figure 6 shows an example of how the k-nearest neighbor algorithm can be used to color name a single slice of CIELAB data at a constant lightness. The input colors, shown on the left as circular points on a grid, are converted to a single color coded color term on the left. A useful property of the k-nearest neighbors is that it can be used to extrapolate from the given color data over the entire input CIELAB range even if those values are out of the assumed sRGB gamut. Table 1 shows a worked tabular example for two different color pairs, in this case the smallest and largest ranked pairs from the last or leftmost block of results shown in Figure 6. The final step in the computation is:

$$S = \frac{\sum A \cdot B}{\sqrt{\sum A^2 \cdot \sum B^2}} \quad \textbf{(1)}$$

Where A is the first categorical vector and B is the second categorical vector and the sums are computed across the size of the vocabulary, in this case 11. Cosine similarity is frequently used in document retrieval[14] but we propose that in combination with a machine color naming algorithm, it can be applied to color. The cosine similarity values range from 1 to completely similar to 0 for completely dissimilar. A difference measure can be computed as $\Delta S = 1-S$.

Computing the color similarity for specific color pairs results in a quantification of within or across categorical differences. For example the top two or smallest difference pairs for Figure 3 have color similarities of 0.93 and 0.98 while the bottom two or largest difference pairs ("yellow/green" and "brown/gray") have color similarities of 0.16 and 0.18. The specific values will vary based on the k-nearest neighbor processing and the optimization of this step is part of an ongoing effort.

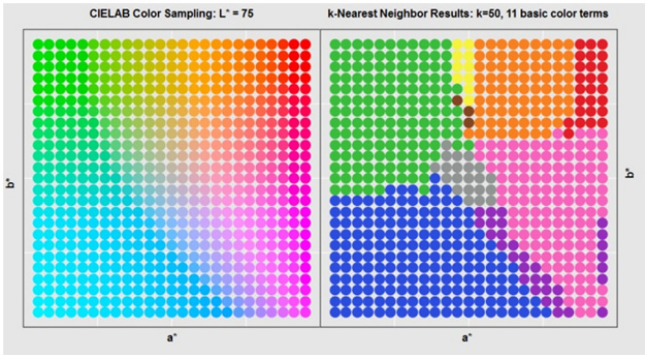

Figure 6. *A visualization of k-nearest neighbors as applied to a single input slice of constant lightness CIELAB data, shown as circular points on a grid on the left. A k of 50 is used in combination with a subset of the Munroe and Ellis color naming database, limited to the eleven basic color terms. The right plot shows the corresponding classifications to a single color coded term. For example "green" is used in the upper left while "blue is applied to the lower left.*

Description	$Pink_1$	$Pink_2$	Pink	Purple
sRGB	253, 1, 121	243, 16, 215	252, 14, 244	132, 84, 200
Black	0	0	0	0
Blue	0	1	0	1
Brown	0	0	0	0
Gray	0	0	0	0
Green	0	0	0	0
Orange	0	0	0	0
Pink	49	45	42	1
Purple	1	4	8	48
Red	0	0	0	0
White	0	0	0	0
Yellow	0	0	0	0
Similarity	0.997		0.207	

Table 1. *Tabular representation of the color similarity calculation for two pairs of colors. The four columns to the left list four different colors with an approximate description and sRGB values. Rows 3 through 13 are the categorical vector as computed using k-nearest neighbors, in this case k=50. The similarity as computed using equation 1 is shown in the last row along with a color representation of the color difference pair.*

Discussion

For the experimental results, it is possible to compute average ranks for each block. That is, the color pair ranking is computed based on the mean rank for the specific of the 18 blocks that the data was presented in. Using averaged rankings per block, the laboratory and web-based versions had a coefficient of determination of 0.60. In addition, sorting the color pairs by average rank shows a large spread in ranks from low average ranks appearing quite similar to high average ranks for colors appearing quite different. For example, both the web-based and laboratory experiments have the pairs with colors that might be described as "green" and "yellow" or "blue" and "purple" has having the largest average ranks. More sophisticated analysis could be used but subjectively this is an initial comparison of the two experiments.

One substantial topic is the derivation of interval scales from the ordinal sorting data, such as using a rank-ordered logit model.[15] A second significant topic is a deeper analysis and testing of the color pairs within blocks that do not appear to be uni-modal. These multi-modal distributions imply large color differences with a magnitude dependent on the sorting attributes weighted. Finally the color similarity measure described in the previous section is one formulation of many possible measures. These measures will differ in the size of the underlying vocabulary and the algorithm used to derive the categorical vector. As more color terms are added the categorical vector becomes larger but always sum to k. Interestingly, even for what is effectively a dimensionality expansion from 3 input dimensions to 11 dimensions in this case, for any given color a much smaller sub-set of terms is used (only 2-3 terms occur for the example shown in Figure 5). The color similarity measure is based on an extensive database of color terms, and as such is a data-intensive model of the color categorization process. However given the complexities of the boundaries shown in Figure 4 this may be the most tractable approach to modeling categorical differences.

A final discussion point relating to the potentially multi-modal distributions is the potential of improving geometric differences metrics by better optimizing the different lightness, chroma and hue weightings. This is not a current focus but publication of the raw color sorting data[16] with this paper allow alternative approaches to modeling large color differences or refinements of the basic similarity measure.

Conclusions

One hundred and sixty two 20 ΔE_{00} color pairs were sorted by color difference on the World Wide Web by 285 observers. They were also sorted by 12 observers in a laboratory and a single sRGB display. A random walk, farthest-point sampling was used to sample colors within the sRGB gamut. A color difference sorting task was used with 18 blocks of 9 pairs of color differences. Within category color differences were sorted to have the least differences while across category color differences were sorted to have the largest differences. A color similarity measure was introduced based on the cosine similarity of categorical vectors representing the colors. The k-nearest neighbors algorithm was used to compute the categorical vectors for an 11 term vocabulary. Example computations were described as well as discussion of the additional work to be performed.

References

[1] Robertson, Alan R., "Historical development of CIE recommended color difference equations", *Color Research & Application,* v. **15** pp. 167–170 (1990).

[2] CIE. Technical report: Industrial colour-difference evaluation. CIE Publication No. 116. Vienna: Central Bureau of the CIE; (1995).

[3] Luo MR, Cui G, Rigg B., "The development of the CIE 2000 colour-difference formula: CIEDE2000", *Color Research & Application,* v. **26** pp. 340 –350 (2001).

[4] Xu H, Yaguchi H, Shioiri S., "Correlation between visual and colorimetric scales ranging from threshold to large color difference", *Color Research & Application*, v. **27**, pp. 349–359 (2002).

[5] R.S. Berns and B. Hou, "RIT-DuPont Supra-Threshold Color-Tolerance Individual Color-Difference Pair Dataset", *Color Research and Application*, v. **35**, pp. 274-283 (2009).

[6] Z. Wang and H. Xu, "Invesitgations of suprathreshold color-difference tolerances with different visual scales and different

perceptual correlates using CRT colors", *J. Opt. Soc. Am. A*., v. **25** pp. 2908-2917 (2008).

[7] Pointer, M. R., and G. G. Attridge. "Some aspects of the visual scaling of large colour differences." *Color Research & Application* 22.5 (1997): 298-307

[8] Guan, Shing–Sheng, and M. Ronnier Luo. "A colour-difference formula for assessing large colour differences." *Color Research & Application* 24.5 (1999): 344-355.

[9] C. Witzel and K.R. Gegenfurtner, "Categorical sensitivity to color Differences", *Journal of Vision*, v. **13**, http://www.journalofvision.org/content/13/7/1 (2013).

[10] Bird, Chris M., et al. "Categorical encoding of color in the brain." *Proceedings of the National Academy of Sciences* 111.12 (2014): 4590-4595.

[11] Eldar, Yuval, et al. "The farthest point strategy for progressive image sampling." *Image Processing, IEEE Transactions on* 6.9 (1997): 1305-1315.

[12] Moroney, N. M., and G. B. Beretta. "Validating large-scale lexical color resources." *Midterm Meeting of the International Colour Association (AIC)*, (2011)..

[13] Rajaraman, Anand, and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2012.

[14] C. D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press. (2008).

[15] Allison, Paul D., and Nicholas A. Christakis. "Logit models for sets of ranked items." *Sociological methodology* 24.1994 (1994): 199-228.

[16] http://inventoland.net/data/published/color_difference_sorting.html