Reverse-Engineering Computational Models for Human Visual Perception using Crowdsourcing

Karan Goel, 2011EE50555

November 18, 2015

Abstract

Crowdsourcing is a popular data-collection paradigm, exploited in particular to collect data for computer vision applications such as segmentation and object detection. In this paper, we explore the idea of using crowdsourcing to carry out research for color perception in humans. An interesting aspect of this is trying to incorporate human models of perception in understanding the kinds of perception mistakes made by humans, as well as generalities in cognitive human behavior when assigned a vision task. In particular, we describe past research in color perception and highlight the challenges faced by color scientists. We also describe past work in crowdsourcing and perception, and highlight lessons that can allow vision researchers to carry out experiments on crowdsourcing platforms. We develop a simple proof-of-concept model for the task of color constancy in humans, using a crowdsourced dataset.

1 Introduction

At the outset, the objective of this paper is to demonstrate the utility of using crowdsourcing to facilitate the development of computational models for the human visual cortex. The traditional method of research on the visual system is to carry out experiments with a limited number of subjects in a highly specialized, laboratory setting. This paper aims to advocate an alternate methodology for this setting; in particular, we will focus our attention on the problem of developing a computational model of color perception in humans. In particular, this paper is organized in the following way:

- In Section 2, we describe color vision, and highlight the challenges faced by color scientists in developing perception models. Most of these challenges stem from lack of empirical data.
- In Section 3, we describe prior research in crowdsourcing, as well as describe in which this research can be extended to our task of collecting data for color perception.
- In Section 4, we carry out experiments using the Cornell Intrinsic Images Dataset. These experiments lead us to develop a simple perceptual model for color constancy for an average human observer.

2 Color Perception

This section lays out the various aspects of color perception that will be relevant to our discussion. In particular, we look at color vision in humans, and discuss the various ways in which researchers have developed computational models for color, and the challenges faced by them.

2.1 Color Vision in Humans

Color vision is the ability of living organisms to distinguish and perceive different wavelengths of monochromatic light. Our understanding of color vision has evolved as multiple scientific disciplines have contributed to this grasp of color perception. Here, we will restrict ourselves to a discussion of the current scientific understanding of color perception.

Color perception in humans is a subjective phenomena, and visual processing of color happens via the presence of *cone cells* in the fovea centralis. The Young-Helmholtz theory first proposed the idea of trichromatic color vision - the idea that humans have photoreceptors for different wavelengths of light. The theory hypothesized the presence of 3 kinds of photoreceptor cells - sensitive to short (S, $420 - 440$nm), medium (M, $530 - 540$nm) and long (L, $560 - 580$nm) wavelengths. This sensitivity was empirically demonstrated in 1956, and is a widely established fact at this point. (Gouras [2009])

Another important theory for color perception is the opponent process theory, which postulates that a color is perceived based on how it contrasts with an *opposing* color. It is conjectured that the visual cortex processes yellow in opposition to blue, and red in opposition to green. This process of using color contrast to determine the color in question, has also been shown to be a valid processing step in color perception by the human visual system. (Gouras [2009])

Kuehni [2008] suggests that the current state of our understanding of color perception is unsatisfactory, in part due to the lack of a large dataset based on some common standard. Most studies carried out by researchers are not identical in terms of the conditions under which they are carried out, which causes a headache in creating, and comparing color models. The idea of using crowdsourcing for this sort of data collection can contribute to bridging this gap – online data collection is generally well documented, and the methodology is easy to share in the form of an online application.

Before moving on to talk about computational models for color perception, we will lay out a set of terms that are necessary for the rest of this discussion.

Chromaticity is the brightness independent quality of a color, composed of 2 parts – *hue* and *purity/saturation*. Hues are the different monochromatic wavelengths of light that the human eye can perceive – the unique hues are considered to be red, blue, green and yellow, although this can vary depending on the particular color model. (Kaiser and Boynton [1996]) The purity of a color on the other hand, is linked to the shade of a particular hue and its distance from the white point in the color space. Examples of different purities are light blue, azure blue, etc. The *luminance* of a color is its perceived brightness, while the *lightness* of a color is the ratio of its brightness to that of a white point with the same luminance.

2.2 Color Spaces

A color space is an organization of color in some specific fashion, where distances between different points in the color space reflect some knowledge of human perceptual sensitivity. In terms of formulation of color spaces, the aim is to come up with a *perceptually uniform* color space, that mimics how humans perceive colors. Distances in the color space should adequately reflect this sensitivity. (Ford and Roberts [1998])

An early, and canonical attempt to formulate a color space came in the form of the CIE-1931 XYZ color space. To understand this color space, we will first talk about the set of experiments that lead to its development. These experiments were carried out by Wright and Guild in the 1920s – they used 17 subjects in total, and asked them to mix 3 fixed-wavelength primary colors – red, blue, green – to match the monochromatic color shown to them on a screen. The screen size was fixed to have an angular diameter of 2°, which was the assumed diameter of the human fovea, where most of the cone cells are concentrated. (Fairman et al. [1997]) The experimental setup is shown in Figure 1.[1]

The data collected in this study was used to define a set of RGB *color-matching* functions, which in turn were used to construct the CIE-XYZ color space. The reason behind this is that as stated above, the experiment was carried out with RGB primary colors of fixed wavelengths. This data was then extrapolated to come up with the amount of each color that needs to be matched to generate *any* wavelength of one's choice. However, the disadvantage of this RGB space is that it depended on the values of the actual RGB colors used in the experiment. Thus, to generalize this space, the CIE-XYZ space was created, where Y was chosen to be the luminance of the color (calculated as a linear combination of R, G and B). (Kerr [2010])

In addition, all of this work assumes the existence of a so-called *standard observer*, a world-average observer for whom the CIE-XYZ color space was assumed to be perceptually uniform. The standard observer's perception is described in terms of the RGB color matching functions, described above.

Moving forward, the CIE-XYZ space was used to construct a non-additive space called the the xyY space (where Y is the luminance of the color in the XYZ space). The x and y coordinates in this space define the chromaticity of the color, and this is shown in Figure 2 for a fixed luminance. (Kerr [2010])

Note that the experiments carried out by Wright and Guild required some very specific conditions to hold. The observer was made to change colors on a bipartite screen that projects color on a fixed angular diameter of the eye – al-

[1] Courtesy of Genevieve Orr.

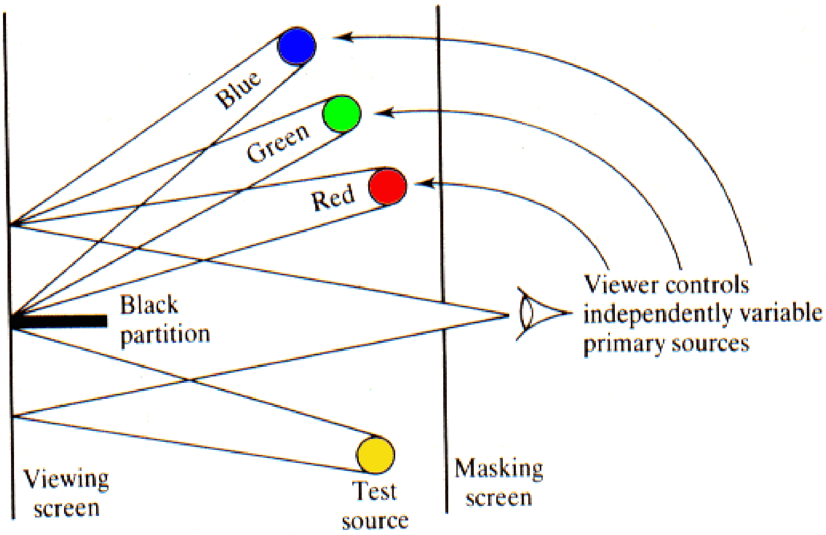


Figure 1: The experimental setup used in the original experiments of Wright and Guild.

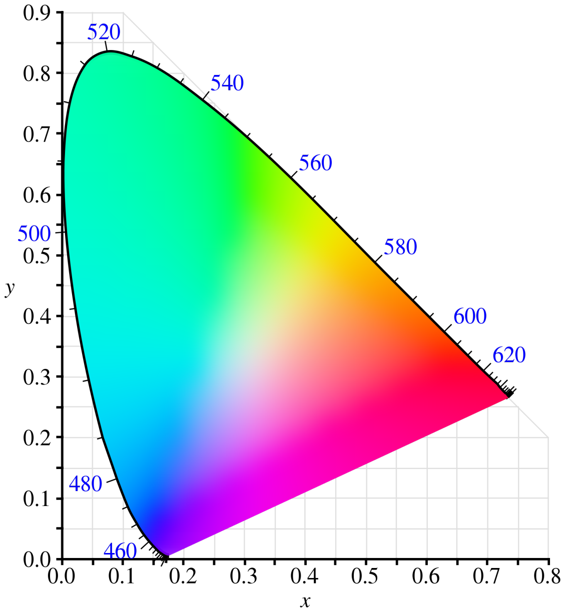


Figure 2: The canonical CIE-XYZ color space, shown as the x-y plane of the xyY color space. The colors on the boundary are pure spectral colors (wavelengths shown). The white point is around the center of the figure (not shown). Courtesy of Wikimedia Commons.

though it has now been shown that the fovea has an angular diameter of 10°. In addition, the surroundings were carefully calibrated to ensure that illumination or other reflective effects did not make the results noisy. It will be impossible to address some of these concerns in a crowdsourced setting, since we have no access to the subjects who are providing us experimental data, nor access to their environment. However, it is true that the entire perceptual modeling of the CIE-XYZ space depended on the empirical data collected. (Kuehni [2008])

The CIE-XYZ color space has since been followed by the 1976 CIE-LUV (L*, u*, v*) color space, which was intended to address some problems in the perceptual uniformity of the CIE-XYZ space. In particular, MacAdam showed in experiments that we can identify ellipses in the CIE-XYZ space, where the chromaticities within the ellipse are indistinguishable to the human eye. (Robertson [1977]) The CIELUV color space is simply a non-linear transformation of the CIE-XYZ space to rectify these ellipses (to make them circular). (Parkyz et al. [1999], Wyszecki and Fielder [1971])

This leads us to the final color space of interest; the CIE L*a*b* space, which is once again, an attempt at more perceptually uniform space. The CIE L*a*b* is similar in spirit to the CIE-LUV space; L* corresponds to the lightness of the color, a* is the difference between its green and red components, while b* is the difference between its yellow and blue components. We will not discuss the more recent color spaces, due to their increased complexity; they also do not deviate significantly from this approach, aside from performing corrections to specific parts of the color space, based on new evidence collected over the years.

3 Crowdsourcing Perception Experiments

So far, we have discussed traditional approaches for collecting data and creating computational models for color perception. An important observation to reiterate is that much of the work has relied on experimental data, and the difficulty of collecting this data has made it hard to truly capture human color perception in a widespread sense.

This section will begin by describing some of the ways in which crowdsourcing has been used in prior work for collecting data on human perception. We will then make a case for the use of crowdsourcing for understanding color perception, and extend this to general vision tasks.

3.1 Prior Work on Gauging Human Perception

To understand the potential of crowdsourcing for our color task, we will first look at some of the ways in which crowdsourcing has been used by researchers in collecting data for visual perception tasks.

3.1.1 Study 1: Graphical Perception

Heer and Bostock [2010] focus on the task of using crowds to determine perception based on a variety of graphical stimuli. Their work aims to first replicate the results of an older study carried out in a formal setting, to demonstrate the viability of using crowdsourced data, after which they demonstrate the viability of this approach to draw new conclusions.

In their setting, users were given different kinds of charts, and then asked to give quantitative judgments on the data that was encoded in those charts. For example, users were asked to judge the percentage difference between 2 points in a bar graph. The visual encoding was changed across experiments by varying the position, shape, size, etc. of the charts to judge how users respond to changes in these variables.

The authors were able to collect 3400+ judgments for this dataset, using 50 subjects for every single visual stimuli that they generated. The use of 50 subjects is nearly unheard of in laboratory studies. They found that their results matched and extended earlier work, and the quality of the data collected by them *did not suffer*, despite having little control over what subjects did, beyond the quality control mechanisms that they introduced.

Another task they focus on is to replicate an alpha contrast experiment carried out by earlier work (once again in a laboratory setting). Users were asked to vary the background luminance contrast on some scatter plots, and their behavior was observed on changing the density of the scatter plots. This experiment is especially valuable since both monitor display settings, and environmental illumination had a potential effect on results. This is similar to the case of color perception, where these factors are likely to effect the outcome of the experiment. However, they were able to also collect information about the user's monitor settings, and found that the information could potentially be used to debias the results.

In total, they collected 1400+ judgments, and found that their results actually *generalize better* than laboratory experiments – earlier work had reported unexpected results that went against common wisdom, which their work was able to rectify.

In both cases, they make use of quality control schemes, that allow them to check if (a) subjects were serious participants in their study; (b) subjects were able to come up with responses to very obvious tasks correctly; (c) subjects had display settings that would allow them to participate in their study in a (mostly) non-noisy fashion.

Kong et al. [2010] extend this work, by considering the case of treemap visualizations, and come up with a set of perceptual guidelines for constructing these visualizations, while also taking into account viewer ease of use. In particular, they come up with thresholds on rectangular aspect ratios for these visualizations. The flexibility of crowdsourced experimentation allows them to conduct their experiments by varying several independent aspects of the visualizations, and gauge the response from users. This degree of flexibility is almost impossible in a non-crowdsourced setting.

3.1.2 Study 2: 3D Shape Perception using Line Drawings

Unlike color perception, the perception of shape is a visual task that is similar across different viewers. However, Cole et al. [2009] consider an interesting cognitive task – determining how well 2D line drawings can depict 3D shapes. They ask subjects to place surface normals (perpendicular lines) on the 2D drawing provided to them; these normals allow them to judge whether the subject is able to correctly identify the contours of the 3D shape. In all, they collect 275,000 judgments and analyze them to understand the reasons that subjects make mistakes.

Using this data, they are able to isolate localized contours that create confusion, and they provide recommendations on how to better depict 3D shapes using line drawings in a minimal fashion, while still retaining lines that provide key perception guides to viewers.

3.1.3 Study 3: Color Naming

Munroe [2010] collected the largest color naming dataset using crowdsourcing, containing over 3 million responses. These colors were uniformly sampled from the 3D RGB color cube, and respondents were unconstrained. Heer and Stone [2012] make use of this dataset to come up with a probabilistic model for color naming, also creating distance metrics that predict how similar 2 colors are, as well as modeling saliency of colors in the CIE L*a*b* color space.

Their paper demonstrates an interesting use of crowdsourcing to understand the relationship between visual perception and linguistic expression. For instance, they show that the saliency between green and blue is low, with a great deal of confusion in naming different shades of these colors. High saliency indicates that a color is perceived similarly by most subjects, and also indicates that subjects are less likely to resort to different names, even if multiple shades of that hue were shown to her. This in turn indicates lower discerning power for that particular hue and its different shades.

A notable problem with the dataset, pointed out in their paper, is that it was collected by considering each color swatch against a white background, which may bias the results towards that background luminance. The authors suggest varying this background to other shades of grey, as well as black, so that models are robust to these changes in background luminance. Their suggestion demonstrates a key point in designing visual experiments – if the experiments are in an uncontrolled setting, then noise introducing factors must be varied, to make sure that they don't bias the results.

Takeaway: Crowdsourcing, while not able to replicate the laboratory setting perfectly, generates orders of magnitude more data at much lower cost, with potentially more variation, but far less control. In Study 1, it ended up replicating and extending older results that were earlier carried out in a more controlled setting. In all studies, the study had access to a wider population, were able to add more variation to their study. Problems in response quality were made up for by the extra data that was collected. In addition, the papers found that the *ease and speed of carrying out these tasks online greatly decreased experimenter burden.*

3.2 Quality Mechanisms and Interface Issues

Heer and Bostock [2010] in Study 1 above describe the importance of quality mechanisms when performing crowdsourcing tasks. Kosara and Ziemkiewicz [2010] also discuss how the quality of responses for perceptual tasks can be improved on crowdsourcing platforms; the important mechanisms are summarized below

- *Pre-Screening*: Subjects can be pre-screened by giving out custom qualifications to those workers who have contributed to some study in the past. This is useful in case online studies of this kind need to be carried out regularly. Only those workers that meet the qualification requirement are admitted as subjects to the study.

- *Instructions*: An important factor in ensuring good quality work is to provide clear instructions for the perceptual task at hand. In particular, it is essential that the instructions do not bias subjects towards any particular kind of stimulus or a response for a stimulus that is presented to them. For instance, suppose we are designing a color matching experiment. In this case, we must clearly specify that the worker needs to mix some primary colors to create the color that is provided to them. Using instructions, we can also specify/request subjects to change their monitor settings to some standard, if we would like to ensure that.

- *Initial Screening*: Screening subjects allows us to filter those that may be spammers, or may not be equipped to participate in the study. A good example of this is that most studies on color may want to isolate/prevent color-blind subjects from participating. In addition, it is common on crowdsourcing platforms to ask workers to perform some sample tasks, to judge their quality against some ground-truth answers. This can be useful to ensure that only the highest quality subjects are allowed to participate in the study. To ensure fair testing, the tests must not have variability of perception *i.e.* subjects should not be rejected because of their differing perception of the task, rather the task should be simple enough that everyone would give the same response to it. For instance, workers can be asked if 2 identical color swatches are different, if 2 (very) different color swatches are similar.

- *Calibration*: Given the observations from prior work, it seems that a good necessary step for these experiments is a calibration stage. The idea of calibration would be to (a) normalize the subject's environment by encouraging her to make changes to her visual settings; (b) ask the subject to answer some calibration tasks, that allow us to understand the effect that her screen settings may have on her responses. Computer monitors are unable to represent the full gamut of human color perception, and may do so in different ways. A color may be seen differently depending on how the screen is configured. Getting responses to calibration tasks allows us to quantify these differences and use them to denoise our results.

- *Task Presentation*: Task presentation involves presenting the task to the subject in a way that is non-ambiguous and allows the user to answer the questions quickly and accurately. Importantly, prior work has found that crowdsourced subjects should not be given extremely long tasks, since they may abandon them in the middle, leading to a wastage of data. Instead, it may be necessary to split the questions/tasks that need to be asked, so that subjects are given a short set of questions in a single task. Subjects can of course attempt multiple tasks, but abandoning them midway will not cause significant disruption to the study in that case.

- *Post-hoc Processing*: Post-hoc processing of data is required to ensure that the quality of the subjects' responses were good. Unlike in a lab setting, where subjects are constantly supervised, online subjects are left to their own devices, and may therefore suffer from fatigue or boredom while carrying out the study. In this case, it is necessary to identify these subjects, and perhaps remove their data, to ensure integrity of results. Common ways in which quality is ensured is to look at all worker responses to a particular task (there will typically be multiple) and remove those responses that deviate wildly from the median.

3.3 Meeting Suggested Guidelines

Studies on the nature of human visual perception (McLeod [2014]) generally adhere to strict conventions for data collection. There are several aspects to data collection that are important in this regard: (a) the data must be collected in a calibrated environment, without factors that may perturb results; (b) the subjects must meet some qualifying criteria.

The CIE established a set of guidelines for carrying out color perception experiments in 1976, titled *CIE Guidelines for Coordinated Research on Color-Difference Evaluation* (Robertson [1978]). They provide a 4 step process which any study should attempt to follow. It is important to analyze whether these guidelines can be followed in a crowdsourced setting, and if so, to what extent. We analyze the first two steps below using suggestions from McLeod [2014],

- *Step 1:* The first step suggests different methodologies of data collection that can be utilized by the researcher. Among these are the *Rich-Billmeyer-Howe* method, where pairwise similarity of color swatches is tested. Each pair is tested multiple times on the same subject to ensure that the colors are consistently matched (or not). Another method, the *constant color difference* comparison compares a color pair with a standard greyscale pair, and asks which pair has a greater color difference. To adapt this sort of testing to crowdsourcing, McLeod [2014] suggests using triplet comparisons, where a reference color is compared to 2 other colors and the subject is asked to choose the color most similar to the reference. This is most similar to the *visual scaling* method suggested in the CIE guidelines. As long as the colors being tested fall within the gamut of the subject's device, this testing can be carried out without any issues.

- *Step 2:* The second step suggests that the study be conducted by varying different parameters, which may influence the results – sample size, sample separation in color space, monocular or binocular observation, illumination level of the color swatch, surrounding illumination, etc. Notice that some of these parameters may be hard to test for, given that we do not have control over the subjects. For instance, it is impossible for us to ensure that subjects carry out the test with monocular or binocular vision. To test the effect of illumination, we can vary the brightness of the screen. We can also potentially make use of the phone's light sensor to measure the environment lighting, and use that to measure effect on the subject's perception.

Note that while this section focused on the guidelines established for the specific case of color perception, most visual perception tasks will require some set of conditions that must be satisfied in order for the data collection to be considered useful. These conditions must be encoded into the interface, and any other issues that come up must also be addressed.

4 Some Experiments with Data

In this section, we will look at a proof-of-concept model for the task of color constancy (described below), using a crowdsourced dataset.

4.1 Dataset Description

The Cornell Intrinsic Images dataset Bell et al. [2013] is a very large dataset collected from crowdsourced workers by researchers at Cornell University. The dataset was established with the objective of generating intrinsic image decompositions of images provided by Flickr users. Workers were asked to look at a pair of points in an image, and judge which was of a darker color. The perception of workers mattered greatly in this experiment, since workers were not asked to judge the RGB value of the pixel; rather, they were asked to ignore lighting/shadow and other illumination, and judge with respect to the underlying color of the surface. They were allowed to specify the pair of points to be equal if the points were too ambiguous. In addition, they were also asked to provide a level of confidence for their answer.

The concept of *color constancy* relates to exactly this notion of human perception – colors are recognized as being identical even under vastly differing illumination. The set of experiments below aims to test this claim of color constancy by making use of the dataset above. In particular, we will answer the following,

- Does color constancy hold even when the color pairs being compared is very dissimilar in terms of their values?
- Can we come up with a proof-of-concept model for color constancy using this data?
- What other possible inferences can be drawn from this data?

This dataset is also interesting because it provides researchers the flexibility to assess responses to real-world situations. All the data is collected on real images, and is therefore likely to elicit responses similar to how a subject would respond if he was actually in the scene. This of course is contingent on the fact that the image is able to sufficiently capture coloring distinctions in the scene.

4.2 Analysis

The first question we would like to address is whether there exists a threshold for color constancy; is there a point after which it becomes difficult for subjects to identify that 2 colors are in fact identical (only under different levels of illumination). The intuitive suspicion is that as the pair of points become increasingly different in terms of their luminance, the ease of carrying out this comparison should lessen. To carry out this analysis, we will consider only those judgements for which the established answer is that they are color constant.

For this particular set of judgements, we plot a normalized histogram as follows. For each pair of points considered, we calculate their Euclidean distance under the sRGB color model. Note that this calculation is not perceptually uniform, since the sRGB system is not perceptually uniform. However, it should suffice as a heuristic. We then bin the number of times that a worker was able to correctly identify that the pair of points are identical in color, normalized by the total number of responses in that bin.

The Euclidean distance is simply,

$$d(c_1, c_2) = \sqrt{(R_1 - R_2)^2 + (G_1 - G_2)^2 + (B_1 - B_2)^2}$$

The plot is shown in Figure 3. In Figure 3a, the number of comparisons that were used to generate this plot are 2,899,016(!) across 5230 images, a mind-boggling number that cannot possibly be collected in *any* lab-setting. Figure 3a confirms our hypothesis; bins with large distances tend to have much lower probability of correct identification of identical colors.

We can also see that a similar hypothesis is confirmed by Figure 3b, although the values seem to deviate wildly after a distance of 150. This is because of the lack of enough data for distances beyond that value, so the values vary widely, as confirmed by Figure 3c.

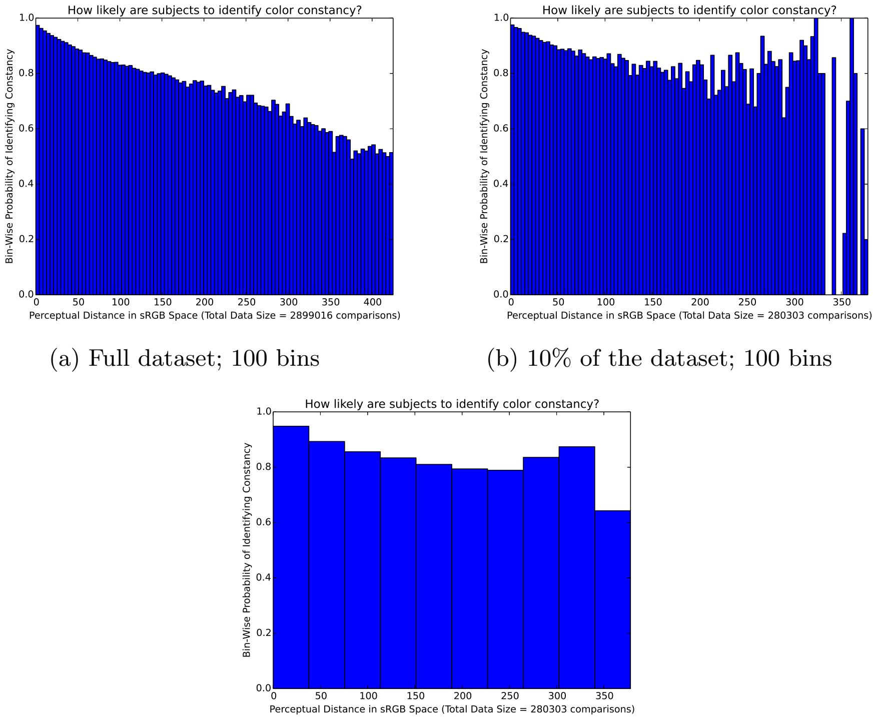
The next analysis we will attempt is to try to fit a simple model to this data. Let C be a random variable that is 1 when a subject correctly identifies a pair as color-constant, and 0 if not. Ideally, we would like to have a model of the following form,

$$p(C = 1|c_1, c_2) = f(d(c_1, c_2))$$

where $d(c_1, c_2)$ is some distance metric, in the color space where c_1 and c_2 are defined.

We first transform the sRGB annotation of each point to the perceptually uniform CIE L*a*b* space, since distances in this space will more accurately reflect human perception. This transformation requires us to first map the sRGB point to the XYZ space and then to the L*a*b* space.

The transformation is,



(a) Full dataset; 100 bins

(b) 10% of the dataset; 100 bins

(c) 10% of the dataset; 10 bins

Figure 3: Histograms which demonstrate how the likelihood of a worker correctly identifying color-constancy, falls with increasing perceptual distance between the points in the sRGB space.

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.4124 & 0.3576 & 0.1805 \\ 0.2126 & 0.7152 & 0.0722 \\ 0.0193 & 0.1192 & 0.9505 \end{bmatrix} \begin{bmatrix} R_l \\ G_l \\ B_l \end{bmatrix}$$

where R_l, G_l, B_l are a linearized version of the sRGB values (constrained between 0 and 1). The linearization formula is omitted for brevity. We then perform a non-linear transformation to the L*a*b* space. Due to the complexity of the transformation, the details are omitted. As a sanity check, we perform the same histogram binning for the L*a*b* space as we did in the sRGB case. This time, the distance metric that we use is the CIE-1976 distance, which once again corresponds to the Euclidean distance,

$$d(c_1, c_2) = \sqrt{(L_1^* - L_2^*)^2 + (a_1^* - a_2^*)^2 + (b_1^* - b_2^*)^2}$$

Figure 4 shows how the appearance is quite similar to the sRGB case.

Instead of creating bins of equal width, we will instead opt to create bins of equal size (100,000 comparisons each). This will allow us to understand exactly how reliable our model is in each distance interval. Our model generation process will be as follows: we will enforce a uniform prior on each bin by adding 100 points for both $C = 0$ and $C = 1$. We can also use more complicated priors

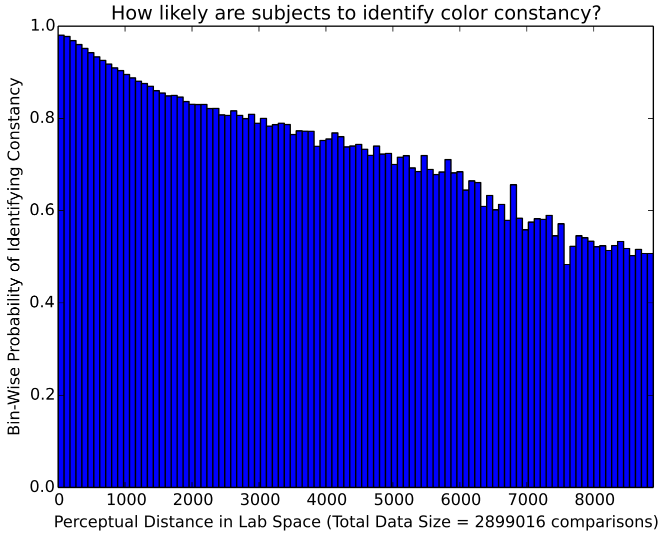


Figure 4: A 100 bin histogram which demonstrates how the likelihood of a worker correctly identifying color-constancy, falls with increasing perceptual distance between the points in the L*a*b* space.

when generating these sorts of models; however, we will keep this simple for the purpose of this discussion. A simple extension would be to have a prior that approaches 0 for large distances, and approaches 1 for small distances. However, using a weak uniform prior also allows us to learn something interesting about the data.

Each bin gives us a posterior probability for that distance interval; we will assume that the mean value of the bin represents it. Figure 5 shows the histogram that is generated for this case. Note how most of the distances (¿ 3000) are truncated due to lack of data. We will now simply perform polynomial regression on this histogram.

Performing regression gives us a quadratic fit (R-squared value = 0.999), with the following equation,

$$p(C = 1|c_1, c_2) = 0.9915 - 1.103 \times 10^{-4} d(c_1, c_2) + 1.619 \times 10^{-8} d(c_1, c_2)^2$$

The fitted regressor is shown in Figure 6.

While this discussion contained a somewhat rudimentary use of the dataset, the data also contains information about how confident each subject feels about his/her answer, as well as information about which color is lighter or darker (ignoring surface reflectance). It is possible, for instance, to adapt a simple Maximum-Likelihood procedure, as outlined by Whitehill et al. [2009] or Welinder and Perona [2010] to estimate each subject's perceptiveness in this task. This can be done by considering that each worker has some underlying perceptiveness parameter, and finding the maximum likelihood estimate of that param-

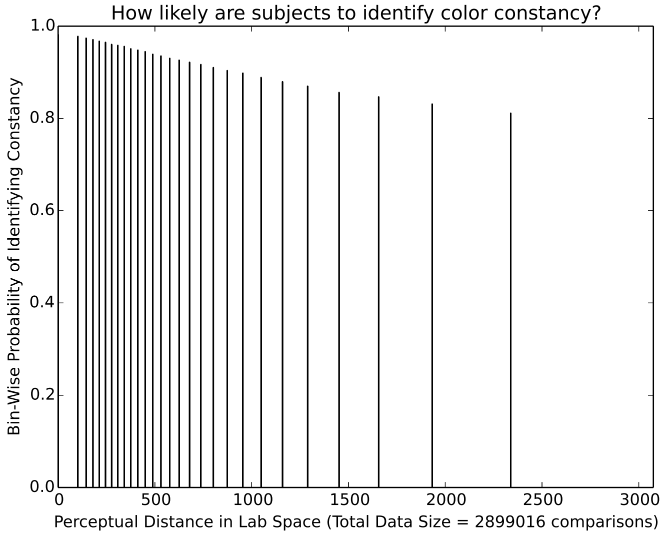


Figure 5: Equally sized histograms (size = 100,000) in the L*a*b* space.

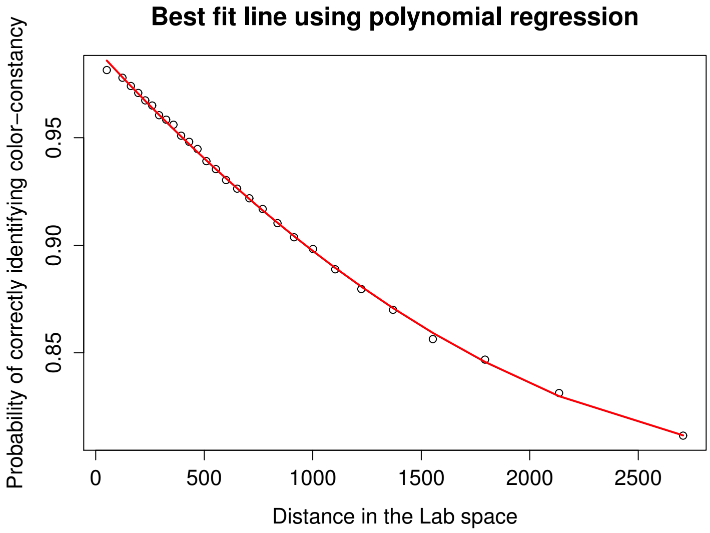


Figure 6: Quadratic regressor fitted to model probability of correct identification of identical colors, as a function of distance in the L*a*b* color space.

eter, under a model that captures the difficulty of discerning color-constancy for some pair of points. This effectively lets us capture observer variability, which is something that has not been possible in laboratory settings (due to the small population size).

5 Extending to General Vision Tasks

So far in this paper, we have focused on the specific area of color perception in humans. We demonstrated the efficacy of crowdsourcing for this task, by considering a novel approach to calculating color-constancy in a truly world-standard observer. This task is a somewhat low-level vision task, since it involves limited use of cognitive processes compared to other high level vision tasks.

Crowdsourcing also allows us to determine human efficiency and aptitude for high level vision tasks. For instance, recent work by Sharma et al. [2015] demonstrated that humans cannot count more than 25 faces in an image without making mistakes – thus, there is clearly some underlying threshold for object counting, which they attempted to explain by a simple model. In addition, they leveraged this observation to improve on the state-of-the-art for face recognition in images. Understanding human visual perception can thus lead to a direct impact on real-world systems in terms of both performance and cost.

Other recent work in vision pointed out that human accuracy for a variety of simple vision tasks is quite different. Humans have different accuracy, when asked to draw a bounding box around some object, or when asked to verify if some drawn box is tight around a given object. Thus each task brings with it a different computational challenge when modeling it. (Russakovsky et al. [2015])

These examples serve to illustrate how crowdsourcing can be extremely valuable in not only modeling a low level vision task, like color perception, but also to arbitrarily increase the cognitive complexity of the vision task, to effectively collect and model high level tasks.

6 Conclusion

This paper has aimed to address the stagnation of research in color perception by means of a novel suggestion – the usage of crowdsourced data in improving data collection, and the generation of computation color perception models. Through the course of this paper, we have identified ways in which crowdsourcing has been of service to researchers interested in understanding human perception in the past, and have identified lessons that can allow us to extend this usage to future work. More importantly, we have seen how this can be extended to the task of color perception. While we had no control over the data that was collected, in terms of the lessons that we outlined (aside from post-hoc processing and the quality control mechanisms used by Bell et al. [2013]), we did show even then, how scaling up data can positively effect the ease of generating computational models.

In addition, we believe that concerns about whether crowdsourcing is appropriate for the collection of this data – which is often restricted to collection in a laboratory setting – are unfounded. While a computer screen is incapable of representing the full gamut of colors that can be perceived by the human eye, we can still determine exactly what colors our subjects can see in the sRGB gamut that is available on monitor screens, and use this information to embed our data in the CIE L*a*b* space. In addition, we believe that it is possible to calibrate users to a large extent in line with CIE guidelines, and that the scalability that crowdsourcing provides can allow us to observe responses to changes in parameter settings, which makes the data collection process both more robust, as well as more representative.

As demonstrated by the experiments carried out, it is possible to come up with a simple computational model for color constancy using crowdsourced data. In fact, as our experiments demonstrated, averaging of the data seems to be taking place, since using 10% of the data does not give us the same quality of results as using the full dataset. However, we do note that there are several confounding factors that were not considered in our analysis, which would be useful to both separate and analyze – for instance, subjects who are asked to judge color for 2 points located on a common surface can do so more easily, since they know that points on a common surface are likely to share the same color. It is important to take note of these issues when creating the interface, as well as generating the data on which judgments are to be collected through crowdsourced subjects. It would also be more fruitful to make use of data where complete information about the scene – in terms of its lighting and luminance components – is known.

Finally, most work in vision and crowdsourcing does not leverage the insights that cognitive scientists have developed while working in their field over the years. As a final conclusion to this paper, a strong recommendation is to incorporate perceptual mechanisms of modeling into these vision computational systems, both with the end-goal of improving these systems, as well as contributing new and improved data to the cause of understanding these perceptual mechanisms, as well as the human visual cortex.

References

Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Opensurfaces: A richly annotated catalog of surface appearance. *ACM Transactions on Graphics (TOG)*, 32(4):111, 2013.

Forrester Cole, Kevin Sanik, Doug DeCarlo, Adam Finkelstein, Thomas Funkhouser, Szymon Rusinkiewicz, and Manish Singh. How well do line drawings depict shape? In *ACM Transactions on Graphics (ToG)*, volume 28, page 28. ACM, 2009.

Hugh S Fairman, Michael H Brill, and Henry Hemmendinger. How the cie 1931 color-matching functions were derived from wright-guild data. *Color Research & Application*, 22(1):11–23, 1997.

Adrian Ford and Alan Roberts. Colour space conversions. *Westminster University, London*, 1998:1–31, 1998.

Peter Gouras. Color vision in humans. http://webvision.med.utah.edu/book/part-vii-color-vision/color-vision/, 2009.

Jeffrey Heer and Michael Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI*

Conference on Human Factors in Computing Systems, pages 203–212. ACM, 2010.

Jeffrey Heer and Maureen Stone. Color naming models for color selection, image editing and palette design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1007–1016. ACM, 2012.

Peter K Kaiser and Robert M Boynton. Human color vision. 1996.

Douglas A Kerr. The cie xyz and xyy color spaces. *Issue*, 2010.

Nicholas Kong, Jeffrey Heer, and Maneesh Agrawala. Perceptual guidelines for creating rectangular treemaps. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):990–998, 2010.

Robert Kosara and Caroline Ziemkiewicz. Do mechanical turks dream of square pie charts? In *Proceedings of the 3rd BELIV'10 Workshop: BEyond time and errors: novel evaLuation methods for Information Visualization*, pages 63–70. ACM, 2010.

Rolf G Kuehni. Color difference formulas: An unsatisfactory state of affairs. *Color Research & Application*, 33(4):324–326, 2008.

Ryan Nathaniel McLeod. A proof of concept for crowdsourcing color perception experiments. 2014.

Randall Munroe. Color survey results. http://blog.xkcd.com/2010/05/03/color-survey-results/, 2010.

Du-Sik Parkyz, Jong-Seung Parky, and Joon Hee Hany. Image indexing using color histogram in the cieluv color space. 1999.

Alan Robertson. Cie guidelines for coordinated research on color-difference evaluation. *Color Research & Application*, 3(3):149–151, 1978.

Alan R Robertson. The cie 1976 color-difference formulae. *Color Research & Application*, 2(1):7–11, 1977.

Olga Russakovsky, Li-Jia Li, and Li Fei-Fei. Best of both worlds: human-machine collaboration for object annotation. In *CVPR*, 2015.

Akash Das Sharma, Ayush Jain, Arnab Nandi, Aditya Parameswaran, and Jennifer Widom. Surpassing humans and computers with jellybean: Crowd-vision-hybrid counting algorithms. In *HCOMP*, 2015.

Peter Welinder and Pietro Perona. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 25–32. IEEE, 2010.

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, pages 2035–2043, 2009.

Günter Wyszecki and GH Fielder. New color-matching ellipses. *JOSA*, 61(9): 1135–1152, 1971.