Selective Color Naming in Complex Natural Images

Gabriella Oudsema

Statement of Work:

As the sole team member, I am responsible for the novel contents of this project (with the exception of survey responses, which were provided by my helpful classmates). References to past studies and previously collected datasets are clearly marked in the report and code.

Introduction

Color naming refers to programmatically turning image data, such as the RGB values of a pixel, into the a color label a person would use to describe that image, such as “blue” or “green” (van de Weijer, Schmid, & Verbeek, 2007). Typically this process involves the use of image data and human input in providing labelled images for training or reference data. Color naming is part of the field of computer vision, but is also closely tied to a subfield of psychology known as psychophysics. The aim of psychophysics is to examine the relationship between external stimuli and people’s internal perceptions; for instance, the relationship between a particular image and the color it is perceived to be (Goldstone & Hendrickson, 2010). Since computer vision is essentially seeking to replicate human vision, there is a great deal of potential for interdisciplinary research work. Although involving human participants for labelling and research brings its own unique ethical concerns, such as anonymity and properly compensating labelers, with careful consideration a psychophysics element can produce novel insights and avenues of study in computer vision.

With color naming in particular, the phenomenon of categorical perception should be considered — essentially, the idea that having categories like color names in mind when looking at an image affects how people perceive it (Goldstone & Hendrickson, 2010). A study by Witzel and Gegenfurtner (2016), for example, looked at the categories of “red” and “brown”, and found participants were more easily able to distinguish small differences in shades from *different* categories (i.e. a shade considered red from a shade considered brown) than shades that *shared* a category (i.e. two differing shades of red or two differing shades or brown). Named categories

draws deeper distinctions for human observers, in a way machines will not necessarily be able to duplicate unless this is deliberately acknowledged during programming.

Another aspect of human perception that complicates computer vision is color constancy: the tendency for people to see an object as being one consistent color even under different lighting conditions. For example, a white flower at sunset might be tinted orange by the light, while the same flower in deep shade will be tinted blue, but because our mind "compensates" for different lighting, a person would still perceive and identify the flower as white. However, such lighting conditions could cause ver different readings for computer vision programs.

An innovation by van de Weijer, Schmid, and Verbeek (2007) in their influential study addresses the issue of variable lighting. Typically, obtaining labelled training data for color naming tasks was done by asking participants to label color chips in a well-lit laboratory setting (van de Weijer, Schmid, & Verbeek, 2007). These authors instead pulled labelled photos from the internet and statistical methods to determine query relevance to develop a reference for color names. Because online photos have more diversity in lighting conditions than the paint chips laboratory conditions, this allows the labels to be more accurate to real-world images, which are also taken in a variety of lighting conditions.

This paper seeks to address a third aspect of human perception and apply it to a computer vision model: judgements of importance. In complex images, there are many elements that a person may not notice or choose not to describe because they seem unimportant or tangential. Machines can analyze every part and pixel of an image, but sometimes more description is not useful. Narrowing down descriptors to a brief, relevant summary is an especially important part of creating an information retrieval system that allows images to be searched for based on their

memorable characteristics. In efforts to make digital collections accessible, Othman, Wook, and Qamar (2020) attempted to use color categories that summarized the overall impression of an image — bright, pastel, dull, pale, or dark — and determine if participants reliably agreed which images belonged to which categories. While bright was relatively clear, participants dis not always have a clear idea of which images should be considered "dull" or "pale" (Othman, Wook, & Qamar, 2020). While there is a merit in providing a category like "bright" that can describe several different colors in an image, the benefit of using established color names is that there is likely a higher level of agreement about what constitutes "red" — even if at times it overlaps with pink or orange.

There are two main goals in this study: first, to determine which — if any — strategies by an automated computer program can correctly replicate the labels of human participants; and second, to examine the most and least successful strategies to see what can be learned about the underpinnings of human perception that led to certain labels being chosen for an image.

Methods

A selection of 20 images depicting flowers were taken from 102 Category Flower Dataset by Maria-Elena Nilsback and Andrew Zisserman (*Visual Geometry Group,* n.d.). Images were selected by hand to ensure a mix of colors and degrees of complexity (here defined as how many colors the image seemed to contain in total), while avoiding images with unusual lighting to limit extraneous variables. The selected images can be viewed in the Appendix.

To obtain human labels, a survey was produced in Google Forms. For question, one photo was presented, and participants were asked to "select **only 2 colors** to describe this image

to someone else who is not looking at the image". Possible answers were the 11 basic color names in English(van de Weijer, Schmid, & Verbeek, 2007): red, orange, yellow, green, blue, purple, pink, white, gray, black, and brown. Participants were masters students enrolled in a data science program at a Midwestern university, and a total of 18 respondents completed the survey. Results were anonymized for data storage. For each photograph, the two most popular color labels were taken to be the "human labels" for comparison. All the color responses associated with a photograph were taken to create a set of "plausibly human" labels for that photo.

Four additional sets of labels were produced programmatically using various strategies. First, images were read using the OpenCV python library to iterate through the pixels of an image. Once RGB values were obtained for a pixel, the values were compared to the reference table produced by van de Weijer, Schmid, and Verbeek (2007). This table consisted of combinations of RGB values and the associated probability that a pixel of that value should be labelled red, blue, black, and so on (see Figure 1). The table's values for R, G, and B channels each started at 3.5 and went up in increments of 8, with the highest possible value being 251.5. To determine which table line should be referenced for a particular pixel of a survey image, a binary search strategy was used for R, G, and B channels to find the nearest value in the reference table, and pandas filtering was used to locate the corresponding row. The color name with the highest probability was assigned to the pixel.

	r	g	b	black	blue	brown	gray	green	orange	pink	purple	red	white	yellow
0	3.5	3.5	3.5	0.293958	0.020885	0.037604	0.071886	0.141474	0.064986	0.058097	0.125870	0.059671	0.006976	0.118593
1	11.5	3.5	3.5	0.330787	0.028917	0.050061	0.072332	0.123315	0.062952	0.056099	0.114169	0.058952	0.005640	0.096776
2	19.5	3.5	3.5	0.375413	0.028445	0.055683	0.072660	0.096563	0.067473	0.054926	0.091975	0.071449	0.005915	0.079498

Figure 1

This process was repeated to get a label for each pixel, and then a percentage of how many pixels in the survey image were labelled with each color. The color names with the top two percentages were then selected to get two labels. This strategy of deriving labels from the image will be referred to as strategy one, or the “whole image” strategy.

The remaining three strategies made use of image segmentation, specifically employing OpenCV’s watershed algorithm to distinguish the foreground and background of an image. With this information, the program looped through an image’s pixels in much the same way as strategy one, but instead of calculating the percentage of the total image labelled by each color, it calculated what percentage of the foreground what percentage of the background were covered by each color, separately. Strategy two, or the “foreground only” strategy, took the two largest percentages in the foreground category and used those to label each picture. Strategy three, the “foreground and background” strategy, took the label with the highest percentage of coverage in the foreground and the label with the highest percentage coverage in the background to apply two labels to each picture. Finally the fourth strategy used “background only” and took the two labels associated with the highest percentages of coverage in the image’s background.

To measure the accuracy of labels, a score out of four was applied to each set of programmatically-generated labels. Each computer label would receive two points for matching

one of the two human labels associated with an image; if both labels generated by a strategy matched one of the human labels, a strategy's score for the image in question would be a full four. If it did not match either of the human labels, a computer label could still get one point if it matched at least one person's response on the survey, essentially indicating the label was "plausibly human". Thus, a score of two would indicate both labels produced by that strategy for that photo were human-like, but did not correspond to the most popular or agreed-upon human answers. A label that did not match the human labels or any of the human answers would receive zero points.

Results

The mean scores for each strategy are shown in Figure 2. The first "whole image" strategy on average scored 2.95, with a median score of 3. The second "foreground only" strategy had a mean score of 2.8 and a median score of 3. The third "foreground and background" strategy had an average score of 2.95 and a median score of 3. The fourth "background only" strategy has a mean score of 2.5 and a median score of 3. For each strategy's distribution of scores, the minimum was 0, and the maximum was 4.

Figure 2

Discussion

On average, strategy one and three were the best at replicating human labelling. This suggests suggesting that the other "foreground only" and "background only" strategies fail to capture the crux of an image that people look for when providing a summary description. Strategy two, "foreground only", did however perform better than strategy four, "background only", suggesting when people must pick only 2 descriptive colors they focus on foreground. Background colors get disregarded as less important.

All strategies had means and medians above the "plausibly human" baseline, indicating each strategy did fairly well in matching the human labels and none was unusable. That the minimums, maximums, and the medians were all the same suggests no singular strategy was universally accurate. In one particular photo used in this study (image_01354; see Appendix), two distinct flowers of different colors took up most of the image, and people unanimously chose the colors of those flowers as the two summary labels while completely ignoring the green leaves and black shadows as unimportant; however, every computerized strategy scored a zero on this picture. Part of this appears to be because the "white" flower was being labelled "gray" or "blue" by the computer (likely due to lighting, as this flower is slightly in shade), but another part is that the first, third, and fourth strategies all included green as an important descriptor even though no human did. Thus, human agreement is not necessarily a good predictor of which images will be difficult for the programmatic strategies.

The number of colors present in the image may be a better metric, or potentially the comparative areas of the foreground and background. In the set of 20 images used for this study,

some of the flowers were closer to the camera and took up more of the frame than others. This could explain why the labels of some images seemed to come entirely from one area of the picture.

In a broader sense, this may be because people focus on different elements of different types of images. For instance, a person's description of a landscape painting would differ greatly from their description of a product photo; background colors are much less relevant in a photo of a product on a table than they would be in a painting where much of the artistry is in how the background is depicted. This, in turn, has implications for which strategy is able to most accurately label the image's main colors. The strength in human judgements seems to be the ability to adapt strategies based on the layout and characteristics of the image being described. This might be applied in future by categorizing images — for example, by determining what percentage of the image is taken up by the foreground — before choosing a labelling strategy to apply.

However, a complicating factor is the watershed algorithm used to distinguish foreground and background. For some images, the segmentation was neat and accurate, but in others there was a great deal of noise in outlining the foreground and background. See Figure 3 below for examples.

Future research should attempt to replicate these labelling with different image segmentation techniques to determine before the differences in labelling success can confidently be attributed to color differences in the foreground and background of an image. It is possible more precise foreground extraction would alter the outcomes of the strategies outlines above, thus leading to different success scores. Further investigation can also be done with different

types of images datasets to determine if foreground and background segmentation is a widely useful strategy or if only certain domains (e.g. product photos for online commerce) find success in extracting relevant color labels with these methods.

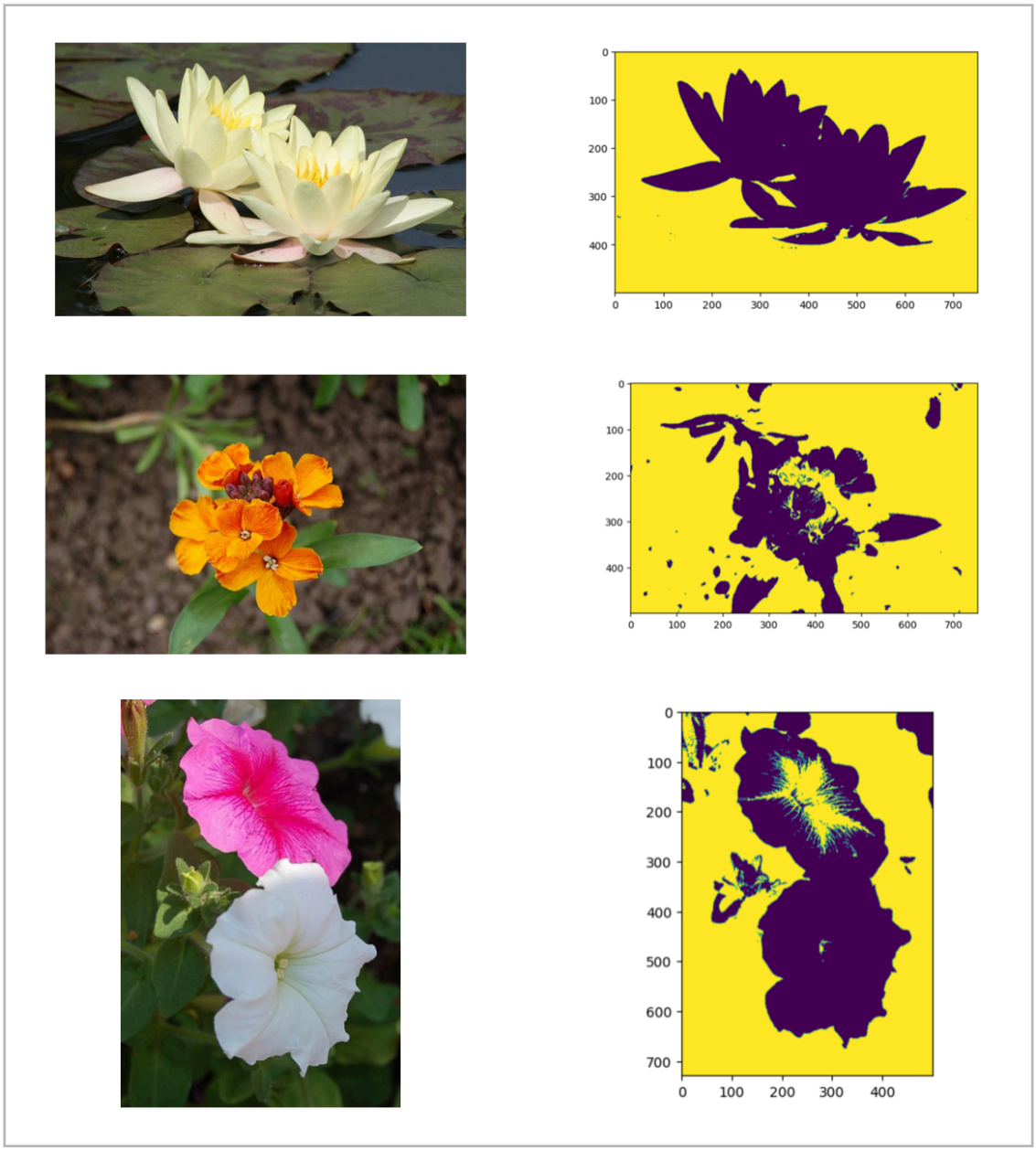

Figure 3.
The above images on the left are the original survey pictures, and the images on the right represent the foreground and background separation, with varying degrees of success. Yellow represents background and the darker purple represents foreground.

Bibliography

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews. Cognitive Science*, *1*(1), 69–78. https://doi.org/10.1002/wcs.26

Othman, A., Wook, T., Qamar, F. (2020). *Categorizing Color Appearances of Image Scenes Based on Human Color Perception for Image Retrieval* (Vol. 8, pp. 161692–161701). https://doi.org/10.1109/ACCESS.2020.3020918

Visual Geometry Group—University of Oxford. (n.d.). Retrieved April 18, 2023, from https://www.robots.ox.ac.uk/~vgg/data/flowers/102/index.html

van de Weijer, J., Schmid, C., & Verbeek, J. (2007). Learning Color Names from Real-World Images. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. https://doi.org/10.1109/CVPR.2007.383218

Witzel, C., & Gegenfurtner, K. R. (2016). Categorical perception for red and brown. *Journal of Experimental Psychology: Human Perception and Performance*, 42(4), 540–570. https://doi.org/10.1037/xhp0000154

Appendix

image_00001

image_00333

image_00524

image_00539

image_00762

image_00968

image_01335

image_01354

image_01377

image_01705

image_01726

image_01807

image_02190

image_02442

image_02468

image_02524

image_02582

image_02606

image_02889

image_07919