Received: 16 June 2022 | Accepted: 18 October 2022

DOI: 10.1002/aps3.11505

APPLICATION ARTICLE

Applications in Plant Sciences

An automated pipeline for supervised classification of petal color from citizen science photographs

Rachel A. Perez-Udell | **Andrew T. Udell** | **Shu-Mei Chang**

Department of Plant Biology, University of Georgia, 2502 Miller Plant Science, 120 Carlton St., Athens, Georgia 30602, USA

Correspondence

Rachel A. Perez-Udell, Department of Biology, University of North Georgia, 151G Health and Natural Sciences, 159 Sunset Drive, Dahlonega, Georgia 30533, USA.
Email: rachel.perez-udell@ung.edu

Shu-Mei Chang, Department of Plant Biology, University of Georgia, 2502 Miller Plant Science, 120 Carlton St., Athens, Georgia 30602, USA.
Email: smchang@uga.edu

Abstract

Premise: Petal color is an ecologically important trait, and uncovering color variation over a geographic range, particularly in species with large distributions and/or short bloom times, requires extensive fieldwork. We have developed an alternative method that segments images from citizen science repositories using Python and *k*-means clustering in the hue-saturation-value (HSV) color space.

Methods: Our method uses *k*-means clustering to aggregate like-color pixels in sample images to generate the HSV color space encapsulating the color range of petals. Using the HSV values, our method isolates photographs containing clusters in that range and bins them into a classification scheme based on user-defined categories.

Results: We demonstrate the application of this method using two species: one with a continuous range of variation of pink-purple petals in *Geranium maculatum*, and one with a binary classification of white versus blue in *Linanthus parryae*. We demonstrate results that are repeatable and accurate.

Discussion: This method provides a flexible, robust, and easily adjustable approach for the classification of color images from citizen science repositories. By using color to classify images, this pipeline sidesteps many of the issues encountered using more traditional computer vision applications. This approach provides a tool for making use of large citizen scientist data sets.

KEYWORDS

citizen science, flower color, *Geranium*, image segmentation, *Linanthus*, supervised learning

Variation in floral pigmentation is an ecologically important trait that is frequently studied in the context of plant response to biotic and abiotic factors, demographic changes, population structure, and community interactions (Endler, 1977). It is particularly intriguing when a species shows nonrandom variation, such as clines, patches, and other gradational patterns, across a geographic range. A well-known case study of corolla color variation among populations of *Linanthus parryae* (A. Gray) Greene in western North America has been used as a prime example for the population genetics concept of isolation by distance described by the renowned geneticist Sewall Wright (Wright, 1943, but see alternative conclusions by Schemske and Bierzychudek, 2007). Additional examples of flower pigmentation revealed natural selection following secondary contact hybrid zones (*Phlox* L., Hopkins et al., 2012) and balancing selection imposed by pollinators (*Ipomoea* L., Fry and Rausher, 1997; *Raphanus* L., Irwin and Strauss, 2005; and *Clarkia* Pursh, Eckhart et al., 2006), among others. Despite the rich evolutionary and ecological insight provided by patterns of variation across geographical scales, there are relatively few studies that offer detailed description. This is partly because characterizing natural variation at a broad geographic level can be difficult, as it not only requires a great amount of time and effort, but also presents other challenges inherent to field studies. Short bloom times in spring ephemerals, inclement weather during field seasons, long distances between populations, scarcity of

funding for fieldwork, and travel restrictions all pose additional barriers to the conventional field surveys used to collect such data. As a result, researchers have sought alternative approaches to alleviate some of these challenges.

Recent studies using easily accessible, and often large, databases generated by ever-increasing digitization efforts (e.g., SERNEC for herbarium specimens, https://sernecportal.org/portal/) present an exciting alternative to field surveys and allow studies to expand in both spatial and temporal scales (Willis et al., 2017; Espinosa and Castro, 2018; Koski et al., 2021). One form of digitized data is the image records found in photograph repositories such as iNaturalist (https://www.inaturalist.org/) and the Global Biodiversity Information Facility (GBIF; https://www.gbif.org/), which stores the subset of iNaturalist images deemed as "research grade." With the popularity of smartphones and the wide coverage of wireless networks, these virtual collections taken by citizen scientists include photos illustrating floral pigmentation of fresh samples collected from a large number of georeferenced locations, thereby sidestepping the problem of degrading pigments in dried herbarium samples. Additionally, popular mobile apps like iNaturalist and Seek by iNaturalist (https://www.inaturalist.org/pages/seek_app), with support from community curation and artificial intelligence algorithms aiding species identification, are steadily increasing the number and quality of these image records. Indeed, several recent studies clearly showed the power of extracting the occurrence data from these citizen science collections (Li et al., 2019; Mancinelli et al., 2021; Mesaglio et al., 2021).

Obtaining morphologic data from these records, however, often requires manually sifting through large numbers of photographs—a task that is both time-consuming and sometimes inaccurate. While feasible, the time and effort required for manual evaluation can potentially delay research progress and may introduce biases or inconsistency due to researcher subjectivity on traits like pigmentation levels (Webster et al., 2002; Emery et al., 2017). Furthermore, photographs taken in the field without any color standard pose additional challenges for their use in scientific research. For example, it is well established that humans have variable ability to perceive and distinguish saturation of colors (Cooper et al., 1991; Webster et al., 2002; Emery et al., 2017). The manual cataloging of images into different color bins could introduce unwanted noise into the data and interfere with subsequent analysis. Furthermore, in traditional red-green-blue (RGB) color analysis, changes in illumination, such as shade or shadows, can alter the R, G, and B values in a discontinuous and nonlinear manner (Chavolla et al., 2018), making it difficult to compare the color data from images collected in the field or other environments without controlled lighting.

To improve the evaluation process for utilizing photographs taken by citizen scientists, we developed a simple computer vision methodology to automate the classification of petal pigmentation in large photograph data sets using the hue-saturation-value (HSV) color system. This method couples the Anaconda distribution of Python, a programming language that is easily customizable for mathematics and image processing, and Jupyter Notebooks, an integrative development environment, to generate a pipeline with four scripts. It uses a relatively low amount of computer storage, using just over half a gigabyte to store ~5000 records. It is well suited for large floral structures that display either a gradient in color variation or separate, discrete colors, and works for both a priori and de novo color classification. We include two examples, one with a continuous color gradient in petals and the other with discrete binary color morphs, to illustrate the range of usage.

METHODS

The analysis pipeline we have developed downloads and analyzes photographic records using four Python scripts (*Downloader*, *Color Cluster Visualizer*, *Data Collector*, and *Classifier*; Figure 1) formatted as Jupyter Notebooks (Kluyver et al., 2016). The notebooks, a quick start guide, and example data sets are freely available at GitHub (https://github.com/atudell/Color-Cluster-Kit). The example data sets provided include the iNaturalist image records for two flowering plant species, although our pipeline can also work with similar photographic data from other sources.

The pipeline described consists of four scripts. The first script, *Downloader*, downloads and names images onto the user's computer. The second script, *Color Cluster Visualizer*, examines pixel colors of a subset of images and identifies major color clusters. The same clustering method is used in the third script, *Data Collector*, as a form of image segmentation to isolate the flower from the background for each image and to return summarized pixel values describing its colors. The fourth script, *Classifier*, puts images into color bins based on user-specified rules (Appendices S1–S8; see Supporting Information with this article).

To demonstrate how this pipeline can be used for both continuous and discrete color variation, we showcase implementations using *Geranium maculatum* L. and *Linanthus parryae*, respectively. We use the HSV model to describe the color (H), the amount of gray in the color (S), and the brightness of the color (V) (Smith, 1978).

Data acquisition

In this study, we used digitized photographs downloaded from iNaturalist. Photographs were retrieved using *Downloader*, which automates the downloading and naming of photographs. As written, *Downloader* works with iNaturalist but can be modified for use with any other database. Mechanically, *Downloader* loads a specific web page for the displayed image and interacts with the HTML to download it. Minimal changes to the URL scheme would be needed to obtain similar results with other data archives, provided

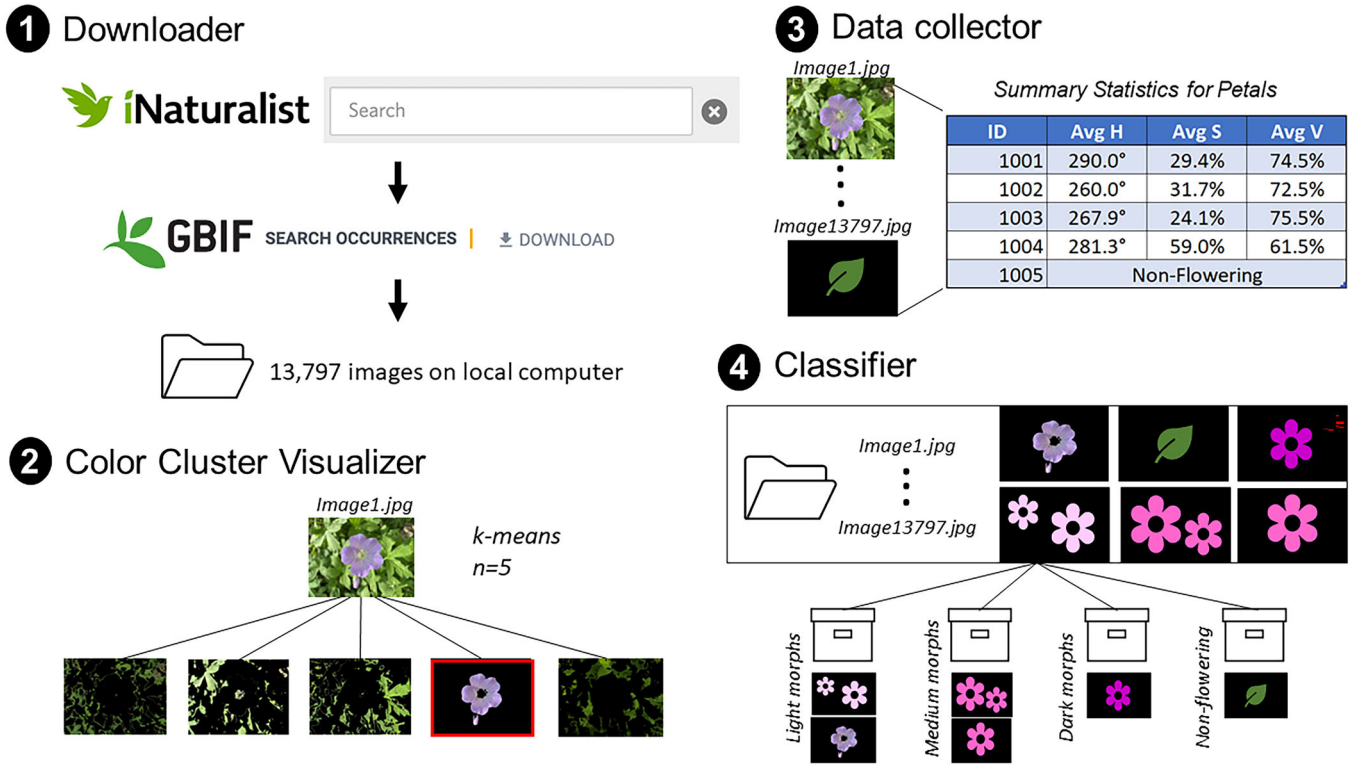

FIGURE 1 Workflow of the pipeline. (1) The *Downloader* script downloads a list of images in a CSV file as well as the images in the JPEG format. Images are named by their sample ID. (2) The *Color Cluster Visualizer* script creates image clusters within photographs for visualization. Coarse- and fine-tuning the parameter *k* segments the flower to a greater degree from its background at the expense of computational power. Additionally, this script gives the average H, S, and V for each identified cluster and may be used to derive an estimate of the HSV range. (3) The *Data Collector* script repeats the *Color Cluster Visualizer* script for all photographs and outputs a CSV file with the summary statistics only for clusters that fall within the range defined by the information obtained from *Color Cluster Visualizer*. If no cluster exists for a photograph within the defined range, the output is "non-flowering." (4) The *Classifier* script bins the summary statistics according to user input.

those archives use a similar structure. Because the method determines color on a pixel-by-pixel basis, this pipeline is flexible in terms of the overall photograph quality. For example, as neither shape nor prominence of the key feature is critical in our color clustering, our pipeline is able to extract usable data from flower images that are blurry or off center. Additional data qualities can be specified depending on the requirements of individual research projects (see examples below).

Data preparation

The second script, *Color Cluster Visualizer*, helps the user determine the number of clusters needed to distinguish the color(s) of the target feature from the background. It also derives an estimate for the typical HSV ranges for these segments as part of the output. The script uses *k*-means clustering, a statistical method that groups like-data points, to explore the pixel properties within each image. Based on a user-specified parameter *k* (the number of clusters), the pixel colors are placed into *k* clusters. The script then outputs *k* partial images, each highlighting only the pixels corresponding to one particular cluster and its average H, S, and V values, as highlighted in Figure 2.

One of the main goals of the *Color Cluster Visualizer* script is to find the optimal value for *k*, defined as the smallest cluster number that allows clear separation of the target from the background. A *k* value that is too large will increase the computational time without adding additional benefit to the output, but one that is too small will not properly segment the target from the background. In this pipeline, the optimal *k* needs to be determined using a small representative subset of images that encapsulates the color variation of the target feature. The *Color Cluster Visualizer* script will be run iteratively, with the *k* initially changing in increments of 5 as coarse adjustment and then fine-tuning in smaller increments at the user's discretion. An optimal *k* results in all pixels of the target feature (a flower in this case) falling entirely within one or a set of clusters. After iterating through the subsample of photos, a *k* value should be estimated.

The second main goal of the *Color Cluster Visualizer* script is to determine the range of HSV encapsulating color variation in the target feature that is captured by more than one cluster. For example, if a flower is captured in multiple clusters, one petal in its own cluster and the rest in a

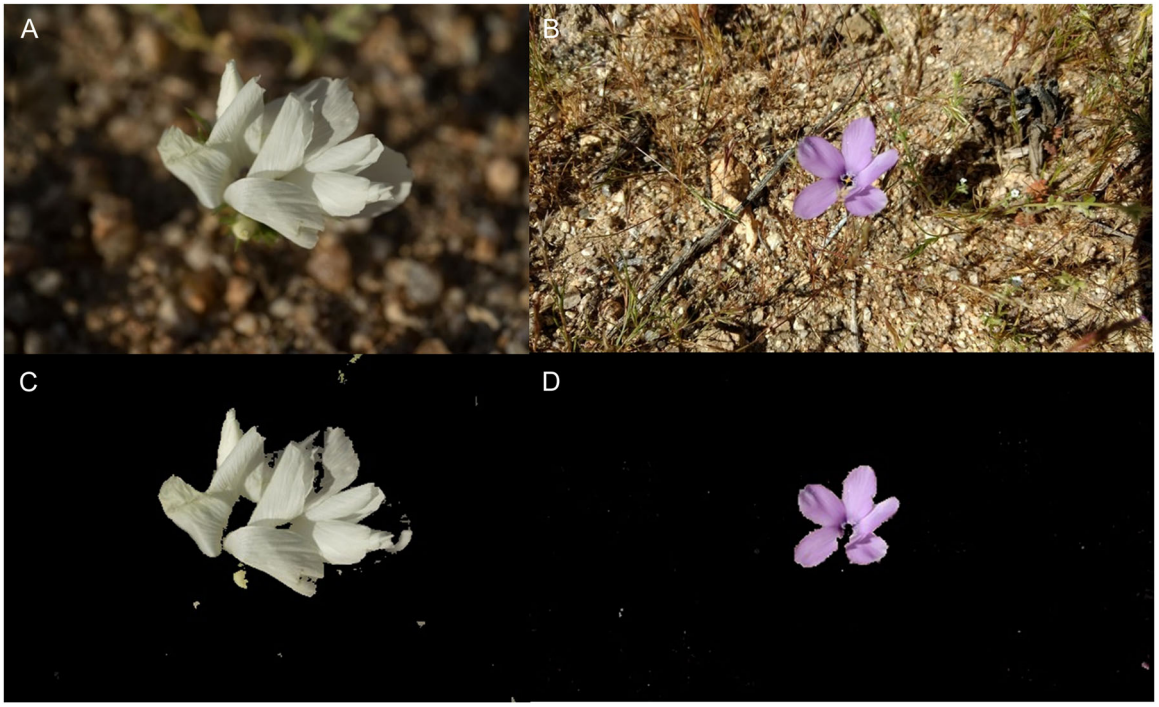

FIGURE 2 Cluster examples. Original photos (A, B) versus isolated color cluster of interest (C, D) from white (A, C) and blue (B, D) *Linanthus parryae* flowers.

separate cluster, the HSV values in all clusters in which the feature appears should be noted and factored into the estimated HSV range used in the *Data Collector* script.

The third script, *Data Collector*, performs the same *k*-means clustering for every photograph in the data set and aggregates clusters whose averages fall within the HSV range determined in the *Color Cluster Visualizer* step. Under ideal conditions, the HSV range will ensure that only the pixels associated with the feature of interest are captured. Consequently, images with no pixels captured are inferred as missing the feature of interest, and, in our case, are labeled as "non-flowering." The output returns a set of summary statistics for each photograph that describes the color profile of the captured pixels. The decision of which summary statistics to output is species specific. Average H, S, and V values provide a good baseline description of the colors and have been set as the default in the script, but the script may be modified to report any other summary statistics to aid in color classification. For example, the color continuity found in *G. maculatum* suggested that a good way to bin the petal color is by the average values of S, whereas the distinct blue and white flowers in *L. parryae* only required a count of pixels in the blue color range. We present this in more detail in the Results section. For users who are strictly interested in quantifying color variation, this script can serve as the end of the pipeline, with the output providing enough information for further analysis of the natural range of colors exhibited by the target feature. In some circumstances, however, an additional step to classify the raw color data into bins is desirable.

Data cataloging

The final script, *Classifier*, sorts the photographs into their respective clusters, or bins, for subsequent analysis. The details of this script are dictated by the user's wishes on how to classify their data for later analyses and will, therefore, need to be developed by individual users. Here, we illustrate how we customize the final step to specifically fit our two examples. When color bins are established a priori, a classification scheme may be developed by a simple algorithm with a predefined set of rules. For example, *L. parryae*, being a true example of an unambiguous classification, lends itself to simple blue and white bins; however, *G. maculatum*, whose color represents a continuum, does not naturally favor any particular classification method. Regardless, researchers artificially coercing flowers into color bins may find it useful to allow comparisons to manually collected data. Fortunately, our pipeline provides enough flexibility to similarly assign colors into bins. The sample pipeline applications given in the Results section provide two basic classification schemes. The classification scheme for *G. maculatum* utilizes the saturation value as a percentile of the sample population, while the scheme applied for *L. parryae* looks for a large number of blue pixels in the image. These strategies for placing images into different bins were developed with preliminary examination of the images or prior knowledge of the color distribution in target features. Once the script is finished running, the resulting output file will include both the raw color data and the classification.

Geranium maculatum specifics

We obtained the records for *G. maculatum* from iNaturalist using the designation of "research-grade," defined as from a wild location, georeferenced, and with verified identification (Appendix S9). A total of 5049 medium-resolution photographs were downloaded using our *Downloader* script. Additional curation removed 47 records that were posted outside of the possible flowering time (i.e., between September and February), were registered in locations missing environmental data from WorldClim (https://www.worldclim.org/), or were registered in improbable locations (i.e, the ocean). The photographs required a total of 576 MB of space, which is relatively small for an average consumer-grade laptop, enabling any machine to conduct this type of analysis.

Geranium maculatum is a species with continuous variation in floral color ranging from light pink to magenta and deep purple. In previous studies, we developed a physical color swatch that encompassed the range of observed petal colors to allow easy scoring of this continuous variation during field surveys. To create a scoring system for the iNaturalist images comparable to the swatch method, we set the parameters for color clustering in the pipeline to replicate the swatch scales and preselected three color bins for classification (light, medium, and dark) (Figure 3). Using a Samsung Galaxy S10 camera (Samsung, Seoul, South Korea), we took 143 images of greenhouse-grown *G. maculatum* flowers whose color had been designated using the swatch method. The greenhouse images were then processed using the *Color Cluster Visualizer* script, and an estimated value of $k = 25$ was found to best separate flower(s) from their respective backgrounds. In addition, the HSV averages of the cluster(s) in which the flower(s) appeared were noted and used to estimate an HSV range (H: 246–314°, S: 5.9–100%, V: 0–100%). This initial step revealed that the variety of colors found in *G. maculatum* fell within a small hue range and most of the variation in colors observed was from color saturation.

The estimates for k and HSV were entered into the *Data Collector* script once the averages of H, S, and V were obtained, and the script was run using photographs downloaded from iNaturalist. Any photographs that did not return a cluster within the HSV range were labeled as "no flowers." Given the number of observations and the value of k, the script took approximately 10 hours to complete and return a file detailing each observation and its summary statistics. To calibrate our results, we used the greenhouse images to develop a simple binning system where flowers with an S value below 27.5% were classified as "Light," those with S values from 27.5% to 36.1% were classified as "Medium," and those 36.2% and greater were classified as "Dark." We then applied this simple system in the *Classifier* script and made the final color classification for each observation in the iNaturalist data set.

Linanthus parryae specifics

Linanthus parryae displays two flower color morphs: one with a blue corolla and one with a white corolla (Figure 4). The pipeline for this species first describes the HSV space of the blue seen in *L. parryae* and then determines a metric with which to classify a flower. Records from iNaturalist were obtained in the same manner as for *G. maculatum*. The *Downloader* script collected 282 photographs, utilizing 38.9 MB of space (Appendix S10). All research-grade photographs of this species contained at least one flower, and some contained both color morphs. We removed the images displaying more than one color morph and retained 224 images in our final data set, each having either a white or blue corolla.

Validations

We validated our results from the reported pipeline for the following three aspects. First, for the classification of the

FIGURE 3 Petal morphs of *Geranium maculatum* showing (from left to right) light, medium, and dark petals. Note: Although the petal color of this species shows continuous variation, the images shown represent the artificial bins used in the study examples.

FIGURE 4 Petal morphs of *Linanthus parryae* showing (from left to right) a blue morph, a white morph, and a morph with white petals and a blue center.

presence/absence of flowers, we manually examined a random subset of the images classified by the pipeline as "non-flowering" or "flowering" to determine the pipeline's accuracy. Second, because the petal color of *G. maculatum* is a continuous trait varying between pink and purple, it is difficult to directly validate the outcome of the *Data Collector* script (see "Data Preparation," above). Instead, we focused on the classification of the *G. maculatum* petal colors into Light, Medium, and Dark bins, as described in the "Data Cataloging" section. To do so, all three authors of this paper independently evaluated a subsample of 150 images using the same swatch scales used for field surveys. We had two goals in this validation step: (i) to compare the results between manual evaluations and the pipeline classification, and (ii) to evaluate consistency between researchers in determining color classification when using the same color swatch. Finally, we similarly examined a subset of the *L. parryae* images to determine the accuracy of the pipeline. Because the floral coloration in *L. parryae* is discrete, the manual examination was able to separate images into discrete bins.

RESULTS

In the continuous color example, we observed that *G. maculatum*'s petal pigmentation ranges from light pink to dark purple. Preliminary analysis showed that this petal color variation is largely explained by the change in S value, with the H and V values showing relatively less variation among individuals (Appendix S11). In our discrete color example, *L. parryae* presents either blue or white petals that differ in all three metrics of H, S, and V.

Application of the pipeline: Example 1 – petal pigmentation of *Geranium maculatum*

Although the petal color of *G. maculatum* varies in a continuous fashion (Appendix S11), we found that the key metric that covaried with our *G. maculatum* swatch scales was the saturation value, with relatively minimal variation in the H and V values. From this pilot study, we chose S = 27.5% and S = 36.1% as the criteria to separate our Light, Medium, and Dark bins. These values were chosen arbitrarily to correspond to existing categories on the swatch used in our previous field studies.

Out of the 5004 input photographs, 818 (16.3%) occurrences were classified as non-flowering. We selected the HSV values to purposefully make the algorithm more conservative (i.e., favoring the classification of flowering images as non-flowering) to minimize the introduction of false positives into the data set at the expense of the final sample size. Our manual validations found that 87.0% of the "non-flowering" and 92.8% of the "flowering" images were classified accurately (Appendix S12). Of the 36 records incorrectly classified as flowering, 30 photographs contained leaf litter in the background whose shades of gray caused the script to falsely label them as light flowers.

Of the 4186 records labeled as flowering, 2874 were classified as Light, 832 as Medium, and 480 as Dark. Our manual validation of a subset of 150 images showed that we unanimously agreed on a classification for approximately 73% of the images, highlighting the inherent human variability in identifying nuances in color variation, as well as the likely variation in computer monitor display (Figure 5A). Of the observations for which there was unanimous agreement using manual scoring, the algorithm gave the same score 85.3% of the time (Figure 5B). Among the observations that did not receive a unanimous vote, the algorithm always agreed with at least one of the scorers.

Application of the pipeline: Example 2 – petal pigmentation of *Linanthus parryae*

The algorithm required for a binary color classification in *L. parryae* is simpler than the *G. maculatum* case; we only needed to accurately recognize one morph and assign the

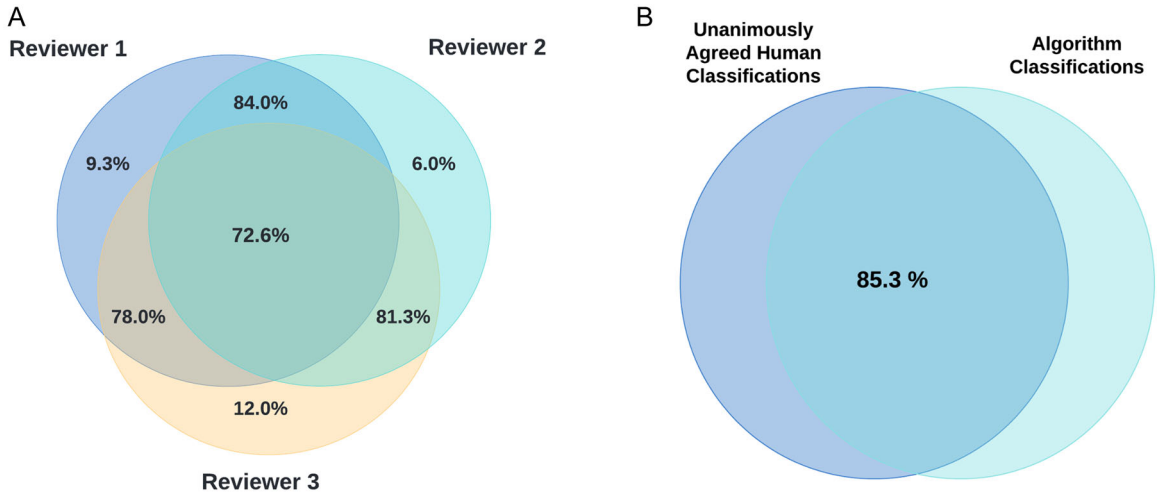

FIGURE 5 Venn diagrams representing agreements among the reviewers and the algorithm in color classification. (A) Reviewer agreement on the color classification of *Geranium maculatum*. Each circle represents the classifications of an individual reviewer. (B) Algorithm and reviewer agreement. The left circle represents the classifications for which all three reviewers agreed on a classification, and the right circle represents the computer classification. It is worth noting that at least one of the reviewers agreed with the classification by the algorithm for all images.

remainders to the other morph. We chose blue because of the rocky white background in most photographs. We used 50 images of blue flowers to derive an estimation of its HSV color range for the *Data Collector* script. To accommodate the fact that a small portion of white flowers contain blue pigments in their corolla tubes (as shown in Figure 4), we modified the *Data Collector* script to exclude photographs with less than 3000 blue pixels. This number was empirically determined based on the actual values observed in the images and only meant as a cut-off measure specific to this species. We chose the absolute number, rather than a percentage, of pixels in the images, because the photos varied in resolution and the flowers varied in their prominence within the photos. This value successfully distinguished the white flowers with the blue corolla tube from blue flowers because images of the latter contain a significantly larger number of blue pixels. Our pipeline predicted the presence of a blue flower with a precision of 91.5%. There were no images in this data set without flowers.

DISCUSSION

We present in this paper an efficient, reliable, and automated color classification system that can allow researchers to tap into the wealth of citizen-contributed images available in public databases that are currently underutilized. The use of a computer algorithm greatly reduces the time and effort required to score large data sets and also provides a much more consistent method of classification than conventional manual scoring. The color identification and classification methods used in our pipeline are flexible, can generate discrete and unambiguous classifications, and can be used with floral features that display continuous variation. Finally, this method also allows users to easily modify the analysis scripts to fit their needs for subsequent analyses. These qualities make the method reported here a great improvement from the existing hand classification method for documenting and classifying floral pigmentation at large geographical scales.

Our method provides an easy interface with citizen science data and returns respectable results. While citizen scientists offer a trove of curated and georeferenced records, the pictures are often difficult for a computer to decipher because they are taken from various angles, with variation in the subject's prominence within the image, with variable numbers of species present, using a variety of camera types, and under many different lighting conditions. Performing image segmentation using only color alleviates most of these concerns. Simply by searching the image for pixels within a target color range and ignoring other traditional hallmarks of computer vision, such as shape and texture (Nilsback and Zisserman, 2008; Guru et al., 2010), we can sidestep the issue of images taken from various angles and distances. In addition, by breaking down an image into its pixel components and analyzing the color mathematically, our pipeline avoids the intrinsic biases or inconsistencies introduced when colors are manually scored, as was demonstrated by the moderate level of agreement among the three authors scoring the same set of images.

The conversion of color data into the HSV space in our methodology addresses a major concern for color analysis using iNaturalist images, which are often taken under variable lighting conditions. This issue makes the traditional RGB color space less suitable because the values of R, G, and B vary widely in photographs taken under shady conditions (Chavolla et al., 2018). Other existing floral classification systems avoid this problem by using curated data sets that employ high-quality images taken with prominent blooms

and/or fruit from consistent angles and reliable lighting conditions (Blasco et al., 2009; Singh et al., 2011; Rosyani et al., 2018). Our approach enables images taken in the field, such as those from iNaturalist, to be used for analysis by using the HSV color space. In the HSV color space, changes in illumination primarily affect V values and minimally impact H and S values (Chavolla et al., 2018). As a result, our approach allows images to be used even when taken under different light conditions and at different angles.

Another advantage of our method is its simplicity and flexibility compared to other approaches commonly used in computer visualization. Many other machine learning approaches attempt to classify floral images using texture and shape in addition to color. Consequently, these methods are often computationally expensive, difficult to understand and implement for non-experts, and overly complex for simple color classifications (Nilsback and Zisserman, 2008; Guru et al., 2010). Cutting-edge computer vision approaches, such as neural networks, generally also require the manual drawing of bounding boxes for large training and testing data sets to properly calibrate deep learning (Venegas et al., 2021). Finally, for especially large data sets, specialized (and often expensive) equipment, such as multiple graphics processing units (GPUs), may be required (Van Horn et al., 2018). Although these other approaches may produce complex measurements of target features, the stringent requirements and complexity of these methods make them less accessible to many researchers.

Our pipeline simplifies analysis by focusing only on the color range of the target feature; it therefore requires minimal manual labor and can be performed quickly on consumer-grade computers using large numbers of medium-resolution images. Although we showed two simple methods of predetermined classification in our examples, the data generated from the first three scripts can be readily fed into other more complex classification methods such as support vector machines or random forests (Cortes and Vapnik, 1995; Breiman, 2001) with a predetermined number of color bins. If there is no prior knowledge or preference for the number of bins, one can use unsupervised machine learning techniques such as OPTICS or DBSCAN (Ester et al., 1996; Ankerst et al., 1999) to create color bins using the HSV data collected in the *Data Collector* script. Discussion of these approaches is beyond the scope of this paper, but information can be found in Kanagala and Krishnaiah (2016). Moreover, this method may be extended to identify any arbitrary number of morphs by deriving HSV estimates for multiple colors and running images iteratively through the pipeline.

Even though our method provides many strengths for working with citizen scientist data, it does have some limitations. For example, this method works best for large, conspicuous blooms that display high contrast relative to the background. Small, inconspicuous blooms, or blooms that blend in with their background, would generally require additional image segmentation and machine learning. However, in some cases, creative workarounds can be established, as illustrated in our *L. parryae* example in which we used blue pixels, showing that it is possible to use the flexibility of our scripts against this limitation.

A second limitation is that, by ignoring shape, our algorithm cannot distinguish between separate bodies with the same color and will aggregate them together. Consequently, the presence of a second species within the same color range, other flowers of the same species but with different color profiles, or even another object in the photo with a similar color will skew the resulting data. For example, in the case of *G. maculatum*, if both dark and light morph flowers were present in the same image, the algorithm would combine and average out their colors, giving a result that does not accurately convey the colors of the individual flowers. Similarly, our algorithm will average differences in coloration patterns present within the same flower, such as dark veins or a dark inner corolla, even if human judgement often places a higher degree of priority on the edges.

Certain aspects of the image classification by our algorithm prove to be messier than hand classification, as reflected in the misclassification rate of 9–10% in both species analyzed. In the case of *G. maculatum*, we selected a more stringent HSV range in order to avoid false positives (e.g., the photographer's hand or reddish leaves), but this resulted in mistakenly excluding ~10% of images. While not a low error rate, this is an acceptable compromise because (1) the initial number of images was large enough to sustain this level of reduction and (2) the demand in time and effort to manually classify a large number of images is high. However, the same error rate may not be acceptable when the starting sample size is already small (e.g., rare species). In these situations, we recommend adding a "triaging" step at the start of the workflow to separate images with flowers from those without before using our pipeline.

Lastly, our algorithm is not able to account for other factors that can affect the quality of citizen scientist photographs such as the presence of additional objects in photographs (e.g., people, trash, keys fobs), a lack of standardization of cameras, and photomanipulation, which can potentially skew color perception. For example, discrepancies in how various cameras capture color can distort the results, and many smartphone cameras, while incredibly accessible to citizen scientists, also allow the option of applying filters to an image that may disrupt the ability to capture the natural color of a plant. To this extent, we provide recommendations for taking photographs for this type of usage in Appendix S13 to help guide the creation of new data sets or broaden the viability of existing citizen scientist data sets.

AUTHOR CONTRIBUTIONS

R.A.P. developed the idea for this project and the pipeline, cowrote the Python scripts, and contributed to the writing of the manuscript. A.T.U. cowrote the Python scripts, including developing the *k*-means approach for the pipeline, and contributed to the writing of the manuscript. S.M.C.

contributed to the development of the idea, provided the biological details during pipeline development, and contributed to the writing of the manuscript. All authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors thank all the contributors to iNaturalist for their invaluable documentation. Chazz Jordan and Riley Thoen (University of Georgia) tested an earlier version of this pipeline and helped improve its development. Summer Blanco's (University of Georgia) creative contribution to the illustration of Figure 1 is highly appreciated. Comments from Ashley Early, Patrick Smallwood, Summer Blanco, and Norris Armstrong (University of Georgia) significantly improved an earlier version of this paper. A grant to R.A.P. from the University of Georgia Libraries Graduate Student Open Access Fund covered the article publication charge for this article.

DATA AVAILABILITY STATEMENT

The scripts, Jupyter notebooks, quick start guide, and example data sets used in this study are freely available at GitHub (https://github.com/atudell/Color-Cluster-Kit). Annotated scripts and raw data for *Geranium maculatum and Linanthus parryae* are provided in the Supporting Information.

ORCID

Rachel A. Perez-Udell http://orcid.org/0000-0003-0625-8742
Shu-Mei Chang http://orcid.org/0000-0002-6005-2238

REFERENCES

Ankerst, M., M. M. Breunig, H. Kriegel, and J. Sander. 1999. OPTICS: Ordering points to identify the clustering structure. *ACM SIGMOD Record* 28: 49–60.

Blasco, J., S. Cubero, J. Gomez-Sanchis, P. Mira, and E. Molto. 2009. Development of a machine for the automatic sorting of pomegranate (*Punica granatum*) arils based on computer vision. *Journal of Food Engineering* 90(1): 27–34.

Breiman, L. 2001. Random forests. *Machine Learning* 45: 5–32.

Chavolla, E., D. Zaldivar, E. Cuevas, and M. A. Perez. 2018. Color spaces: Advantages and disadvantages in image color clustering segmentation. *In* A. Hassanien and D. Oliva [eds.], Advances in soft computing and machine learning in image processing, 3–22. Springer International Publishing, Berlin, Germany.

Cooper, B. A., M. Ward, C. A. Gowland, and J. M. McIntosh. 1991. The use of Lanthony New Color Test in determining the effects of aging on color vision. *Journal of Gerontology* 46(6): 320–324.

Cortes, C., and V. Vapnik. 1995. Support-vector networks. *Machine Learning* 20: 273–297.

Eckhart, V. M., N. S. Rushing, G. M. Hart, and J. D. Hansen. 2006. Frequency-dependent pollinator foraging in polymorphic *Clarkia xantiana* ssp. *xantiana* populations: Implications for flower colour evolution and pollinator interactions. *Oikos* 112(2): 412–421.

Emery, K. J., V. J. Volbrecht, D. H. Peterzell, and M. A. Webster. 2017. Variations in normal color vision. VI. Factors underlying individual differences in hue scaling and their implications for models of color appearance. *Vision Research* 141: 51–65.

Endler, J. A. 1977. Geographic variation, speciation and clines. Princeton University Press, Princeton, New Jersey, USA.

Espinosa, F., and M. P. Castro. 2018. On the use of herbarium specimens for morphological and anatomical research. *Botany Letters* 165(3–4): 361–367.

Ester, M., H. Kriegel, J. Sander, and X. Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* 96: 226–231.

Fry, J., and M. D. Rausher. 1997. Selection on a floral color polymorphism in the tall morning glory (*Ipomoea purpurea*): Transmission success of the alleles through pollen. *Evolution* 51(1): 66–78.

Guru, D. S., Y. H. Sharath, and S. Manjunath. 2010. Texture features and KNN in classification of flower images. *International Journal of Computers and Applications* 1: 21–29.

Hopkins, R., D. A. Levin, and M. D. Rausher. 2012. Molecular signatures of selection reproductive character displacement of flower color in *Phlox drummondii*. *Evolution* 66: 469–485.

Irwin, R. E., and S. Y. Strauss. 2005. Flower color microevolution in wild radish: Evolutionary response to pollinator-mediated selection. *American Naturalist* 165(2): 225–237.

Kanagala, H. K., and V. V. J. R. Krishnaiah. 2016. A comparative study of k-means, DBSCAN and OPTICS. Proceedings of the 2016 International Conference on Computer Communication and Informatics (ICCCI). https://doi.org/10.1109/ICCCI.2016.7479923

Kluyver, T., B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, et al. 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows. *In* F. Loizides and B. Schmidt [eds.], Positioning and power in academic publishing: Players, agents and agendas, 87–90. IOS Press, Amsterdam, the Netherlands.

Koski, M. H., D. MacQueen, and T. Ashman. 2021. Floral pigmentation has responded rapidly to global change in ozone and temperature. *Current Biology* 30: 4425–4431.e3.

Li, E., S. S. Parker, G. S. Pauly, J. M. Randall, B. V. Brown, and B. S. Cohen. 2019. An urban biodiversity assessment framework that combines an urban habitat classification scheme and citizen science data. *Frontiers in Ecology and Evolution* 7: 277.

Mancinelli, G., R. Bardelli, and A. Zenetos. 2021. A global occurrence database of the Atlantic blue crab *Callienectes sapidus*. *Scientific Data* 8: 111.

Mesaglio, T., A. Soh, S. Kurniawidjaja, and C. Sexton. 2021. 'First known photographs of living specimens': The power of iNaturalist for recording rare tropical butterflies. *Journal of Insect Conservation* 25: 905–911.

Nilsback, M., and A. Zisserman. 2008. Automated flower classification over a large number of classes. Proceedings of the Sixth Indian Conference on Computer Vision, Graphics and Image Processing. https://doi.org/10.1109/ICVGIP.2008.47

Rosyani, P., M. Taufik, A. A. Waskita, and D. H. Apriyanti. 2018. Comparison of color model for flower recognition. Proceedings of the 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE). https://doi.org/10.1109/ICITISEE.2018.8721026

Schemske, D. W., and P. Bierzychudek. 2007. Spatial differentiation for flower color in the desert annual *Linanthus parryae*: Was Wright right? *Evolution* 61(11): 2528–2543.

Singh, S., D. Dhyani, A. K. Yadav, and S. Rajkumar. 2011. Flower colour variations in gerbera (*Gerbera jamesonii*) population using image analysis. *Indian Journal of Agricultural Sciences* 81(12): 1130–1136.

Smith, A. R. 1978. Color gamut transform pairs. *Computer Graphics* 12(3): 12–19.

Van Horn, G., M. O. Aodha, Y. Song, Y. Cui, S. Sun, A. Shepard, H. Adam, et al. 2018. The iNaturalist species classification and detection dataset. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (IEEE), 8796–8778.

Venegas, P., F. Calderon, D. Riofrió, D. Benítez, G. Ramón, D. Cisneros-Heredia, M. Coimbra, et al. 2021. Automatic ladybird beetle detection using deep learning models. *PLoS ONE* 16(6): e0253027.

Webster, M. A., S. M. Webster, S. Bharadway, R. Verma, K. Jaikumar, G. Madam, and E. Vaithilingham. 2002. Variations in normal color vision. III. Unique hues in Indian and United States observers. *Journal of the Optical Society of America* 19(19): 1951–1962.

Willis, C. G., E. R. Ellwood, R. B. Primack, C. C. Davis, K. D. Pearson, A. S. Gallinat, J. M. Yost, et al. 2017. Old plants, new tricks: Phenological research using herbarium specimens. *Trends in Ecology & Evolution* 32(7): 531–546.

Wright, S. 1943. Isolation by distance. *Genetics* 28: 114–138.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. The *Downloader* script annotated for *Geranium maculatum*, built for downloading the images cataloged in the raw data set in Appendix S9.

Appendix S2. The *Color Cluster Visualizer* script annotated for *Geranium maculatum*, built to provide a visual representation of *k*-means clustering on an image.

Appendix S3. The *Data Collector* script annotated for *Geranium maculatum*, built to collect color summary statistics for each photograph within the *G. maculatum* data set.

Appendix S4. The *Classifier* script annotated for *Geranium maculatum*, built to assign a color classification to each photograph within the *G. maculatum* data set.

Appendix S5. The *Downloader* script annotated for *Linanthus parryae*, built for downloading images cataloged in the raw data set in Appendix S10.

Appendix S6. The *Color Cluster Visualizer* script annotated for *Linanthus parryae*, built to provide a visual representation of *k*-means clustering on an image.

Appendix S7. The *Data Collector* script annotated for *Linanthus parryae*, built to collect color summary statistics for each photograph within the *L. parryae* data set.

Appendix S8. The *Classifier* script annotated for *Linanthus parryae*, built to assign a color classification to each photograph within the *L. parryae* data set.

Appendix S9. A CSV file containing *Geranium maculatum* raw data downloaded from iNaturalist. The columns of interest are *k*-means data providing the raw output of Appendix S3.

Appendix S10. A CSV file containing *Linanthus parryae* raw data downloaded from iNaturalist. The columns of interest are *k*-means data providing the raw output of Appendix S7.

Appendix S11. Hue (H) and saturation (S) values for *Geranium maculatum*, with the resulting classifications. (Top) Each dot represents one image and is colored in the average HS value (V was held stable at 94%) of the flower cluster from that image. (Bottom) The resulting classifications of light, medium, or dark. Note the difference in the *y*-axis of the two figures; the first figure shows the entire range of possible hue in the HSV color space (0–179) while the second figure is restricted on the range of hue found in *G. maculatum* (approx. 120–160). The H and S values used in these graphs are the converted values used in the Python library OpenCV due to memory efficiency. To convert these values into true color space, use Hue*2 and Saturation/255.

Appendix S12. Verification of the accuracy of flowering versus non-flowering classifications in *Geranium maculatum*.

Appendix S13. Recommendations for photographing floral structures for submission to iNaturalist and other citizen science repositories.

How to cite this article: Perez-Udell, R. A., A. T. Udell, and S.-M. Chang. 2023. An automated pipeline for supervised classification of petal color from citizen science photographs. *Applications in Plant Sciences* 11(1): e11505. https://doi.org/10.1002/aps3.11505