

K-Means Clustering

In this assignment, you will write a Python script that uses K-means clustering to determine the best set of RGB colors to use for a small number of JPG images. Think of the RGB colors as 3D features, each valued 0-255. Since they are scaled the same, there is no need for transforming the data.

In addition to the Python script, you will also write a brief report that documents what you did, what observations you made (if any), and what results you produced. Use a formal tone for the writing and be concise. Submit two files on Canvas called `assignment3_yourname.py` and `assignment3_yourname.pdf`. This is to avoid any confusion when downloading and grading your work.

Python Script Requirements

1. Install and use scikit-image to read the jpg image from the given jpeg file and again when saving the reduced color image to file.
2. Your script must include a K-means function (your own implementation, not the one from sklearn) and a main function that loops thru the images and the K values to be tested, calls the K-means function to obtain a reduced color image as well as data for plots, and creates those plots.
3. Run K-means until the largest change in any centroid RGB color is less than 1. That is, stop if $\Delta \mu = \max(|\mu_R - \mu'_R|, |\mu_G - \mu'_G|, |\mu_B - \mu'_B|) < 1$ where μ denotes the last known centroid and μ' denotes the new centroid about to be created. To make sure you don't wait "forever", add an additional stopping criterion based on max 24 iterations.
4. When convergence is reached, change the pixel values to their nearest centroid RGB color.
5. Return the image as well as the sequence of $\Delta \mu$ values and the number of pixels assigned to each centroid RGB color to the main program.

Report Requirements

The report must include a summary of the assignment and a brief description of the K-means algorithm. You must run the algorithm on the four images given to you, namely, Baboon, Rocket, Smokey, and Truck, using three different values of K, namely, 4, 16, and 32. Feel free to try other K values for fun, e.g., 2, 3, and 10. Note that K does not have to be a power of two.

Include a 2 x 2 figure that shows the original image and the three images obtained by clustering. Also include a plot of the sequence of $\Delta \mu$ values and histograms of centroid RGB color pixel counts normalized to sum to 100.

Grading Rubric

1. Python script (50 pts)

(25 pts) K-means implementation as described above.

(25 pts) Main function that loops thru the images and processes the output from K-means.

2. Report (50 pts)

(10 pts) Brief summary of what the assignment is about. Include a description of the K-means clustering algorithm.

(40 pts) Images, plots and observations. Comment on which images look acceptable and which do not. For the latter, describe the problem.

Aim for 1-2 pages for the text plus 2 pages per image including plots. Help the reader digest visualizations and plots. Point them to what's interesting. Explain what you make them look at.