Image Compression using K-means Clustering

INTRODUCTION

This report presents the results of applying the K-means clustering algorithm for image compression. The K-means clustering algorithm is an unsupervised learning method that aims to partition a dataset into K distinct clusters based on the similarity between data points. In this context, the algorithm is used to reduce the number of colors in an image by grouping similar colors together and replacing them with the mean color of the group.

METHODOLOGY

We implemented K-means clustering algorithm to compress the image. The following steps were performed:

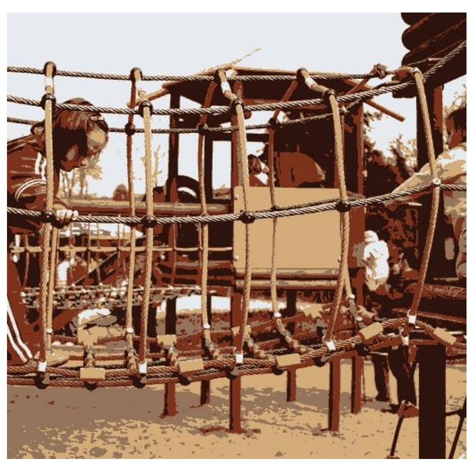
1. Load the image and normalize the pixel values to the range [0, 1].
2. Reshape the image to a 2D array with each row representing a pixel's color.
3. Apply the K-means algorithm with a varying number of clusters (K) to find the optimal color palette.
4. Replace each pixel's color with the nearest cluster center.
5. Calculate the Mean Squared Error (MSE) between the original and compressed images.

The K-means clustering algorithm was implemented using NumPy.
The iterative algorithm consists of two main steps:

Expectation (E) and Maximization (M). In the E-step, each data point is assigned to the closest cluster center. In the M-step, the cluster centers are updated by calculating the mean of all data points assigned to the respective cluster. This process is repeated until convergence or a predefined maximum number of iterations is reached.

OBSERVATIONS

K = 2



K = 5

K = 10 K = 20

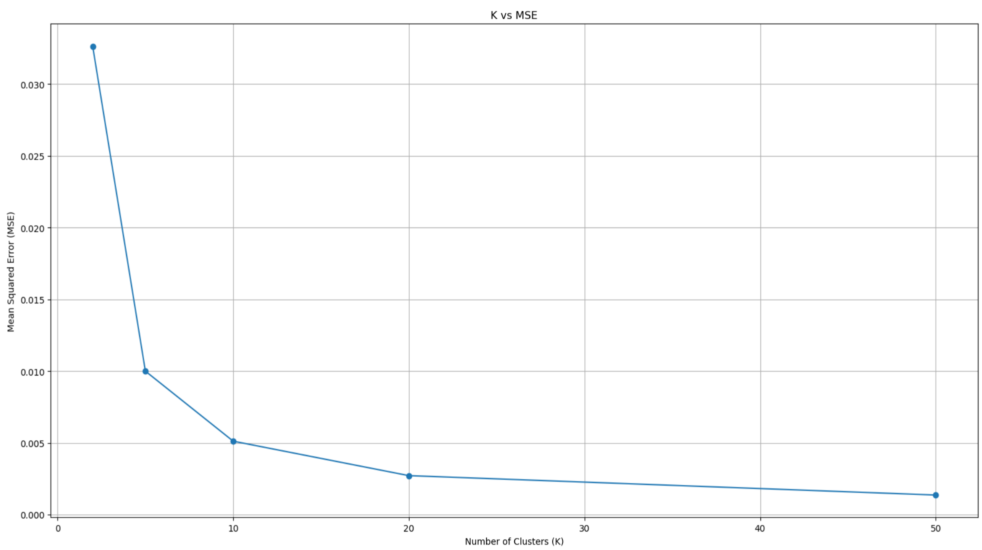
K = 50

The image compression was performed for different values of K, ranging from 2 to 50. The following are the observations:

As the number of clusters (K) increased, the Mean Squared Error (MSE) between the original and compressed images decreased. This is expected as a larger number of clusters allows for better representation of the color space, hence reducing the error.
The execution time for K = 50 was significantly larger compared to K = 2. This is because the algorithm has to calculate distances and update more cluster centers as the number of clusters increases.

- **MSE values:**
  k = 2: 0.032644859985228876
  k = 5: 0.009996235693971359
  k = 10: 0.005114179103103933
  k = 20: 0.002711608547425942
  k = 50: 0.0013591163221544726

  The plot of Mean Squared Error as a function of the number of clusters is shown below:



- **Elbow Behavior of K vs MSE Plot:**
  The K vs MSE plot shows an "elbow" behavior, which represents the point where the reduction in MSE becomes less significant as K increases. The elbow point can be used as a heuristic to choose an optimal value of K, balancing the trade-off between compression quality and computational cost.

ANALYSIS

The decrease in MSE as K increases can be attributed to the algorithm's ability to better represent the color space with more clusters. However, as the number of clusters increases, the computational complexity and execution time of the algorithm also increase. As a result, there is a trade-off between the compression quality and the computational cost.

In the images, it can be observed that as K increases, the quality of the compressed image improves, and the difference between the original and compressed images becomes less noticeable. This can be attributed to the fact that more colors are available to represent the original image, leading to a better approximation of the original color distribution.

CONCLUSION

The K-means clustering algorithm has been successfully applied for image compression, with the quality of the compressed image improving as the number of clusters increases. However, the computational complexity and execution time also increase with the number of clusters. The results demonstrate the trade-off between compression quality and computational cost in image compression using K-means clustering. The elbow behavior observed in the K vs MSE plot suggests that a reasonable compromise between image quality and computational complexity can be achieved by selecting an appropriate number of clusters K.

Overall, K-means clustering provides an effective method for image compression that can be tailored to meet specific requirements in terms of quality and computational resources.