K-Means Clustering on Images

Applied the K-Means Clustering for Image Compression on Koala.jpg and Penguins.jpg images. Using the KMeans.java as a template implemented the KMeanJava.py. The KMeansJava.py is taking huge amount of time processing the Image Compression. Used the NumPy implementation of the RGB cluster map in KMeans.py and its better than the raw implementation. Using the convergence of Means and iterations limit for KMeansJava.py whereas only the convergence of means in KMeans.py

For each image tabulated the Compression Ratios for different values of K (No. of Clusters). Repeated the Experiment with different initializations using the following seeds for the random number generators. Using the **Compression Ratio** as per the below Wikipedia link https://en.wikipedia.org/wiki/Data_compression_ratio

$$\text{Compression Ratio} = \frac{\text{Uncompressed Size}}{\text{Compressed Size}}$$

K = [2, 5, 10, 15, 20] & Seeds = [8191, 131071, 524287, 6700417, 2147483647].

Koala.jpg Image:

K	2	5	10	15	20
seed1(8191)	7.494587	4.489728	4.577587	4.739231	4.398577
seed2(131071)	15.42352	4.430071	4.15497	4.718098	4.984781
seed3(524287)	51.5298	4.453443	4.407441	4.569015	4.931045
seed4(6700417)	10.69954	4.321337	4.508835	4.608086	4.670489
seed5(2147483647)	13.24475	4.410578	4.372395	4.588832	4.644928
Avg. Comp. Ratio	**19.67844**	**4.421031**	**4.404246**	**4.644652**	**4.725964**
Var. Comp. Ratio	**325.7681**	**0.003975**	**0.026037**	**0.006128**	**0.056468**
Avg. Run Time	**19.9872**	**40.4052**	**74.0838**	**106.7062**	**142.0876**

The better K value is K = 15 as the Variance Compression Ratio is very less for that.

Penguins.jpg Image:

K	2	5	10	15	20
seed1(8191)	13.05772	9.123415	6.74326	6.413124	6.647481
seed2(131071)	11.66607	7.815316	6.492346	6.31539	6.881547
seed3(524287)	38.13664	7.558107	7.031404	6.859276	6.699814
seed4(6700417)	14.31897	6.462517	6.15595	6.550631	6.522233
seed5(2147483647)	14.4292	7.222238	6.49853	6.641975	6.698833
Avg. Comp. Ratio	**18.32172**	**7.636319**	**6.584298**	**6.556079**	**6.689982**
Var. Comp. Ratio	**99.16248**	**0.759777**	**0.084893**	**0.035541**	**0.013364**
Avg. Run Time	**19.9034**	**40.886**	**74.7394**	**108.8438**	**140.6488**

The better K value is K = 20 as the Variance Compression Ratio is low compared to other values.

Yes, there is a tradeoff between image quality and degree of compression. As the degree of compression increases to large values the Color image is degrading to Grey Scale image. However, it still able to retain the Image Segmentation features.