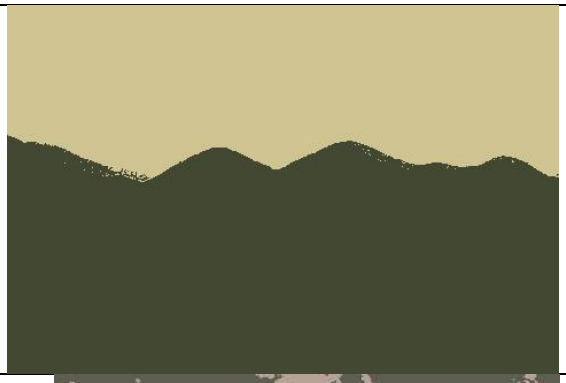
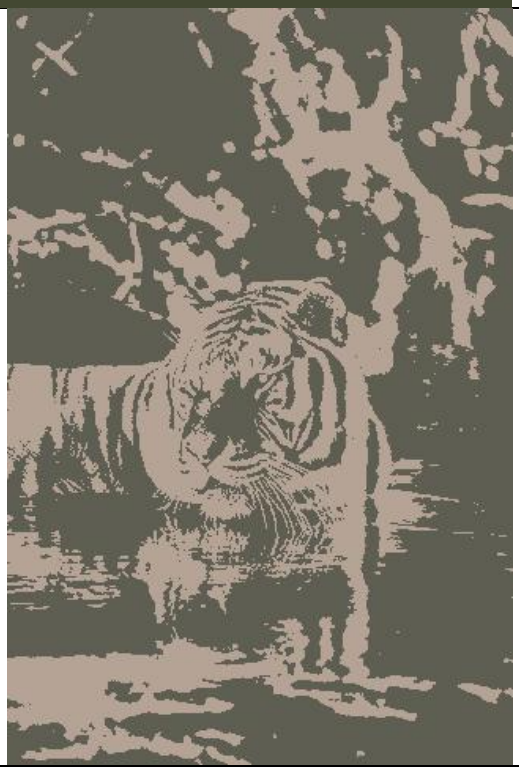
Image Segmentation Using K-means

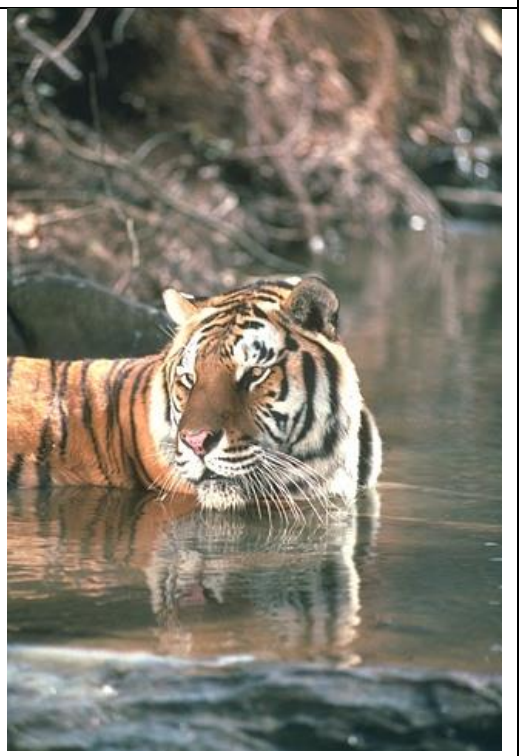
Mohammad Mirzanejad

Color photos with different values of K

K=2







Considering that in the K-means method, K random points are selected every time and this is also taken into account in the implementation. Initially the closest pixels to the selected random point form a cluster according to the RGB component, then the average of each cluster is selected as the new centroid and clustering are done according to the proximity of the pixels to the new centroids. Finally the closest pixels to each cluster center are determined and placed in the same cluster, so that the initial selection does not affect clustering.

A method to determine the best K for clustering is the Silhouette method. In this method, for each data (pixel), the average distance of each data with other data in the same cluster is calculated. Then the average distance of each data with the data in all other clusters except the cluster in which it is located is calculated and the smallest one is selected. The following value is calculated for each record of data.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

The value of s(i) is always between 1 and -1 and the closer this value is to 1 , the more ideal it is. The value 1 indicates that the data is located in the appropriate cluster, the value 0 indicates that the data is located in the border region that can belong to the adjacent cluster, and the value -1 indicates that the data does not belong to the current cluster. It belongs to .the farthest cluster from the current cluster. The obtained values are between 1 and -1 and being closer to 1 indicates the most suitable cluster. A greater distance indicates a greater distance from the current cluster belonging to clusters that are further away in terms of proximity. In this method, the k-means algorithm is executed for different k , and then the above average s(i) is executed for all the pixels and its average is calculated. That round of execution or K that yields the highest averages represents the best K.

Changing the threshold value if it is too much and causes the algorithm to terminate before convergence. In fact, it causes the center of the specified cluster not to have a suitable distance with other pixels and the center of the cluster and consequently the clusters in it do not have a suitable average distance. Hence, there will be data in the cluster that are not related to that cluster. On the other hand, if we consider the threshold value to be zero, the algorithm may never converge because at the final stages the cluster centers may fluctuate .between a certain value and never be fixed.

To evaluate the model, the easiest method is to refer to the labeled data and measure the correctness of the clustering performance, which is not possible in many cases. One of the clustering evaluation methods is to estimate the number of formed clusters. There are three well-known methods for estimating the number of clusters: 1- Silhouette method 2- Elbow Method and 3- GAP Statistic.