



Histogram Based Initial Centroids Selection for K-Means Clustering

Bhavani Srirangam and N Subhash Chandra

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

January 11, 2022

Histogram Based Initial Centroids Selection for K-Means Clustering

S. Bhavani¹, Dr. N. Subhash chandra²

¹*Research Scholar, Dept. of CSE, CVR College of Engineering, Hyderabad, Telangana State, India.*

²*Professor, CVR College of Engineering, Hyderabad, Telangana State, India.*

bhavanisrirangam122@gmail.com

subhashchandra.n.cse@gmail.com

Abstract

K-Means clustering algorithm is one of the most popular unsupervised clustering algorithms which can be used for segmentation to analyze the data. It is an algorithm based on centroids, where the distances are calculated to assign a point to a cluster. Each cluster is associated with a centroid. The selection of initial centroids and the number of clusters play a major role to decide the performance of the algorithm. In this context, many researchers worked on, but they may not reach a goal to cluster the images in minimum runtime. Existing histogram based initial centroid selection methods are used on grayscale images only. Two methods, i.e., Histogram based initial centroids selection and Equalized Histogram based initial centroids selection to cluster colour images have been proposed in this paper.

The colour image has been divided into R, G, B, three channels and calculated histogram to select initial centroids for clustering algorithm. This method has been validated on three benchmark images and compared to the existing K-Means algorithm and K-Means++ algorithms. The proposed methods give an efficient result compared to the existing algorithms in terms of run time.

Keywords:

Equalized Histogram, Histogram, Initial Centroids, K-clusters, K-Means clustering, K-Means++ clustering.

1. Introduction

Image processing is defined as the analysis of an image with the application of complex algorithms. In computer vision, the procedure of dividing an image into different partitions is referred to as image segmentation. Computer vision has given a broad analysis of images using different segmentation methods. The ultimate goal of image segmentation is to convert the representation of an image into more usable, easier to process and analysable form. This paper presents image segmentation using clustering technique. A cluster refers to a grouping of data points aggregated together because of particular similarities. In clustering technique the data points are aggregated based on the distance measure in an n dimensional co-ordinating system. Each feature of an image can be considered as a dimension. Here as we are processing colour images, the intensity levels of red, green and blue values of each pixel are taken as dimensions or features of the images. K-Means clustering is the most simplest and popular unsupervised or semi supervised machine learning algorithm. Even though K-Means is said to be unsupervised clustering technique, we need to assume the K value that is the number of clusters initially and hence sometimes it is considered as semi supervised machine learning algorithm. In K-Means clustering, assuming the K value (number of clusters) and selecting initial points for centroids incorporate a significant role. For the selection of K , that is, the number of clusters, a popular method called "Elbow Method" gives optimum value, but the algorithm has to run several times and then only K value can be decided. There are many methods for initial cluster centre choice like "random data points", "K-Means++". The general procedure to determine the best partition and optimal number of clusters is by validation measures like Sum of Squared Error (SSE) [12], Silhouette Score [11], Calinski_Harabasz_Score [13], Davies_Bouldin_Score [1], Clustering Fitness and Run Time [11]. If the optimum number of clusters could be decided, the next step is to get the best center points or centroids for all clusters.

The goal of this paper is to propose a systematic centroids selection for K-Means clustering based on the histogram peaks that is high density data points to be clustered with in a single cluster initially, later the next level density etc. The selection of the centroids is chosen by sorting the histogram. After the selection of centroids, rest of the process is similar to random centroids method.

2. Related Work

In K-Means with random initial centroids method K number of random centroids or initial seeds is selected initially for k number of clusters. The algorithms start calculating the distance between a pixel point and all the centroids, and the pixel will be assigned to the cluster with a minimum distance. Once a new point is assigned, then a new centroid is obtained by taking mean of all data points of that cluster. This will be continued for all the data points. This procedure will be continued until there is no change in the previous centroid and new centroid for all the clusters [3].

D. T. Pham et. al. [11] says that, the optimal K value must be less than that of the objects in the image but this method could be expensive computationally if it is used for large data sets since it requires several times the K-Means algorithm to be applied before it can identify optimal value for K . To identify the optimal value for K , it is necessary that a set of values to be adopted instead of a single predefined K value.

Haimonti Dutta et. al. [7] presented a semi-supervised K-Means algorithm but the existence of noisy and small clusters in the given data turned to difficult to find the optimal choice of K and says, more sophisticated seeding techniques incorporating with machine learning algorithms are required.

Nameirakpam Dhanachandra et. al. [10] proposed subtractive clustering method on medical images to generate centroids based on the potential value of the data points. Here, author taken the number of clusters, $k=3$.

Zubair Khan et. al. [14] proposed an adaptive histogram-based approach to determine the initial

parameters for K-Means on grey images. Here, authors took initial parameter as a single variable known as grey level is used to assign intensity values to the pixels. It is a 2-step initial parameter estimation procedure to choose proper number of clusters and optimal initial cluster centres will give a better analysis on the data or image from which it can be decided that the K value as well as initial seeding of the algorithm, but the initialization problem of K-Means is used only for grey images. This can be extended for the colour images and the task of grouping the individual peaks in the histograms to represent the true colour and the object boundaries in the image.

Rena Nainggolan et. al. [13] said that, in manual choice of K, the algorithm has to be run many a times in order to get efficient clustering results. There are also a few methods like Silhouette Method, Hierarchical clustering method etc., are available for choice of K.

Raja Kishor Duggirala [12] proposed a clustering using fuzzy-logic; any one can identify if data objects are belonging fully or partially to the clusters based on their membership. K-Means Hybridized with the FCM (Fuzzy C-Means) to improve the performance. Hybrid algorithm of KM and FCM is showing an efficient performance in terms of execution time of the CPU, Clustering Fitness (CF) and Sum of Squared Error (SSE).

Bernad Jumadi Dehotman Sitompul et. al. [2] proposed a clustering method of deciding initial centroids for K-Means algorithm based on the minimum Sum of Squared Error and is able to improve the clustering result and enhanced Davies Bouldin Index (DBI) value obtained with less effort. Here, the authors used numerical data like Seeds Dataset and suggested future work for categorical and image data.

D. Arthur, [5] proposed K-Means++ method, where the centroids to be far away from each other, directing to better results than the random method of initialization. In this method, initially the first cluster centre has been chosen at random from data points, then for the next centre, each data point of the nearest cluster centre is chosen using squared distance method by using the probability formula (1). This step will be repeated until k number of centres has been chosen. The rest of the process is like random centroids method.

$$C_i = \frac{D(x)^2}{\sum_{x \in X} D(x)^2} \quad (1)$$

Md. Zakir Hossain et. al. [9] proposed a dynamic K-Means clustering algorithm. This algorithm first finds a threshold value using the dataset, based on this threshold value the algorithm decides further to continue to cluster or stop clustering.

Kristina P. Sinaga et. al. [8] proposed an unsupervised K-Means clustering algorithm; we know that entropy is basically used for dissimilarity and hence based on entropy, the number of clusters were found where there is no need of giving number of clusters priori.

Chunhui Yuan et. al. [3] proposed four types of K-value selection algorithms; they are Elbow Method, Gap Statistic, Silhouette Coefficient, and Canopy which used to cluster the Iris data set to find the optimal K value and then cluster the result of the data set. If we use incorporate large scale data sets both time and space complexity will be more to run the Gap Statistic algorithm. The processing complexity can be very large; hence the Silhouette Coefficient algorithm may not be used for large data sets. Especially for multidimensional data sets, Canopy algorithm recommended as the better option. For real-time multimedia data containing composite attributes of information and for practical implementation, there is a necessity of thorough inspection of pros and cons of each method or to improve the efficiency of the algorithm based on different parameters.

3. Proposed Methods

This section presents our approach to find efficient centroids selection with the application of Histogram based K-Means algorithm on 3 different images.

3.1. Pre-Processing of Image

Downloaded images are resized and each image is converted into a data frame in order to apply K-Means algorithm.

3.2. K-Means algorithm

A collection of data items (x_1, x_2, \dots, x_n) , where each item is a vector of m -dimension, K-Means clustering goals to deduct k ($\leq n$) clusters $C = \{C_1, C_2, \dots, C_k\}$ by dividing n items, in such a way to minimize the sum of square error (SSE). Formally, the objective is defined with the equation (2).

$$\arg \min C \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2 = \arg \min C \sum_{i=1}^k |C_i| \text{Var } C_i \quad (2)$$

Where, μ_i is the mean of points in C_i .

3.3. Histogram of an image

An image histogram can be defined as a histogram that presents a pictorial presentation of the pixel intensity distribution. [4]. It shows the no. of pixels per each intensity level. By observing histogram of an image, an observer will be able to critic the whole intensity spread immediately. The pseudocode for Histogram is presented as follows.

```
Algorithm: Histogram
Input: Grey image/Single Channel of an Image
Output: Histogram of the given image
Step 1: Declare an n dimensional array of histogram with 256 levels
Step 2: Read the shape of image into a variable
Step 3: For each row
Step 4:     For each column
Step 5:     Increment corresponding histogram level
            count in an n dimensional array of
            histogram
Step 6: Return an n dimensional array of histogram
```

3.4. Histogram Based K-Means

The given colour image split into the R, G, B channels and for each channel separate Histogram is generated. Each histogram is sorted in descending order to get high density of intensity. Based on the given k value, those many or several centroids selected from the sorted list of histograms. The selected centroids are given as input to the K-Means algorithm as initial centroids. The pseudo code for the Histogram Based K-Means is presented as follows.

```
Algorithm: Histogram Based K-Means
Input: Input Image, Number of Clusters
Output: Clustered Image
Step 1: Split Image into B, G, and R channels
Step 2: Call Histogram for each channel
Step 3: Sort histograms of each channel in descending order.
Step 4: Declare an n dimensional array of Initial Centroids
Step 5: For 0 to k clusters
Step 6:     Append centroids for each channel from sorted
            histograms to Initial Centroids
Step 7:     Call K-Means with Initial Centroids
Step 8:     Output the Clustered Image
```

3.5. Equalized Histogram of an image

Histogram equalization of an image is defined as a technique of adjusting intensities of the image in order to increase the contrast [6]. Let h is the given image represented in the form of an $R \times C$ matrix of pixel intensities of type integers vary between 0 to $L-1$. L is the number of feasible integer intensity value, up to 256. The normalized histogram (probability of each level of intensity) of h with a bin for each of the intensity level is defined by P . Hence, P_m is defined with the equation (3).

$$P_m = \frac{\text{number of pixels with intensity } m}{\text{total number of pixels}} \quad (3)$$

Where, $m=0, 1, 2, 3, \dots, L-1$.

The equalization of histogram for the image g can be defined using the equation (4).

$$g_{ij} = \text{floor}((L - 1) \sum_{m=0}^{h_{ij}} P_m) \quad (4)$$

Where $\text{floor}()$ is used to round to the nearest lower integer value. This has to be equated to the transformation of pixel intensity values, l in h using the equation (5).

$$T(l) = \text{floor}((L - 1) \sum_{m=0}^l P_m) \quad (5)$$

The pseudocode for Equalized Histogram is presented as follows.

```
Algorithm: Equalized Histogram
Input: Grey image/Single Channel of an Image
Output: Equalized Histogram of the given image
Step 1: Call Histogram of channel
Step 2: Read the shape of image into a variable
Step 3: Calculate Probability Distribution Function of
        Histogram
Step 4: Calculate Cumulative Distribution Function of
        Histogram
Step 5: Define transformation function
Step 6: For each row
Step 7:     For each column
Step 8:         Apply transformation function and store in
                resulting an n dimensional array as
                Equalized Histogram of channel
Step 9: Return Equalized Histogram of channel
```

3.6. Equalized Histogram based K-means

The given colour image split into the RGB channels and for each channel separate Equalized Histogram is generated. Each Equalized Histogram is sorted in descending order to get high density and intensity values. Based on the given k value, those several numbers of centroids selected from the sorted list of Equalized Histograms. The selected centroids are given as input to the K-Means algorithm as initial centroids. The pseudocode for the Equalized Histogram Based K-Means is presented as follows.

```
Algorithm: Equalized Histogram Based K-Means
Input: Input Image, Number of Clusters
Output: Clustered Image
Step 1: Split Image into B, G, and R channels
Step 2: Call Equalized Histogram for each channel
Step 3: Sort histograms of each channel in descending order.
Step 4: Declare an n dimensional array of Initial Centroids
```

- Step 5: For 0 to k clusters
- Step 6: Append centroids for each channel from sorted histograms to Initial Centroids
- Step 7: Call K-Means with Initial Centroids
- Step 8: Output the Clustered Image.

4. Result and Discussion

The proposed system is experimented using <https://colab.research.google.com>, an open source for python programming.

4.1. Performances Analysis

The present proposed models are experimented on 3 different images and Runtimes of all clustering technique are measured. Comparison of results of Runtime in seconds of each clustering technique is listed in Table 1, the corresponding line graphs are also presented in figure 1, figure 2, and figure 3. Original images are presented in figure 4, figure 5, and figure 6. The clustering results for 'k' values 2 and 8 are presented in tables from 2 to 4.

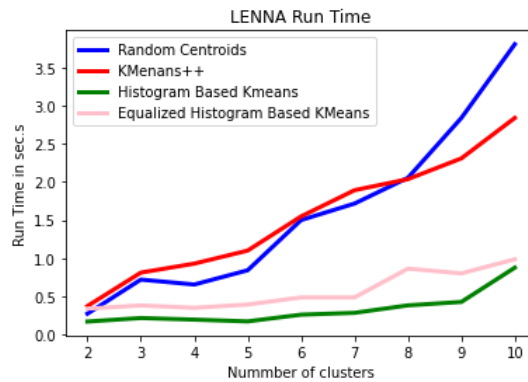


Figure 1: Run Time in Seconds for Lena Image

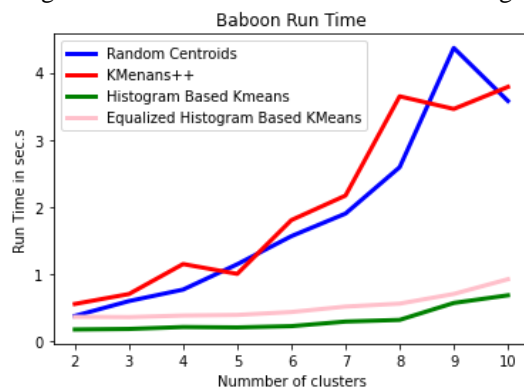


Figure 2: Run Time in Seconds for Baboon Image

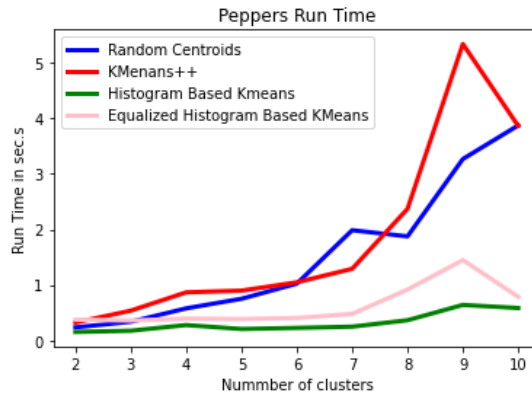


Figure 3: Run Time in Seconds for Peppers Image

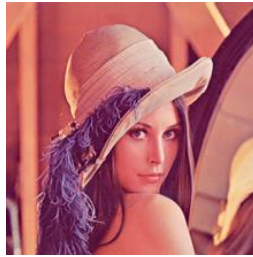


Figure 4: Original Lena Image

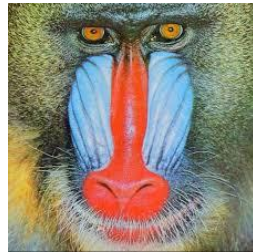


Figure 5: Original Baboon Image



Figure 6: Original Peppers Image

Table 1: Comparison of Runtimes in Seconds

K	Lena Image				Baboon Image				Peppers Image			
	K-Means with Random Centroids	K-Means++	Histogram based K-Means	Equalized Histogram based K-Means	K-Means with Random Centroids	K-Means++	Histogram based K-Means	Equalized Histogram based K-Means	K-Means with Random Centroids	K-Means++	Histogram based K-Means	Equalized Histogram based K-Means
2	0.273	0.373	0.17	0.339	0.378	0.556	0.176	0.365	0.235	0.322	0.152	0.374
3	0.72	0.811	0.216	0.381	0.599	0.704	0.184	0.36	0.335	0.536	0.174	0.357
4	0.655	0.929	0.195	0.35	0.77	1.15	0.213	0.384	0.577	0.866	0.276	0.393
5	0.842	1.1	0.172	0.393	1.147	1.004	0.207	0.393	0.75	0.897	0.205	0.385

6	1.5	1.548	0.26	0.486	1.565	1.804	0.226	0.436	1.022	1.047	0.225	0.404
7	1.716	1.891	0.284	0.486	1.901	2.169	0.294	0.517	1.986	1.29	0.247	0.477
8	2.054	2.036	0.382	0.863	2.591	3.646	0.319	0.561	1.875	2.371	0.363	0.916
9	2.838	2.308	0.427	0.801	4.366	3.458	0.571	0.705	3.269	5.343	0.639	1.448
10	3.804	2.838	0.876	0.986	3.577	3.786	0.685	0.924	3.875	3.865	0.583	0.785

Table 2: Lena Image Results

K	K-Means with Random Centroids	K-Means++	Histogram based K-Means	Equalized Histogram based K-Means
2				
8				

Table 3: Baboon Image Results

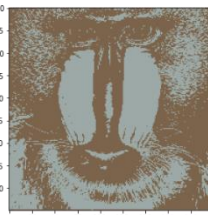
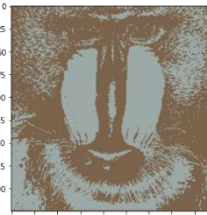
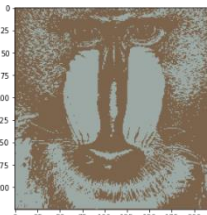
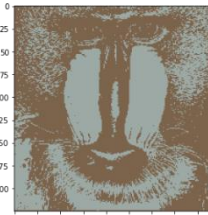
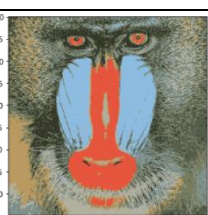
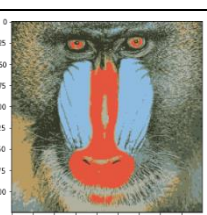
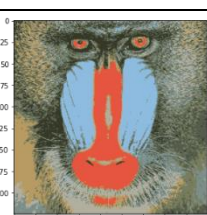
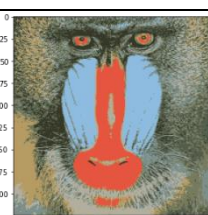
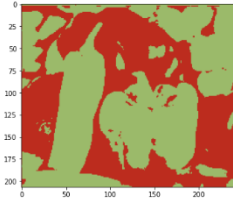
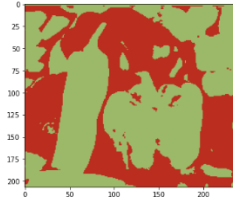
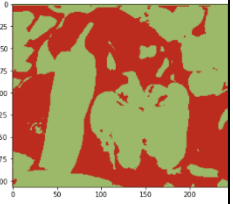
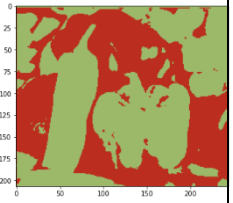
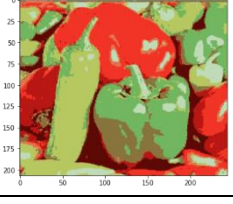
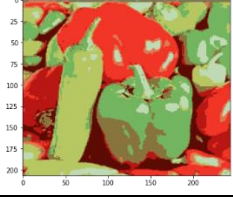
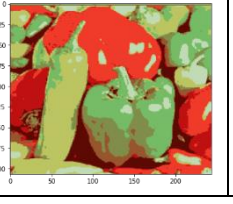
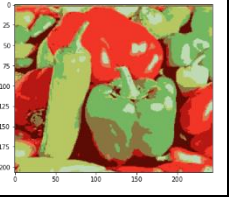
K	K-Means with Random Centroids	K-Means++	Histogram based K-Means	Equalized Histogram based K-Means
2				
8				

Table 4: Peppers Image Results

K	K-Means with Random Centroids	K-Means++	Histogram based K-Means	Equalized Histogram based K-Means
2				
8				

5. Conclusion

Clustering is playing a vital role in image segmentation which used in many applications. The most commonly used K-Means clustering which takes randomly generate initial centroids is not reaching the local optima. The proposed Histogram based selection of initial centroids to overcome the above drawback. The present paper also proposed equalized histogram-based selection of initial centroids to improve the performance of algorithm. But, in the above analysis, the equalized histogram method is not performing appreciate results. This can be enhanced by considering spatial values as another dimension in clustering.

It is observed that Histogram based, and Equalized Histogram based K-Means giving better performance in the view of Run Time comparatively with Random centroids K-Means and K-Means++. Histogram based K-Means is taking less runtime as compared to the Equalized Histogram based K-Means, but in future, it can be extended that Equalized Histogram based K-Means may perform better in other kind of parameters like Sum of Squared Error, Silhouette Score etc., with the application of spatial dimensions.

References

- [1] Ahmet Esad TOP, F. Şükrü TORUN & Hilal KAYA, PARALLEL K-MEANS CLUSTERING WITH NAÏVE SHARDING FOR UNSUPERVISED IMAGE SEGMENTATION VIA MPI, Mühendislik Bilimleri ve Tasarım Dergisi 8(3), 791 – 798, 2020, e-ISSN: 1308-6693, Journal of Engineering Sciences and Design DOI: 10.21923/jesd.748209.
- [2] Bernad Jumadi Dehotman Sitompul, Opim Salim Sitompul and Poltak Sihombing, Enhancement Clustering Evaluation Result of Davies-Bouldin Index with Determining Initial Centroid of K-Means Algorithm, Journal of Physics: Conference Series, Volume 1235, The 3rd International Conference on Computing and Applied Informatics 2018 18–19 September 2018, Medan, Sumatera Utara, Indonesia, 1-6.
- [3] Chunhui Yuan & Haitao Yang, “Research on K-Value Selection Method of K-Means Clustering Algorithm”, (2019), J. 2. 226-235. 10.3390/j2020016, doi: <https://doi.org/10.3390/j2020016>.
- [4] Data Mining – Concepts and Techniques - Jiawei Han & Micheline Kamber, Morgan Kaufmann Publishers, 3rd Edition, 2012.
- [5] David Arthur and Serei Vassilvitskii, 2007, k-means++: The Advantages of Careful Seeding, In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, SIAM, pp. 1027-1035.
- [6] Digital Image Processing, Rafael C. Gonzalez and Richard E. Woods, Fourth edition. Pearson

Education, 2018.

- [7] Haimonti Dutta, Rebecca J. Passonneau, Austin Lee, Axinia Radeva, Boyi Xie, David Waltz and Barbara Taranto, Learning Parameters of the K-Means Algorithm from Subjective Human Annotation, Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference, 2011, 465-470.
- [8] Kristina P. Sinaga and Miin-Shen Yang, Unsupervised K-Means Clustering Algorithm, IEEE Access, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
- [9] Md. Zakir Hossain, Md. Nasim Akhtar, R.B. Ahmad and Mostafijur Rahman, A dynamic K-means clustering for data mining, Indonesian Journal of Electrical Engineering and Computer Science, Vol. 13, No. 2, February 2019, pp. 521~526 ISSN: 2502-4752, DOI: 10.11591/ijeecs.v13.i2.pp521-526
- [10] Nameirakpam Dhanachandra, Khumanthem Manglem, Yambem Jina Chanu, Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm, Procedia Computer Science, Volume 54, 2015, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2015.06.090>, 764-771.
- [11] Pham D, Dimov, Stefan & Nguyen Cuong, Selection of K in K -means clustering, Manufacturing Engineering Centre, Cardiff University, Cardiff, UK 2004, 103-119, <https://doi.org/10.1243/095440605X8298>
- [12] Raja Kishor Duggirala, Segmenting Images Using Hybridization of K-Means and Fuzzy C-Means Algorithms, Introduction to Data Science and Machine Learning, Keshav Sud, Pakize Erdogmus and Seifedine Kadry, IntechOpen, (July 10th 2019), 1-27, DOI: 10.5772/intechopen.86374. Available from: <https://www.intechopen.com/chapters/68050>.
- [13] Rena Nainggolan, Resianta Perangin-angin, Emma Simarmata, and Feriani Astuti Tarigan, , Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method, Journal of Physics: Conference Series, 2019, doi: 10.1088/1742-6596/1361/1/012015, 1-6.
- [14] Zubair Khan, Jianjun Ni, Xinnan Fan & Pengfei Shi, An improved K-means clustering algorithm based on an adaptive initial parameter estimation procedure for image segmentation. International Journal of Innovative Computing, Information and Control, 2017, 1509-1525.
- [15] <https://www.kdnuggets.com/2017/03/naive-sharding-centroid-initialization-method.html>
- [16] https://en.wikipedia.org/wiki/K-means_clustering
- [17] https://en.wikipedia.org/wiki/Image_histogram