

## **Analysis of K-Means Clustering Based Image Segmentation**

Priya P, Dony A.D'souza

PG Scholar, Department of ECE, VTU University, MITE Moodbidre, Karnataka  
Assistant Professor Department of ECE, VTU University, MITE Moodbidre, Karnataka

---

**Abstract:** This paper presents a new approach for image segmentation by applying k-means algorithm. In image segmentation, clustering algorithms are very popular as they are intuitive and are also easy to implement. The K-means clustering algorithm is one of the most widely used algorithm in the literature, and many authors successfully compare their new proposal with the results achieved by the k-Means. This paper proposes a color-based segmentation method that uses K-means clustering technique. The k-means algorithm is an iterative technique used to partition an image into k clusters. The standard K-Means algorithm produces accurate segmentation results only when applied to images defined by homogenous regions with respect to texture and color since no local constraints are applied to impose spatial continuity. At first, the pixels are clustered based on their color and spatial features, where the clustering process is accomplished. Then the clustered blocks are merged to a specific number of regions. This approach thus provides a feasible new solution for image segmentation which may be helpful in image retrieval. The experimental results clarify the effectiveness of our approach to improve the segmentation quality in aspects of precision and computational time. The simulation results demonstrate that the proposed algorithm is promising.

**Keywords:** K-means Algorithm, Clustering, local minimum, global minimum, Segmentation.

---

### **I. Introduction**

Images are considered as one of the most important medium of conveying information. Understanding images and extracting the information from them such that the information can be used for other tasks is an important aspect of Machine learning. One of the first steps in direction of understanding images is to segment them and find out different objects in them. Thus image segmentation plays a vital role towards conveying information that is represented by an image and also assists in understanding the image. Image segmentation is the process of dividing the given image into regions homogenous with respect to certain features, and which hopefully correspond to real objects in the actual scene. Segmentation plays a vital role to extract information from an image to create homogenous regions by classifying pixels into groups thus forming regions of similarity. The homogenous regions formed as a result of segmentation indwell pixels having similarity in each region according to a particular selection criterion e.g. Intensity, color etc. Segmentation plays an important role in image understanding, image analysis and image processing. Because of its simplicity and efficiency, clustering approaches were one of the first techniques used for the segmentation of (textured) natural images [1]. After the selection and the extraction of the image features [usually based on color and or texture and computed on (possibly) overlapping small windows centered around the pixel to be classified], the feature samples, handled as vectors, are grouped together in compact but well-separated clusters corresponding to each class of the image. The set of connected pixels belonging to each estimated class thus defined the different regions of the scene. The method known as k-means [2] (or Lloyd's algorithm). The applications of Image segmentation are widely in many fields such as image compression, image retrieval, object detection, image enhancement etc.

### **II. Image Segmentation**

The main idea of the image segmentation is to group pixels in homogeneous regions and the usual approach to do this is by 'common feature. Features can be represented by the space of color, texture and gray levels, each exploring similarities between pixels of a region. Segmentation [1] refers to the process of partitioning a digital image into multiple regions (sets of pixels). The goal of segmentation is to simplify and change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. The result of image segmentation is a set of regions that collectively cover the entire image, or a set of contours extracted from the image. Each of the pixels in a region is similar with respect to some characteristic or computed property, such as color, intensity, or texture [3], [4]. The segmentation is based on the measurements taken from the image and might be grey level, color, texture, depth or motion. Image segmentation techniques are categorized into three classes: Clustering, edge detection; region growing. Some popular clustering algorithms like k-means are often used in image segmentation [5] adjacent regions are significantly different with respect to

the same characteristic(s). Segmentation is mainly used in medical imaging, Face recognition, Fingerprint recognition, Traffic control systems, Brake light detection, and Machine vision.

### III. Clustering

Clustering refers to the process of grouping samples so that the samples are similar within each group. The groups are called clusters. Clustering is a data mining technique used in statistical data analysis, data mining, pattern recognition, image analysis etc. Different clustering methods include hierarchical clustering which builds a hierarchy of clusters from individual elements. Because of its simplicity and efficiency, clustering approaches were one of the first techniques used for the segmentation of (textured) natural images [6]. In partitional clustering; the goal is to create one set of clusters that partitions the data in to similar groups. Other methods of clustering are distance based according to which if two or more objects belonging to the same cluster are close according to a given distance, then it is called distance based clustering. In our work we have used K-means clustering approach for performing image segmentation using Matlab software. A good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. The quality of clustering result depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns. Image Segmentation is the basis of image analysis and understanding and a crucial part and an oldest and hardest problem of image processing. Clustering means classifying and distinguishing things that are provided with similar properties [7]. Clustering techniques classifies the pixels with same characteristics into one cluster, thus forming different clusters according to coherence between pixels in a cluster. It is a method of unsupervised learning and a common technique for statistical data analysis used in many fields such as pattern recognition, image analysis and bioinformatics.

#### 1.1. K Means Overview

There are always K clusters. There is always at least one item in each cluster. The clusters are non-hierarchical and they do not overlap. Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'centre' of clusters [9]. K-means clustering in particular when using heuristics such as Lloyd's algorithm is rather easy to implement and apply even on large data sets. As such, it has been successfully used in various topics, ranging from market segmentation, computer vision and astronomy to agriculture. It often is used as a preprocessing step for other algorithms, for example to find a starting configuration. In statistics and data mining, k-means clustering is a method of cluster analysis which aims to partition  $n$  observations into  $k$  clusters in which each observation belongs to the cluster with the nearest mean. Figure1 shows the flow chart of k-means algorithm which is relatively efficient and applicable only when mean is defined.

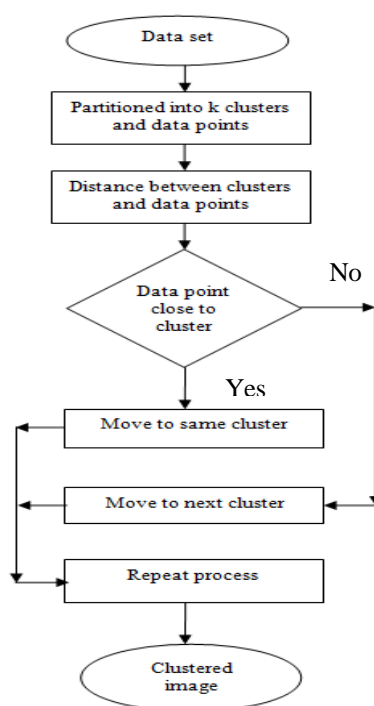


Figure1. Flow chart for K-Means Algorithm

## 1.2. K-Means Clustering

K-means (Macqueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering (assume  $k$  clusters) fixed a priori. The main idea is to define  $k$  centroids, [7] one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to recalculate  $k$  new centroids as barycenters [8] of the clusters resulting from the previous step. After we have these  $k$  new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the  $k$  centroids change their location step by step until no more changes are done. In other words centroids do not move anymore. K-Means clustering generates a specific number of disjoint, flat clusters. K-Means method is numerical, unsupervised, non-deterministic and iterative. Hierarchical clustering is also widely employed for image segmentation [12], [13]. The most popular method for image segmentation is k-means clustering.

## IV. Feature Extraction

The issue of choosing the features to be extracted should be guided by the following concerns. The features should carry enough information about the image and should not require any domain-specific knowledge for their extraction. They should be easy to compute in order for the approach to be feasible for large image collection and rapid retrieval. An image is partitioned into  $4 \times 4$  block, a size that provides a compromise between texture granularity, computation time and segmentation coarseness [8] As a part of preprocessing, each  $4 \times 4$  block is replaced by a single block containing the average value over the  $4 \times 4$  block. To segment an image into objects, some features are extracted from each block. Texture features are extracted using Haar Wavelet Transform. After obtaining features from all pixels on the image, perform k-means clustering to group similar pixel together and form objects. Feature extraction has been done using MATLAB Image Processing tool. The advantage of k-means algorithm is that it works well when clusters are not well separated from each other, which is frequently encountered in images. However k-means requires the user to specify the initial cluster centers. Image clustering consists of two steps, the former is feature extraction and the second part is grouping. For each image in the database, a feature vector capturing certain essential properties of the image is computed and stored in a feature base. Clustering algorithm is applied over this extracted feature to form the group. In terms of performance the algorithm is not guaranteed to return a global optimum. The quality of the final solution depends largely on the initial set of clusters, and may, in practice, be much poorer than the global optimum. Since the algorithm is extremely fast, a common method is to run the algorithm several times and return the best clustering found.

## V. Cluster Algorithm

K-Means uses a two-phase iterative algorithm to minimize the sum of point-to-centroid distances, summed over all  $k$  clusters: The first phase uses *batch updates*, where each iteration consists of reassigning points to their nearest cluster centroid, all at once, followed by recalculation of cluster centroids. This phase occasionally does not converge to solution that is a local minimum, that is, a partition of the data where moving any single point to a different cluster increases the total sum of distances. This is more likely for small data sets [10, 11]. The batch phase is fast, but potentially only approximates a solution as a starting point for the second phase. The second phase uses *online updates*, where points are individually reassigned if doing so will reduce the sum of distances, and cluster centroids are recomputed after each reassignment. Each iteration during the second phase consists of one pass though all the points. The second phase will converge to a local minimum, although there may be other local minima with lower total sum of distances. The problem of finding the global minimum can only can be solved in general by an exhaustive (or clever, or lucky) choice of starting points, but using several replicates with random starting points typically results in a solution that is a global minimum.

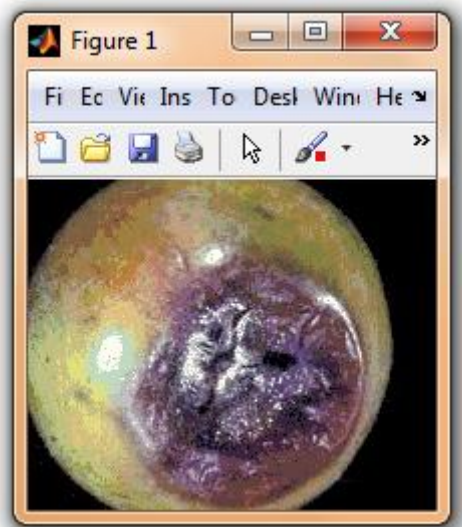
## 1.3. K Means Function

K-means is a clustering algorithm, which partitions a data set into clusters according to some defined distance measure. Images are considered as one of the most important medium of conveying information. Understanding images and extracting the information from them such that the information can be used for other tasks is an important aspect of Machine learning. An example of the same would be the use of images for navigation of robots. One of the first steps in direction of understanding images is to segment them and find out different objects in them. To do this, we look at the algorithm namely K-means clustering. It has been assumed that the number of segments in the image is known and hence can be passed to the algorithm [9], [10]. K-Means algorithm is an unsupervised clustering algorithm that classifies the input data points into multiple classes based on their inherent distance from each other. The algorithm assumes that the data features form a

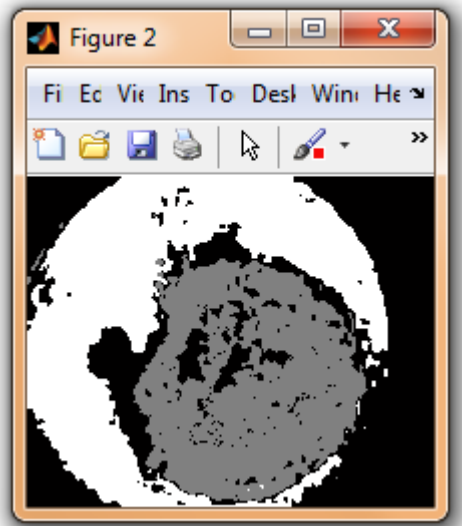
vector space and tries to find natural clustering in them. The functions of k-means are as follows.  $IDX = kmeans(X, k)$  partitions the points in the  $n$ -by- $p$  data matrix  $X$  into  $k$  clusters. This iterative partitioning minimizes the sum, over all clusters, of the within-cluster sums of point-to-cluster-centroid distances. Rows of  $X$  correspond to points, columns correspond to variables. The function  $kmeans$  returns an  $n$ -by-1 vector  $IDX$  containing the cluster indices of each point. By default,  $kmeans$  uses squared Euclidean distances [8, 9]. When  $X$  is a vector,  $kmeans$  treats it as an  $n$ -by-1 data matrix, regardless of its orientation.  $[IDX, C] = kmeans(X, k)$  returns the  $k$  cluster centroid locations in the  $k$ -by- $p$  matrix  $C$ .

## VI. Experimental Result

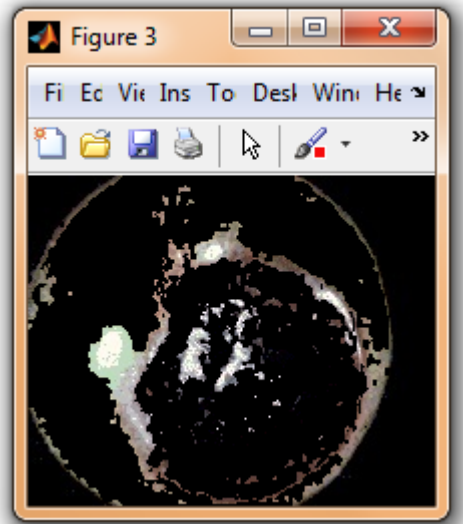
Figure 2 represents the various outcomes of k-means clustering algorithm. Figure 2(a) represents the original image 'rotapple.jpg'. Figure 2(b) represents the image labelled by its cluster index and figure 2(c), figure 2(d), and figure 2(e) shows the objects in cluster 1, 2 and 3 respectively. However image segmentation is a key step for understanding the image, which is a natural manner to obtain high level semantic.



(a)



(b)



(c)



(d)



(e)

Figure2. K-Means Clustering Observations

## VII. Conclusion

We have successfully implemented k-means clustering algorithm. For smaller values of k the algorithms give good results. For larger values of k, the segmentation is very coarse; many clusters appear in the images at discrete places. This is because Euclidean distance is not a very good metric for segmentation processes. Different initial partitions can result in different final clusters. Hence it is necessary to re-run the code several number of times for same and different values of k in order to compare the quality of clusters obtained. The result aims at developing an accurate and more reliable image which can be used in locating tumors, measure tissue volume, face recognition, finger print recognition and in locating an object clearly from a satellite image and in more[9]. The advantage of K-Means algorithm is simple and quite efficient. It works well when clusters are not well separated from each other. This could be happen in web images. We proposed a framework of unsupervised clustering of images based on the colour feature of the image. It minimizes intra-cluster variance, but does not ensure that the result has a global minimum of variance.

## References

- [1] J. A. Hartigan and M. A. Wong, "A *K-Means Clustering Algorithm*", Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, No. 1 (1979), pp. 100-108
- [2] S. P. Lloyd, "*Least squares quantization in PCM*", IEEE Trans. Inf.Theory, vol. IT-28, no. 2, pp. 129–136, Mar.1982.
- [3] J. Shi and J. Malik, "*Normalized cuts and image segmentation*", IEEE Trans. Pattern Anal.Mach. Intell., vol. 22, no. 8, pp. 888–905, Aug.2000.
- [4] S.Mary Praveena and Dr.IlaVennila, "*Optimization Fusion Approach for Image Segmentation using K-Means Algorithm*", International Journal of Computer Applications (0975 – 8887) Volume 2 – No.7, June 2010
- [5] M. Brindha and M. Annie Flora, "*A Label Field Fusion Using ICM Optimization in Image Segmentation*", IJMIE, volume 2, issue 12, ISSN: 2249-0558, December 2012.
- [6] Julian Besag, "*On the Statistical Analysis of Dirty Pictures*",Journal of the Royal Statistical Society, Series B (Methodological), Vol. 48, No. 3. (1986), pp. 259-302.
- [7] J.P. Braquelaire and L. Brun, "Comparison and optimization of methods of color image quantization," IEEE Trans. Image Process., vol. 6, no. 7, pp. 1048–1952, Jul. 1997.
- [8] H. Stokman and T. Gevers, "Selection and fusion of color models for image feature detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 3, pp. 371–381, Mar. 2007.
- [9] H. Stokman and T. Gevers, "Selection and fusion of color models for image feature detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 3, pp. 371–381, Mar. 2007.
- [10] Seber, G. A. F. *Multivariate Observations*. Hoboken, NJ: John Wiley & Sons, Inc., 1984.
- [11] Chinki Chandhok, "*a novel approach to image segmentation using artificial neural networks and k-means clustering*", International Journal Of Engineering Research and Applications,ISSN:2248-9622,www.ijera.com, vol.2,issue 3,may-jun 2012,pp.,274-279.
- [12] T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, & A.Y.Wu (2002) "An efficient k-means clustering algorithm: Analysis and implementation" Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.881-892.
- [13] M.Betke, N.C Makris, "*Fast Object Recognition in Noisy Images Using Simulated Annealing*", Intl. Conf on Computer Vision, pp. 523, Jun.20-23,1995.