



**RESEARCH ARTICLE**

**AUTOMATIC INITIALIZATION of IMAGE CLUSTERING ALGORITHMS**

Merve ARSLAN<sup>1,\*</sup>, Recep DEMİRÇİ<sup>2</sup>

<sup>1</sup>Gazi University, Graduate School of Natural and Applied Sciences, Computer Engineering, Ankara,  
[merve.arslan6@gazi.edu.tr](mailto:merve.arslan6@gazi.edu.tr), ORCID:0000-0002-2867-6198

<sup>2</sup>Gazi University, Technology Faculty, Computer Engineering, Ankara, [rdemirci@gazi.edu.tr](mailto:rdemirci@gazi.edu.tr), ORCID:0000-0002-3278-0078

*Receive Date:14.10.2022*

*Accepted Date: 15.11.2022*

**ABSTRACT**

Clustering is partition of a data set into subsets where each item in assigned subset is similar and different from that of other subsets. K-means and fuzzy c-means (FCM) algorithms are frequently used for clustering of color image. On the other hand, randomly determination of initial cluster centers is one of the most important problems of both algorithms since results to be obtained vary according to initial values of cluster centers. Thus, obtaining different results at each run time reduces reliability of algorithms. One of a typical solution is that number of iterations is increased in order to obtain an accurate result. However, it increases computation cost. A novel solution for initial cluster centers has been proposed in this study where octree algorithm was used. Color images were initially quantized in certain numbers of color vectors depending on level of octree algorithm. Then, means of each quantized color vector set were obtained. The pixel numbers of each pre-subset were sorted and assigned as initial cluster centers. Consequently, cluster centers are selected automatically. As positions of quantized vectors in color space are fixed, a deterministic algorithm has been attained.

**Keywords:** *K-means, Fuzzy c-means, Octree, Color Quantization.*

**1. INTRODUCTION**

Image segmentation is considered the first stage of image processing area. It is defined as dividing an image into sub-regions, each of which contains meaningful features. Since it acts as a pre-processor in image processing studies, performing correct segmentation is very important for carrying out of further image analysis. K-means and fuzzy c-means approaches are frequently used to cluster images. Although both algorithms stated are highly preferred for color classification, their performance was insufficient in some cases. The random assignment of initial cluster centers of related algorithms and determination of number of clusters by user affect performances of algorithms. Since the initial cluster centers are randomly allocated, outputs of algorithms could be different for each run time. Thus, it is undesired properties of the algorithms.

Segmentation process is used in many areas such as health, traffic, industry, object recognition, and face recognition. For example, K-means algorithm is a frequently used segmentation algorithm in computer vision although it also has weaknesses. Determination of cluster number, K before algorithm starts is subjective as it depends on user judgment. K-means algorithm is extremely sensitive to initial conditions as it produces different results based on initial cluster centers [1]. Fuzzy c-means algorithm is also very sensitive to initial cluster centers and number of clusters. There is no generally accepted method for initializing the algorithm. Kim et al. proposed a dominant color-based approach. Idea behind the related solution is that dominant colors in image could not be belonging to the same cluster. So, if dominant colors in image is determined in advance, they could be assigned for initial centers [2]. Dörterler et al. dealt with the problem of randomly generating center values for K-means algorithm. It is combined with metaheuristic algorithms (DEA: differential evolution algorithm and HSA: harmony search algorithm) to increase performance of clustering success. For example, parameters used for diagnosis of heart disease were selected with metaheuristic algorithms. It is aimed to increase the clustering success of K-means algorithm [3].

Separation of revealing region in skin image for psoriasis diagnosis is an example of segmentation in field of health. In the study carried out by Juang and Wu, K-means algorithm was employed to complete diagnosis procedure for psoriasis disease [4]. Segmentation of brain MRI images with K-means algorithm for detection of tumors was realized. In the method proposed by Hrosik et al., the firefly algorithm was used to find optimal centroids. K-means algorithm and firefly algorithm was combined to improve performance of process [5]. In a study conducted by Nitta et al., peaks in histogram of grayscale image were detected and then assigned as initial cluster centers of K-means algorithm. Consequently, random assignment of initial values of cluster centers was partially prevented. Additionally, the suggested algorithm was tested with MR images [6].

A segmentation process was applied to fish images for disease detection and animal behavior. K-means algorithm was employed to accurately separate fish image from background. In a study conducted by Yao et al., number of peaks in histogram of gray level image was used as number of cluster. The first cluster centers were determined with Otsu thresholding method [7]. An automatic vehicle license plate recognition technique was recently developed for intelligent transportation systems where K-means algorithm was used for segmentation of license plate image. Accordingly, characters on license plates were recognized by using a convolutional neural network (CNN) [8].

An initialization method for fuzzy c-means algorithm was proposed by Tan et al. to select the first cluster center and determine the number of clusters. The proposed hierarchical approach consists of two levels. Firstly, partition module in which splitting technique is used to divide image into small homogeneous region is implemented. In the second stage, unification technique is applied, and merging procedure is completed to find the number of clusters and initial cluster centers. In merging procedure, Manhattan distance between two close clusters is calculated. An iterative operation is performed by comparing Manhattan distance with a predetermined threshold value [9].

Recently, noisy color images were segmented by means of fuzzy c-means algorithm. Initially, noise level was estimated and then eliminated. Finally, segmentation process was realized with fuzzy c-means algorithm and block matching [10]. Segmentation of gray-scale images can be reasonably

realized by using Otsu and Kapur threshold methods. However, since there are three channels such as red, green, and blue, it is a problem to unify results obtained from different channels. In the method proposed by Demirci et al., threshold values were calculated for each color channel. The color space is divided into small cubes. The remaining pixels in each cube are included in the same group. With the suggested method, color images are automatically classified into eight classes [11].

A study combining fuzzy c-means and color deconvolution method was carried out to increase segmentation outcome. The peaks in image histogram are determined. If the number of peaks is greater than two, layer segmentation based on fuzzy c-means is performed. If the number of peaks is two or less, color deconvolution method is applied. Therefore, peaks in histogram determine the method to be used. Moreover, mouse skin images were used to test the proposed algorithm [12]. In fact, image segmentation is a color quantization process. Aim of color quantization is to reduce the number of colors in an image by accumulating similar colors in the same cluster. Subsequently, methods used in color quantization are grouped under two headings: clustering and splitting [13]. On other hand, octree algorithm is one of the well-known color quantization techniques. Park and Kim have studied K-means algorithm and octree color quantization method. Different palettes with 8, 16, 32, 64, 128, or 256 colors were investigated [14].

Recently, an entropy-based technique for initialization of K-means algorithm has been proposed by Chowdhury et al. In the related study, partition coefficient, classification entropy, partition index, and separation index measures of data set were used to calculate number of clusters. Furthermore, performance of suggested algorithm depends on threshold value that is selected by user [15]. Also K-means algorithm initialization strategy proposed by Cao et al. is based on neighborhood concepts. Thus, two levels were defined as intra-cluster similarity and inter-cluster similarity. As all vectors in data set are processed, complexity of the scheme is  $O(n^2)$  [16]. Performance analyzes of various K-means initialization techniques were completed by Çelebi et al. Eight popular initialization methods including Forgy's method, Jancey's method, and maximin have been compared. As a result, it was stated that Forgy's method, MacQueen's second method, and maximin method were slow in terms of convergence [17]. Segmentation procedures using Ham, Otsu, and Kapur thresholding algorithms were realized by Kılıçaslan et al. Unlike traditional methods, a segmentation process was implemented based on color space. In the related study, color ranges were initially determined by thresholding approaches, then color space was divided into 8 sub-cubes where each cube corresponds to a cluster [18].

In this study, a new method to automatically determine initial cluster centers of K-means and fuzzy c-means algorithms has been proposed. In the suggested procedure, color image to be segmented was initially quantized with octree algorithm and then means of each subset were obtained. Accordingly, one dimensional histogram of color image was attained by using pixel numbers assigned to particular subset. Once histogram data was sorted, the biggest subset was assigned as initial center of first cluster. Then, the process was repeated depending on desired cluster number of K-means and fuzzy c-means algorithms.

## 2. CLUSTERING ALGORITHMS

Clustering is an algorithm that involves partition process of given data points. Thus, clustering procedures are used to classify each data point into specific subgroups. Data points or vectors that are assigned to the same group should have similar features, while data vectors in different groups should have highly dissimilar features. On other hand, vector quantization is a lossy encoding method used for storing and transmitting data. Vector quantization is used to represent original data with fewer data points. It is desired to minimize data loss while quantizing. The similarity rate of the data around the same cluster center should be high, and the similarity rate between the data belonging to different cluster centers should be low [19]. Color image is a collection of pixels that are spatially distributed in plane. On the other hand, each pixel in image has three components that represent coordinates in RGB color space. Therefore, a color image is also data points or vectors in color space. When it is desired to quantize color images, a vector or data point clouds in RGB space are taken as input. At the end of clustering or color classification, a color image is represented with fewer colors [20]. When the size of color image increases, the computation cost of quantization process also grows.

### 2.1. K-means Algorithm

In digital environment, data set is basically colocation of vectors or data points in relevant space. Labeling of every data point under any of subsets could be defined as clustering. K-means algorithm, which is one of traditional approaches was developed by MacQueen in 1967 [21]. The goal of the algorithm is to minimize the objective function which is defined as follow:

$$J = \sum_{i=1}^n \sum_{j=1}^K \|x_i - c_j\|^2 \quad (1)$$

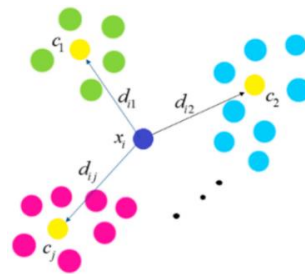
where  $n$  represents the number of data points or vectors,  $K$  is the number of clusters or subsets,  $x_i$  is data point or data vector,  $c_j$  is centroid of cluster. Furthermore, Euclidean distance between centroids of clusters and data points is calculated as:

$$d_{ij} = \|x_i - c_j\| \quad (2)$$

The number of clusters is determined by user and the centroid of cluster centers are initialized randomly. Mathematical target of K-means algorithm is to minimize the objective function defined in Eq.1. The specified goal is achieved by moving the centroids of clusters in related space as shown in Figure 1. Thus, positions of centroids of clusters are randomly initialized at the beginning and then the updated with proper algorithm. Consequently, fundamental steps of K-means algorithm are:

- Step 1: Determine number of clusters ( $K$ ),
- Step 2: randomly select initial cluster centers,
- Step 3: calculate Euclidean distances between centroids of clusters and all data points and classifies each point into the category of the nearest cluster center,
- Step 4: recalculate the mean of each cluster as a new cluster center,

Step 5: if there is no change in the cluster centers or the specified number of iterations is reached, terminate the algorithm; otherwise, it is repeated from step 3.



**Figure 1.** K-means algorithm.

### 2.2. Fuzzy C-means Algorithm

K-means is one of hard clustering algorithms since a data point is classified into only one of categories. Euclidean distance between a data point and the nearest cluster center is evaluated. The relationship between the particular data point and other classes is not considered. So, there is a kind of binary relation. On other hand, the data point to be classified may belong to more than one class in fuzzy c-means (FCM) algorithm which was established by Bezdek in 1981 [22]. The class to which the data point belongs is determined by membership degrees. Accordingly, the objective function of the FCM algorithm is defined as:

$$J = \sum_{i=1}^n \sum_{j=1}^K u_{ij}^m \|x_i - c_j\|^2 \quad (3)$$

where  $n$  represents the number of data points,  $K$  is the number of clusters,  $x_i$  is data points,  $c_j$  is centroid of cluster. The degree of membership between data point,  $x_i$  and cluster center,  $c_j$  calculated as:

$$u_{ij} = \frac{1}{\sum_{k=1}^K \left( \frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \quad (4)$$

where  $m$  is fuzziness index greater than 1. It is also chosen by user. In K-means algorithm, a data point belongs to only one cluster while in fuzzy c-means algorithm, the membership value is distributed for all classes. Accordingly, the sum of membership values of data point,  $x_i$  must be 1 as follows:

$$\sum_{i=1}^K u_{ij} = u_{i1} + u_{i1} + \dots + u_{ij} = 1 \quad (5)$$

Furthermore, the cluster centers are updated by means of:

$$c_j = \frac{\sum_{i=1}^n u_{ij}^m x_i}{\sum_{i=1}^n u_{ij}^m} \quad (6)$$

Nevertheless, the fuzzy c-means algorithm has the same problems as the K-means algorithm since the centroids of clusters are randomly determined at the beginning. So, the performance of the algorithm depends on initial parameters.

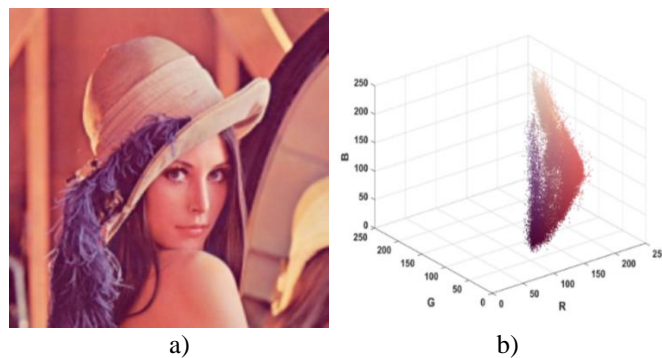
### 2.3. Octree Algorithm

Octree algorithm was developed in 1988 [23]. It is one of the vector quantization methods in which the number of vectors representing data set is reduced to certain numbers. Color images that have three channels are collections of pixels scattered in spatial plane as shown in Figure 2(a). On the other hand, pixels in color image are also corresponding to a point in RGB color space. Accordingly, a color image shown in Figure 2(a) builds point clouds or data points as shown in Figure 2(b). Each channel in color image is coded with 8 bits memory area, color images may have 16,777,216 different colors. Since each channel range is limited with 255, pixels in color image take position in a 255x255x255 sized cube. Octree algorithm is based on a tree structure as shown in Figure 3. Cells are divided into 8 subsets at each level. Nodes in tree structure are numbered in Morton order [24] as shown in Figure 3 [25]. The simple principle of octree algorithm is based on slicing or partition of three-dimensional space into smaller cubes as shown in Figure 3. Each sub-cube holds different point clouds. The points which are included in a particular sub-cube are assigned to the same cluster or class. So, their labels will be the same as well. The octree algorithm is realized with different levels. If single level is used, 8 sub-cube or clusters is obtained. When second level is employed, each sub-cube is sliced again, and consequently, the whole space is partitioned into 64 clusters as shown in Figure 3. If third level is chosen, 512 subsets are obtained.

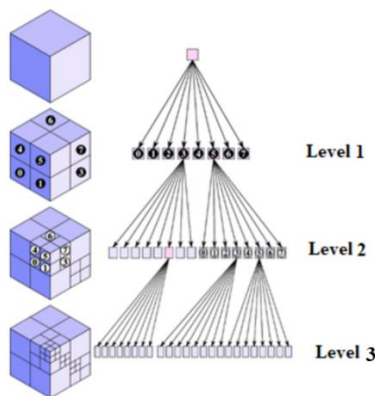
Color quantization is reduction in the number of colors in color image. As pixels in color image are presented by points or vectors in RGB color space, the octree algorithm could be easily used for clustering of color images. The cluster index is simply attained from red, green, and blue values of pixels as shown in Figure 4. Firstly, color values are converted to an 8-bit level. Then, the most significant bits of color channels (R: Red, G: Green, B: Blue) are combined at each level to form binary codes. The cluster index is achieved by using binary codes of each color component. If the most significant bits are used, a first-level quantization is performed. For example, if a pixel has 90,148 and 118 values for red, green, and blue channels, respectively, its cluster index becomes 2 as shown in Figure 4. On the other hand, when second level is preferred, its cluster index meets 10

(2x5). Consequently, every pixel in color image is assigned into any of relevant sub-cube or relevant sub-set.


When Lena image given in Figure 2(a) is quantized octree algorithm by using first level, output shown in Figure 5(a) has been obtained. The total number of colors has been reduced into 8 as could be seen in Figure 5(b). Furthermore, cluster index of every pixel in image has been produced. So average of each clusters was assigned for every pixel in the same cluster. Subsequently, color reduction was achieved. Of course, information loss is indispensable. However, peak-signal-to noise level (PSNR) between original Lena image and quantized Lena image are 21.65, 21.85 and 24.35 for red, green, and blue channels, respectively. Moreover, the total number of pixels in image and the total number of pixels in every subset are known, one-dimensional color histogram of color Lena image has been obtained as shown in Figure 6. The horizontal axis represents cluster index whereas vertical axis indicates normalized pixel numbers of each cluster [26].



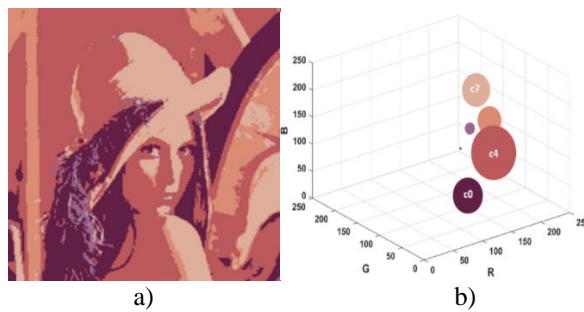
**Figure 2.** Lena a) original b) distribution in RGB color space.



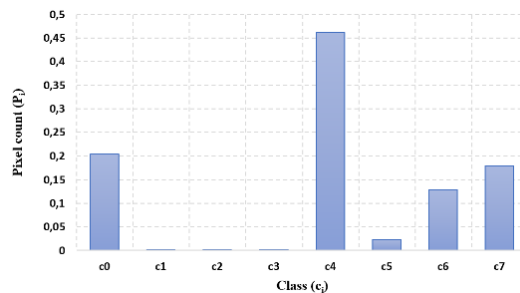
**Figure 3.** Octree algorithm [25].

			7	6	5	4	3	2	1	0
 (R:90,G:148,B:118)	R	90	0	1	0	1	1	0	1	0
	G	148	1	0	0	1	0	1	0	0
	B	118	0	1	1	1	0	1	1	0
Class index			$(010)_2$	$(101)_2$	$(001)_2$	$(111)_2$	$(100)_2$	$(011)_2$	$(101)_2$	$(000)_2$
			1. level	2. level	3. level	4. level	5. level	6. level	7. level	8. level

**Figure 4.** Octree algorithm for color quantization [26].



**Figure 5.** Lena a) quantized with octree, level 1 b) quantized color distribution.



**Figure 6.** One-dimensional color histogram of quantized Lena.

#### 2.4. Determination of Cluster Centers with Octree Algorithm

The purpose of K-means algorithm is to classify pixels in image. Mathematically, it is minimization process of the objective function given in Eq. 1. The number of data points,  $n$  in Eq. 1 corresponds to the number of pixels in image. Therefore, the computation cost of K-means algorithm increases with the size of images. Furthermore, since it is an iterative algorithm, in each iteration, all the pixels in image need to be processed again so that the objective function given in Eq. 1 converges to minimum. As the positions of centroids are updated in each iteration, they could be interpreted as moving



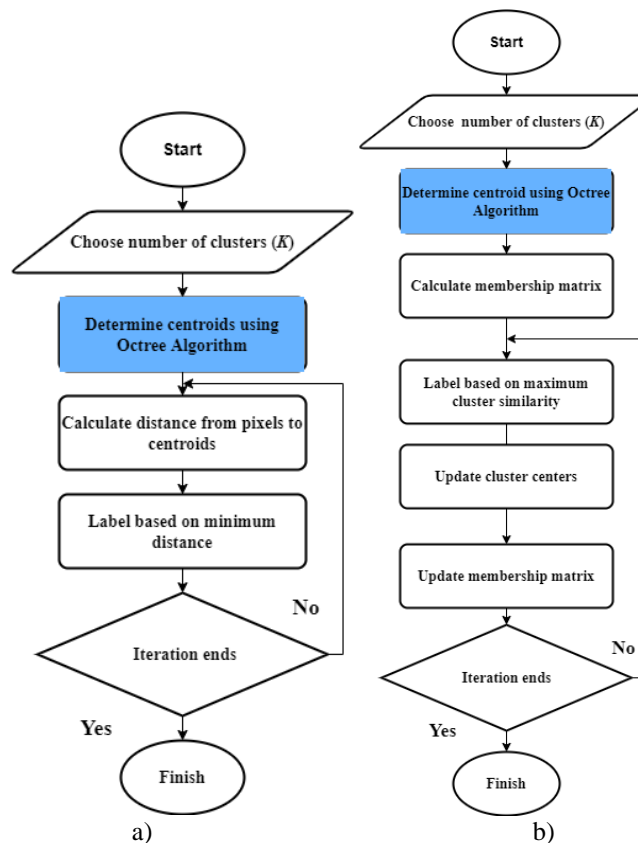
particles within RGB color space. Nevertheless, when the objective function is reached into minimum value, the positions of centroids are fixed. In other words, motion of centroids is stopped. If the centroids are located near to fixed or optimal positions, the iteration numbers or computation cost could be reduced.

The distribution of Lena image in RGB color space is shown in Figure 2(b) while its quantized version with octree algorithm is given in Figure 5(b). As first-level quantization is used, only eight balls with different spherical masses exist in Figure 5(b). Although diameters or masses are different, the centroid positions of each ball are fixed. Additionally, the numbers of balls represent the color vectors or cluster numbers obtained with octree algorithm.

The idea behind this study is based on employment of the color vectors or cluster centers obtained with octree algorithm for K-means algorithm and fuzzy c-means algorithms. The positions of quantized color vectors with octree algorithm could be set as initial positions of centroids of each algorithm. Consequently, the iteration numbers to reach minimum value of objective functions could be reduced. Thus, the computation cost is also reduced. It is logical that the positions of centroids which makes the minimum of the objective function must be somewhere within point clouds shown in Figure 2(b). On the hand, the positions of quantized color vectors with octree algorithm are already within point clouds shown in Figure 5(b). Therefore, it is reasonable that the quantized vectors could be set as initial centroids of K-means algorithm and fuzzy c-means algorithms.

Depending on level considered for octree algorithm, the number of cluster could be 8, 64, 512, and so on. If the number of cluster desired for K-means algorithm and fuzzy c-means algorithms are the same with that of octree algorithm, there would be no problem. However, if the cluster numbers are not the same, one dimensional color histogram given in Figure 6 help us to choose proper centroid. For example, cluster number for K-means algorithm or fuzzy c-means algorithms is set as 3 classes. Initially, one dimensional color histogram is sorted from the largest to minimum. Then the corresponding first three vectors or colors are assigned as centroids. It is clear that one dimensional color histogram holds the number of pixels assigned into related cluster. Also, diameters of globes in Figure 5(b) are proportional to number of pixels. Therefore, it means that some of pixels in image have already been labeled. Consequently, computation cost to reach minimum value of objective function is reduced. Consequently, flowcharts of suggested initialization strategy for K-means and fuzzy c-means algorithms are in Figure 7(a) and Figure 7(b), respectively.

Proposed method has eliminated the random assignment of initial cluster centers of K-means and fuzzy c-means algorithms. The octree algorithm use pixel values as shown in Figure 4. The developed algorithm does not require any threshold value. Once quantization process is completed without any recursive computation. Initialization of K-means algorithm and fuzzy c-means algorithms is realized. Since the number of pixels in color image is  $n$ , time complexity of proposed method is  $O(n)$ . On the other hand, time complexity of neighborhood concepts is  $O(n^2)$  [16]. Thus it could be concluded that the developed algorithm is fast.



**Figure 7.** Octree-based initialization a) *K*-means b) fuzzy *c*-means.

### 3. EXPERIMENTAL RESULTS and DISCUSSIONS

A user interface was designed in visual C# environment to test the suggested algorithms. Lena image of 346x346 size and Landscape image of 256x192 size were used for experiments. The desired number of cluster for *K*-means and fuzzy *c*-means algorithms was chosen as three for Lena image. Hence, the first level octree algorithm would be enough to create centroids for *K*-means and fuzzy *c*-means algorithms since the number of quantized color vectors with octree will be eight. Experiments were realized with two concepts. In the first category, the centroid of *K*-means and fuzzy *c*-means algorithms were randomly initialized while they were automatically assigned with octree technique in second strategy. Figure 8(a) shows the output of *K*-means where the centroids:  $C_0(233,45,244)$ ,  $C_1(155,91,100)$ ,  $C_2(22,219,2)$  were randomly selected. The means of pixels in each cluster are positioned in RGB color space as shown in Figure 8(b). Furthermore, the diameters of each globe in proportional with number of pixels in each class. The objective function converges to its minimum values at 6<sup>th</sup> iteration as shown in Figure 8(c).

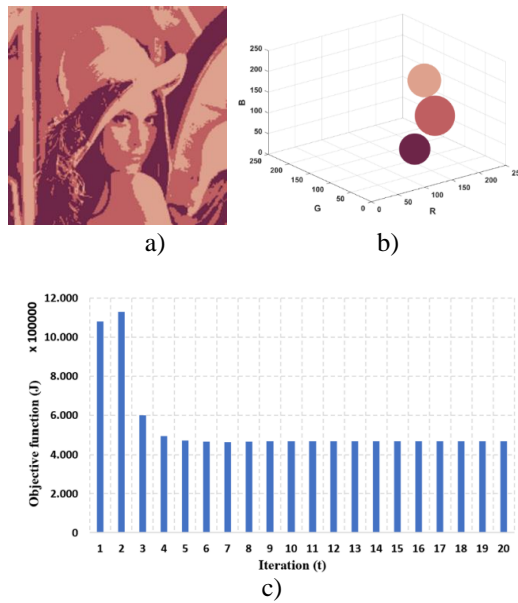
The color quantization of Lena image with first level octree algorithm was already given in Figure 5 where the size of balls in Figure 5(b) is proportional to number of pixels in related cluster. For example, cluster,  $c_4$  is a larger set in terms of pixels count. So it is reasonable to assign it as the first centroid of K-means. With using similar approach, clusters,  $c_0$  and  $c_7$  were allocated for second and third centroids for K-means algorithm. Consequently, an automatic procedure has been obtained. As the outputs of the octree algorithm are fixed, the positions of centroids will also be stationary. Thus, the randomness has been avoided. Figure 9(a) shows the output of K-means algorithm initialized with octree algorithm where the centroids are  $C_0(187,89,92)$ ,  $C_1(99,32,70)$ , and  $C_2(223,175,156)$ . The positions of centroids obtained with octree algorithm are shown in Figure 9(b). Moreover, variation of objective function is illustrated in Figure 9(c). It is obvious that the objective function converges to its minimum at third iteration. It is known that in each iteration, the whole image pixels are processed. When the size of image gets larger, the time consumed for each loop gets longer as well.

Figure 10(a) shows output of conventional fuzzy c-means algorithm where the centroids:  $C_0(222,101,40)$ ,  $C_1(227,107,206)$ , and  $C_2(161,223,125)$  were randomly selected. Final positions of each cluster centers are shown in Figure 10(b) as balls. Furthermore, the objective function converges to its minimum value at  $9^{th}$  iteration as shown in Figure 10(c). On the other hand, octree based Fuzzy c-means algorithm suggested in this study was also tested with Lena image. The centroid obtained with octree algorithm:  $C_0(187,89,92)$ ,  $C_1(99,32,70)$ , and  $C_2(223,175,156)$  were assigned as cluster centers for fuzzy c-means shown in Figure 7(b). Figure 11(a) shows output of fuzzy c-means algorithm initialized with octree algorithm while the final positions of color vectors are shown in Figure 11(b). It is obvious that the objective function converges to its minimum at second iteration as given in Figure 11(c). When the size of images to be clustered gets larger, computation cost becomes higher. Thus, it could be concluded that the overall performance of fuzzy c-means algorithm has been improved with octree algorithm.

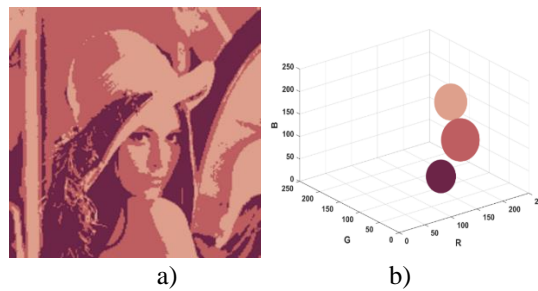
Proposed algorithm was also tested with Landscape image shown in Figure 12(a). The desired number of cluster for K-means and fuzzy c-means algorithms was chosen as three. Since the number of quantized color vectors with octree algorithm is 8, the first level is suitable. Figure 13(a) shows the output with first level while quantized color vectors are given in Figure 13(b). The first three color vectors as shown in Figure 14, which have high frequencies, have been assigned as initial cluster centers. In experiments, initially random approach was tested. Figure 15(a) shows output of K-means algorithm where centroids:  $C_0(154,125,156)$ ,  $C_1(108,54,55)$ , and  $C_2(122,118,11)$  were randomly selected while Figure 15(b) shows distribution in color space of image classified. Also, it could be seen that objective function in Figure 15(c) converges to minimum in  $19^{th}$  iteration. Histogram of quantized Landscape image is shown in Figure 14 where color vectors:  $c_0$ ,  $c_3$ , and  $c_1$  were selected as cluster centers. Figure 16(a) shows the output of K-means algorithm initialized with octree algorithm where the centroids are  $C_0(37,52,59)$ ,  $C_1(71,150,216)$ , and  $C_2(30,112,179)$ . It is clear that objective function given in Figure 16(c) started to converge at  $8^{th}$  iteration.

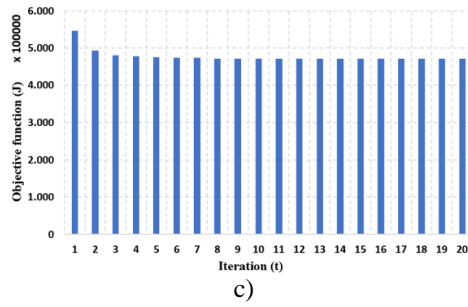
Additionally, Figure 17(a) shows output of conventional fuzzy c-means algorithm where centroids:  $C_0(242,230,154)$ ,  $C_1(95,243,27)$ ,  $C_2(182,151,62)$  were randomly selected. Final positions of cluster centers are shown in Figure 17(b). The objective function begins to converge at iteration 11 as could be seen in Figure 17(c). Landscape image was also tested with the fuzzy c-means algorithm proposed

in this study. Cluster centers obtained by octree algorithm:  $C_0(37,52,59)$ ,  $C_1(71,150,216)$ , and  $C_2(30,112,179)$  are assigned. The result obtained is given in Figure 18(a) and final positions of color vectors are displayed in Figure 18(b). Objective function converges at the 5<sup>th</sup> iteration as could be seen in Figure 18(c).

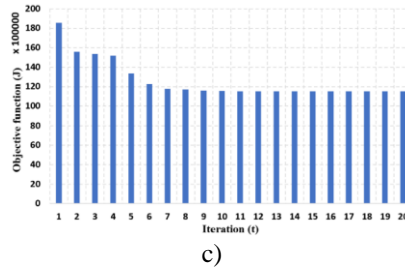
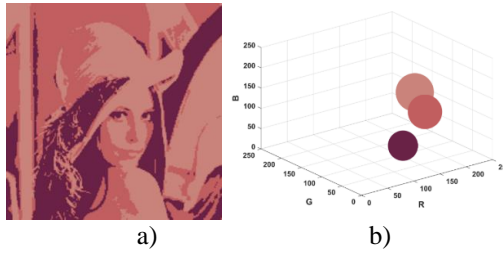


**Figure 8.** Lena: K-means (randomly initialized):  $C_0(233,45,244)$ ,  $C_1(155,91,100)$ ,  $C_2(22,219,2)$   
a) clustered b) color distribution c) objective function.

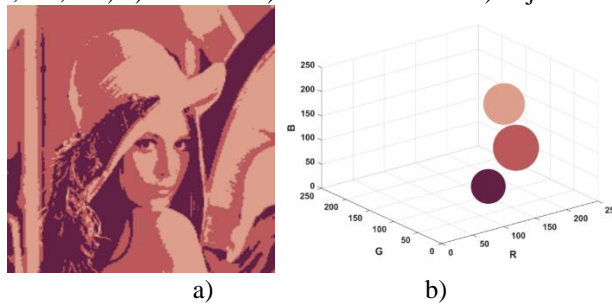


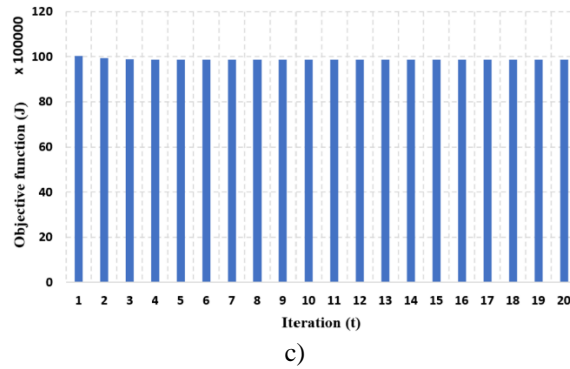


**Figure 9.** Lena: K-means (octree-based):  $C_0(187,89,92)$ ,  $C_1(99,32,70)$ ,  $C_2(223,175,156)$   
a) clustered b) color distribution c) objective function.

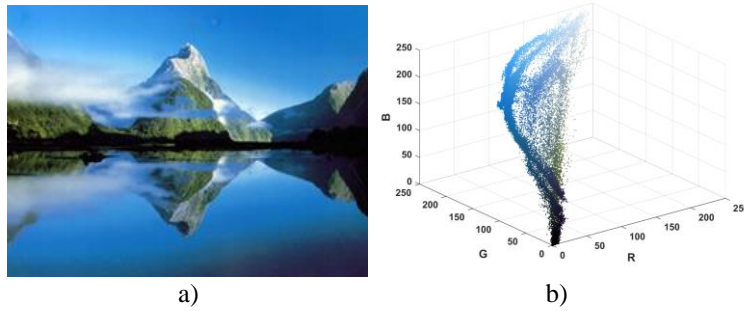


**Figure 10.** Lena: Fuzzy c-means (randomly initialized ):  $C_0(222,101,40)$ ,  $C_1(227,107,206)$ ,  $C_2(161,223,125)$  a) clustered b) color distribution c) objective function.

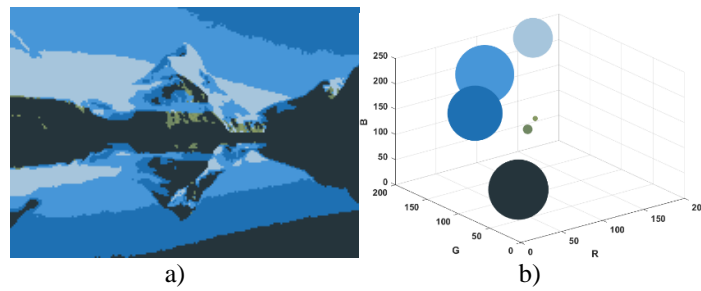




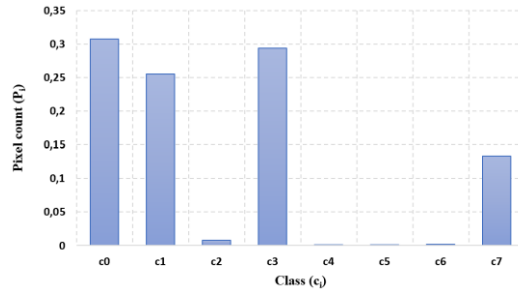
**Figure 11.** Lena: Fuzzy c-means (octree-based):  $C_0(187,89,92)$ ,  $C_1(99,32,70)$ ,  $C_2(223,175,156)$   
a) clustered b) color distribution c) objective function.



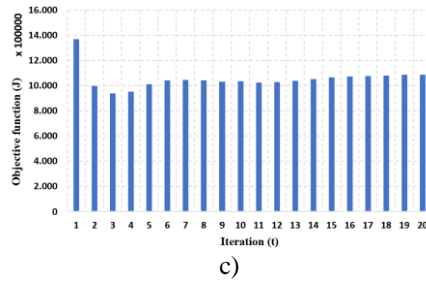
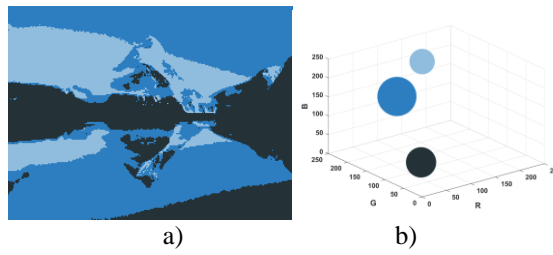
**Figure 12.** Landscape a) original b) distribution in RGB color space.



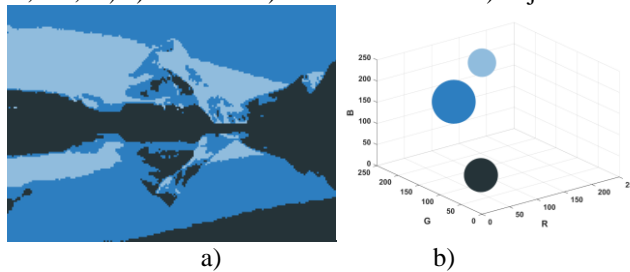
**Figure 13.** Landscape a) quantized with octree, level 1 b) quantized color distribution.

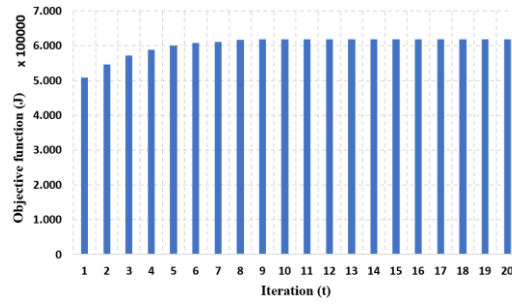


**Figure 14.** One-dimensional color histogram of quantized Landscape.



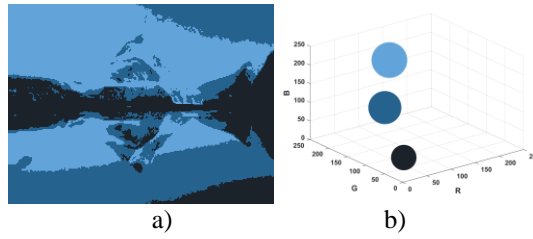
**Figure 15.** Landscape: K-means (randomly initialized):  $C_0(154,125,156)$ ,  $C_1(108,54,55)$ ,  $C_2(122,118,11)$  a) clustered b) color distribution c) objective function.





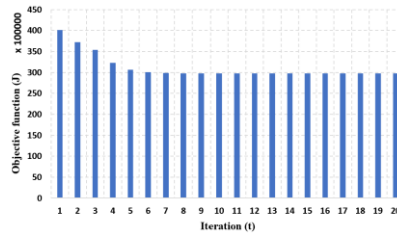
c)

**Figure 16.** Landscape: K-means (octree-based):  $C_0(37,52,59)$ ,  $C_1(71,150,216)$ ,  $C_2(30,112,179)$  a) clustered b) color distribution c) objective function.



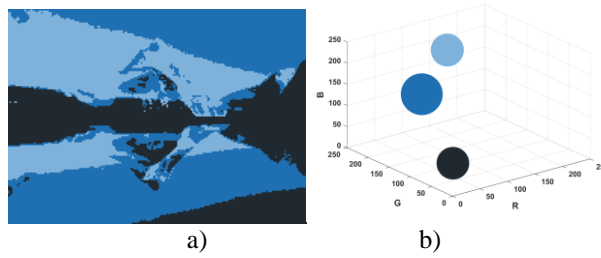
a)

b)



c)

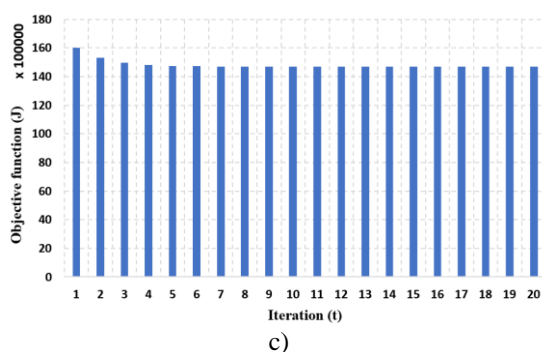
**Figure 17.** Landscape: Fuzzy c-means (randomly initialized):  $C_0(242,230,154)$ ,  $C_1(95,243,27)$ ,  $C_2(182,151,62)$  a) clustered b) color distribution c) objective function.



a)

b)





**Figure 18.** Landscape: Fuzzy c-means (octree-based):  $C_0(37,52,59)$ ,  $C_1(71,150,216)$ ,  $C_2(30,112,179)$   
a) clustered b) color distribution c) objective function.

In computer science, data could be interpreted as collection of vectors. On the other hand, vector quantization is reduction of the number of vectors that represent any data set. Consequently, processing time and memory requirements are reduced. Thus, image clustering could also be considered color quantization. Therefore, it is desired that similarity between original image and quantized image is high. Table 1. and Table 2. show PSNR results for Lena and Landscape. According to tables, it is obvious that PSNR results obtained with octree-based approaches are better than those obtained with traditional techniques.

**Table 1.** PSNR results: Lena.

	K-means			Fuzzy c-means		
	R	G	B	R	G	B
Randomly	23.094	22.534	22.868	21.800	20.318	22.039
Octree-based	23.136	22.521	22.822	23.207	22.360	22.689

**Table 2.** PSNR results: Landscape.

	K-means			Fuzzy c-means		
	R	G	B	R	G	B
Randomly	19.117	19.908	17.948	16.410	19.866	17.714
Octree-based	19.348	20.144	18.420	18.469	20.006	18.267

#### 4. CONCLUSION

Although they are frequently preferred for image segmentation process, results obtained from K-means and fuzzy c-means algorithms vary according to initial values of cluster centers. The related drawback is usually compensated by increasing the number of iterations. However, it causes serious computation cost. Automatic assignment of initial cluster centers helps to reduce computational cost. In this study, octree algorithm, also known as a color quantization technique, was employed to determine the initial centroids from K-means and fuzzy c-means procedures. Consequently, an

automatic assignment procedure for K-means and fuzzy c-means algorithms has been developed. The suggested algorithm has also been confirmed with experimental results.

#### **ACKNOWLEDGMENT**

The authors did not receive any financial support in the research and preparation of this article.

#### **REFERENCES**

- [1] Isa, N. A. M., Salamah, S. A., and Ngah, U. K. (2009). Adaptive fuzzy moving k-means clustering algorithm for image segmentation. *IEEE Transactions on Consumer Electronics*, 55(4), 2145-2153.
- [2] Kim, D. W., Lee, K. H., and Lee, D. (2004). A novel initialization scheme for the fuzzy c-means algorithm for color clustering. *Pattern Recognition Letters*, 25(2), 227-237.
- [3] Dörterler, S., Dumlu, H., Özdemir, D., Temurtaş, H. (2022). Hybridization of k-means and meta-heuristics algorithms for heart disease diagnosis. *New Trends in Engineering and Applied Natural Sciences* (55-72. ss).
- [4] Juang, L. H., and Wu, M. N. (2011). Psoriasis image identification using k-means clustering with morphological processing. *Measurement*, 44(5), 895-905.
- [5] Hrosik, R. C., Tuba, E., Dolicanin, E., Jovanovic, R., and Tuba, M. (2019). Brain image segmentation based on firefly algorithm combined with k-means clustering. *Studies Informatics and Control*, 28(2), 167-176.
- [6] Nitta, G. R., Sravani, T., Nitta, S., and Muthu, B. (2020). Dominant gray level-based k-means algorithm for MRI images. *Health and Technology*, 10(1), 281-287.
- [7] Yao, H., Duan, Q., Li, D., and Wang, J. (2013). An improved k-means clustering algorithm for fish image segmentation. *Mathematical and Computer Modelling*, 58(3-4), 790-798.
- [8] Pustokhina, I. V., Pustokhin, D. A., Rodrigues, J. J., Gupta, D., Khanna, A., Shankar, K., and Joshi, G. P. (2020). Automatic vehicle license plate recognition using optimal k-means with convolutional neural network for intelligent transportation systems. *Ieee Access*, 8, 92907-92917.
- [9] Tan, K. S., Lim, W. H., and Isa, N. A. M. (2013). Novel initialization scheme for fuzzy c-means algorithm on color image segmentation. *Applied Soft Computing*, 13(4), 1832-1852.
- [10] Gamino-Sánchez, F., Hernández-Gutiérrez, I. V., Rosales-Silva, A. J., Gallegos-Funes, F. J., Mújica-Vargas, D., Ramos-Díaz, E., and Kinani, J. M. V. (2018). Block-matching fuzzy c-means

clustering algorithm for segmentation of color images degraded with Gaussian noise. *Engineering Applications of Artificial Intelligence*, 73, 31-49.

- [11] Demirci, R., Güvenç, U., and Kahraman, H. T. (2014). Görüntülerin renk uzayı yardımıyla ayrıştırılması. *İleri Teknoloji Bilimleri Dergisi*, 3(1), 1-8.
- [12] Hussein, S. (2021). Automatic layer segmentation in H&E images of mice skin based on colour deconvolution and fuzzy c-mean clustering. *Informatics in Medicine Unlocked*, 25, 100692.
- [13] Pérez-Delgado, M. L. (2019). The color quantization problem solved by swarm-based operations. *Applied Intelligence*, 49(7), 2482-2514.
- [14] Park, H. J., and Kim, K. B. (2015). Improved k-means color quantization based on octree. *Journal of The Korea Society of Computer and Information*, 20(12), 9-14.
- [15] Chowdhury, K., Chaudhuri, D., and Pal, A. K. (2021). An entropy-based initialization method of K-means clustering on the optimal number of clusters. *Neural Computing and Applications*, 33(12), 6965-6982.
- [16] Cao, F., Liang, J., and Jiang, G. (2009). An initialization method for the K-Means algorithm using neighborhood model. *Computers and Mathematics with Applications*, 58(3), 474-483.
- [17] Celebi, M. E., Kingravi, H. A., and Vela, P. A. (2013). A comparative study of efficient initialization methods for the K-means clustering algorithm. *Expert System with Applications*, 40(1), 200-210.
- [18] Kılıçaslan, M., Tanyeri, U., İncetaş, M. O., Girgin, B. Y., and Demirci, R. (2017). Eşikleme Tekniklerinin Renk Uzayı Tabanlı Kümeleme Yönteminin Başarısına Etkisi. In *1st International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT 2017)*, Tokat, Türkiye (pp. 107-110).
- [19] Dursunoğlu, N. (2006). Image compression using vector quantization, M.Sc. thesis, Yıldız Technical University, Graduate School of Science and Engineering, İstanbul, 4-7.
- [20] Kalkan, M. B. (2019). Run-length encoding and segmentation based image compression, M.Sc. thesis, Gazi University, Graduate School of Natural and Applied Sciences, Ankara.
- [21] MacQueen, J. (1967). Classification and analysis of multivariate observations. In *5th Berkeley Symposium Mathematical Statistics and Probability* (pp. 281-297).
- [22] Bezdek, J. C. (2013). *Pattern recognition with fuzzy objective function algorithms*. Springer Science and Business Media.

- [23] Gervautz, M., and Purgathofer, W. (1988). A simple method for color quantization: Octree quantization. In *New Trends in Computer Graphics* (pp. 219-231). Springer, Berlin, Heidelberg.
- [24] Morton, G. M. (1966). A computer oriented geodetic database and a new technique in file sequencing.
- [25] Laurmaa, V., Picasso, M., and Steiner, G. (2016). An octree-based adaptive semi-Lagrangian VOF approach for simulating the displacement of free surfaces. *Computers and Fluids*, 131, 190-204.
- [26] Kılıçaslan, M. (2020). Content based image retrieval by using color histogram, Ph.D. thesis, Gazi University, Graduate School of Natural and Applied Sciences, Ankara, 129s.