The Study of the Unsupervised Classification Method Using the K-means Algorithm by a Proposition of a Simple Initialization Technique

Rahma Ouchani(✉), Mohammed Merzougui, and M'barek Nasri

Applied Mathematics, Signal Processing and Computer Science Laboratory (MATSI), Higher Institute of Technology (ESTO), Mohammed First University (UMP), BV Mohammed VI B.P. 524, Oujda 60000, Morocco
rahmaouchani01@gmail.com

Abstract. Image segmentation is a fundamental step in image processing, The existing segmentation techniques are numerous, but they are generally grouped into three main approaches which are the contour approach, the region approach and the cooperative approach. The segmentation by region approach aims to group the pixels with the same characteristics into homogeneous regions. It is characterized by the measure of uniformity of the regions constructed into the image. Generally, we distinguish three families of algorithms for the region approach: classification methods, region growth methods, and methods that divide or merge regions according to the chosen criterion. Our study is based on the unsupervised classification method using the K-means algorithm. This is the most known and used clustering algorithm, because of its simplicity of implementation. We know that this algorithm still suffers from several drawbacks such as The number of classes must be fixed at the beginning, The result depends on initial draw or starting values of the centers of the classes, so a main problem in the initialization phase. For this we propose to improve this behind by a simple initialization technique based on the optimization of the distance between the objects inside each class to obtain a set of compact and clearly separated clusters.

Keywords: Image segmentation · Classification methods · K-means algorithm

1 Introduction

Image segmentation is one of the most important techniques of image processing, is a fundamental step in the field of image processing we can mention as examples, object recognition, computer vision, medical imaging, and image tracking and analysis, etc. Image segmentation is a low-level image processing task whose main objective is to partition the digital image into different segments or regions containing similar characteristics in terms of color, intensity or texture so that the image can be easily understood for analysis. There are various image segmentation techniques, are generally grouped into three main approaches which

M. Ben Ahmed et al. (Eds.): SCA 2022, LNNS 629, pp. 617–625, 2023.
https://doi.org/10.1007/978-3-031-26852-6_57

are the contour approach, the region approach and the cooperative approach. We have tried to propose a classification of these methods according to the following scheme:

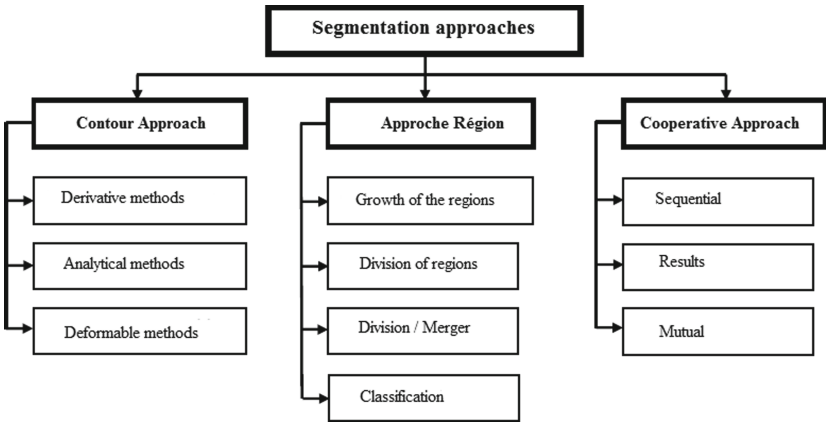

Fig. 1. Approaches to image segmentation.

1.1 Contour Approach

In the contour approach, we consider that the primitives to be extracted are the contrast lines separating regions of different and relatively homogeneous gray levels, or regions of different texture. In practice, it is a question of recognizing the transition zones and locating the boundary between the regions as well as possible. There are several methods such as derivative methods, analytical methods and deformable methods.

1.2 Region Approach

Image segmentation into homogeneous regions is based on the intrinsic properties of the regions. The choice of these properties determines what is called segmentation criteria. These criteria can be the gray level value [1], the color [2], the texture [3], or a combination of several information [4]. From an algorithmic point of view, segmentation consists in assigning to each pixel of the image a label of belonging to a given region. It is defined mathematically by Zucker in the following way [Zucker, 1976] : Segmenting an image I into N regions is equivalent to partitioning it into N subsets R_1, R_2.....R_n such that:

1. $\mathrm{I} = \cup_i \, Ri$
2. R_i consists of ∀i related pixels.
3. $\mathrm{P}(R_i) =$ true i.
4. $\mathrm{P}(R_i \cup R_j) =$ false for all i, j , with R_i and R_j adjacent in I.

The first condition indicates that each pixel of the image must belong to a region R_i and the union of all regions forms the whole image. The second condition is related to the structure of the regions, it defines a region as a set of pixels that must be related. The third condition expresses that each region must respect a uniformity predicate. The last condition implies that the non-fulfillment of this same predicate for the meeting of two adjacent regions. We distinguish four types of methods, Growth of regions [5], Division of regions, Division/Merge and Segmentation by classification. The method used in this article is the Classification Segmentation:

This method consists in grouping and classifying the pixels of an image into classes according to their properties. Each point in the image is associated with a vector of attributes. The classification is then performed on these attribute vectors in order to achieve a limited number of homogeneous regions within the image. It is a procedure in which similar pixels of an image are identified and grouped into a single class. There are two main types of classification:

• Supervised Classification: It is based on the knowledge of each class defined by a probabilistic approach. It is based on the learning of discriminating properties on a sample of already classified data.

• Unsupervised Classification: Automatic separation of the image into clusters without any a priori knowledge of the classes. It is based on a measure of distance between the attribute vectors. The most common algorithms for this category are K-means, Isodata, and Fuzzy c-means.

1.3 Cooperative Approach

Cooperative region-contour segmentation combines the two segmentation techniques, region-based segmentation and contour-based segmentation, taking advantage of the complementary nature of the information on these two approaches. Thus, segmentation by region-contour cooperation can be expressed as a mutual aid between these two concepts are used to improve the final segmentation result.

2 Related Work

Many works have been done in the field of image segmentation using different methods, such as threshold value [6] based segmentation method, edge based segmentation method and region based segmentation method (classification method). The K-means algorithm is one of the simplest and most effective clustering algorithms, many researches implemented to improve the initialization of the center. Some of the existing recent works that show better results are cited.

- Simon Tongbram and all proposed a new method uses an improved subtractive algorithm [7] which is based on the distance relationship between data points and cluster center and provides a more accurate center than the conventional subtractive clustering method. and the centroids of the modified SC algorithm

are used in the k-means algorithm to segment the image. experimental results validate a good performance of the method. - Alan Jose and all proposed a new method to generate the center of the cluster by reducing the mean square error of the final cluster with very little increase in execution time.
- Dhanachandra Nameirakpam and Khumanthem Manglem Singh proposed to segment the image using k-clustering algorithm, using subtractive cluster [8] to generate the initial centroid value. At the same time, partial contrast stretching is used to improve the quality of the original image and median filter is used to improve the segmented image. The proposed clustering algorithm has better segmentation.
- K. A. Abdul Nazeerand all proposed the k-means algorithm which combines a systematic method consisting of two approaches. The first approach is to find the initial centroid and another to assign the data point to the clusters. The proposed algorithm reduced the time complexity without sacrificing the cluster accuracy.

In this paper, we propose an image segmentation algorithm based on k-means algorithm with a simple centroıdes initialization technique, the experimental results validate the superiority of the proposed method and its performance over the methods.

3 K-means Clustering

Clustering method is one of the unsupervised learning methods in which a set of essential features is separated into uniform groups; K-means clustering technique is a widely used approach that has been applied to solve large-scale image segmentation tasks. There are different types of clustering: hierarchical clustering, Fuzzy C-means clustering, K-means clustering. The K-means method is one of the most commonly used clustering techniques for various applications [9]. The K-means clustering algorithm [10,11] is widely used due to its simple process and fast computation. The image is divided into clusters, i.e., “k” clusters of data. The centroid of a cluster is represented by a point such that the sum of the distances between all other data points in the cluster and this point is the smallest. The choice of initial cluster centers is very important because it prevents the clustering algorithm from producing incorrect decisions. The most common initialization procedure chooses the initial cluster centers randomly from the input data [12]. K-means algorithm is an algorithm based on the initial centroids of clusters, it is used to separate similar data into clusters.

The operation of the algorithm is:

Chooses k data values as initial cluster centers, then finds the distance between each cluster center and each data value and assigns it to the nearest cluster, updates the averages of each cluster, repeat this process until the criterion does not match. K-means clustering aims to divide the data into k clusters in which

each data value belongs to the cluster with the closest mean [13].

1. Initialize number of cluster k and centre.
2. For each pixel of an image, calculate the Euclidean distance d, between the center and each pixel of an image using the relation given below.

$$d = \|p(x, y) - c_k\| \tag{1}$$

3. Assign all the pixels to the nearest centre based on distance d.
4. After all pixels have been assigned, recalculate new position of the centre using the relation given below.

$$c_k = \frac{1}{k} \sum_{y \in c_k} \sum_{x \in c_k} p(x, y) \tag{2}$$

5. Repeat the process until it satisfies the tolerance or error value.
6. Reshape the cluster pixels into image.

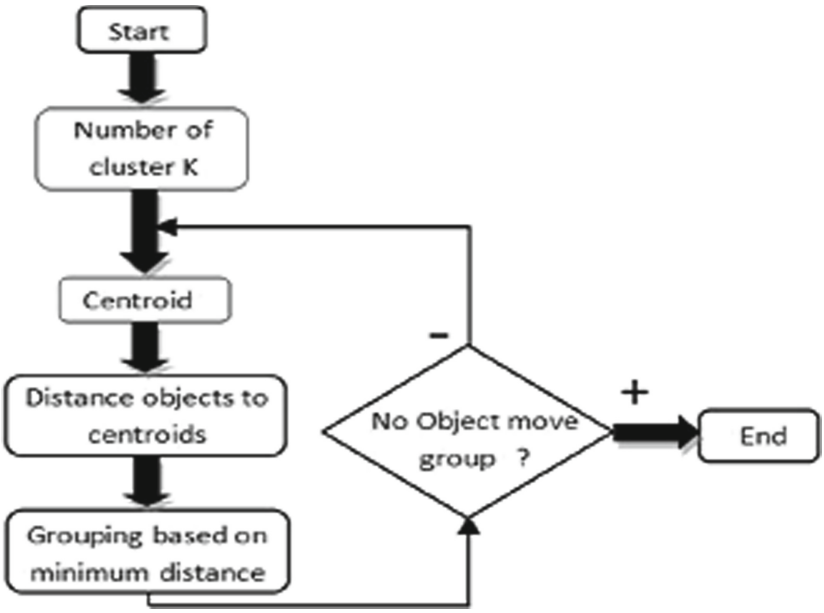

Fig. 2. K-means algorithm process

Advantages of the Algorithm:

– The k-means algorithm is very popular due to the fact that it is very easy to understand and implement.

– Its conceptual simplicity and speed
– Applicable to large data, and also to any type of data (even textual), by choosing a good notion of distance.

Disadvantages of the Algorithm:

– The number of classes must be fixed at the beginning,
– The result depends on the initial drawing of the class centers,
– The clusters are built in relation to non-existent objects (the middles).

4 Proposed Algorithm

However, the main limitation of this method is the dependence of the results on the values (initial centers). To each initialization corresponds a different solution (local optimum) which can in some cases be very far from the optimal solution (global optimum). A naive solution to this problem consists in launching the algorithm several times with different initialization and retain the best grouping found. The use of this solution remains limited because of its cost and because a better partition can be found in a single run.

In this paper we present a simple initialization technique of the k-means with the aim of maximizing the separability and compactness of the groups.This technique consists in determining according to the value of k given a spectrum of image color starts from the value of the brightest pixel to the darkest in order to build a set of groups K of color well determined to choose the initialization centroïdes, not in a totally random way but more exactly, in order to reduce the number of iterations and the numerous repetitions of the algorithm launch.

1. Initialize number of cluster k
2. creation of the distance matrix of the color spectrum.
3. choice of the centroïdes from the created spectrum (k groups)
4. For each pixel of an image, calculate the Euclidean distance d, between the center and each pixel of an image using the relation given below.

$$d = \|p(x, y) - c_k\| \tag{3}$$

5. Assign all the pixels to the nearest centre based on distance d.
6. After all pixels have been assigned, recalculate new position of the centre using the relation given below.

$$c_k = \frac{1}{k} \sum_{y \in c_k} \sum_{x \in c_k} p(x, y) \tag{4}$$

7. Repeat the process until it satisfies the tolerance or error value.
8. Reshape the cluster pixels into image.

5 Experimental Results

We segmented the images using k-means clustering algorithm, with the new technique to generate the initial centroid. We can conclude that the proposed clustering algorithm has better segmentation. In this method the number of distance calculations as well as the time reduced and the accuracy also improved. If the number of pixels in the image is increased, the execution time also increases.

The results that are obtained by using k-means clustering shown in below figures.

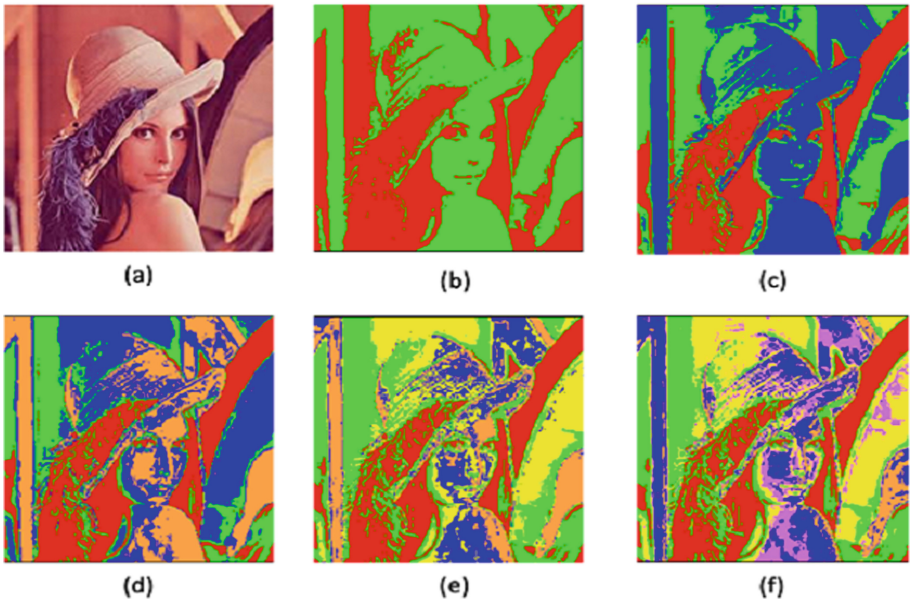

Fig. 3. Image segmentation results of real images: (a) original image, (b) k = 2, (c) k = 3, (d): k = 4, (e): k = 5, (f): k = 6.

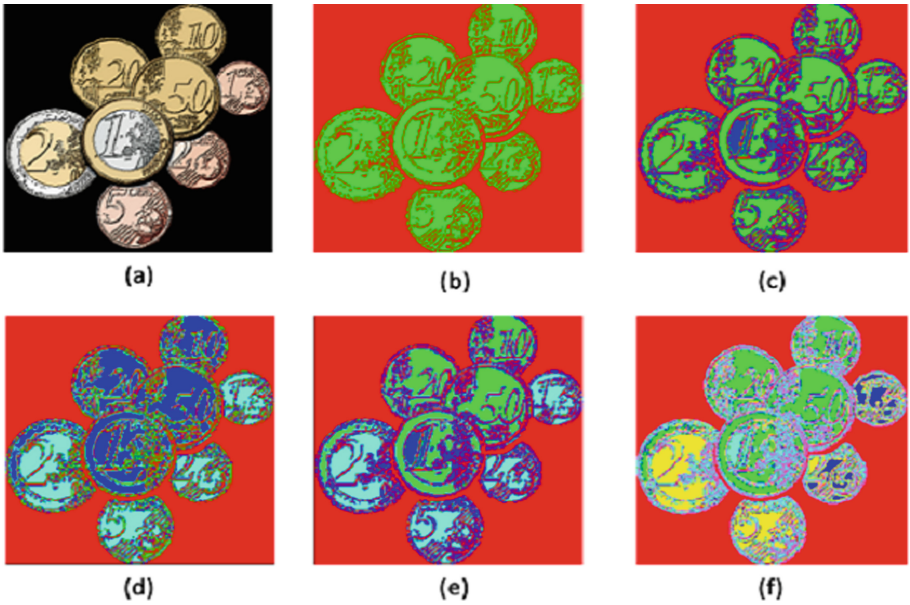

Fig. 4. Image segmentation results of real images: (a) original image, (b) k = 2, (c) k = 3, (d): k = 4, (e): k = 5, (f): k = 6.

6 Conclusion

This paper proposes a new image segmentation method based on k-means clustering methods using a new initialization technique. This technique consists in determining according to the value of k given a spectrum of image color starts from the value of the brightest pixel to the darkest in order to build a set of groups K of color well determined to choose the initialization centroïdes.We can conclude that we have successfully implemented the K-means clustering by this new technique which gives good results.

References

1. Sezgin, M., Sankur, B.: Survey over image thresholding techniques and quantitative performance evaluation. J. Electron. Imag. **13**(1), 146–165 (2004)
2. Lucchese L., Mitra, S.K.: Color image segmentation: A state-of-the-art survey. In: Proceedings of the Indian National Science Academy (INSA-A), Vol. 67-A(2), pp. 207–221 (2001)
3. Materka, A., Strzelecki, M.: Texture Analysis Methods - A review. Technical University of Lodz, Institute of Electronics, Bruxelles (1998)
4. Tremeau, A., Borel, N.: A region growing and merging algorith to color segmentation. Pattern Recogn. **30**(7), 1191–1204 (1997)

5. Li, C., Kao, C.Y., Gore, J.C., Ding, Z.: Minimization of region-scalable fitting energy for image segmentation. IEEE Trans. Image Process. **17**(10), 1940–1949 (2008)
6. Alamr, S.S., Kalyankar, N.V., Khamitkar, S.D.: Image segmentation by using threshold techniques. Comput. Sci. **2**(5), 83–86 (2010)
7. Tian, L., Han, L., Yue, J.: Research on Image Segmentation based on Clustering Algorithm. In: Int. J. Signal Process. Image Process. Pattern Recogn. **9**, 1–12 (2016)
8. Tongbram, S., Shimray, B.A., Singh, L.S.: Segmentation of image based on k-means and modified subtractive clustering. Indonesian J. Electr. Eng. Computer Sci., **22**(3), 1396 1403 ISSN: 2502–4752 (2021). https://doi.org/10.11591/ijeecs.v22.i3.pp1396-1403
9. Dhanachandra, N., Manglem, K., Jina Chanu, Y.: Procedia Computer Science - December (2015). https://doi.org/10.1016/j.procs.2015.06.090
10. Celebi, M.E., Kingravi, H.A., Vela, P.A.: A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Syst. Appl. **40**(1), 200–210 (2013). https://doi.org/10.1016/j.eswa.2012.07.021
11. Alhawarat, M., Hegazi, M.: Revisiting k-means and topic modeling, a comparison study to cluster arabic documents. IEEE Access **6**, 42740–42749 (2018). https://doi.org/10.1109/ACCESS.2018.2852648
12. Panda, S.: Color Image Segmentation Using K-means Clustering and Thresholding Technique. In: IJESC (2015)
13. Shinde, S., Tidke, B.: Improved K-means Algorithm for Searching Research Papers. Int. J. Comput. Sci. Commun. Netw. **4**, 197–202 (2014)