National Conference on "Advanced Technologies in Computing and Networking"-ATCON-2015
Special Issue of International Journal of Electronics, Communication & Soft Computing Science and Engineering, ISSN: 2277-9477

Document Image Segmentation Using K-Means Clustering Technique

Manish T Wanjari **Vikas K. Yeotikar** **Keshav D. Kalaskar** **Dr. Mahendra P. Dhore**

Abstract — Clustering technique is active research field in machine learning. In Document Image Segmentation, clustering technique is one of the most famous, simple and easy to implement technique. The K-means clustering technique is most widely used technique in the literature and many researchers compare their proposed work with the results achieved by the K-means. Clustering is the process of grouping samples so that the samples are similar within each group. This paper deals with document image segmentation using K-means clustering technique. The Otsu method is also implemented. Some of the advances clustering techniques are also discuss in this paper.

Key Words — Document Image Analysis (DIA), Document image segmentation (DIS), Segmentation, Clustering, K-means algorithm

I. INTRODUCTION

Document Image segmentation is the process of dividing a document image into multiple parts. It is used to identify objects of other relevant information in digital document image. There are different methods to perform document image segmentation such as Threshold method, Clustering method and Texture method. Document Image Segmentation plays an important role in document image understanding, image analysis and image processing. [1, 2] The aim of clustering is to partition a dataset into several disjoint subsets in a manner that elements in a subset are more similar to each other than to elements in other subsets.

Document images are considered one of the most important ways of providing information. Understanding document images segmentation and extracting the required information from a document image to create homogeneous regions by dividing pixels into groups thus forming regions of similarity, can be used for other task is an important aspect of computer learning. One of the processes in direction understanding document images is to segment them and find out different objects in them. Document image segmentation is the process of partitioning the given document image into homogeneous regions with respect to certain features, and which hopefully correspond to real object in the actual scene [3]

Clustering is the process of grouping samples so that the samples are similar within each group these groups are called clusters. Clustering is document image analysis; clustering approaches were one of the first techniques used for the segmentation of document images. [2] In partitional clustering, the aim is to create one set of clusters that partitions the data in to similar groups. In our work we have used K-means clustering approach for performing document image segmentation using matlab.

Basically, clustering algorithm can be categorized into Partitioning technique, Hierarchical technique, density-based technique and grid-based technique. Partitioning algorithm that constructs various partitions and then evaluate some criterion is described. Hierarchical algorithms that create hierarchical decomposition of the instances using some criterion covered. The density-based algorithms that are based on connectivity and density functions. And also covered the grid-based technique, which are based on a multiple-level granularity structure, these are the recent advance in clustering technique discussed.

II. DOCUMENT IMAGE SEGMENTATION

The fundamental idea of the document image segmentation is to group pixels and the usual approach to a common feature extraction. Feature can be represented by the space of gray level exploring similarities between pixels of a region. Segmentation refers to the process of partitioning a digital image into multiple regions (sets of pixels). The aim of segmentation is to simplify and change the representation of an image into something that is more meaningful and easier to analyze. Document image segmentation is typically used to locate object and boundaries (line, curve, etc) in document images. The result of document image segmentation is set of regions that collectively cover the entire document images, or a set of contours extracted from the document images. [4, 5]

Document image segmentation techniques are categorized into three parts such as Clustering, edge detection, region growing. Some popular clustering algorithms like K-means are often used in document image segmentation [6] adjacent regions are significantly different with respect to the same characteristics. Segmentation is basically use in medical imaging, face recognition, fingerprint recognition, Traffic control system, Brake light detection and Machine vision.

III. CLUSTERING

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters) so that the data in each subset (ideally) share some common trait – often proximity according to some define distance measure. Data clustering is the common technique for statistical data analysis, which is used in various fields, including machine learning, pattern recognition, image analysis and bioinformatics. The computational task of classifying the data set into k clusters is often referred to as **k-clustering**.

A. K-Mean Clustering

In our work we have used K-mean clustering for performing document image segmentation using Matlab. A good clustering method will produce high quality cluster with high intra-class

similarity and low inter class similarity. The quality of clustering result depends in both the similarity measures used by the methods and its implementation. Clustering means classifying and distinguishing things that are provided with similar properties. [7] Clustering technique classifies the pixels with same characteristics into one cluster thus forming different cluster according to coherence between pixels in a cluster. In a method of unsupervised learning and a common technique for statistical data analysis used in many fields such as pattern recognition and document image analysis.

An appropriate space is essential to document image segment which form the basis for the basic image processing technique. Clustering is a way to separate groups of objects. K-Means clustering treats each object as having a location in a space. It finds partitions such that objects within each cluster are as close to each other as possible. K-means requires that the number of clusters to be partitioned should be specified and also a distance metric to quantify how close to object to each other.

B. K-Mean Algorithm

K-means is one of the simplest and easiest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume K cluster) fixed apriori. The main idea is to define k centers one for each clusters. These centers 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 center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter 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 center. A loop has been generated. As a result of this loop, we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally this algorithm aims at minimizing an objective function know as squared error function given by

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (||x_i - v_j||)^2$$

Where, $||x_i - v_j||$ - is the Euclidean distance between x_i & v_i
c_i - is the number of data points in i^{th} cluster
c - is the number of cluster centers [8]

The algorithm performs as follows

i) Initialize K centroids
ii) Until a stop criteria is not satisfied
    i. Calculate the distance between all element in the datasets and K-centroids. Elements closer to centroids form cluster.
    ii. Centroids are updated (assume the cluster values)

The basic advantages of this algorithm are

i) It is easy to implement
ii) It is simple and unsupervised learning algorithm.

The properties of k-mean algorithm

i) There are always k clusters
ii) There is always at least one item in each cluster
iii) The clusters are non hierarchical and they do not overlap
iv) Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the center of clusters.

[9] K-means is a simple algorithm that has been adapted to many problem domains.

C. Global K-mean Algorithm

According to Likas global K-mean clustering algorithm does not depend upon the initial parameter values and utilize the K-mean algorithm as a local search procedure that constitutes a deterministic global optimization method. This technique proceed in an incremental way of attempting to optimally include one new cluster center at each stage instead of randomly selecting starting value for all cluster.[10, 11, 12]

IV. CLUSTERING TECHNIQUE

Clustering techniques are basically categorized into partitional, hierarchical, density-based, and grid-based clustering technique. [13, 14, 15] Hierarchical clustering techniques make a cluster tree by means of heuristics splitting or merging procedures. And other Partitional clustering techniques divide the input data into specified in advance number of clusters.

A. Partitional Technique

The Clustering problem has been addressed in many contexts and by researchers in many disciplines. Recent frameworks in the domain of document image analysis use this tool in applications such as document image segmentation by pixel grouping. Among the existing methods, the partitional clustering algorithms are the most appropriate for such applications. Partitional clustering algorithm is based on the probability density function estimation. [16]

Partitioning technique is categorized into two subcategories such as Centroid and the Medoids algorithms. The centroid algorithm represents each cluster by using the gravity centre of instances. The Medoid algorithms represent is cluster by means of the instances closest to the gravity centre. The most well known centroid algorithm is k-mean. The k-mean method partitions the data set into k subsets such that all points in a given subset are closest to the same centre. Basically, the k-mean algorithm has some important properties.

i) It is efficient in processing large data set.
ii) It often terminates at a local optimum.
iii) The clusters have spherical shapes.
iv) It is sensitive to noise.

The algorithm described above is classified as a batch method because it requires that all the data should be available in advance. However there are variants of the k-means clustering process, which gets around this limitation. Choosing the proper initial centroids is the key step of the basic k-means procedure.

Traditional clustering approaches generate partitions, in a partition; each pattern belongs to one and only one cluster. Hence the clusters in a hard clustering are disjoint. [17]

B. Hierarchical Technique

The hierarchical techniques group data instances into a tree of cluster. There are two major techniques include this category. One is agglomerative technique, which forms the clusters in a bottom up sequence until all data instances belong to the same cluster. The other is divisive technique, which spits up the data set into smaller cluster in a top down sequence until each cluster contains only one instance. Both divisive and agglomerative algorithm can be represented by dendrograms. But agglomerative and divisive algorithms are called for their quick termination.

However both techniques suffer from their inability to perform adjustments once the splitting and merging decision is made. Techniques has some of the advantages

i) Does not require the number of clusters to be known in advance.

ii) Computes a complete hierarchy of clusters.

iii) Good result visualizations are integrated into the methods

iv) A flat partition can be derived afterwards

Hierarchical Clustering techniques use various criteria to decide locally at each step which clusters should be joined.

C. Density-based Technique

Density-based clustering algorithm try to find clusters based on density of data point in a region. The density based clustering is that for each instance of a cluster the neighborhood of a given radius has to contain at least a minimum number of instances. One of the most well-known density based clustering algorithm is the DBSCAN. DBSCAN separate data points into three classes such as core points (points that are at interior of cluster), border points (points that is not a core point) and noise points. (Points that is not core and border point)

D. Grid-based Technique

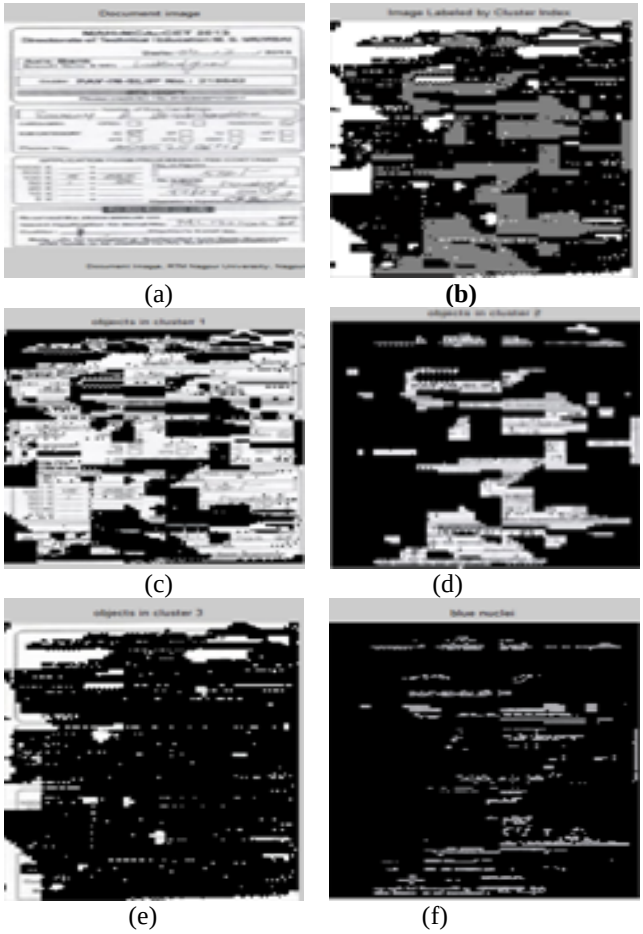
Grid-based clustering algorithms first quantize the clustering space into a finite number of cells and then perform the required operations on the quantized space. Cells that contain more than certain number of points are treated as dense and the dense cells are connected to form the cluster. [18]

V. EXPERIMENTAL RESULTS

(a)

(b)

Fig. 1. a) Original Grayscale document image b) Otsu Method



(a) **(b)** (c) (d) (e) (f)

Fig.-2 (a) Original Grayscale image (b) Label every pixel in the document image using the result from K-means clustering (c) Objects in cluster-I (d) Objects in cluster-II (e) Objects in cluster-III (f) Segmenting the nuclei into a separate document image.

Figure 1 a) Shows the original grayscale document image and in b) result shows the segmenting image by using Otsu method applying on the document image, getting the different clustering.

Figure 2 a) Shows the original grayscale document image b) For every object in original input document image, k-means returns an index corresponding to a cluster. The cluster center result from k-means used, Label every pixel in the document image with its cluster index. By using pixel label can separate object in original document image by which shows result in c), d)

& e). f) Programmatically determine the index of the cluster containing the objects because k-means will not return the same cluster index value every time. Using the cluster center value which contains the mean a* & b* value for each cluster and segmenting the nuclei into a separate document image.

CONCLUSION

We have successfully implemented the K-means clustering technique. For smaller values of k the algorithms give good result. For larger values of k, the segmentation is very course; many clusters appear in the document images at discrete places. We got good results for document images in MATLAB Software using K-means clustering technique. The benefit of K-means clustering that its simplicity, speed and efficient which allows us to run on large amount of datasets. Otsu method is more proper for images that their objects are distinguished from their background and for document images.

ACKNOWLEDGMENT

The Authors are thankful to the University Grants Commission, New Delhi for supporting this work as a part of Major Research Project.

REFERENCES

[1] S. P. Lloyd, —"Least squares quantization in PCM," *IEEE Trans. Inf. Theory, vol. IT-28*, no. 2, pp. 129–136, Mar.1982.
[2] Ms. Chinki Chandhok, Mrs. Soni Chaturvedi, Dr. A. A Khurshid, "An Approach to Image Segmentation using K-Means Clustering Algorithm", International Journal of Information Technology (IJIT), Volume – 1, Issue – 1, August 2012 ISSN 2279 – 008X.
[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] M. Mignotte, c. Collet, P. Perez, and P. Bouthemy, "Sonar image segmentation using a hierarchical MRFmodel", IEEE Trans. Image Process. Vol. 9. No 7, pp. 1216-1231, oct. 2006.
[5] F. Destremps, J. F. Angers, and M. Mignotte,"Fusion of hidden morkov random field model and its Bayesian estimation", IEEE Trans. Image Process. Vol. 15. No 10, pp. 2920-2935, oct. 2006
[6] J. A. Hartigan, "Clustering algorithms", New Yark Wiley, 1975.
[7] Wang, Xiao-song, Huang Xin-yuan and Fu Hui,"The study of color free image segmentation", In: Second International Workshop, Computer science and Engineering WCSE 09, Sch of Inf. Beijing Forestry Univ. Beijing China (2009)
[8] Tapas Kanungo, David M. Mount, Nathan S. Netynyahu, Chritine D. Piatko, Ruth Silverman and Angela Y. Wu., " An Efficient K-means Clustering Algorithm: Analysis and Implementation". M. Young, *The Techincal Writers Handbook.* Mill Valley, CA: University Science, 1989.
[9] Jose F. Costa and Jackson G. de Souza, "Image Segmentation through clustering Based on Natural computing techniques", Federal University of Rio grande do Norte Brazil.
[10] P. Scheunders, "Wavelet thresholding of multivalued images," IEEE Trans. Image Proc. 13, pp. 475– 483, 2004.
[11] P. Scheunders and S. De Backer, "Wavelet denoising of multicomponent images, using a noise-free image," in Proc. IEEE Internat. Conf. Image Proc. ICIP, 2006.
[12] N.Valliammal & Dr.S.N.Geethalakshmi, " Leaf Image Segmentation Based On the Combination of Wavelet Transform and K Means Clustering", *(IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 1, No. 3, 2012, p.p.37-43*
[13] R. Xu. D. Wunsch II, "Survey of Clustering algorithm", IEEE Transactions on Neural Networks, 16, 2005, 645-678.
[14] A. Jain, M. Murty, P. Flynn, "Data Clustering: A Review", ACM Computing Surveys, 31, 1999, 264-323.
[15] O. Hall, I. Barak, J. C. Bezdek, "Clustering with a generically optimized approach", IEEE Transactions Evo. Computations, 3, 1999, 103-112.
[16] Danielle Nuzillard and Cosmin Lazar, "Partitional Clustering Techniques for Multi-Spectral Image Segmentation", Journal of Computers, Vol. 2, No. 10, December 2007, p.p. 1-8.
[17] A. Jain, M. Murty, P. Flynn, "Data Clustering: A Review", ACM Computing Surveys, 31, 1999, 264-323.
[18] S.B. Kotsiantis, P. E. Pintelas, "Recent Advances in Clustering: A Brief Survey", University of Patras Educational Software Development Laboratory Hellas.

AUTHOR'S PROFILE

Manish T. Wanjari is a Research Scholar pursuing Ph. D. in Computer Science. He is currently working as Project Fellow under UGC Sponsored Major Research Project in the Department of Computer Science, SSESA's, Science College, Congress Nagar Nagpur.
Email Id- mwanjari9@gmail.com

Vikas K. Yeotikar is a Research Scholar Pursuing Ph. D. in Computer Science. He is currently working as Lecturer in Department of Computer Science, SSESA's, Science College, Congress Nagar Nagpur.
Email Id- vkyeotikar@gmail.com

Keshao D. Kalaskar is Associate Professor & Head, Department of Computer Science, Dr. Ambedkar College, Chandrapur. He is having teaching experience of more than 20 years at UG level. He is Pursuing Ph. D. in Computer Science.
Email Id- keshao_kalaskar@yahoo.co.in

Dr. Mahendra P. Dhore is Associate Professor in Computer Science, Department of Electronics & Computer Science, RTM Nagpur University, Nagpur. He is having teaching experience of more than 18 years at UG & PG level. His research areas include Digital Image Processing, Document Image Analysis, Mobile Computing & Cloud Computing. He is Member of IEEE, IAENG, IACSIT, IETE, and ISCA.
Email Id – mpdhore@ieee.org
mpdhore@rediffmail.com