Improved Fuzzy C-means for Document Image Segmentation

Hasanen S. Abdullah
University of Technology
Computer sciences Department
E-mail: qhasanen@yahoo.com

Ammar H. Jassim
University of Baghdad/ College of Science for women
Department of computer Science
E-mail: ammar_hussein_2004@yahoo.com

Abstract

Interest in the automatic analysis and segmentation of document images has been increased during the recent years. Also, the document segmentation plays an important role in document analysis, since every day, thousands of documents including government files, technical reports, books, Newspapers, magazines, etc, need be processed and provided an intelligent access to its contents both the text and non-text components. Lots of time, money and effort will be preserving whenever it can be executed automatically. Hence, this paper introduced a new document image segmentation approach based on suggested improved fuzzy C-means (IFCM) that focus on segmented the text, images and background pixels from the scanned document images using the statistical features of regions pixels, collected areas and then clustered in text and non-text areas. In this approach, a document image is segmented to several non-overlapping regions via a novel recursive clustering technique relies on the statistical features of each pixel with its neighborhoods. The performance of this method is evaluated by examining a variety of complex document images such as newspaper layouts and artificially segmented the text and images. Also the performance has been recorded in terms of quantitative and qualitative measures. The experimental performance results are promising and encouraging without the need of any assistant techniques for pixel segmentation, unlike many techniques of this class. Since they prove the feasibility and practicality of IFCM and can provide near-optimal solutions to document layout analysis problems. They achieved accuracy rate 95.21% and recall rate around 97.51% on a set of 390 documents that confirms the robustness of suggested algorithm. However, the overall precision is higher due to the different evaluation metrics. Although the pixel wise evaluation allows for more accurate improvement, this evaluation metrics reflects the objective of the IFCM.

Keywords: Fuzzy C-means (FCM); Fuzzy C-means Algorithm, Clustering Algorithms, Cluster Center, Clustering, Document Image Segmentation, Document Image Analysis, Document Image, Document Image Segmentation.

Introduction

Nowadays, a wide variety of information is being available and converted into electronic format for efficient storage and processing. This needs handling of documents using image analysis techniques. The document analysis techniques decompose the document image into different consistent items which represent the consistent components of the documents image such as text, graphics and tables, without a prior knowledge of specific format.

Document images are frequently generated from physical documents via digitization using scanner devices or digital cameras. Various documents, such as newspapers, and magazines, contain very complex layout. Automatic analysis of a document with complex structure and layout is considered a difficult task and not within the capabilities of the current document layout analysis systems.

The extraction of data from documents requires human effort and time, thus in office automation, the document segmentation and analysis have a main role, especially in intelligent manipulation application. So, automatic information extraction from the document image has become a more important issue. Hence, the segmentation is considered a significant phase in document image processing and analysis. Major applications need to provide intelligent access to documents including text and non-text components like pictures, graphics, tables etc. Therefore, correct document image segmentation is necessary for document image analysis, information retrieval and other document image analysis applications.

Literature Review

The following section show the briefly overview of some related works in the document image segmentation area.

Melissa C. and Alexandra B. A. (2014) suggested an approach to distribute pixels in document images into four essential classes (text, image, background and graphics) using for classification the support vector machine with a novel low dimension feature descriptor based on textual properties.

El-Omari N. K. T., Omari A. H., Al-Badarneh O. F., Hussein Abdel-jaber, (2012) they present a new method for segmentation and classification document images using an artificial neural network algorithm based on the detected and recognize the colors number in the document images. The neural network is designed to recognize some color patterns from labeled document. And, unlabeled document are classified based on these color. This approach aims to segment the original document images content to homogeneous and consistent four regions: graphics, picture, text, and background. They obtain 89% as segmentation accuracy.

Bukhari S. S. (2012) presents a method based on a discriminative learning for page segmentation, where trained a self-tunable feed forward multilayer perceptron classifier for distinguishing between text and non-text connected components. He obtained 89.54% as segmentation accuracy.

Amy Winder, Tim Andersen and Elisa H. Barney Smith, (2011) this work add capability to segment documents images that containing text, pictures, and graphics in open source OCR engine like OCRopus. By improving the OCRopus RAST algorithm to analyze and recognize non-text blocks so that documents have mixed content can be analyzed and recognized the regions in addition to documents with text-only. Also, an approach for classifying the text and nontext regions which was developed and performed for Voronoi algorithm to allow the users to implement OCR on documents image processed by this approach. The average segmentation accuracy of RAST algorithm is about 80%. The accuracy of Voronoi algorithm is averaged around 70% accuracy.

Mustafa S. E. (2011) primarily applies Markov random field, Run Length Encoding, wavelet transform, and Hough transform to segment photo, text and edge/line regions in gray and color scale documents. He obtained 85% as segmentation accuracy rate.

Bukhari S. S., Shafait F., and Al Azawi M. I. A., (2010) presented segmentation approach introducing connected component based classification, thereby not requiring block segmentation formerly. Here they train the self-tunable MLP classifier using context information and shape as a feature vector to differentiate connected components for text and non-text. They achieved 95.01% as segmentation accuracy.

Jamal S., Reza S., and Saeed M., (2009) offer a new approach for extract textual areas of an image which utilized dual-tree discrete wavelet transform and fuzzy classifier. For classification they extended text extraction scheme to classify document images components into text, picture, and background. Three class fuzzy classifiers with a morphological post processing operation are used for this purpose. They obtained 86.75% as accuracy Rate.

Document Image Segmentation

Document images segmentation is defined as an approach to subdividing the document areas into the text and non-text regions and is an important and emergent concept in document image analysis and understanding. Automation of document segmentation and analysis contains region extraction, identification of region type and finally each region is processed separately. Document image segmentation does the work of segmenting and identifying the type of region. So as to process each region, document should be segmented then submit to the respective system for more processing. For instance, text regions are processed using the OCR system which converts text region to machine-readable form and the non-text regions are preserved for processing such as enhancement, compression, recognition, and storage etc (Priyadharshini N. and Vijaya MS., 2013).

Document image segmentation is considered most challenging issue in documentation area and has been studied extensively in the last few decades. (Manish T. W., Keshao D. K. and Mahendra P. D., 2015). Therefore, they has been studied extensively in the last few years by a number of research teams began to develop existing system and designing new systems to process and extract related information from the document image automatically. They focus on developed and improved new techniques for document image segmentation

Document segmentation to text and non-text elements considered an essential preprocessing step before the document image analysis and character recognition operation. Therefore, this will affect the next stage of these systems in case of poor segmentation (D.Shobana and M.Phil, 2012; Yong Y., Shuying H., 2007). Therefore, this paper presents an approach to segment and identify the text and non text area in document images using an IFCM and features analysis. This approach is robust and efficient to segment adequately the textual and graphical components of a document.

Document image segmentation is done to analyze the layout and content of the document image. Various document segmentation techniques have been suggested and generally classified into three categories: the top-down, bottom-up and the hybrid approach. A top-down mechanism repeatedly segments the document image to smaller regions till further it cannot be segmented such as run-length smearing algorithm, Fourier transforms, projection profile methods etc (Ritu G., Gaurav H. and Santanu C., 2008). A bottom-up approach starts by merging the pixels into characters. After that the characters are merged to construct words until the entire document regions are merged. Methods based on this approach are run-length smoothing, connected component analysis, neural networks, and region-growing methods. A hybrid Approach is a combination of both the top down and the bottom up approach. Some hybrid based methods are the texture based and the Gabor filters. The advantage of utilizing the top-down approach is, high speed processing and the weakness is, it cannot process table, irregular layout documents and forms (Priyadharshini N.et al., 2013).

Feature Extraction

The suggested approach uses the text, image, and background characteristics of objects. These characteristics referred to as features, which are extracted to identify and recognize text, image, and background objects. The proposed observation include simple and uncomplicated statistical features, that is, mean, standard deviation and intensity of block pixel gray scale to discriminate those objects from one another. These statistics were exploited and utilized as the feature values of pixel colors. For each pixel the features vector are compute representing the observation using sliding window with size 3*3 moving around each pixel of filtered image. Document image is transformed to a set of features known as a feature space. Pixel converted to the corresponding gray scale vector (tuple of three values mean, standard deviation, and intensity) as represented in Figure 1.

The principles lie on the observation that the color of pixel is lighter than color of background in the gray scale level. In addition, every pixel's feature values belong to the same object block is relatively close to those of its neighbors.

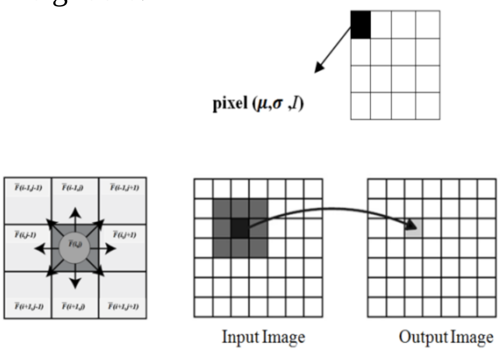


Figure 1. Grayscale Color Map Pixel to Features Vector.

Each pixel will converted to a corresponding features vector (*M, S, I*). The value of mean and variation leads us to conclude the following observations:

1. Image pixels colors are lighter than color of background.
2. Pixels that differ a little in standard deviation value or mean are belonging to the same object.

Based on these observations, the pixel of document could be segmented to text, image, and background pixel as follows:

1. Pixels representing a textual area have higher feature values and high standard deviation than their background. Standard deviation of the text block pixel was almost higher than the mean in gray scale.
2. Pixels representing a picture area have low mean and standard deviation.
3. Pixels representing a background area close to the zero standard deviation and had high mean (bright) level.

Such observation will be used to segment the three objects.

Image Segmentation based Clustering Algorithms

Clustering is the process that divides the data to groups of similar objects. Representing and present the data by smaller number of clusters Leads to lose a certain fine details of original data, but achieves simplification. Clustering is unsupervised learning technique; therefore the class labels are not well-known in advance. The clustering quality is considered by its ability to find some or all hidden patterns. Generally clustering methods can be classified into grid-based, model-based, partitioning, hierarchical and density-based methods (Vijay J., Mandar S. and Avinash S., 2014). From these types of clustering techniques, the partitioning based clustering algorithms chosen to segment images. The advantage of implementing the partitioning algorithm relies on clustering technique is constructs different partitions and evaluates them depended on some criterion. The different types of the partitioning clusters are the K medoids, K Means and fuzzy clustering (Bhagwati C. P. and G.R.Sinha, 2010; Nameirakpam D., Khumanthem M. and Yambem J. C., 2015).

Image Segmentation based Fuzzy C-means Algorithm

The FCM algorithm try to minimizes the objective function $J_m(u, v)$ with related to membership functions $u_{j,k}$ and the centroids v_k . The FCM algorithm assigns input pixels to the fuzzy clusters without labeling. Unlike hard clustering techniques like k-means clustering technique which enforce input pixels to be belong absolutely to one class, the FCM permits input pixels to be belong to the multiple clusters with variable degrees of the membership value.

Hence, Because of flexibility, The FCM is a soft segmentation algorithm which recently has been utilized for images segmentation applications. But, the main disadvantages of FCM are computational complexity and performance of algorithm was degraded significantly when noise is increased (Hamed S. and Hadi S., 2012; Esh N., Yogesh B. and Gaurav K. T. , 2012).

There are two steps for each clustering iteration in algorithm. The first step is to compute the membership function in spectral domain; the second step is mapping membership information for each pixel to spatial domain then calculates the spatial function from that (Sofien T., Najoua Ben A., and Hamid A., 2005). The outline of the FCM algorithm illustrated by the steps in algorithm 1 (Xiaofeng L. , Li S., Sumin S., Kang H., Songyu Y. and Nam L., 2013; Youngeun A., Jongan P., Younghun L., Gukjeong K. and Jonghun C., 2013):

Algorithm 1: FCM Segmentation Algorithm
Input: Image
Output: Segmented Image
Begin
Step 1: Parameters values Initialization (iteration 0)
Step 2: Scan image line by line for constructing the vector X containing the entire gray level of input image.
Step 3: Initialize randomly the centers of classes vectors V(0) starting the iteration t=1 to end of the algorithm.
Step 4: Compute membership matrix U(t) of the element $u_{j,k}$ using:

$$u_{i,k} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{i,k}}{d_{j,k}}\right)^{\frac{2}{m-1}}} \quad (1)$$

$u_{i,k}$ *is a matrix of size* $(c * n)$
Step 5: Calculate the vector $V(t) = [v_1, v_2, \ldots, v_c]$ using:

$$v_i = \frac{\sum_{k=1}^{n} u_{i,k}^{m} x_k}{\sum_{k=1}^{n} u_{i,k}^{m}} \quad (2)$$

Step 6: Convergence test: if $\left\| V^{(t)} - V^{(t-1)} \right\| > \varepsilon$ increment the iteration t by one, after that return to the step 4, Else, Go-To step 7. ε is a chosen positive threshold.
Step 7: End.

Improved FCM for Document Image Segmentation

This proposed approach for document image segmentation was based on concept of FCM clustering technique and spatial feature space. Hence, taking into consideration the spatial relationship information of the pixel with its neighborhoods. Adding this information of neighborhood pixels, the clustering will get better. Using the features for each pixel that extracted according to sliding window in section (*Feature Extraction*). These features were clustered by applying the proposed FCM, each one belongs to its region. The combination of statistical features characterization and fuzzy clustering has some advantage for the proposed approach. The Segmentation can be achieved through clustering the feature vectors. Designed the proposed algorithm to cluster feature vectors into three classes, each class belong to one region in segmented document image.

Design the Improved FCM for Document Image Segmentation

The basic idea of proposed algorithm is that instead of dealing with pixel value as in traditional FCM method, they use a matrix of vectors containing the same number of image size but with 3 columns each one for mean, standard deviation and intensity.

The proposed algorithm which is used in this system composed with extracted features shown in the following section.

The objective of proposed algorithm is to modify the objective function of standard FCM using the neighborhood information of each pixel in input document image for clustering these pixels according to region's type. Therefore the mahalanobis distance was compute of pixel x_i with its neighbor:

$$d_{i,l} = arg\{min\|F_i - F_l\|\} \quad (3)$$

$$where\ F_i\ are\ features\ of\ central\ pixel\ i$$

$$where\ F_l\ are\ neighbor\ features\ of\ pixel\ i\ \ ,L = [1,2,\ldots.8]$$

Select the pixel with minima distance with the central pixel and compute the distance between the selected pixels with the cluster centers:

$$d_{l,k} = arg\{min\|F_l - v_k\|\} \quad (4)$$

$$where\ F_l\ is\ the\ best\ neighborhood\ feature\ for\ pixel\ x_i$$

And the objective function will become:

$$J_{FCM}^{m} = \sum_{i=1}^{N}\sum_{j=1}^{c} u_{i,j}^{m}\|x_i - v_j\|^2 . d_{l,k} \quad (5)$$

Where N is the number of pixels, *m* is exponent coefficient that controls the fuzziness of the resulting partition, x_i is the i^{th} pixel in the image, v_j is the center of the j^{th} cluster and $u_{i,j}$ represents the membership of x_i belonging tov_j.

A pixel i, j is closer to specific cluster center when it's the closer neighbor are closed to specific cluster center. Hence, the objective function value is minimized when the values of membership allocated to pixels whose features are close to cluster center is high and the membership values allocated to pixels whose features are far from the cluster center is low.

The suggested method can be described by the flowchart in Figure 2.

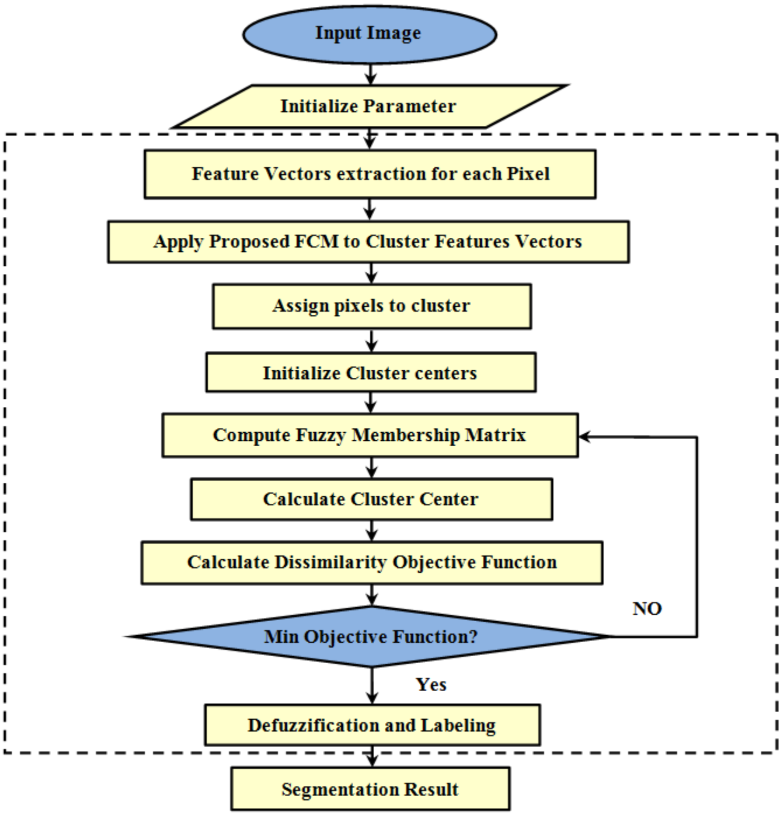


Figure 2. Segmentation Steps based on Improved FCM Approach.

The Improved FCM Algorithm (IFCM) for Document Image Segmentation

The suggested algorithm adds the heuristic information to objective function to guide the clustering process which increase the speed of FCM by fast convergence, decrease the effect of noise and saving times. So, the modified FCM algorithm is used to specify the membership degree for each pixel characterized by 3 statistical features.

The proposed modeling information method using the IFCM algorithm combined with the statistical features can be summarized by algorithm 2.

Algorithm 2: Improved FCM Segmentation Algorithm
Input: Document Image
Output: Segmented Document Image
Begin

Step 1: FCM Initialization Phase

Randomly Initialize the membership matrix U^0 (fuzzy partition matrix), Set loop counter k = 0 Number of cluster C=3, Degree of fuzziness m, Number of Iteration, Tolerance of error ε, Matrix of features vectors extracted from input image, and the stopping condition. *n* is the length of the image data.

Randomly initialize the centers of the prototype cluster vectors V(0) of size (c*3) containing the centers of each classes.

Step 2: FCM Construction Phase

– For K< max-iteration Do steps 3,4 and 5

Begin

Step 3: Update Cluster Center (prototypes, Fuzzy centroids) matrix V

Compute the matrix V(t) which composed of 3 columns v_i using:

$$v_i = \frac{\sum_{k=1}^{n} u_{i,k}^{m} F_k}{\sum_{k=1}^{n} u_{i,k}^{m}} \qquad (6)$$

And calculate the objective value J_{FCM}^{m} according to Eq. (3.16)

Step 4: Update partition matrix *U*

For each pixel and for every cluster, Calculate/compute the membership matrix U(t) (membership values in the matrix) of element $u_{i,k}$ using Eq. (3.18):

$$u_{i,k} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|F_k - v_i\|}{\|F_k - v_j\|}\right)^{\frac{2}{m-1}}} \qquad (7)$$

In the modified method, F_k and v_i are vectors of size (1*3).

Step 5: Test convergence: If the value of dissimilarity function J_{FCM}^{k} between successive iterations is less than stopping condition ($\|J_{FCM}^{k} - J_{FCM}^{k-1}\| < \varepsilon$) (the variance of cluster centers meets a criteria), then stop; otherwise, increment the iteration k (set k=k+1) and return to step 3, ε is a chosen positive threshold, i.e. $\varepsilon = 0.01$. (Till termination)

– Until completion the maximum number of iterations. Also it can set a convergence precision as condition for a loop terminates.

End-For

Step 6: Defuzzification and segmentation Phase.

When the algorithm has been convergence, the defuzzification process of the maximum membership is applied for converting the partition matrix to the segmentation results (Label each pixel with its cluster index.). Image segmentation after defuzzification using

$$C_{i,j} = \arg\{max(u_{i,j,k})\} \ , K = 1,2,3$$

Step 7: End.

IFCM Implementation and Discuss Experimental Results

In preparation for implementation, the deployment platforms and tools (namely outset) was selected and that will be used. The outsets are JAVA Object-oriented programming, preprocessing generally consists of a sequence of image-to-image processing such as noise elimination. It will not increase our knowledge of the documents structure, other than help to analyze it. the suggested algorithm are implemented using JAVA code via Eclipse version Kepler and tested on Intel Core i7-2.7 GHz CPU, 16GB Memory, Windows 7, 64 bit OS.

Very generic dataset was used for training and testing the IFCM. These combined dataset, which is part of the page framework, was the basis of the ICDAR 2009, PRIME website, UW-III and used own collection. The dataset contained 390 document images for segmentation methods, containing newspapers, magazine with complex layouts and scientific papers. Thus, these documents contain images and text, which they have been used to evaluate the IFCM.

This dataset was collected randomly from 4 categories and collected 390 images divided to 3 groups; each one contains 130 images and constructs ground truth for each document images. Using LabelMe commercial software, which is open annotation tool to build ground truth datasets. Precisely concentrate on selecting a variety of document images to test the performance evaluation of IFCM for segmentation purpose.

Figure 3 shows some of the selected samples of document image from Dataset UW-III with its ID and the ground truth used in this paper.

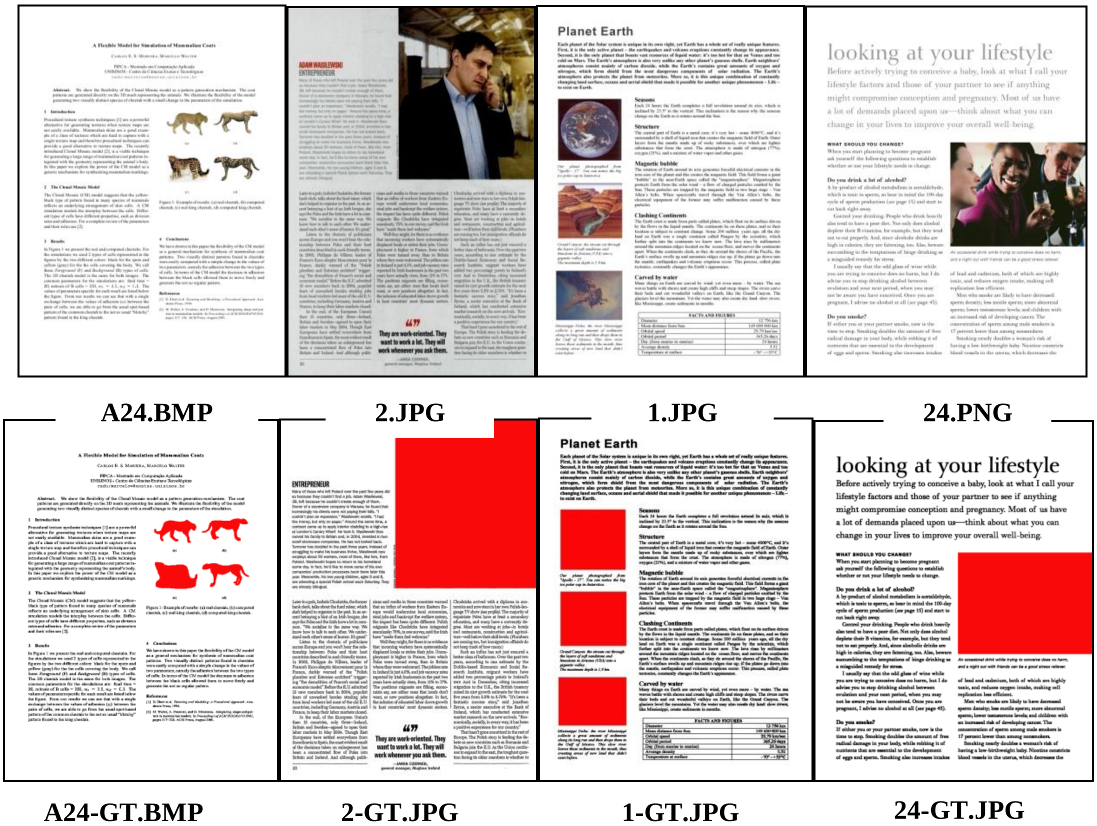


Figure 3. Some Document Image Samples with Corresponding Ground-Truths.

This section analyzed various types of document images to illustrate the performance evaluation of designed IFCM and discussed some of obtained implementation results.

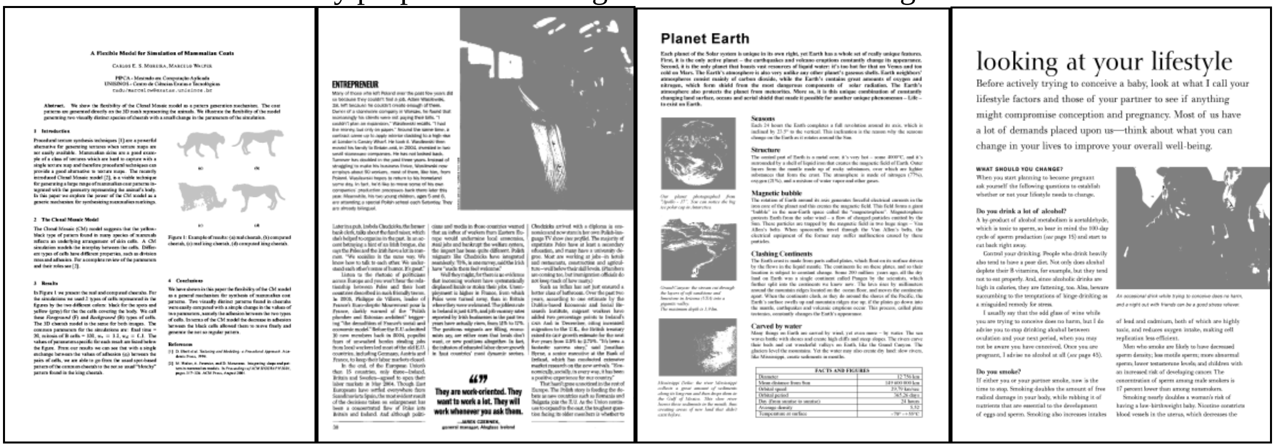
Performance of IFCM is evaluated using per-pixel accuracy. This is referring to the pixels in the document image that are accurately segmented: that is, whose region/class label matches the region determined by the ground truth zones labels. Un-segmented pixels are counted as incorrect. The IFCM presented in this paper are evaluated by using eleven metrics, such as Peak Signal to Noise Ratio (Known by acronym PSNR), Mean Absolute Error (Known as MAE), the Mean Square Error (MSE) and the Signals to Noise Ratio (Known as SNR) to evaluate the segmentation quality. Test and ground truth document images are used to compute the performance metrics. These metrics are computed separately and performance analysis is recorded. The results obtained by testing the IFCM over a number of selected document images that are subdivided into three main categories; (i) High quality document images, (ii) Low quality document image, (iii) noisy document image and uses these document image groups for performance measurement of IFCM.

To implement the proposed IFCM algorithm, firstly convert the selected input document image to gray scale image to performed segmentation. Distance measure used in this proposal for calculating the spatial difference is Mahalanobis measurements. And several parameters in the proposed algorithm are required and must be initialize these important parameters, also initialize the fuzzy membership matrix. After that the algorithm performs the main processes until entries in membership matrix (fuzzy c-partition) do not change significantly or condition criteria satisfied. Finally, at the point when the designed algorithm becomes converged, the defuzzification procedure takes place to transform fuzzy image to crisp segmented image.

In all experiments when test the proposed IFCM algorithm, the iterations number is set equal to 40, termination criterion is set to 0.0001, and the influence on the centre pixel conducted through 8-neighborhood spatial pixels ((3×3) window is taken).

Conditions for segmentation method were: fuzziness index m= 2, stop condition ε= 0.0001, number of clusters c= 3.

The results obtained by proposed IFCM algorithm are shown in Figure 4.



A24-FCM.BMP **2-FCM.JPG** **1-FCM.PNG** **24-FCM.PNG**

Figure 4. Results of IFCM based Segmentation Technique.

Table (1) show the average computational time or cost for IFCM algorithm, which is computed as time required for the designed algorithm to reach its best convergence result.

Table 1: IFCM Segmentation Computational Cost Time

Input Image	Image Resolution	Image Size	Elapsed Time (ms)
24.png	2161x2776, 300 dpi, 32 Bit depth	2.18MB	149078
54.png	2008x2833, 300 dpi, 32 Bit depth	1.01MB	140501
A24.bmp	739x1123, 72 dpi, 24 Bit depth	2.54MB	96184
A26.bmp	739x1123, 72 dpi, 24 Bit depth	2.54MB	82123
1.jpg	2479x3508, 300 dpi, 24 Bit depth	1.46MB	179772
100.png	600x400, 72 dpi, 24 Bit depth	90.2KB	30036
2.jpg	2303x3136, 300 dpi, 32 Bit depth	1.95MB	175154
3.jpg	2303x3136, 300 dpi, 24 Bit depth	1.82MB	179749
380.png	260x279, 72 dpi, 24 Bit depth	22.5 KB	18033
381.jpg	470x386, 96 dpi, 24 Bit depth	51.1 KB	22176
		Mean Average	107280.6

The qualities of the segmented experiment document images are analyzed using the measurement value of PSNR and MSE because they have high estimation to human perception of reconstruction quality; a higher PSNR would denotes that the segmented image is of higher quality, more similarity between the two images, in another words the higher PSNR value the better segmentation results. The lower MSE values show the better performance in presence of under and over segmentation by the corresponding image segmentation method algorithms; a higher MSE means a bigger difference between the ground truth and segmented image (lower image similarity).

We also used the SNR and MAE; higher values of MAE mean the image has low quality. The smaller MAE or the larger SNR value means the higher of image quality. Table (2) tabulates the segmentation performances gotten from segmented document images based on previous mentioned metrics.

Table 2: Showed PSNR & MSE for IFCM Clustering

IMAGE	PSNR	MSE	SNR	MAE
A24.BMP	52.6062	0.3568	0.0183	0.0947
1.JPG	54.0221	0.2576	0.1132	0.0756
2.JPG	53.7279	0.2756	0.0298	0.0872
3.JPG	55.3129	0.1913	0.1161	0.0595
24.PNG	54.3274	0.2401	0.1144	0.0702
54.PNG	54.6511	0.2228	0.1154	0.0658
100.PNG	58.5123	0.0916	0.1197	0.0253
A26.BMP	55.7656	0.1724	0.1185	0.0476
380.PNG	58.7585	0.0835	0.1274	0.2389
381.JPG	58.9488	0.0759	0.1298	0.2135
Mean Av.	55.66328	0.19676	0.10026	0.09783

Table (2) depicts the average value of PSNR is maximum and show the MSE is minimum. These results clearly show that IFCM are given better results.

For better evaluation of IFCM, multiple measures have been used to assess the segmentation result, So as to further confirmation the performances of segmentation listed in table (2). The accuracy (correctly segmented area) is not enough when it's used, because the segmentation may possibly also cover the region that is not in ground truth. So, the suggestion is use the following measures for performance evaluation of IFCM, these metrics namely **SSIM-Index, MS-SSIM-Index, precision**, **Sensitivity**, **Specificity** and **F-score** with **Accuracy**. As showed in table (3).

The value of SSIM index is between -1 and 1, and value 1 is indicating the two images are identical. Greater values of SSIM indicate greater image similarity. SSIM measure similarity with the greater accuracy and consistency than the MSE and PSNR, but have greater computational cost. Therefore, main focus is on the MSE and PSNR considering their commonness and the SSIM Index due to its high performance. MS-SSID It assumes values in [-1, 1]; Segmentation methods with high value of MS-SSIM, assures that the segmentation method is of a good quality.

Precision and recall (Sensitivity is true positive rate) are attractive as measures of the segmentation quality because they are precise and sensitive to over and under-segmentation, the low value of precision relate to the over-segmentation, while low value of recall relate to the under-segmentation. Over-segmentation is characterized by high recall but low precision, and that under-segmented images correspond to high precision, but low recall. The sensitivity has been defined as pixels percentages that are segmented correctly as positive regions.

Precision gives information about the validity of segmentation result and recall gives information about the correctly identified segmented pixels in an image. A comparison of two segmentations can only yield high values of precision and recall.

Segmentation mistake can be due to the missing regions which are appeared in ground truth but missing in the segmented image or added regions which are parts in segmented image but not appeared in the ground truth. Therefore, sometimes missing regions affected more than added regions, therefore the algorithms are preferred to maximize recall or sensitivity and precision. In this situation, metrics that reward sensitivity could be a good choice.

Higher value of precision and recall denote a good performance by the segmentation method. Under segmentation is resulted in the segmentation method when the recall value is low and over segmentation is resulted when value of precision is low. The F-score can be defined as a combination of precision with recall to give a single statistical measure for the segmented document image, where an F-score achieves worst value at 0 and best value at 1 (lie between 0 and 1). F-score high value indicates a perfect segmentation and a good value of precision and sensitively.

The specificity is defined as pixels percentage that is segmented correctly as negative regions. Specificity assures that the segmentation method is of a better quality and has higher perfection. Highest specificity approximately equal to zero means that the segmented pixels are in the domain of interested regions and the segmented image has fewer false positive counts.

The Segmentation accuracy also known as validity and can be obtained by calculating pixels percentages that are correctly segmented as foreground or background in the image.

Table 3: Qualitative Analysis of IFCM Method

IMAGE	SSIM Index	MS-SSIM Index	Precision	F-Score	Sensitivity	Specificity	Accuracy
A24.BMP	0.7535	0.97787	0.9420	0.9424	0.9428	0.1236	0.8819
1.JPG	0.8552	0.89305	0.9625	0.9693	0.9762	0.5740	0.9432
2.JPG	0.9128	0.94909	0.9828	0.9844	0.9860	0.8367	0.9617
3.JPG	0.9029	0.91489	0.9779	0.9782	0.9784	0.8633	0.9524
24.PNG	0.8725	0.91305	0.9247	0.9609	0.9999	0.2405	0.9264
54.PNG	0.9102	0.92450	0.9657	0.9825	0.9998	0.6170	0.9573
100.PNG	0.9845	0.99393	0.9976	0.9971	0.9966	0.9607	0.9845
A26.BMP	0.9128	0.99072	0.9647	0.9818	0.9994	0.5346	0.9656
380.PNG	0.9675	0.97938	0.9869	0.9901	0.9935	0.8591	0.9733
381.JPG	0.9697	0.97991	0.9888	0.9928	0.9969	0.8386	0.9799
Mean Av	0.9042	0.951639	0.96936	0.97795	0.98695	0.64481	0.95262

Segmentation performances based on used metrics that have been obtained from segmented document images of 24.png, 1.jpg....., and overall 382 document images are tabulated in Table (4) and (5). These tables shows the average values of used metrics for the results obtained using 390 document images.

Table 4: Quantitative Analysis of Designed IFCM Algorithm on 390 images

PSNR	MSE	SNR	MAE	Required Time/Sec.
55.09575	0.21537	0.1105	0.062751	10.75

Table 5: Qualitative Analysis of Designed IFCM Algorithm on 390 images

SSIM Index	MS-SSIM Index	Precision	F-Score	Sensitivity	Specificity	Accuracy
0.9075	0.95998	0.96135	0.96818	0.9751	60.98	95.21

The performance evaluation indices will be conducted by compute the amount of matching between the blocks (text and image) detected by the designed algorithm and blocks in the ground-truth (intersection sets of pixel in resultant image and ground truth).

Depending on the average segmentation performance that obtained for the whole 390 images, Table (5) shows the experimental results of segmentation using IFCM model was excellent. The IFCM satisfies the best results in segmenting the text, image and background in document images with accuracy of segmentation equal to 95.21% (depend on pixels error). So that, from table (5), it can be notice that the percentage of average values on all 390 segmented document images of sensitivity is 97.51 % also for F-Score 96.81 % . These results reveal that the IFCM yields to higher results in sensitivity, specificity and can segment the document image in three classes of text, image and background very well.

Another test on the document images in prepared dataset was done; the prepared dataset has been divided into three datasets. Each set contains 130 document images of approximately one type in terms of quality. Then IFCM was performed on each dataset. The evaluated results using the IFCM are presented in the table (6).

Table 6: Evaluate the IFCM on Three Datasets

Dataset	PSNR	MSE	Recall	Precision	Accuracy
Dataset1	56.36297	0.16752	94.33	92.71	96.64
Dataset2	53.75026	0.19245	92.83	92.55	94.75
Dataset3	51.36425	0.20353	84.76	83.8	94.23

To verify the performance of IFCM, two different evaluation scenarios have been applied. First one is evaluated the segmentation algorithms with each other as depicted before and the second one, compare the IFCM with the recent methods based on literature survey.

The results of the proposed algorithms and survey are shown in table 7. The proposed techniques are used pixel accuracy rate as a performance metric. Different issues were encountered when evaluate and compare the different systems that make it difficult to ensure a satisfactory comparison of both systems. It's not reasonable that the systems are compared, unless using identical environment for the same work experience and dataset on the experience. For example the IFCM use high, middle and low quality with complex background document image when evaluate it.

Table 7: Comparative results of IFCM Algorithm with Some latest Algorithms in Literature

Method	Segmentation Accuracy
Bukhari et al., (2010)	95.01%
Bukhari S. S., (2012)	89.54%
Amy W. et al., (2011) Voronoi algorithm	70%
Amy W. et al., (2011) RAST algorithm	80%
Leptonica [21]	92.39%
El-Omari N. et al., (2012)	89%
Jamal S. et al., (2009)	86.75%
Mustafa S. E. (2011)	85%
Proposed IFCM	95.21%

Conclusion

The suggested algorithm is an efficient clustering method designed for automatically grouping and labeling each pixel with the cluster that contains its features vector into different homogeneous regions (text, image and background) for all types of gray and color document images like books, and newspapers, based on the mean, standard deviation and intensity-levels of the pixel and its neighbors. Hence, these features are a good set to utilize in pixels clustering to identify the text, image and background regions. Furthermore, the initialization value of cluster centers based on the proposed observation, decreasing the time of convergence considerably and declining significantly the number of iterations.

A variety of simple, complex, color, and grey-scale documents are used to evaluate the suggested technique. Experimental results obtained from IFCM algorithm are satisfactory, giving better segmented document and indicate that the algorithm works with an average of 95.21% segmentation accuracy on both color and gray-level documents images. For this reason, it gives an opportunity to use with several different types of documents images applications where the other methods cannot provide this feasibility. Our experiments show that significant enhancements can be achieved with proposed algorithm compared with current state algorithms.

References

Melissa C. and Alexandra B. A. (2014). Texture Sparseness for Pixel Classification of Business Document Images. *International Journal on Document Analysis and Recognition (IJDAR), Volume 17, Issue 3,* pp: 257-273.

El-Omari N. K. T., Omari A. H., Al-Badarneh O. F., Hussein Abdel-jaber, (2012). Scanned Document Image Segmentation Using Back-Propagation Artificial Neural Network Based Technique. *International Journal of Computers and Communications, Volume 6, Issue 4.*

Bukhari S. S. (2012). Generic Methods for Document Layout Analysis and Preprocessing. *PhD dissertation, Technical University of Kaiserslautern.*

Amy Winder, Tim Andersen and Elisa H. Barney Smith, (2011). Extending Page Segmentation Algorithms for Mixed-Layout Document Processing. *International Conference on Document Analysis and Recognition.*

Mustafa S. E., (2011). Page Layout Analysis and Classification for Complex Scanned Documents. *Master thesis Kate Gleason College of Engineering Rochester INSTITUTE OF TECHNOLOGY Rochester, New York.*

Bukhari S. S., Shafait F., and Al Azawi M. I. A., (2010). Document image segmentation using discriminative learning over connected components. *in Proc. 9th IAPR Workshop on Document Analysis Systems, ISBN: 978-1-60558-773-8, pp. 183-190.*

Jamal S., Reza S., and Saeed M., (2009). Document Image Segmentation Using Fuzzy Classifier and the Dual-Tree DWT. *14th International CSI conference (CSICC2009), ISBN: 978-1-4244-4261-4, pp: 385 - 391, July 1-2.*

Priyadharshini N. and Vijaya MS., (2013). Document Segmentation and Region Classification Using Multilayer Perceptron. *IJCSI International Journal of Computer Science Issues, Volume 10, Issue 2, No 1, ISSN (Print): 1694-0814.*

Manish T. W., Keshao D. K. and Mahendra P. D.**,** (2015). "Document Image Segmentation using Region Based Methods. *International Journal of Computing Science and Information Technology, ISSN: 2278-9669, Volume 3, Issue 3, 01-08.*

D.Shobana and M.Phil, (2012). Layout Based Information Retrieval from Document Images. *IOSR Journal of Computer Engineering***,** *ISSN: 2278-0661, Volume 4, Issue 4, pp: 31-35.*

Yong Y., Shuying H., (2007). Image Segmentation by Fuzzy C-Means Clustering Algorithm with a Novel Penalty Term. *Computing and Informatics, Volume 26, Issue 1, pp: 17-31.*

Ritu G., Gaurav H. and Santanu C.**,** (2008). A Hierarchical Analysis Scheme for Robust Segmentation of Document Images using White-Spaces. *National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), Gujarat Section.*

Vijay J., Mandar S. and Avinash S., (2014). Color Image Segmentation Using K-Means Clustering and Otsu's Adaptive Thresholding. *International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume 3, Issue 9.*

Bhagwati C. P. and G.R.Sinha, (2010). An Adaptive K-means Clustering Algorithm for Breast Image Segmentation. *International Journal of Computer Applications (0975-8887), DOI: 10.5120/1467-1982, Volume 10, Issue 4, pp: 35-38.*

Nameirakpam D., Khumanthem M. and Yambem J. C., (2015). Image Segmentation using *K*-means Clustering Algorithm and Subtractive Clustering Algorithm. *Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015), Procedia Computer Science, DOI: 10.1016/j.procs.2015.06.090,Volume 54, pp: 764-771.*

Hamed S. and Hadi S., (2012). A Modified Fuzzy C-Means Clustering with Spatial Information for Image Segmentation. *International Journal of Computer Theory and Engineering, Volume 4, Issue 5.*

Esh N., Yogesh B. and Gaurav K. T. , (2012). Enhancement of Fuzzy C-Means Clustering using EM Algorithm. *International Journal of Computer Applications (0975 – 8887) Volume 43, Issue 13.*

Sofien T., Najoua Ben A., and Hamid A., (2005). Generalized Hough Transform for Arabic Printed Optical Character Recognition. *The International Arab Journal of Information Technology, Volume 2, Issue 4.*

Xiaofeng L. , Li S., Sumin S., Kang H., Songyu Y. and Nam L., (2013). Parallel Hough Transform-Based Straight Line Detection and Its FPGA Implementation in Embedded Vision. *Sensors ISSN: 1424-8220, Volume 13, Issue 7.*

Youngeun A., Jongan P., Younghun L., Gukjeong K. and Jonghun C., (2013). Rearranged Hough Transform based on Image Retrieval Technique. *Advanced Science and Technology Letters, ISSN: 2287-1233, Volume 29, pp: 88-91.*

Bloomberg, D. S. Leptonica: An Open Source C Library for Efficient Image Processing and Image Analysis Operations. http://code.google.com/p/leptonica/.