International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Special Issue 2, April 2014)
National Conference on Computing and Communication-2014 (NCCC'14)

Detection and Classification of Hard Exudates in Human Retinal Fundus Images Using Clustering and Random Forest Methods

T.Akila[1], G. Kavitha[2]

Department of Electronics Engineering, Madras Institute of Technology Campus, Anna University, Chennai – 600 044, India.

E-mail : [1]akilathandavan@gmail.com, [2]kavithag_mit@annauniv.edu

***Abstract*—Diabetic Retinopathy (DR) is a vascular disorder where the retina is damaged because fluid leaks from blood vessels into the retina. One of the primary lesions of diabetic retinopathy is exudates, which appear on retinal images as bright patches with various borders. In this work an image processing framework is presented to automatically detect and classify the presence of hard exudates in the human retinal fundus images. A total of 50 images have been used to detect the hard exudates from the Messidor database. Digital image processing methods help to extract the location and level of abnormalities in retinal fundus images. The contrast adaptive histogram equalization is used for preprocessing stage and Fuzzy C-Means (FCM) and K-means clustering algorithms are applied to segment the exudates in abnormal images. A set of features such as the standard deviation, mean, energy, entropy and homogeneity of the segmented regions are extracted and fed as inputs into random forest (RF) classification to discriminate between the normal and pathological image. The proposed method achieved 92.94% accuracy for early detection of DR.**

***Index Terms*—Hard Exudates, k-means clustering, Fuzzy c means, Random Forest.**

I. INTRODUCTION

Retina is the innermost coat at the back of the eye and light sensitive layer of tissue. It sends visual messages through the optic nerve. The retina serves as the film in a camera. When Light strike on the retina which initiates a cascade of electrical and chemical events that activate nerve impulses. These are sent to different visual centers of the brain through the fibers of the optic nerve. In the developed countries, the diabetic retinopathy is the most common cause of vision loss. According to the fact, diabetes is a rapidly growing disease among large part of the population. When the disease is detected in its early stages, laser photocoagulation can slow down the progression of DR. The retinal fundus of diabetic patients needs to be examined at least once a year, to ensure that treatment is received on time.

Digital image processing techniques help to extract the location and size or the level of abnormalities in retinal images. DR is the main cause of new cases of blindness among adults aged 20–74 years. During the first 20 years of the disease, approximately all patients with type 1 diabetes and greater than 60% of patients with type 2 diabetes have retinopathy. In the older-onset group, in which other eye diseases were common, due to DR one-third of the cases have blindness. It occurs when the increased glucose level in the blood damages the capillaries. As a result of this damage, the capillaries leak blood and fluid on the retina [14]. The visual effects of this leakage are features, such as hard exudates, microaneurysms, cotton wool spots or venous loops, hemorrhages, of DR [15].

Exudates are accumulations of lipid and protein in the retina. Typically they are bright, reflective, white or cream colored lesions. They indicate increased vessel permeability and a connected risk of retinal edema. They are a marker of fluid accumulation in the retina. When they present close to the macula center they form sight threatening lesions [2]. Most of the time they are seen together with microaneurysms.

The aim of the present work is to process digital color retinal images to automatically detect and classify hard exudates. Image segmentation is widely used in grouping, exploratory pattern-analysis, decision-making, and machine-learning situations for medical images. These algorithms are used to automate process of segmentation of the hard exudates in retinal fundus images. K-means algorithm follows a simple and easy way to classify a given document set through a certain number of clusters. The K-Means clustering method has low complexity [11]. Fuzzy clustering is more natural than hard clustering. It is used to highlight salient regions, extracts relevant features and finally classifies those regions using random forest classifier.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Special Issue 2, April 2014)
National Conference on Computing and Communication-2014 (NCCC'14)

In this work, the abnormal retinal images are subjected to two clustering algorithms. K-Means and Fuzzy C-Means algorithms are used. Abnormalities are extracted using these algorithms. From the segmented regions, features are extracted and fed as input to the development of a supervised classification technique that is random forest. Better results are given as input to RF classifier. Random forest consists of a number of single predictor trees such that each tree is trained over randomly and independently selected samples from the training data [21]. This data represents the same distribution of whole training data. The random forest depends on the strength of its individual trees and the correlation between them.

II. METHODOLOGY

A. Preprocessing

Contrast-Limited Adaptive Histogram Equalization (CLAHE) [15] was applied for contrast enhancement to improve the image quality. CLAHE operates on small regions in the image. Each small region's contrast is enhanced with histogram equalization. This image processing technique can improve the local contrast of the image and have a more uniform image so the edges of the exudates are enhanced.

B. Segmentation

Image segmentation is used to partition an image into meaningful regions. They were HE candidate regions that had to be classified afterwards. Segmentation was accomplished using the histogram properties of the second color component of the preprocessed image. The goal of segmentation is to simplify the representation of an image and provide meaningful information which is easier to analyze. This paper use two clustering based techniques that are K-means clustering and FCM clustering.

K-Means or Hard C-Means clustering is basically a partitioning method applied to analyze data and treats observations of the data as objects based on locations and distance between various input data points [11]. Figure 1 shows that the block diagram of hard exudates detection. Clusters can be chosen randomly, manually, or based on some conditions. Distance between the cluster centre and pixel is calculated by the absolute difference or squared between a cluster centre and pixel. The difference is normally based on intensity, pixel color, location, texture and, or a weighted combination of these factors. The initial set of clusters is important to the quality of the final result of the clustering method. The algorithm is extremely fast, a collective method is to run the obtainable.

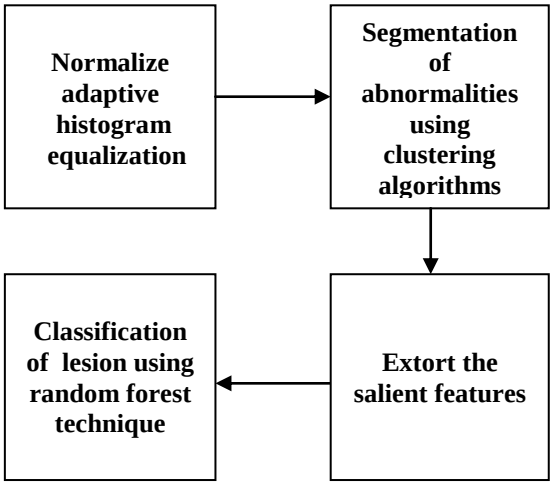


Fig 1. Block diagram of hard exudates detection

In a dataset, a desired number of clusters K and a set of k initial starting points, the desired number of distinct clusters and their centroids are found by the K-Means clustering algorithm. A centroid is the point whose co-ordinates are obtained by means of computing the average of each of the co-ordinates of the points of samples assigned to the clusters.

Clustering Method is an iterative technique that is used to partition an image into clusters [16, 17]. The following method can be employed to find the cluster centers

1. Compute the intensity distribution (also called the histogram) of the intensities.
2. Initialize the centroids with k random intensities.

$$C(i) = \text{argmin}\|X(i)-\mu_j\|^2 \qquad (1)$$

3. Cluster the points from the centroid intensities based on distance of their intensities.
4. Calculate the new centroid for each of the clusters.

Where i iterates over the all the intensities, k is a the number of clusters to be found, μ_j are the centroid intensities and j iterates over all the centroids (1). The algorithm is very fast, a common method is to run the algorithm several times and return the best clustering found.

Fuzzy C-Means is a soft segmentation algorithm. It allows pixels belong to multiple clusters with varying degree of membership [18]. This technique preserves lot of information from the original image than other segmentation methods. K-means and the Fuzzy C-Means are two successful region-based approaches. FCM can be obtained from the k-means algorithm by a little modification [20].

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Special Issue 2, April 2014)
National Conference on Computing and Communication-2014 (NCCC'14)

The algorithm is an iterative clustering technique that produce an optimal c partition by minimizing the weighted within group sum of squared error objective function H

$$H=\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^{q}\, d^{2}\,(x_k, v_i) \qquad (2)$$

Where X = {x1, x2, .. , x_n} is the data set in the p-dimensional vector space, n is the number of data items, c is the number of clusters with $2 \leq c < n$, cluster centre v_i, u_{ik} is the degree of membership of x_k in the i^{th} cluster, v_i is the prototype of the centre of cluster i, q is a weighting exponent on each fuzzy membership and d^2 (x_k, v_i) is a distance measure between object x_k. A solution of the object function H (2) can be obtained through an iterative process [23]. The following steps can be employed to find the object function:

- Select a number of clusters
- Assign indiscriminately to each point coefficients
- Repeat until the algorithm has meet at a point
- Determine the centroid for all cluster
- Compute each point coefficients of being in the clusters

C. Feature extraction

From the segmented region, features are extracted. Extracted features consist of standard deviation, mean, energy, entropy and homogeneity. The features show distinct variation between normal and abnormal images and it is given as input to classifier.

D. Classification using Random Forest

The features obtained using K-means algorithm and fed as input to RF classifier. K-means shows better variations compared to fuzzy c-means. Random forests are an ensemble learning method for classification that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. A RF consists of multiple layers of nodes, with each layer entirely connected to the next one. Each node is a neuron with a nonlinear activation function except for the input nodes. A random forest is a classifier consisting of a collection of tree structured classifiers {h(**x**, Θ_k), k=1...} where the {Θ_k} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input **x.**

RF utilizes a supervised learning technique called back propagation for training the network. RF is a modification of the standard linear perceptron and can distinguish data that are not linearly separable.

The following steps describe how the random forest is built in this work:

1. From the original data draw ntree bootstrap samples.
2. Select a random subspace of ntrial features and detect the best split at each node.
3. Detect the threshold and feature which leads to the best split by maximizing the Information Gain (IG) index.
4. Continue until the maximum depth (mdepth) is reached, or, there only remains a few numbers of samples - nminleaf - in a node. A node that cannot be further split forms a leaf.
   This implementation of random does not only check ntrial number of features and it also tries as many features to detect ntrial features with Information Gain more than zero. Whenever the algorithm finds ntrial features with IG > 0, it stops searching the space and selects the split threshold of the feature with highest IG. If there is a case that all the features are checked but there is less number of features than ntrial with IG > 0, it still makes the split. A node is considered as a leaf if none of the features have IG > 0.
5. There is no way to determine the best number for ntrial and it depends on the nature of the classification problem, it has been confirmed that this number must be much fewer the number of features. To satisfy this constraint, ntrial has been set to the rounded value of the squared root of the number of features.

Given an ensemble of classifiers $h_1(x)$, $h_2(x)$,...,$h_K(x)$, and with the training set drawn at random from the distribution of the random vector X,Y, explain the margin function as

$$mg(X,Y) = av_k\, I(h_k(X)=Y) - \max_{j \neq Y} av_k\, I(h_k(X)= j) \qquad (3)$$

Where I(•) is the indicator function [21]. For the right class exceeds the average vote for any other class the margin measures the extent to which the average number of votes at X, Y. The generalization error is specified by

$$PE^* = P_{X,Y}(mg(X,Y) < 0) \qquad (4)$$

Where the subscripts X,Y indicate that the probability is over the X,Y space. In random forests, $h_k(X) = h(X, \Theta_k)$. For a large number of trees, it follows from the tree structure and strong law of large numbers. The Random forests are a truly random statistical method. The random forest classifier runs efficiently on large databases.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Special Issue 2, April 2014)
National Conference on Computing and Communication-2014 (NCCC'14)

It can handle thousands of input variables without variable detection. Random forest is easy to set parameters and simple to classify.

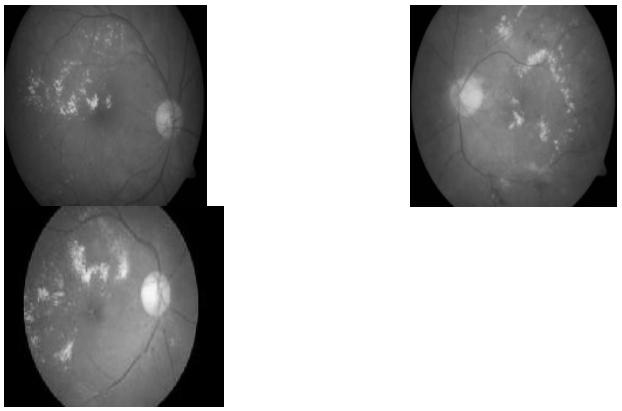
III. Results and Discussion

Figure 2 shows the result obtained using k-means and fuzzy c-means clustering methods. The retinal images are segmented and each image is characterized by its corresponding segmented region. Extracted regions are discriminated as exudates or non exudates. Figure 2(a) shows the grayscale of the input images. Figure 2 (b) shows the preprocessed input images.

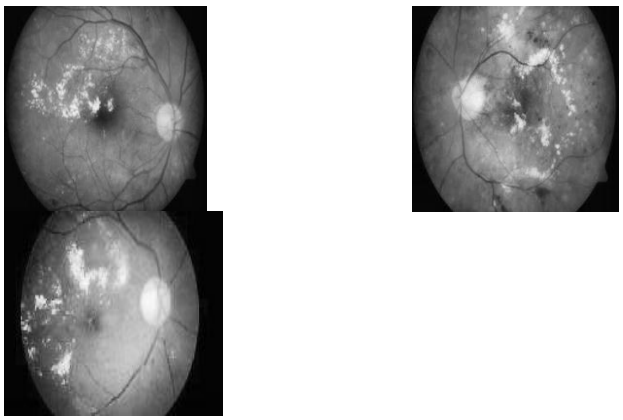
This is accomplished by extracting a set of features for each region and then the regions are classified based on the generated feature vectors. Figure 2(c) shows the exudates detection of retinal fundus images using artificial intelligence technique based clustering segmentation. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. When perform analysis of complex data one of the major problems stems from the number of variables involved. Figure 2(d) shows the segmentation of hard exudates using FCM clustering technique. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy. Statistical features for the hard exudates and their values have been determined and given in the table 1.

The mean is the average value which defines the general brightness of the image. The standard deviation is known as the square root of the variance and defines the contrast. Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. The energy measures something about how gray levels are distributed. Homogeneity returns a value that measures the closeness of the distribution of elements in the gray level. Table 1 contains the feature extraction of normal and abnormal images using k means clustering. Table 2 contains the feature extraction of normal and abnormal images using fuzzy c-means clustering.

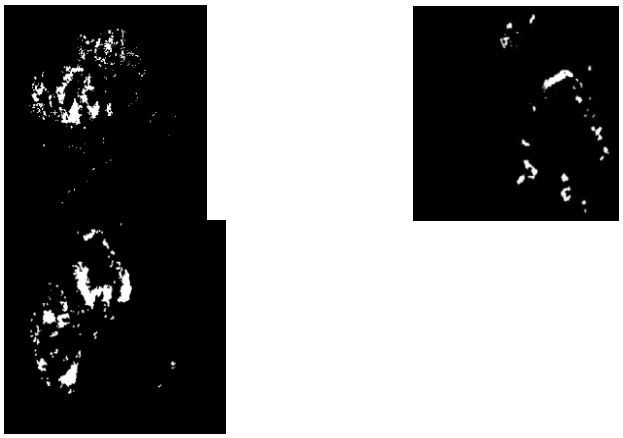
The difference between the feature extraction of normal and abnormal image values using K-means clustering are higher than the feature extraction of image values using FCM clustering.



(a)

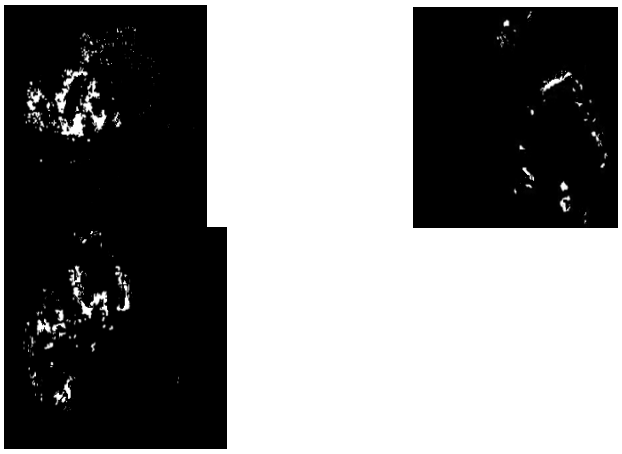


(b)



(c)

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Special Issue 2, April 2014)
National Conference on Computing and Communication-2014 (NCCC'14)



(d)

Fig. 2 (a) Grayscale image (b) Preprocessing image (c) Segmentation of hard exudates using k- means clustering (d) Segmentation of hard exudates using FCM clustering

Feature sets play one of the most important roles in a detection system. A best feature set should signify characteristic of a class that helps distinguish it from other classes.

TABLE 1.
FEATURE EXTRACTION OF NORMAL AND ABNORMAL IMAGES USING K-MEANS

K-means (N=50)			
Features	**Normal image values**	**Abnormal image values**	**Differences**
Entropy	0.010832	0.010367	0.000104
Homogeneity	0.718865	0.725455	0.00659
Energy	1	0.992693	0.007307
Mean	0.004666	0.003956	0.00071
Standard deviation	0.235728	0.34111	0.105382

TABLE 2.
FEATURE EXTRACTION OF NORMAL AND ABNORMAL IMAGES USING FCM

FCM (N=50)			
Features	**Normal image values**	**Abnormal image values**	**Differences**
Entropy	0.01247	0.012523	0.000053
Homogeneity	0.72245	0.726656	0.004206
Energy	0.999469	1	0.000531
Mean	0.004732	0.004801	0.000078
Standard deviation	0.279445	0.301971	0.022526

A patient was classified as abnormal if the presence of exudates is found else it is classified as normal. This random forest classifier has a sensitivity of 88.8%, specificity of 94% and accuracy of 92.94% from the calculation of true positive, false positive, true negative and false negative. Specificity was always high because the number of true negatives was much higher than the number of false positives. The positive predictive value was regarded as a more informative measure. The random forest classifier is not very sensitive to outliers in training data and easy to set parameters. It offers an experimental method for detecting variable intersections.

IV. CONCLUSION

In this work exudates are extracted and classified using two clustering techniques and RF classifier. Results shows that FCM produces close results to K-Means clustering but from the obtained results the K-Means algorithm is better than FCM algorithm. The random forest classifier produces good classification of abnormalities. Accuracy is found to be 92.94%. Hence this framework could be used to assist the ophthalmologist to grade the retinal diseases.

REFERENCES

[1] A. D. Fleming et al., "Automated detection of exudates for diabetic retinopathy screening," PMB, vol. 52, pp. 7385–7396, August 2007.

[2] C.JayaKumari, R.Maruthi,"Detection of Hard Exudates in Color Fundus Images of the Human",Retina International Conference on Communication Technology and System Design 2011.

[3] Jaeger, H,"The echo state approach to analyzing and training recurrent neural networks; (Tech. Rep. No. 148). Bremen", German National Research Center for Information Technology, 2001.

[4] D. Vallabha, R. Dorairaj, K. Namuduri, and H.Thompson, "Automated detection and classification of vascular abnormalities in diabetic retinopathy," Proceedings of Thirty- Eighth Asilomar Conference on Signals, Systems and Computers, vol. 2, pp. 1625-1629, November 2004.

[5] Eero Salli, Hannu J Aronen, Sauli Savolainen, Antti Korvenoja, AriVisa, "Contextual clustering for analysis of functional MRI data", IEEE transactions on medical imaging, 20(5), 403-414, 2001.

[6] M. Niemeijer et al., "Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis," IOVS, vol. 48(5), pp. 2260–2267, May 2007.

[7] S. S. Basha, K. S. Prasad, "Automatic Detection of Hard Exudates in Diabetic Retinopathy Using Morphological Segmentation and Fuzzy Logic", IJCSNS International Journal of Computer Science and Network Security, Vol. 8, No. 12, 2008.

[8] C. I. S´anchez, R. Hornero, M. I. Lopez, M. Aboy, J. Poza, D. Abasolo, "A novel automatic image processing algorithm for detection of hard exudates based on retinal image analysis", Medical Engineering and Physics, Elsevier, Vol. 30,pp. 350-357, 2008.

[9] Albayrak, S.—Armasyali, F.: Fuzzy C-Means Clustering on Medical Diagnostic System. Proc. Int. XII Turkish Symp. on Artif. Intel. NN, 2003.

[10] Akara Sophia , Bunyarit Uyyanonvara and Sarah Barman," Automatic Exudate Detection from Non-dilated Diabetic Retinopathy Retinal Images Using Fuzzy C-means Clustering" Sensors , Vol.9, pp.2148-2161,2009.

[11] Oyelade., O. J, Oladipupo., O. O, Obagbuwa., I. C.: Application of k-Means Clustering algorithm for prediction of Students' Academic Performance, International Journal of Computer Science and Information Security, vol. 7,292-295, 2010.

[12] D. L. Pham., C. Xu., L. Prince.: Current methods in medical images segmentation, Annual review of biomedical engineering, vol.2, 315-337,2000.

[13] Klein, R., Klein, B. E. K., Moss, S. E., Davis. D., and DeMets, D.L., The Wisconsin Epidemiologic Study of Diabetic Retinopathy III, prevalence and risk of diabetic retinopathy when age at diagnosis is 30 or more years. Arch. Ophthalmol. 102(4):527–532, 1984.

[14] Frank, R. N., Diabetic retinopathy. Prog. Retina. Eye Res. 14 (2):361–392,1995.

[15] Acharya, U. R., Ng, E. Y. K., and Suri, J. S., Image modeling of human eye. Artech House, MA, 2008.

[16] R.M. Haralick and L.G. Shapiro, Survey image segmentation techniques, Comput. Vision Graphics Image Process., vol. 29, pp. 100-132, 1985.

[17] P.K. Sahoo, S. Soltani, A.K.C. Wong and Y.C. Chen, A survey of thresholding techniques, Compute. Vision Graphics Image Process., vol. 41, pp. 233-260, 1988.

[18] M.Abdulghafour, "Image segmentation using Fuzzy logic and genetic algorithms," Journal of WSCG, vol, No.1, 2003.

[19] O. Faust, R. AcharyaU., E. Y. K. Ng,K.-H. Ng, and J. S. Suri, "Algorithms for the automated detection of diabetic retinopathy using digital fundus images: A review," J. Med. Syst., April 2012.

[20] Omid Jamshidi and Abdol Hamid Pilevar," Automatic Segmentation of Medical Images Using Fuzzy c-Means and the Genetic Algorithm" Medical Intelligence Lab, Department of Computer Engineering, December 2012.

[21] Leo Breiman. Random forests. Machine learning,45(1):5–32, 2001.

[22] Yong Yang, Shuying Huang,' Image segmentation by fuzzy c-means clustering algorithm with a novel penalty term, Computing and Informatics, Vol. 26, , 17–31, 2007.

[23] Robert L. Cannon, Jitendra V. Dave,James C. Bezdek,' Efficient Implementation of the Fuzzy c-Means Clusteng Algornthms' IEEE transactions on pattern analysis and machine intelligence. vol. pami-8, no. 2, march 1986.