Genetic Algorithm for Retinal Image Analysis

Jestin V.K.
Karunya University
Coimbatore
India

J.Anitha
Karunya University
Coimbatore
India

D.Jude Hemanth
Karunya University
Coimbatore
India

ABSTRACT
Diabetic Retinopathy is one of the leading causes of blindness. Hard exudates have been found to be one of the most prevalent earliest clinical signs of retinopathy. Thus, identification and classification of hard exudates from retinal images is clinically significant. For this purpose the images from the hospitals were taken as reference. In this work, Genetic Algorithm (GA) for best feature selection from retinal images is proposed. The features that improve the classification accuracy are selected by Genetic Algorithm and termed as optimized feature set. The others that degrade the performance are rejected at the end of specified generation (in this case 100 generations).

General Terms
Image classification, feature selection.

Keywords
Diabetic retinopathy, hard exudates, retinal images, Genetic Algorithm (GA).

1. INTRODUCTION
Diabetic Retinopathy (DR) is one of the leading causes of blindness and vision defects in developed countries [1]. Fundus images permit a high quality permanent record of retinal fundus for detecting early signs of DR and monitoring its progression. Moreover, their digital nature allows automatic analysis to reduce the workloads of the ophthalmologists and the health costs in the screening of the disease [2]. Many of the common eye diseases have a very slow progression and their early detection is difficult. Consequently there is a possibility for the prediction of the ophthalmologists to go wrong. Almost all patients with type 1 diabetes and over 60% of patients with type 2 diabetes will have some degree of retinopathy [3].

Feature extraction is done in order to classify abnormal retinal images. Most of these features are either partially or completely irrelevant or redundant to the classified target. It cannot be known in advance which features will provide sufficient information to discriminate among the classes. It is also infeasible to include all possible features in the processes of classifying the patterns and objects. Feature selection is one of the major tasks in classification problems. The main purpose of feature selection is to select a number of features used in the classification and at the same time to maintain acceptable classification accuracy.

In this work, Genetic Algorithm (GA) is used for optimizing or selecting best features which gives a reduced feature set eventually results in high classification accuracy. Reducing the dimensions of the feature space not only reduces the computational complexity, but also increases estimated performance of the classifiers. GA is biologically inspired and has many mechanisms resembling natural evolution [4] [5] [7]. The main objective is to optimize feature set for the improved classification accuracy and application of GA is experimented for best feature selection which can be further used on retinal image classification. Abnormal retinal images from four different categories are used in this research.

2. PROPOSED METHODOLOGY
The framework for the proposed scheme for optimization of features is shown in Fig 1.

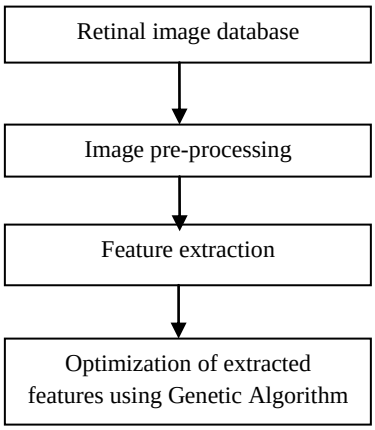

Fig 1. Flow representation of the proposed work

The abnormal retinal images from four different categories are collected from ophthalmologists for disease identification system. The images are initially processed to enhance the contrast in order to accurately detect the anatomical structures. An extensive set of features are extracted from these anatomical structures which ultimately aid in enhancing the accuracy of the automated image classification system.

The rest of this paper is organized as follows: Section 3 deals with the retinal image database and image pre-processing techniques, Section 4 comprises the feature extraction methodologies, Section 5 deals with the Genetic Algorithm

used for optimization of extracted features and Section 6 shows results and discussions.

3. RETINAL IMAGE DATABASE AND IMAGE PREPROCESSING

Most of the test images are collected from Eye Hospitals and STARE database. The image database that used in this work consists of 420 digital retinal images obtained using the imaging camera. The images are stored as colour TIFF images and are 1504×1000 pixels in size for all the objects. The intensity value of all the retinal images ranges from 0 to 255 (for each R, G & B planes). The real time images are collected from four abnormal categories namely Non-Proliferative Diabetic Retinopathy(NPDR), Central Retinal Vein Occlusion(CRVO), Choroidal Neo Vascularisation(CNVM) and Central Serous Retinopathy(CSR).

Pre-processing is an important and diverse set of image preparation program. Retinal images usually have pathological noise and various texture backgrounds, which may cause difficulties in extraction. So it should be removed. The green channel of color retinal images is extracted as an RGB image gives the highest contrast between vessels and background, this channel is a good choice for contrast enhancement. Therefore as a primary step in pre-processing, the green channel of retinal images is extracted. The raw retinal images usually have very low contrast which is signified by the grouping of large peaks in a small area on the histogram plot. The contrast of the retinal images is improved by histogram equalization which brings out details which are not clearly visible in the raw retinal images. A better contrast is obtained by Gabor filtering the resultant image. The second derivative Gabor filtering is used since it distinguishes the background and foreground region besides enhancing the contrast of the image. These methods are applied separately to the red, green and blue components of the RGB color values of the image.

4. FEATURE EXTRACTION

To classify images into four different categories of diseases, it is represented them using relevant features.The purpose of feature extraction is to reduce the original data set by measuring certain properties, or features, that distinguish one input pattern from another pattern. The feature set should be selected such that the between-class discrimination is maximized while the within class discrimination is minimized.

Fourteen features are used in this work among which eleven are based on texture of image (statistical features) and three are disease based features.

4.1 Statistical Features

4.1.1 Mean

It is the mean of pixels in image. The n[th] moment of about mean is

$$\mu_{n=} \sum_{i=0}^{L-1} (z_i - m)^n \, p(z_i) \quad (1)$$

where m is the mean value of z(the gray level)

$$m = \sum_{i=0}^{L-1} z_i \, p(z_i) \quad (2)$$

4.1.2 Standard Deviation

$$SN = \frac{1}{N} \sum_{i=1}^{N} (x_i - x')^2 \quad (3)$$

Where x' is the mean value.

4.1.3 Variance

It is a measure of gray level contrast that can be used to establish descriptors of relative components.

$$\sigma(z) = \mu_2(z) \quad (4)$$

4.1.4 Entropy

$H(z) = -\int p(z) \ln p(z) dz$ is Shannon's entropy of the image window z, and p is the distribution of the gray levels in the considered window.

4.1.5 Contrast

Contrast is determined by the difference in the color and brightness of the object and other objects within the same field of view.

$$\text{Contrast} = \sum_{i,j} |i-j|^2 \, p(i,j) \quad (5)$$

here *i* and *j* are the *i*[th] *j*[th] element of the two dimensional image and *p(i,j)* is the average intensity of all pixel values in the GLCM(Gray Level Co-occurrence Matrix) of the image.

4.1.6 Correlation

The distance among the pixels can be calculated using the correlation distance.

$$d_{rs} = 1 - \frac{(x_r - x_r')(x_s - x_s')^T}{\left[(x_r - x_r')(x_r - x_r')^T\right]^{0.5} \left[(x_s - x_s')(x_s - x_s')^T\right]^{0.5}} \quad (6)$$

where $x_r' = \frac{1}{n} \sum_j x_{rj}$ and $x_s' = \frac{1}{n} \sum_j x_{sj}$

$$\text{Correlation} = \sum_{i,j} \frac{(i - \mu_i)(j - \mu_j) p(i,j)}{\sigma_i \sigma_j} \quad (7)$$

4.1.7 Energy

Returns the sum of squared elements in the GLCM and range will be in [0 1].

$$\text{Energy} = \sum_{i,j} p(i,j)^2 \quad (8)$$

where *i* the number of rows, j is the number of columns and *p(i,j)* is the mean of the GLCM of the image.

4.1.8 Homogeneity

Returns a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal and range will be in [0 1].

$$\text{Homogeneity} = \sum_{i,j} \frac{p(i,j)}{1 + |i-j|} \quad (9)$$

where *i* the number of rows, j is the number of columns and *p(i,j)* is the mean of the GLCM of the image.

4.1.9 Angular Second Moment

This is a measure of local homogeneity and the opposite of Entropy. High values of ASM occur when the pixels in the moving window are very similar.

$$\text{ASM}=\sum_i \sum_j \{p(i,j)\}^2 \quad (10)$$

where *i,j* is the number of rows and columns respectively and *p(i,j)* is the mean of GLCM of the image.

4.1.10 Inverse differential moment

This measure relates inversely to the contrast measure. It is a direct measure of the local homogeneity of a digital image. Low values are associated with low homogeneity.

$$\text{IDM}=\sum_i \sum_j \left\{\frac{p(i,j)}{1+(i-j)^2}\right\} \quad (11)$$

where *i,j* is the number of rows and columns respectively and *p(i,j)* is the mean of GLCM of the image.

4.1.11 Skewness

It is a measure of skewness of the histogram.

$$\text{?}_3(z) = \sum_{i=0}^{L-1} (z_i - m)^3 p(z_i) \quad (12)$$

where z_i is the i^{th} gray level and $p(z_i)$ is the mean of the GLCM of the image.

4.2 Disease based features

4.2.1 Area

It gives the area of the disease spread and objects in the eye .

$$\rho(i,j) = \frac{1}{M-N} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} b_i \ (i-m, j-n) \quad (13)$$

which supports M,X,N region for every (i,j) point of the image.

4.2.2 Minimum intensity

It gives the minimum intensity of the abnormality present in the retinal image. It is actually a pixel value measurement.

4.2.3 Mean intensity

It gives the mean intensity of the pixels present in the retinal image. It is also a pixel value measurement.

5. OPTIMIZATION OF EXTRACTED FEATURES USING GENETIC ALGORITHM

Genetic algorithm is based on the process of Darwin's Theory of Evolution. By starting with a set of potential solutions and changing them during several iterations the Genetic Algorithm hopes to converge on the most 'fit' solution.

The process begins with a set of potential solutions or chromosomes (usually in the form of bit strings) that are randomly generated or selected. The entire set of these chromosomes comprises a population. The chromosomes evolve during several iterations or generations. New generations (offspring) are generated using the crossover and mutation technique. Crossover involves splitting two chromosomes and then combining one half of each chromosome with the other pair. Mutation involves flipping a single bit of chromosome.

The chromosomes are then evaluated using a certain fitness criteria and the best ones are kept while the others are discarded. This process repeats until one chromosome has the best fitness and thus is taken as the best solution of the problem. These parameters are taken out of trial and error method.

5.1 Algorithm

STEP1: Initialize the random population.

STEP2: Calculate the fitness value of the random population

STEP3: Crossover is done between the fittest individual.

STEP4: Mutation is done between the fittest individual.

STEP5: New population is created.

STEP6: End of the generation.

STEP7: If the generation is not ended, it will calculate fitness value.

STEP8: If the generation is ended, it will calculate fitness individual.

5.2 Initial population

All the fifteen features considered in this research work are represented by a string of binary bits '0' or '1'and it form a fifteen bit string. Each string is called chromosome and each bit is called gene. In this case, 2^{15} potential solutions (chromosomes) are formed because fifteen features are considered here. This entire set is called population.

Then these chromosomes are evaluated based on fitness function. For the problem of feature selection, a chromosome has length d, the total number of features. A '1' stands for selected feature, whereas a '0' stands for a rejected feature.

A fitness function evaluates trained classifiers by considering overall accuracy on known test data, the difference (balance) between class accuracies, and the number of features considered. Balance is measured as the difference between the highest class accuracy and the lowest class accuracies among all cases. The GA seeks higher overall accuracy while avoiding bias among the different classes using a few features for efficient classification. The GA minimizing objective function is:

$$\text{Obj}= C_{mask} * \text{number of unmasked features} + C_{bal} * \text{class accuracy} \quad (14)$$

Where C_{mask}, C_{bal} are the objective function coefficients. The authors' experiments with same values as in [14]: C_{mask}=1.0 and C_{bal} =10.0. Number of unmasked features mentioned in the objective function is the number of selected features and the class accuracy is the average of class accuracies from each class. The values of objective function show minimum and maximum such as 3.00 and 5.50 respectively. So the chromosome with fitness value 3.00 is taken as best chromosome and its features is noted down for further classification process. The GA places maximum emphasis

upon achieving high accuracy by maintaining balance and dimensionality reduction of feature set.

5.3 Crossover

Crossover is done between the fittest individual. Crossover is a genetic operator that combines (mates) two chromosomes (parents) to produce a new chromosome (offspring).the idea behind crossover is that the new chromosome may be better than both of the parents if it takes the best characteristics from each of the parents .Crossover occurs during evolution according to a user-definable crossover probability.

One point

A crossover operator that randomly selects a crossover point within a chromosome then interchanges the two parent chromosomes at this point to produce two new offspring.

Consider the following 2 parents which have been selected for crossover. The"|" symbol indicates the randomly chosen crossover point.

Parent1:11001|010

Parent2:00100|111

After interchanging the parent chromosomes at the crossover point, the following offsprings are produced:

Offspring1:11001|111

Offspring2:00100|010

Two point

A crossover operator that randomly selects two crossover points within a chromosome then interchanges the two parent chromosomes between these points to produce two new offspring. Consider the following 2 parents which have been selected for crossover. The "|" symbols indicate the randomly chosen crossover points.

Parent1:110|010|10

Parent2:001|001|11

After interchanging the parent chromosomes between the crossover points, the following offsprings are produced.

Offspring1:110|001|10

Offspring2:001|010|11

5.4 Mutation

Mutation is a genetic operator that alters one or more gene value in a chromosome from its initial state. This can result in entirely new gene values being added to the gene pool. With these new gene values, the genetic algorithm may be able to arrive at better solution than was previously possible. In this work mutation process is done in different ways such as:

Flip bit-A mutation operator that simply inverts the value of the chosen gene (0 goes to 1 and 1 goes to 0).This mutation operator can only used for binary genes.

Boundary-A mutation operator that replaces the value of the chosen gene with either the upper or lower bound for that gene (chosen randomly).this mutation operator can only be used for integer and float genes.

Non Uniform-A mutation operator that increases the probability that the amount of the mutation will be close to 0 as the generation number increases. This mutation operator keeps the population from stagnating in the early stages of the evolution then allows the genetic algorithm to fine tune the solution in the later stages of evolution. This mutation operator can only be used for integer and float genes.

Uniform-A mutation operator that replaces the value of the chosen gene with a uniform random value selected between the user-specified upper and lower bounds for that gene. This mutation operator can only be used for integer and float genes. Genetic Algorithm can be used in both unconstrained and constrained optimization problems. It can be applied to find the best feature value in this work. Since there are so many combinations, all of them cannot be used and checked out. Hence the best feature value that gives high efficient output can be found out using GA logic.

5.5 Flowchart

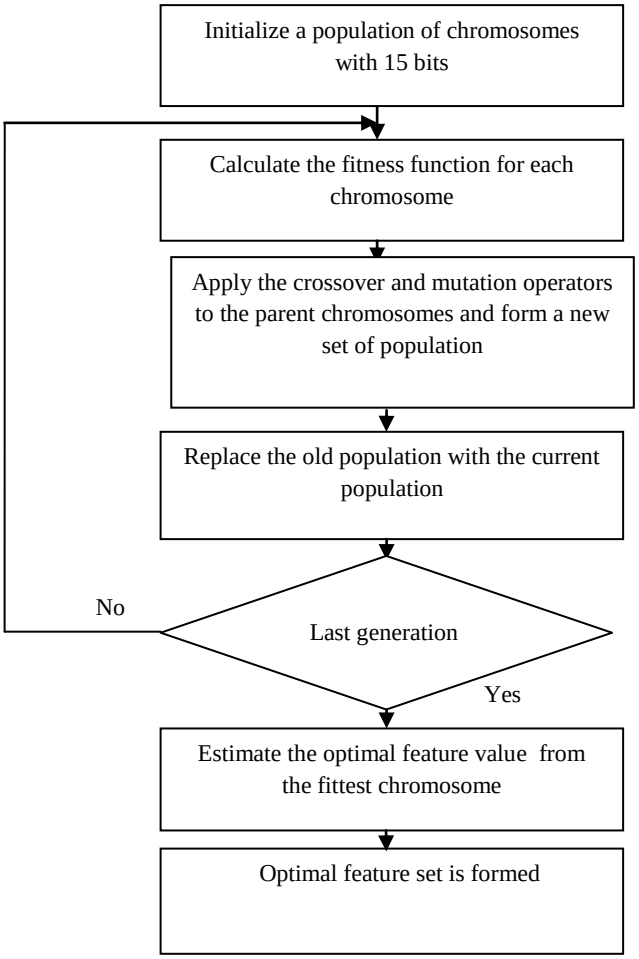

Fig 2. Flow chart for genetic algorithm

6. RESULTS AND DISCUSSIONS

Experiments are conducted on real time retinal images collected from hospital. After pre-processing, the features mentioned in (4) are extracted. Then all these fifteen features are represented by fifteen bit string consisting of binary '1's

and '0's. '1' means selected feature and '0' means unselected feature.

Iterative process is done up to 100 generation and it can be increased or decreased. If it is decreased, number of selected features will be equal or one less than original feature set and efficiency of GA will not be attained. If it is increased to a high value, convergence time period will be more. This whole process is done for specific times and the most repetitive features are selected which yields best optimized features as:

Table 1: Table showing input feature set and optimized feature set

Retinal Image Features	Optimized Retinal Image Features
Mean, Standard Deviation, Variance, Entropy, Contrast, Correlation, Energy, Homogeneity, Angular Second Moment, Inverse Differential Moment, Skewness, Area, Mean Intensity, Minimum Intensity, Euler	Mean, Variance, Entropy, Energy, Homogeneity, Angular Second moment, Inverse Differential moment, Area, Mean Intensity, Minimum Intensity

7. ACKNOWLEDGEMENT

Authors would like to thank Lotus Eye Care Hospital, Coimbatore, India for providing real time retinal images for the research work.

8. REFERENCES

[1] Clara I. Sánchez , María García , Agustín Mayo ,María López , Roberto Hornero , May 2009, "Retinal image analysis based on mixture models to detect hard exudates", Journal for Medical Image Analysis, Journal for Medical Image Analysis, Vol. 13 ,pp.650-658.

[2] Wong li yun, U. Rajendra Acharya, Y.V. Venkatesh, Caroline Chee, Lim Choo Min, E.Y.K. Ng. 2008," Identification of different stages of diabetic retino- pathy using retinal optical images", Information Sciences, Vol. 178, pp.106–121.

[3] R. Acharya U, L.Y. Wong, E.Y.K. Ng, J.S. Suri, 2007, "Automatic identification of anterior segment eye abnormality", Journal of ITBM-RBM, Vol. 28, pp. 35–41.

[4] Il-Seok Oh, Member IEEE, Jin-Seon Lee, and Byung-Ro Moon, Member IEEE, Nov 2004, "Hybrid genetic algorithms for feature selection", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 26, No.11, pp.1424-1437

[5] Enrique J. Carmona , Mariano Rincon , Julian Garcia-Feijoo Jose M. Martinez-De-La-Casa, 2008, "Identification of the optic nerve head with genetic algorithms", Artificial Intelligence In Medicine,Vol.43, pp. 243—259.

[6] Hiroshi Someya, Member, IEEE, April 2011," Theoretical analysis of phenotypic diversity in real-valued evolutionary algorithms with more-than-one element replacement", IEEE Transactions On Evolutionary Computation, Vol.15, No.2, pp.248-266.

[7] Michael L. Raymer, William F. Punch, Erik D. Goodman, Leslie A. Kuhn, And Anil K. Jain, July 2000, "Dimensionality reduction using genetic algorithms", IEEE Transactions On Evolutionary Computation", Vol. 4, No. 2, pp.164-172.

[8] P.M.Narendra, and K. Fukunaga, 1977 "A branch and bound algorithm for feature selection," IEEE Transactions on Computers, 26(9**)**, 917-922 .

[9] I. Foroutan, and J. Sklansky, 1987. "Feature selection for automatic classification of non-Gaussian data," IEEE Transactions on System, Man and Cybernetics, 17,187-198

[10] D.E.Goldberg.,1989."Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley Reading ,MA

[11] J.H. Holland.,1975.., Adaptation in Natural and Artificial System, University of Michigan Press, Ann Arbor, MI

[12] W.Siedlecki and J. Sklansky, 1989. "A note on Genetic algorithms for large-scale feature selection," Pattern Recognition Letters, 10, 335-347

[13] M. Kudo and J. Sklansky, 2000. "Comparison of Algorithms that select features for pattern classifiers," Pattern Recognition, 33, 25-41.

[14] M.R.Peterson, T.E.Doom, and M.L.Raymer, 2004. "GA facilitated knowledge discovery and pattern recognition optimization applied to the biochemistry of protein salvation. In GECCO 2004 proceedings, LNCS,3102, 426-437