Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 85 (2016) 455 – 462

Procedia Computer Science

International Conference on Computational Modeling and Security (CMS 2016)

# Hybrid Genetic Algorithm for Medical Image Feature Extraction and selection

G.Nagarajan[a*], R.I.Minu[b], B Muthukumar [c], V.Vedanarayanan [d] & S.D.Sundarsingh[e]

[a*,c,d,e]*Department of EEE, Professor ,Sathyabama University,Chennai,India*

[b]*Department of CSE, Associate Professor,Jerusalem College of Engineering, Chennai,India*

**Abstract**

For a hybrid medical image retrieval system, a genetic algorithm (GA) approach is presented for the selection of dimensionality reduced set of features. This system was developed in three phases. In first phase, three distinct algorithm are used to extract the vital features from the images. The algorithm devised for the extraction of the features are Texton based contour gradient extraction algorithm, Intrinsic pattern extraction algorithm and modified shift invariant feature transformation algorithm. In the second phase to identify the potential feature vector GA based feature selection is done, using a hybrid approach of "Branch and Bound Algorithm" and "Artificial Bee Colony Algorithm" using the breast cancer, Brain tumour and thyroid images. The Chi Square distance measurement is used to assess the similarity between query images and database images. A fitness function with respect Minimum description length principle were used as initial requirement for genetic algorithm. In the third phase to improve the performance of the hybrid content based medical image retrieval system diverse density based relevance feedback method is used. The term hybrid is used as this system can be used to retrieve any kind of medical image such as breast cancer, brain tumour, lung cancer, thyroid cancer and so on. This machine learning based feature selection method is used to reduce the existing system dimensionality problem. The experimental result shows that the GA driven image retrieval system selects optimal subset of feature to identify the right set of images.


Peer-review under responsibility of the Organizing Committee of CMS 2016

*Keywords:* Feature Extraction, Feature Selection, Texton, Genetic Algorithm


Peer-review under responsibility of the Organizing Committee of CMS 2016
doi:10.1016/j.procs.2016.05.192

## 1 Introduction

For browsing, searching and retrieving relevant images form the large datasets image retrieval system algorithm were used. Due to the advancement in the medical field a huge volume of medical images are generated tremendously in medical centers using many advanced medical equipment's. Thus to aid the expertise in medical field analysis content based medical image retrieval is one of the most revolving research area. So that, for analysis the similarity in diagnosis methodology can be synchronized by comparing the patient's current medical image with the medical database to retrieve the medical diagnosis history.

The aim of feature selection algorithm is to find the relevant of features that produces the best recognition rate and least computational effort. As high dimensionality features can increase system complexity, which also leads to higher recognition rate. There is a need to use some complex feature extraction algorithm which should not be depend on other features. So, the selection of subset of best features is important. For the selection of effective feature on the run we need to device a learning model in the training phase. In this paper instead of use high complicated model, we use minimum description length principle based genetic algorithm (GA). Genetic algorithm is a search heuristic, which routinely used to generate solution for searching problem. In a GA [1-3] for a better solutions, the population of candidate solution (feature vector) are optimized iteratively. The iteration process, which is also called as generation will usually starts from a population of randomly generated candidate solution. In each iteration, the fitness of each candidate solution are evaluated using an objective function, for an optimized solution. Thus new set of population of candidate solution are generated using the objective solution for the next iteration. This algorithm is terminated either a satisfactory fitness level has been achieved for the population or a maximum number of generations has been produced.

## 2 Technical approach

To improve the performance of the proposed hybrid content based medical image retrieval system, a hybrid approach of genetic approach algorithm is used for feature selection. This feature selection method reduces the data dimensionality issues while selecting the optimal features and thus improves system performance. Distance measurement using Chi Square distance is used for assessing the similarity between database images and the query images. The proposed system employs the diversity density based relevance feedback approach for improving the performance of the system. Relevance feedback refines the query image feature selection method for retrieving most similar images to query image. The overall system design of the proposed concept in shown in Fig.1. The whole system are explained in sub section such as feature extraction, feature selection and diverse density based relevance feedback

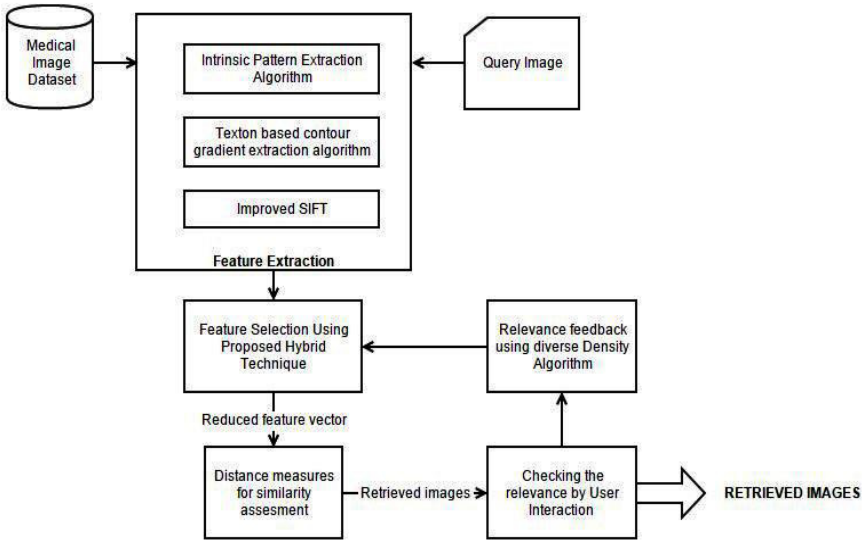



**Fig. 1.** Overall Framework

## 3 Feature extraction

Feature extraction methodologies analyses the medical images to extract the most dominating features that are prime representative of the image. Texture, is one of the key feature used in this prosed idea . The algorithm devised for the extraction of the features are Texton based contour gradient extraction algorithm, Intrinsic pattern extraction algorithm and modified shift invariant feature transformation algorithm.

### **3.1** Intrinsic Pattern Extraction Algorithm

The intensity variation pattern in an image is said to be texture. The texture of an image cannot be analysed from a single pixel point; it requires the neighbouring point's intensity values. To represent the texture pattern of the images an algorithm called Intrinsic Pattern Extraction algorithm, was derived from the basic of Principal Component Analysis (PCA) . The core objective of this approach is to reduce the computation and the size of the identified feature vector, which would represent the intrinsic pattern of the given image. As per Smith [9] tutorial, PCA is a statistical model which is used to identify some of the discrete patterns for a given dataset. From the identified pattern design, PCA implements a linear system, which is derived from applied linear algebra. This linear system is used to identify the potential feature vectors from the pattern design and formalize a tactic to analyse the continuity in data sets.. The abstract algorithm of Intrinsic Pattern is shown in Algorithm 1.

**Algorithm 1: Intrinsic Pattern Extraction Algorithm**
**Input: Medical images from the datasets**
**Output: Positive Intrinsic Pattern feature vector for each input image**

1. Begin
2. Input images are converted to 8 level grey scale image of size (100 x 100)
3. For each group of images , select 5 individual primary image AI = (AI1∪ AI2∪AI3∪AI4∪ AI5)
4. For each AI determine the grey level co-occurrences matrix of size (8x8) for eight grey levels. The resultant matrix would be:
5. X(AI) = X(AI1∪ AI2∪AI3∪AI4∪ AI5)
6. Let X(AI3) be the mean image, then compute X(AI)m the resultant matrix would be of size (64 x 5) X(AI)m = X(AI3) – X(AI1∪ AI2∪AI3∪AI4∪ AI5)
7. Normalize the X(AI)m to square matrix using Square Gram matrix concept, now K Aim would be of size (64 x 64)
8. K AIm = X(AIm).X(AIm)T
9. Compute 64 Eigen values from K AIm matrix
10. Select first five positive Eigen values as Intrinsic Pattern feature vector value for that group of image.
11. End

### 3.2 Texton based contour gradient extraction algorithm:

Intrinsic Pattern extraction algorithm is used to identify some of the component patterns in the pixel intensity value for each images in the medical image datasets. The feature detecting system would be still more effective if the gradient of the edge pixel were also analysed. The fundamentals of gradient are derived from the concept of derivatives. The variation of the functional variable can be mathematically derived through derivatives, whereas, partial derivatives are used to identify the variation of the particular variable in the given function. Gradient is one such concept used to identify the variation in image pixel intensity value at (i,j) location in a two-dimensional space. The gradients are vector values, whose magnitude determines the change of pixel intensity values and the direction of the gradient specifies the direction where the changes take place. To extract syntactic feature from any size of user defined image sets, some kind of spatial filter transformation has to be employed on that image to acquire exact contour gradient of the given image. From the [10], and Zhu et al [11] Texton is one of the evolving concepts derived from texture analysis. Those complex patterns can be analysed by Texton, which is used to develop extensible texture models for an image. In this work, these Texton vector features are analysed and normalized, which can be used for indexing the medical images as per their domain. In short the algorithm is shown below.

**Algorithm 2: Texton based contour gradient extraction algorithm(TCGR)**
**Input: Medical images from the datasets**
**Output: TCGR feature vector for each input image**

1. Begin
2. Input images are converted to 8 level grey scale image of size say (700 x 700)
3. The input image of size (700 x 700) is partitioned into 12 patches of (175 x233)
4. PI = (PI1∪ PI2∪PI3∪PI4∪ PI5∪ PI6∪PI7∪PI8∪ PI9∪PI10∪PI11∪ PI12)
5. The unwanted blank patches are deleted whose homogeneity == 1 and sum of square variance < = 5 PI = (PI1∪ PI4∪PI8∪PI11)
6. The informative image patches are filtered using 13 different rotation invariant SFilter to create a Texton base of the image patch
7. PTI1 = (PTI1∪ PTI2∪PTI3∪PTI4∪ PTI5∪ PTI6∪PTI7∪PTI8∪ PTI9∪PTI10∪PTI11∪ PTI12∪ PTI13)
8. The edges of all the 13 Texton images are identified
9. The edge based histogram is calculated for all the 13 edge detected images.
10. The histogram values are clustered and the clustered centroid are identified
11. The centroid values with more number of data are considered as the TCGR value of the image
12. End

### 3.3 Improved SIFT

Lowe presented SIFT [12], which was successfully used in recognition, stitching and many other applications because of its robustness. SIFT consists of four major stages: scale-space extrema detection, Keypoint localization, orientation assignment and Keypoint descriptor. In this work the interesting points from the image are extracted using Harris corner extractor instead of DoG(Difference of Gaussian). DoG detectors were designed for efficiency and the other properties are slightly compromised. Quantity and good coverage of the image are crucial in recognition applications, where localization accuracy is less important. Thus, Hessian-Laplace detectors have been successful in various categorization tasks although there are detectors with higher repeatability rate. Random and dense sampling also provide good results in this context which confirms the coverage requirements of recognition methods although they result in far less compact representations than the interest points. DoG detector performs extremely well in matching and image retrieval probably due to a good balance between spatial localization and scale estimation accuracy

## 4 Feature Selection

To minimize the computation time and maximize the accuracy better feature selection algorithm is required. Thus the algorithm has to be robust in eliminating the redundant, irrelevant and noisy features. In this paper we have used a hybrid branch and bound technique and artificial bee colony algorithm for the optimal feature selection. The algorithm constructs a binary search tree where the root represents the set of all features and leaves represent subsets features. While traversing the tree down to leaves, the algorithm successively removes single features from the current set of ‘‘candidates’’ The algorithm keeps the information about both the currently best subset and the criterion value it yields. This value is denoted as bound value. The branch and bound algorithm is shown in algorithm 3

| **Algorithm 3: Branch and Bound Feature Reduction Algorithm**<br>**Input: Medical images Feature vector**<br>**Output: Reduced Feature vector** |
|---|
| 1. Fitness function initialization and evaluation |
| 2. Create the consecutive tree level |
| 3. Test the right-most descendant node with the updated evaluation function. |
| 4. When all the descendant are test go to step 6 |
| 5. Descendant node connected by the -edge (and its possible sub-tree) may be cut-off |
| 6. Backtracking |
| 7. Bound value is updated, store the current best subset and go to step 3 |

ABC algorithm starts with generation of a population of binary strings (or bees). Initialize population and evaluate fitness. Select other features from neighbourhood features in the initial population and compare with evaluate fitness .If the selected features not satisfy the fitness function remove that features from the population. Thus all the bees are compared with fitness function and form the optimal feature set. If none of the features not satisfy the fitness function, find the new fitness function. Then continue the searching for selecting the optimal value. The ABC algorithm is shown in algorithm 3. The proposed hybrid approach algorithm union the features of both branch and bound algorithm and artificial bee colony algorithm shown in Algorithm 5.

| **Algorithm 4: ABC Feature Reduction Algorithm**<br>**Input: Medical images Feature vector**<br>**Output: Reduced Feature vector** |
|---|
| 1.Initialize all the features |
| 2. Initialization of ABC parameters (no of cycles, fitness value, probability) |
| 3. Evaluate the fitness value of each individual features. |

4. Employee bee phase
   (a) Construct feature subsets randomly
   (b) Calculate the fitness of the feature subset.
   (c) Calculate the probability of feature subset.
5. Onlooker phase
   (a) Select the feature based on the probability.
   (b) Apply the greedy selection to find the best feature subset.
6. Scout bee phase
   (a) Determine the scout bee and the abandoned solution
   (b) Calculate the best feature subset of the cycle
   (c) Memorize the best optimal feature subset
   (d) Cycle = Cycle + 1
   (e) Until pre-determined number of cycles is reached

**Algorithm 5: Proposed Hybrid Feature Reduction Algorithm**
**Input: Collect N number of features A1, A2, A3, … , AN from feature extractor**
**Output: Use the n features where n<N for distance measurement.**

1. Apply branch and bound algorithm to select the optimal set containing n1 number of features where n1<N.
2. Apply artificial bee colony algorithm to select the best subset containing n2 number of features n2 where n2<N.
3. Find the union of n1 features and n2 features as n features.

## 5 Diverse Density based Relevance feedback

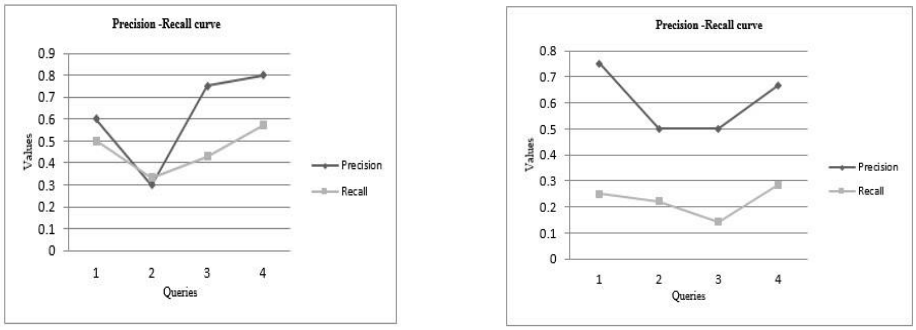
In medical image retrieval based on relevance feedback method, the system first receives a query image that user submits .The system uses Euclidean distance to calculate the similarity between images, to return an initial query results. The query results are marked positive and negative, by using user feedback information. Then based on the user's interest as the feedback refine the query image feature selection again and again. In this paper, the Diverse Density algorithm is used to achieve the relevance feedback.
In DD image content is considered as a set of features and the task of DD is to find the features with the greatest diversity density within the features space. The diversity density is a measure refers to the more positive examples around at that point, and the less negative examples, With DD approach, an objective function called DD function is defined to measure the co-occurrence of similar features from different images. The target of DD is to find features which are closest to all the positive images and farthest from all the negative images. The algorithm is shown in Algorithm 6

**Algorithm 6: Proposed Hybrid Feature Reduction Algorithm**
**Input: positive Images, negative Images.**
**Output: Features which user is interested.**

1. Select the features in the positive images.
2. Calculate the probability that feature set in positive and negative.
3. If the probability is largest then quit, otherwise the Go to Step 4;
4. Maximize the probability function to find other feature set; Repeat step 2;

## 6 Experimental Result

The selection of optimal features from all the extracted features of database images. The hybrid approach of "branch and bound algorithm "and "artificial bee colony algorithm is used for feature selection. We compare the performance of proposed CBMIR system before relevance feedback algorithm and after the relevance feedback algorithm. Diverse Density (DD) algorithm is used for relevance feedback. We conducted two groups experiments, and each group experiment included five different queries and the five different queries from every group may overlap or not. The precision and recall radio of every group which are used to evaluate our system are expressed by the average of each five queries, as is shown in Fig2 (a) and (b).




**(a)** **(b)**

Fig. 2 (a) &(b) Overall Evaluation

## 7 Conclusion

An improved approach for the feature selection method in content based medical image retrieval using hybrid approach of "branch and bound algorithm" and "Artificial bee colony algorithm" was implemented and the relevance feedback also implemented. The implemented system is tested with various test images from medical image data set. For the final retrieval result, although most relevant images were retrieved, there were some obvious relevant images missed by the diverse density relevance feedback algorithm. The relevancy can be improved by using more advanced algorithms

## References

1. B. Bhanu, D. Dudgeon, E. Zelnio, A. Rosenfeld, D. Casaseut, I. Reed (Eds). Special Issue on Automatic Target Recognition, IEEE Transactions on Image Processing 6 (1) (1997).
2. S. Cagnoni, A. Dobrzeniecki, R. Poli, J. Yanch, Genetic algorithmbased interactive segmentation of 3D medical images, Image and Vision Computing 17 (12) (1999) 881–895.
3. B. Bhanu, T. Poggio, (Eds) Special section on Machine Learning in Computer Vision, IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (9) (1994).
4. Sérgio Francisco D, et al., "Improving the ranking quality of medical image retrieval using a genetic feature selection method", Decision support systems, Vol 51,NO 810,2012.
5. Aswini Kumar Mohanty, et al., "A novel image mining technique for classification of mammograms using hybrid feature selection", Neural computer and application ,Vol 22,No 1151,2013.
6. G.Nagarajan, R.I.Minu 2015 'Fuzzy Ontology based Multi-Modal semantic information retrieval ' , Elsevier Science Procedia Computer Science ,Vol 48, Pages 101–106,2015

7. Shunmugapriya S and Palanisamy A., “Artificial Bee Colony Approach For Optimizing Feature Selection” International Journal of Computer Science Issues, Vol. 9, Issue 3, No 3, May 2012.
8. Liang Lei, Jun Peng and Bo Yang, “Image Feature Selection Based on Genetic Algorithm”, International Conference on Information Engineering and Applications (IEA) Vol 8,No 25, 2012.
9. Smith, LI 2002, A tutorial on principal components analysis, Cornell University.
10. Julesz, B 1981, ‘Textons, the elements of texture perception, and their interactions’, Nature, vol.290, no.5802, pp.91–7.
11. Minu, RI & Thyagharajan, KK 2014 ‘Semantic Rule Based Image Visual Feature Ontology Creation’, International Journal of Automation and Computing, Springer Publication, vol.11, no.5, pp.489 - 499.
12. Lowe, DG 2004, ‘Distinctive image features from scale-invariant keypoints’, International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110.

Dr.G.Nagarajan has received his Diploma in Electronic & Communication Engineering from Directorate Of Technical Education 1997. He has received his BE degree in Electrical & Electronic Engineering from Manonmaniam Sundaranar University 2000. He received his ME degree in Applied Electronic Engineering from Anna University 2005. He also received his ME degree in Computer Science Engineering from Sathyabama University 2007 He obtained his Ph.D. degree in Computer Science Engineering from Sathyabama University 2015. He has 15 year of teaching experience. Now, he is working as Professor in department of EEE in Sathyabama University. He has published more than 40 paper in the research areas of Wireless Sensor Network, Intelligent Irrigation Systems, Artificial Intelligent, Ontology Learning, Machine Leaning,IoT and Computer Vision.

Dr..R.I.Minu obtained her B.E., degree in Electronics and Communication Engineering from Bharathidasan University 2004 and received her M.E., degree in Computer Science Engineering from Anna University 2007 . He obtained his Ph.D. degree in Information and Communication Engineering (Computer Science) from College of Engineering Guindy, Anna University 2015. She has 8 year of teaching experience. Now she is working as Associate Professor in the Department of Computer Science and Engineering in Jerusalem College of Engineering. She has published more than 30 papers in National/International Journals and Conferences. She has involved in many UG and PG projects in the area of Image Processing, Internet of Things ,Wireless Sensor Network, Bioinformatic-BigData, E-Learning and Artificial Intelligent.