International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:10, 2014

Survey on Image Mining Using Genetic Algorithm

Jyoti Dua

***Abstract*—**One image is worth more than thousand words. Images if analyzed can reveal useful information. Low level image processing deals with the extraction of specific feature from a single image. Now the question arises: What technique should be used to extract patterns of very large and detailed image database? The answer of the question is: "Image Mining". Image Mining deals with the extraction of image data relationship, implicit knowledge, and another pattern from the collection of images or image database. It is nothing but the extension of Data Mining. In the following paper, not only we are going to scrutinize the current techniques of image mining but also present a new technique for mining images using Genetic Algorithm.

***Keywords*—** Image Mining, Data Mining, Genetic Algorithm.

I. Introduction

ADVANCEMENT in digitization has enhanced the growth of image acquirement and storage technology, add the incredible growth of large and detailed image databases in various fields. The World Wide Web is the largest global image repository. Image mining deals with drawing out intrinsic knowledge, data, image relationship or other patterns from the images. It is used in various fields like medical diagnosis, space research, remote sensing, and industries and even in the educational field. The fundamental challenge in image mining is to determine how this low-level pixel representation encapsulates with raw image or image sequence can be processed proficiently to recognize high-level image objects and its relationship. Image mining has two main themes. The first is mining large collections of images and the second is the combined data mining of large collections of image and associated alphanumeric data. Image mining technique is highly specific because the image databases are predominantly non-relational. An image mining system is often complicated because it employs various approaches and techniques ranging from image retrieval and indexing schemes to data mining and pattern recognition. A good image mining system is expected to provide users with an effective access into the image repository and the generation of knowledge and also patterns beneath the images. Such a system typically encompasses the following functions: image storage, image processing, feature extraction, image indexing and retrieval, and finally patterns and knowledge discovery [1].

Fig. 1 shows the working principle of Image Mining process:

- Quality of images from the collection of images or database needs to be improved. This is done during image preprocessing.
- Important features are extracted.
- After transformation and extraction, mining comes. It can be performed by using Data Mining techniques.
- The outcome is a pattern that is to be evaluated and is interpreted in order to obtain the desired knowledge.

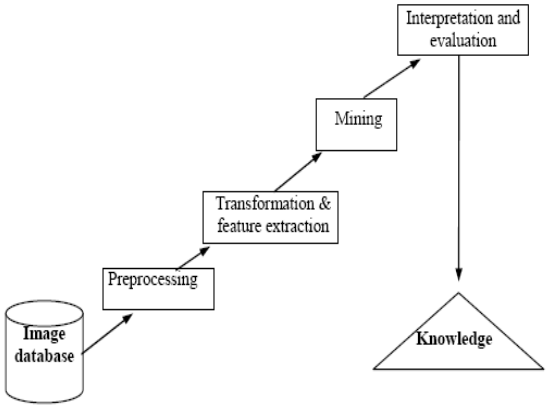

Fig. 1 Image mining process [1]

II. Problem Statement

To obtain the desired knowledge that is present in a given collection of images and alphanumeric data.

III. Methodology

Image Mining is currently in its infancy as a field of research. Images can be a great source of information if made good use. This can be done through the process of image mining. Knowledge discovery from image poses many interesting challenges such as preprocessing the image data set, learning the data and discovering useful image patterns applied to many new applications.

Image Mining= Image Processing + Data Mining

IV. Image Mining Techniques

According to [7], Image Mining can be broadly divided into two categories:

- Domain-specific, application- process of extracting relevant image features into a suitable form for data mining.
- General application-process for generating images pattern for understanding the interaction between high level human perception and low-level image feature which improve the accuracy of the images retrieved from an image database.

The Image Mining techniques are:

Jyoti Dua is with the United Institute of Technology, Allahabad, India (e-mail: jyotidua_ecimt@yahoo.com).

A. Object Recognition

Object Recognition provides a way to identify specific objects within the image depending on the filtered parameter of confidence, size, rotation, etc. Meaningful information can be realized when some objects have been recognized by the machine. The object recognition is based on labeling of objects. In the image, the system should assign correct set of labels to regions as an image contains one or more objects of interest and also a set of labels corresponding to a set of models known to the system. Object Recognition is closely coupled with segmentation. Without segmentation, object recognition is not possible and vice-versa. Some issues that are considered to design an object recognition system they are:

Object Recognition: How should objects are represented in a database? What are the features or attributes must be captured for an object? The object representation should capture relevant information and organize this information without any redundancy that allows easy access.

Feature Extraction: Which features should be extracted, and how they are extracted reliably?

Feature-Model matching: How can features in image can be matched with the model in the database? Effectiveness of feature and efficiency of matching technique depends on matching approach.

Hypothesis formation: It is a heuristic approach to reduce the size of the search space. This uses knowledge of the application domain to assign some probability or confidence measure to different objects in a domain.

Object Verification: The presence of each likely similar object can be verified by using their model. One must examine each hypothesis to verify the presence of an object or ignore it.

The object recognition problem can be classified into two classes:

Two-dimensional: In many applications, images are acquired from a distance sufficient to consider the projection is orthographic. If the object is always in one stable position, they can be considered two-dimensional.

Three-dimensional: If the images of an object can be taken from arbitrary viewpoints, then an object appears very different in its two views. If three-dimensional models are used for object recognition then the perspective effect and the viewpoint of the images have to be considered as well.

The images are segmented to separate object from background.

B. Image Retrieval

Image retrieval means searching or browsing, and retrieving, images from large image database based on any keyword or description of the image, and this has to be done efficiently. Content-based-image retrieval is the application to the image retrieval problem. The term 'content' in this context refers to colors, shapes, textures, or any other information that can be derived from the image. 'Content-based' means that the search will analyze the actual contents of the image rather than the metadata such as keywords, tags, or descriptions associated with the image. Images have many types of attributes which could be used for retrieval, including a particular combination of color, texture or shape features, and arrangement of specific types of objects, named individuals, locations, or events. Each query type represents a higher level of abstraction and each has answered with reference to some external knowledge. This requires a classification of query types into three levels of increasing complexity:

Level 1 comprises retrieval of *primitive* features such as color, texture, shape or the spatial location of image elements. This level of retrieval uses features which are both objective, and directly derivable from the images themselves, without the need to refer to any external knowledge base.

Level 2 comprises retrieving by *deriving* (sometimes known as *logical*) features, involving some degree of logical inference about the identity of the objects depicted in the image. It can usefully be divided further into: retrieval of objects of a given type and retrieval of individual objects or persons.

Level 3 comprises retrieval by *abstract* attributes, involving a significant amount of high-level reasoning about the meaning and purpose of the objects or scenes depicted. Again, this level of retrieval can usefully be subdivided into: retrieval of names, events or types of activity and retrieval of pictures with emotional or religious significance.

There are three fundamental bases in content-based image retrieval namely visual information extraction, image indexing and image retrieval system.

C. Image Indexing

The objective of image indexing is to retrieve similar images from detailed image database for a given query image by comparing their features. The standards of similarity for the images are based on features, namely color, texture, size, location and other description. Image indexing techniques are of two types:

Textual- User approach keywords are given to a particular type of image. These are captioned indexing, keyword addition, standard subject heading, classification etc.

Content based- It is automated indexing. Images are indexed based on their content such as color, shape, spatial relation and texture etc.

D. Image Classification and Clustering

Image classification and clustering are the supervised and unsupervised classification of images into groups. These are two important techniques employed in image mining. Creating classifiers involve a general form of model and training patterns to learn or estimate the unknown parameters of the model. Classification involves supervised learning, in which a teacher provides a category label for each pattern in the training set and a newly encountered image or pattern is to be labeled.

A Bayesian Classifier is an important technique for classification. Bayesian classifiers find the distribution of attribute values for each class in the training data, when given in a new instance 'd', they use the distribution information to estimate, for each class 'c_j', the probability that instanced 'd' belongs to class 'c_j', denoted by $p(c_j|d)$. The class with

maximum probability becomes the predicted class, for instance 'd'.

On the other hand, clustering, which is unsupervised learning, there is no explicit teacher. The aim is to group a given collection of unlabeled images into clusters, without any training. Clustering of images is usually done in the early stages for the process of mining. Color, texture, and shape are a few attributes which, are individually or in combination used as feature attributes for clustering. There are many clustering techniques, agglomerative hierarchical clustering algorithm, online clustering, nearest neighbor clustering, evolutionary clustering algorithms, etc.

The advantage that image classification and clustering provides lies in the better storage and management of images, and also if an optimal indexing scheme is used, the retrieval of images is faster.

K-Means Clustering:

K-means clustering is the most common partitioning algorithm. K-Means re-assigns each record in the dataset to *only one* of the new clusters formed. A record or data point is assigned to the nearest cluster (the cluster which it is most similar to) using a measure of distance or similarity

1. Separate the objects (data points) into K clusters.
2. Cluster center (centroid) = the average of all the data points in the cluster.
3. Assigns each data point to the cluster whose centroid is closer (using distance function.)

K-means Algorithm

1. Place K points in the space of the objects being clustered. They represent the initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. Recalculate the positions of the K centroids.
4. Repeat Steps 2 & 3 until the group centroids no longer move.

The advantage of performing image classification and clustering lies in better image storage, management and fast retrieval.

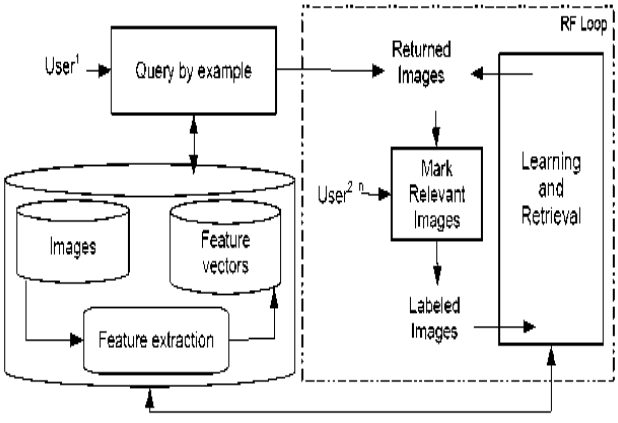

Fig. 2 Classifier for image retrieval [11]

Fig. 3 A set of images classified as Sunset using K-mean clustering approach [9]

E. Association Rule Mining

Association Rules (ARs) describe relationships between items in data sets. Association rules define relationships of the form: A→B.

Association Rules: Interestingness Measures

The most commonly used "interestingness" measures are: Support, Confidence, and Lift

Support:

A measure of the frequency with which an item set occurs in a database:

$$\text{supp}(A) = (\#\text{ records that contain } A) / m$$

If an item set has support higher than some specified threshold, then we say that, the item set is *supported* or *frequent* (some authors use the term *large*).

Support threshold is normally set reasonably low (say) 1%.

Confidence:

A measure, expressed as a ratio, of the support for an AR compared to the support of its antecedent:

$$\text{conf}(A \rightarrow B) = \text{supp}(A \cup B) / \text{supp}(A)$$

We say that we are confident in a rule if its confidence exceeds some threshold (normally set reasonably high, say, and 80%).

Lift:

A measure of the "surprisingness" (*implication intensity*) of a rule, expressed as follows:

$$\text{lift}(A \rightarrow B) = \text{supp}(A \cup B) / \text{supp}(A).\text{supp}(B)$$

Or alternatively: $\text{lift}(A \rightarrow B) = \text{conf}(A \rightarrow B) / \text{supp}(B)$

A rule is "interesting" if lift is greater than 1.

V. Image Mining Algorithm

The four major *image mining* steps are as follows: [2]

1. Feature extraction: Segment images into regions identified by region descriptors (blobs). Ideally, one blob represents one object. This step is also called segmentation.
2. Object identification and record creation: Compare objects of one image with other objects in and label them

with an identity. We call this step the preprocessing algorithm.
3. Create auxiliary images: Generate images from identified objects to interpret the association rules obtained from the following step.
4. Apply data mining algorithm*:* Produce object association rules.

VI. Related Work

A. Algorithm

Genetic Algorithm is an adaptive heuristic search algorithm and optimization technique based on evolutionary algorithms of natural selection and genetics. It is based on the principle of Darwin's theory of evolution-"Survival of the fittest" which states "select best discard rest". The genetic algorithm is based on an analogy with the genetic structure and behavior of chromosomes within the population of individuals. Genes from "good" individuals propagate throughout the process to produce offspring better than either parent.

The basic steps involved are-: [4]

1. Generation of a population of solution
2. Finding the objective function (which is to be maximized or minimized) and fitness function
3. Application of genetic operator:
4. *Selection*: This operator selects chromosomes in the population for reproduction. The fitter the chromosome, better the chances to be selected to reproduce again and again.
5. *Crossover*: This operator randomly chooses a locus and exchanges the subsequences previous and next to that locus between two chromosomes to create two offspring. For example, the strings 10000100 and 11111111 could be crossed over after the third locus in each to produce the two offspring 10011111 and 11100100. The crossover operator roughly mimics biological recombination between two single chromosomes (haploid) organisms.
6. *Mutation*: This operator randomly flips some of the bits in a chromosome. For example, the string 00000100 might be mutated in its second position to yield 01000100. A mutation can occur at each bit position in a string with some probability, usually very small (e.g., 0.001)

B. The basic Genetic Algorithm [10]

1. Start with a randomly generated population of n l−bit chromosomes (candidate solutions to a problem).
2. Calculate the "fitness" $f(x)$ of each chromosome x in the population
3. Repeat the following steps until 'n' offspring have been created:

a) *Select* a pair of the parent chromosomes from the current population, the probability of selection being an increasing function of fitness. Selection is done "through replacement," meaning means the same chromosome can be selected more than once to become a parent

b) With probability 'p_c' (the "crossover probability" or "crossover rate"), "*crossover"* the pair at a randomly chosen point (chosen with uniform probability) to form two offspring. If no crossover takes place, form two offspring that are exact copies of their respective parents.

c) *Mutate* the two offspring at each locus with probability pm (the mutation probability or mutation rate), and place the resulting chromosomes in the new population. (If n is odd, one new population member can be discarded at random)

4. Replace the current population with the new population.
5. Go to Step 2

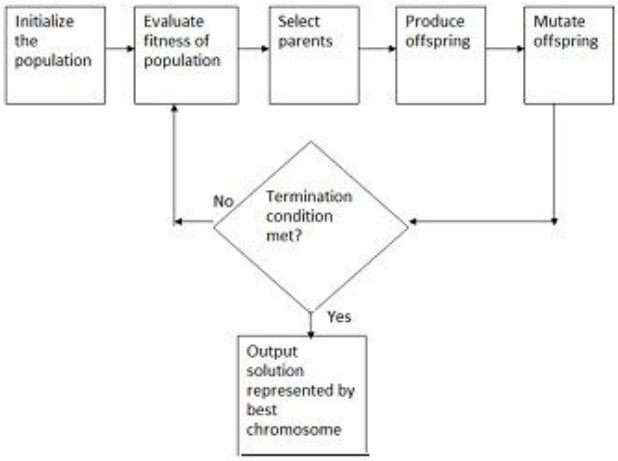

Fig. 4 Genetic Algorithm flowchart [5]

C. Genetic Algorithm on Image Data

According to [6], using different chromosome encoding schemes and fitness functions, Genetic Algorithm has been successfully implemented for various clustering problems. It performs clustering of an input set of data objects such that supervised learning is applied to predict class labels in the second step. It takes as input a set of data objects that have both kinds of attributes that is numeric and label, and a desired number of clusters. The goal of the Genetic Algorithm is to produce clusters of data objects that minimize dispersion of clusters. The Genetic Algorithm uses a fitness function with two components. The first component measures cluster variance using a distance metric and the second component measures the similarity of the labeled attributes of the data objects. A large input data set is preprocessed to make a set suitable for use by the algorithm and which also provides better space and time efficiency. According to [8], in GA, we can implement two alternate preprocessing methods for clustering algorithm. The first preprocessing method uses a random sampling to obtain a data set with fewer points. This reduced data set is then used in evaluating the fitness of the chromosomes. The second preprocessing method uses a summarized form of the input data set. In this method, a grid is formed and the input dataset is applied on it. A single point location and corresponding weight are calculated for each region defined by the grid. The location of the representative point is chosen as the mean value of all the points in the region

and the weight of the representative point is equal to the number of points that it replaces.

D. Pseudo code for Genetic Algorithm

According to [3], we take 'n' number of chromosomes for 'n' number of centroids, 'm' number of Samples 'S_j'and 'Sj^i' is 'j^{th}' sample assigned to 'i^{th}' chromosome. We take maximum iteration to find best "fitness" solution (F). In search space each iteration it simultaneously found objects on the label as L_a and input attribute as I_a. First chromosome is considered as a parent chromosome and in each iteration it builds child chromosomes. Below are GA rule to find cluster construction. Genetic Algorithm is more efficient because final solution is evaluated by frequency of fitness value and samples matching value.

Step 1. Initialize n no. of chromosomes to 'n' no. of centroids.
Step 2. Iteration (I) <=MaxIter
Step 3. Parent chromosomes P_i is selected randomly.
Step 4. Method1<=n
Step 5. Randomly select one sample in multidimensional space.
Step 6. Sample j is assigned to chromosome 'i' by

$$S_j^i=1/n_i \{\log_a (L_a)*\log_a (I_a)\}$$

Step 7. S_j^i value is either 0 or 1i
If S_j^i =1, the sample is matched to 'i^{th}' chromosomes
If S_j^i =0, sample j belongs to some other chromosomes
Step 8. If chromosomes matched, then "fitness" value of sample j to input attribute is calculated by

$$F_j (I_a) =1/ S_j^i (I_a / n_i)$$

Step 9. Fitness value of sample j for the label attribute is

$$F_i (La) =1/ S_j^i (L_a / n_i)$$

Step 10. Steps 4 to 9 are repeated until Method1 reaches n samples in the search space
Step 11. Fitness value for chromosome i is calculated by

$$F_j^i =1/ n_i \{Fj (I_a)*Fj (L_a)\}$$

Step 12. Repeat step 2 to step 10 until reach MaxIter
Step 13. Finally, find maximum fitness value from the solution of each iteration by using frequency as
Step 14.

$$F=\max \{1/n (f|F_i i- S_j^i |^2), 0\}$$

VII. Performance Evaluation

Clustering is an important task having applications in many fields. Heuristic algorithms are used for this task to provide acceptable results, both in terms of solution quality and running time, because all of the nontrivial clustering problem variations are NP-Hard. In K-means algorithm at different runs it produces poor results when the initial centroids are chosen randomly. It is important to realize that the choice of the initial centroid has a huge effect on the final result. K-means algorithm for multiple runs on large data sets does not work and it takes a lot of time to complete. For clustering a very large data sets, such as image data sets, the size of the associated databases makes it necessary to modify the traditional GA because of their slow running times and combinations of input and label attributes. Table I revealed the experimental results on a data set [3].

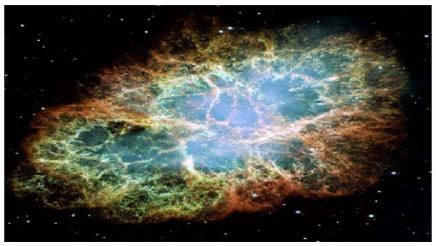

Fig. 5 Image from Google [3]

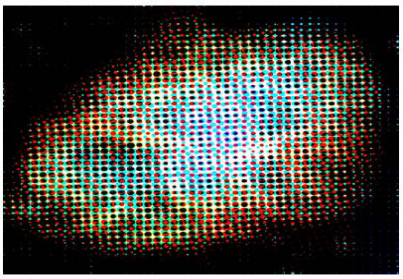

Fig. 6 Using Genetic Algorithm to group red spaces [3]

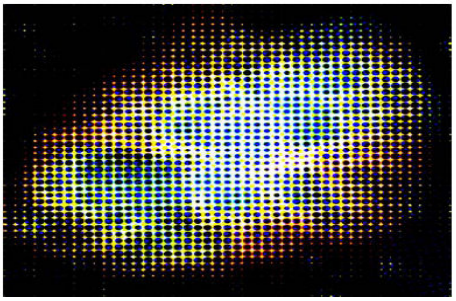

Fig. 7 Using Genetic Algorithm to group blue spaces [3]

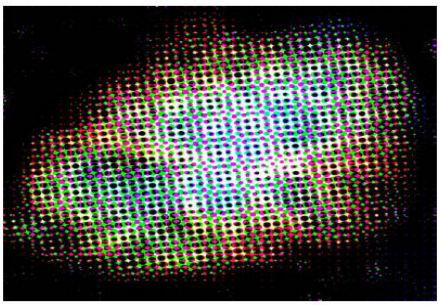

Fig. 8 Using Genetic Algorithm to group green spaces [3]

TABLE I
RESULTS ON ARTIFICIAL DATA SET [3]

Input Set	Algorithm	Running Time	Distance Measurement
3 Centroids 1000 points	GA	230	5.23×10^4
	K-Mean	219	7.89×10^4
5 Centroid 10000 points	GA	457	4.56×10^5
	K-Mean	512	2.34×10^6
7 Centroids 20000 points	GA	890	8.27×10^8
	K-Mean	2023	1.10×10^{10}

VIII. CONCLUSION

For a large number of records and clusters, genetic algorithm shows better result as compared to k-means algorithm. The preprocessing method prevents the creation of representative points for regions that contained less than a certain threshold of points. This refinement removes the negative effect that outlier points have on the clustering quality. It makes the genetic algorithm run faster than k-means algorithm because there would be lesser points in the processed data set.

REFERENCES

[1] Ji Zhang, Wynne Hsu, and Mong Li Lee, "Image Mining: Trends and Development".
[2] Carlos Ordonez and Edward Omiecinski, "Image Mining: A New Approach for Data Mining", 1998
[3] R. Balakrishnan and U. Karthick Kumar, "An Application of Genetic Algorithm with Iterative Chromosomes for Image Clustering Problem", *IJCSI International Journal of Computer Science Issues,* vol. 9, Issue 1, No 1, January 2012
[4] Nitin Gupta, Randhir Singh, Parveen Lehana, "Texture Enhancement of Plants IR Images Using Genetic Algorithm", October *2014 International Journal of Scientific and Research Publication, Vol.04 Issue 10 ISSN 2250-3153*.
[5] Qin Ding and Jim Gasvoda "A Genetic Algorithm for Clustering on Image Data", *International Journal of Information and Mathematical Sciences* 1:1 2005.
[6] A. Demiriz, K. P. Bennett, and M. J. Embrechts, "Semi-supervised clustering using genetic algorithms," R.P.I. Math Report No. 9901, Rensselaer Polytechnic Institute, 1999.
[7] Madhumathi.k, Dr.Antony Seladoss Thanamani, "Image Mining: Framework and Techniques", *Proceedings on International Conf. on Global Innovations in Computing Technology, Vol. 2 Special Issue 1, March 2014*.
[8] W. DuMouchel, C. Volinsky, T. Johnson, C. Cortes, and D. Pregibon, "Squashing flat files flatter," in Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 1999, pp. 6-15.
[9] Ishwar K. Sethi, Ioana L. Coman, Daniela Stan "Mining Association Rules between Low-level Image Features and High-level Concepts"
[10] Pradnya S.Borkar, Dr. A.R. Mahajan, "Types and Application of parallel Genetic Algorithm", *International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4 Issue 4, April 2014,ISSN 2277 128X*.
[11] Alexandre Xavier Falcao, Visual Informatics Laboratory Institute of Computing - University of Campinas, "Image processing using graph".

Jyoti Dua completed her MCA from Ewing Christian Institute of Management and Technology and perusing her M.Tech from United Institute of Technology Allahabad, India. Her area of interests are Algorithms, Data Mining, and Computer Networks.