Algorithms For Feature Selection In Content Based Image Retrieval:A Review

Manju Bala, Krishan Kumar Saluja, Sonika Jindal

M.Tech Student, Associate Professor, Assistant Professor , Department Of CSE
SBS State Technical Campus Ferozepur, India
Email: manju.bala89@gmail.coms, K.salujasbs@gmail.com, sonikamanoj@gmail.com

Abstract- CBIR applies to techniques for retrieving similar images from image databases, based on automated feature selection methods. Feature selection is an important topic in data mining, especially for high dimensional datasets. Feature selection (also known as subset selection) is a process commonly used in machine learning, wherein subsets of the features available from the data are selected for application of a learning algorithm.

In recent years, various heuristic optimization methods have been developed. Many of these methods are inspired by swarm behaviours in nature. In this paper, a new optimization algorithm based on the law of gravity and mass interactions is introduced. In the proposed algorithm, the searcher agents are a collection of masses which interact with each other based on the Newtonian gravity and the laws of motion. A hybrid meta-heuristic swarm intelligence-based search technique, called mixed gravitational search algorithm (MGSA), is employed. Some feature selection parameters are optimized to reach a maximum precision of the CBIR systems. Meanwhile, feature subset selection is done for the same purpose.

Keywords— CBIR,Featue selection,Feature selection Algorithms and Feature selection Apporachess

I. INTRODUCTION

Content-based image retrieval (CBIR) is a technique which uses visual contents to search images from large scale image databases according to user's interests. "Content-based" means that the search will analyze the actual contents of the image rather than keywords, tags or descriptions associated with the image [1]. The term 'content' in this context refers to colors, shapes, textures, or any other information that can be derived from the image itself [2].

A CBIR system performs indexing and retrieval tasks using features computed from images as opposed to using the whole images. These features are numerical values computed by image processing algorithms that capture visual images descriptions and store them in feature vectors. During the retrieval or similarity searching process, the images that are most similar to the query image according to some distance measure (e.g. the Euclidean distance), are returned. Knowing what the most relevant and non-redundant features are, according to the application domain, is very important for improving the accuracy of the similarity search. It is well-known that features obtained from only a single extractor is not always the most appropriate way to characterize images, mainly because more than one class of visual features is needed to represent them. However, using several feature extractors usually produces a large number of features, often

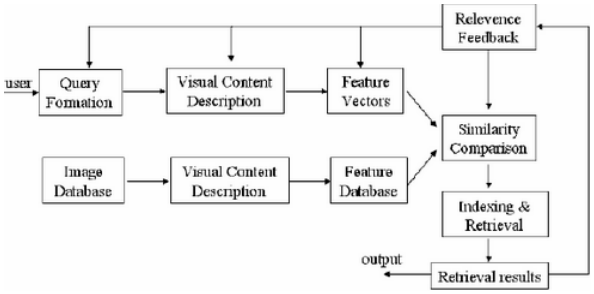

Fig.1 (Block Diagram of CBIR)

Containing correlated and irrelevant information that leads to the dimensionality curse problem [3] and deteriorates the efficiency of the retrieval process. Beyer et al. [4]in showed that increasing the number of features leads to a significant loss of the discriminative power of each feature. Thus, dimensionality reduction methods have been employed to ameliorate the dimensionality curse.

Content-based image retrieval uses the visual contents of an image such as color, shape, texture and spatial layout to represent and index the image.

Fig.1 represents the architecture of Content Based Image Retrieval (CBIR). The visual contents of the images in the database are extracted and described by multi-dimensional feature vectors. The feature vectors of the images in the database form a feature database. To retrieve images, users provide the retrieval system with

example images or sketched figures. The system then changes these examples into its internal representation of feature vectors. The similarities /distances between the feature vectors of the query example or sketch and those of the images in the database are then calculated and retrieval is performed with the aid of an indexing scheme. The indexing scheme provides an efficient way to search for the image database. Recent retrieval systems have incorporated users' relevance feedback to modify the retrieval process in order to generate perceptually and semantically more meaningful retrieval results[5].

A. APPLICATIONS OF CBIR

A wide range of possible applications for CBIR technology has been identified. Potentially fruitful areas include:

1) Medical diagnosis
2) Education and training
3) Digital Libraries
4) Home entertainment
5) Architectural and engineering design
6) Fashion and interior design
7) Journalism and advertising
8) Cultural heritage
9) Geographical information systems (GIS) and remote Sensing Crime prevention [6][7]

II. FEATURE SELECTION

Feature selection is the process of choosing a subset of the original feature spaces according to discrimination capability to improve the quality of data.

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection techniques are a subset of the more general field of feature extraction.[9]

In many machine learning applications, high-dimensional feature vectors impose a high computational cost as well as the risk of “overfitting”. Feature selection addresses the dimensionality reduction problem by determining a subset of available features which is the most essential for classification[10].

A "feature" or "attribute" or "variable" refers to an aspect of the data. Usually before collecting data, features are specified or chosen. Features can be discrete, continuous, or nominal. Generally, features are characterized as:

1) Relevant: These are features which have an influence on the output and their role can not be assumed by the rest.

2) Irrelevant: Irrelevant features are defined as those features not having any influence on the output, and whose values are generated at random for each example.

3) Redundant: A redundancy exists whenever a feature can take the role of another (perhaps the simplest way to model redundancy).

III. FEATURE SELECTION PROBLEM

The features selection module deals with the problem of choosing appropriate features for a given query, where the query is specified by positive and negative examples. Several approaches have been proposed to perform feature selection In the CBIR system , where the feature selection module is to be used, the requirements from the module are as follows:

1) It must select a subset of features that provides the best input for the image selection module. If too many features are selected, the presence of irrelevant features will obscure the ‘signal’ (reduce the signal to noise ratio (SNR) . On the other hand, taking too few features might impair the discrimination.

2) It must reduce the overall computational complexity by reducing the dimensionality of the classification problem.

3) Since we want the feature selection to take place at every query (i.e. within the user loop), it must be efficient. As the number of features is in the thousands, we require the module to have linear time complexity with respect to number of features.

4) As we are dealing with generic high level queries, the module should not expect any a-priori organization of the images in the database to increase its efficiency.

5) The image selection module in [3] is also constrained to have linear time complexity w.r.t. number of features, therefore the feature selector should assume only linear discriminance based classifiers.

6) The module should be able to handle example sets of sizes as small as 5.

As can be seen, [2], [6], [4] do not meet requirement 4, [8] is not designed for the large number of features involved, and [5] cannot handle the small example set sizes which are available. Our present work addresses the feature selection problem in the context of the above mentioned requirements.

Problem of selecting some subset of a learning algorithms input variables upon which it should focus

attention, while ignoring the rest. Feature selection is the process of selecting the best feature among all the features because all the features are not useful in constructing the clusters: some features may be redundant or irrelevant thus not contributing to the learning process[11][12].

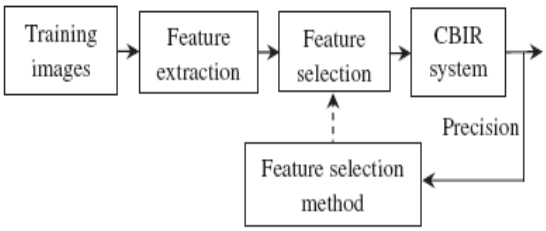

Fig. 2 (Block diagram of the traditional feature selection techniques in the CBIR Systems) [19]

Fig. 2 shows the block diagram of a traditional CBIR system with feature selection. As this figure shows, the feature extraction and the feature selection (FS) are two separate subsystems in the CBIR systems. In the CBIR systems, the feature selection is usually used after the feature extraction and selects the most important subset of features to increase the competence of the CBIR system.

IV. APPROACHES FOR FEATURE SELECTION (FS)

A feature selection algorithm determines how relevant a given feature subset ''s'' is for the task ''y'' (usually classification or approximation of the data). In theory, more features should provide more discriminating power, but in practice, with a limited amount of training data, excessive features will not only significantly slow down the learning process, but also cause the classifier to overfit the training data, as irrelevant or redundant features may confuse the learning algorithm, (Duda Hart, & Stork, 2001).

A feature selection algorithm (FSA) is a computational solution that is motivated by a certain definition of relevance. The purpose of a FSA is to identify relevant features according to a definition of relevance.

feature selection algorithms can be also categorized depending on search strategies used. Thus, the following search strategies are more commonly used:

a) Forward selection: start with an empty set and greedily add features one at a time.

b) Backward elimination: start with a feature set containing all features and greedily remove features one at a time.

c) Forward stepwise selection: start with an empty set and greedily add or remove features one at a time.

d) Backward stepwise elimination: start with a feature set containing all features and greedily add or remove features one at a time.

e) Random mutation: start with a feature set containing randomly selected features, add or remove randomly selected features one at a time and stop after a given number of iterations.

Fig.3 represents the characterization of a feature selection algorithm (FSA).In this shows the three characterization of algorithm i.e Search Organization, Generation of Successors

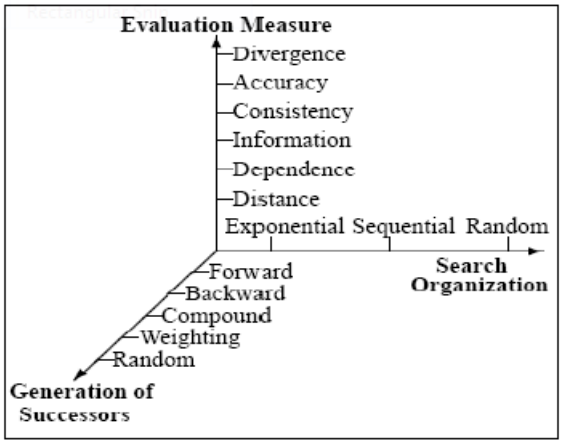

Fig.3 Characterization of a FSA

Generation of Successors and Evaluation Measure.

- Search Organization:

General strategy with which the space of hypothesis is explored. This strategy is in relation to the portion of hypothesis explored with respect to their total number. A search algorithm is responsible for driving the feature selection process using a specific strategy. We consider three types of search: exponential, sequential and random.

- Generation of Successors:

Mechanism by which possible variants (successor candidates) of the current hypothesis are proposed. Up to five different operators can be considered to generate a successor for each state: Forward, Backward, Compound, Weighting, and Random.

- Evaluation Measure:

Function by which successor candidates are evaluated, allowing to compare different hypothesis to guide the search process. Some of the evaluation measures are Probability of error, Divergence, Dependence, interclass distance, Information or Uncertainty and consistency[8,9,11].

V. FEATURE SELECTION ALGORITHMS

Feature Selection (FS) algorithms aim at choosing a reduced number of features that preserves the most relevant information of the dataset. FS is usually applied as a pre-processing step in data mining tasks by removing irrelevant or redundant features (dealing with the dimensionality curse), therefore leading to more efficient (reducing the computational cost and the amount of memory required) and accurate classification, clustering and similarity searching processes.

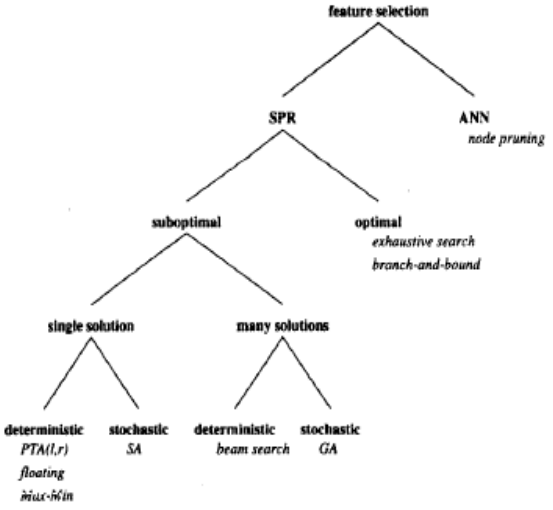

Fig.4 Taxonomy of Feature Selection Algorithms

A taxonomy of all the available feature selection algorithms broad categories is presented in fig.4.In this figure first divide methods into those based on statistical pattern recognition (SPR) techniques and those using Artificial neural networks(ANN).The SPR category is then split into those assured to find the best solution and those that may result in a suboptimal feature set. Another distinction is made between algorithms that are deterministic, producing the same subset on a given problem every time and those that have a random element which could produce different subsets on every run.

The first group of methods begins with a single solution and iteratively add or remove features until some termination Criterion is met. These "sequential "methods can be divided into two categories ,those start with the empty set and add features and those start with the full set and delete features.

Various algorithms which are used for feature selection in content based image retrieval that algorithms and its description is written below:

a) GSA:Gravitational Search Algorithm.The first version of GSA is a search algorithm for real-valued optimization inspired by the Newtonian laws of gravity and motion. In this algorithm introduce optimization algorithm based on the law of gravity. In the proposed algorithm, agents are considered as objects and their performance is measured by their masses. All these objects attract each other by the gravity force and this force causes a global movement of all objects towards the objects with heavier masses. Hence, masses cooperate using a direct form of communication, through gravitational force. The heavy masses – which correspond to good solutions – move more slowly than lighter ones, this guarantees the exploitation step of the algorithm.

In GSA, each mass (agent) has four specifications: position, inertial mass, active gravitational mass, and passive gravitational mass. The position of the mass corresponds to a solution of the problem, and its gravitational and inertial masses are determined using a fitness function.

In other words, each mass presents a solution, and the algorithm is navigated by properly adjusting the gravitational and inertia masses. By lapse of time, we expect that masses be attracted by the heaviest mass. This mass will present an optimum solution in the search space.

The GSA could be considered as an isolated system of masses. It is like a small artificial world of masses obeying the Newtonian laws of gravitation and motion. More precisely, masses obey the following laws:

- Law of gravity: each particle attracts every other particle and the gravitational force between two particles is directly proportional to the product of their masses and inversely proportional to the distance between them, R. We use here R instead of ^{R2}, because according to our experiment results, R provides better results than ^{R2} in all experimental cases.

- Law of motion: the current velocity of any mass is equal to the sum of the fraction of its previous velocity and the variation in the velocity. Variation in the velocity or acceleration of any mass is equal to the force acted on the system divided by mass of inertia.[13,14,15]

b) BGSA: A Binary Gravitational Search Algorithm (BGSA) was introduced by Rashedi et al [17]. for solving binary valued problems. In a binary environment, each dimension takes the value of 0 or 1. Moving in each dimension means that its value changes from 0 to 1 or vice versa. In BGSA, equations of force and velocity are same as continuous version, although the distance of R is computed based on the hamming distance [18].

c) MGSA: Mixed gravitational search algorithm. the objects move in a search space with the dimensions of both types of real and binary variables. A binary GSA was combined with a real GSA in order to simultaneously optimize the input feature subset

selection and the SVM parameter setting. An agent in this complex search space is called 'superagent'. To run GSA in a mixed real and binary search spaces, the equations of force, velocity and movement are calculated separately in real and binary parts. In MGSA, active mass is calculated using the fitness of super-agent and is the same for real and binary parts of the agent. For the real-part of super-agent, the Euclidean distance is used to calculate the distance between the real-part of Super-agents, while for the binary-part, the Hamming distance is replaced. The gravitational constant can be different in real and binary part. For more information, interested readers are referred to [19].

d) GA: The genetic algorithm is a general optimization method that is useful especially for computation-intensive applications. It mimics theevolution process in biology by representing the solution of the problem as genomes. The crossover of good genomes (indicated by small fitness value) tends to yield better results, and a certain probability of mutation allows for exploration of the whole solution space. After many generations of crossover and mutation, the algorithm yields an acceptable solution. In this study, each generation had the same number of features, and the fitness function was defined as the misclassification rate of a tenfold cross-validation procedure. In this procedure, the samples were divided randomly into ten groups, while one group was used as test data; the rest of the samples were used to fit a multivariate normal density function. The test data were classified based on likelihood ratios. After each group had acted as test group exactly once, the fitness function was calculated as the misclassification rate. The smaller the value was, the better the was fitness of the genome [20].

e) SVM: Support vector machine is based on the structural risk minimization principle. The SVM approach enjoys many attributes. It is lesscomputationally intense in comparison to artificial neural networks. It performs well in high-dimensional spaces and also well on both training data and testing data but does not suffer from the small size of training dataset as do other kinds of classifiers since the decision surface of SVM-based classifier is determined by the inner product of training data. The basic idea of SVM is to construct a hyper plane that maximizes the margin between negative and positive examples. The hyper plane is determined by the examples called support vectors that are closest to the decision surface. The decision surface is determined by the inner product of training data, which enables us to map the input vectors through function Φ into a higher-dimensional

inner product space called feature space. The feature space could be implicitly defined by kernel K(x, y)[21][22].

VI. CONCLUSION

In this paper, we reviewed the CBIR system and feature selection and its approaches.Various algorithms are described in this paper. All these algorithms are used for feature selection in content based image retrieval. GSA is constructed based on the law of Gravity and the notion of mass interactions. The GSA algorithm uses the theory of Newtonian physics and its searcher agents are the collection of masses. The GA algorithm, SVM algorithm etc. are present in this paper . The algorithms to different data particularities and thus the danger in relying in a single algorithm. This point in the direction of using new hybrid algorithms or combinations thereof for a more reliable assessment of feature relevance. There are a variety of methods in literature of performing feature selection. But, most feature selection methods assume that data are expressed with values without imprecision and uncertainty.

ACKNOWLEDGMENT

The authors would like to thank Dr.Krishan Kumar Saluja and Mrs. Sonika Jindal from Shaheed Bhagat Singh, State Technical Campus, Ferozepur, India for their valuable assistance in formulating this work.

REFERENCES

[1] Amandeep Khokher et al. / (IJAEST) International Journal of Advanced Engineering Sciences and Technologies Vol No. 9, Issue No. 2, 207 – 211.

[2] Sagar Soman, Mitali Ghorpade, Vrushali Sonone and Satish Chavan. Article: Content Based Image Retrieval using Advanced Color and Texture Features. IJCA Proceedings on International Conference in Computational Intelligence (ICCIA2012) iccia(9):-, March 2012. Published by Foundation of Computer Science, New York, USA.

[3] F. Korn, B. Pagel, C. Faloutsos, On the 'dimensionality curse' and the 'self-similarity blessing', IEEE Transactions on Knowledge and Data Engineering 13 (1) (2001) 96–111.

[4] Sérgio Francisco da Silva a,□, Marcela Xavier Ribeiro b, João do E.S. Batista Neto a, Caetano Traina-Jr. a, Agma J.M. Traina a. Improving the ranking quality of medical image retrieval using a genetic feature selection method. S.F. da Silva et al. / Decision Support Systems 51 (2011) 810–820

[5] Long, Fuhui, Hongjiang Zhang, and David Dagan Feng. "Fundamentals of content-based image retrieval." Multimedia Information Retrieval and

Management. Springer Berlin Heidelberg, 2003. 1-26.

[6] A Review of Content–Based Image Retrieval Systems in Medical Applications – Clinical Benefits and Future Directions Henning Miller, Nicolas Michoux, David Bandon and Antoine Geissbuhler Division for Medical Informatics, University Hospital of Geneva Rue Micheli–du–Crest 21, 1211 Geneva 14, Switzerland.

[7] John Eakins Margaret Graham University of Northumbria at Newcastle "Content-based Image Retrieval " documents/jtap 039.doc Octomber 1999.

[8] Ms. Anchal A. Solio, CSE, SGBAU/ Sipna C.O.E.T. Amravati, India Dr. Siddharth A. Ladhake, CSE, SGBAU/ Sipna C.O.E.T. Amravati, India. A Review of Query Image in Content Based Image Retrieval. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 4, April 2013

[9] http://en.wikipedia.org/wiki/Feature_selection#Subset_selection

[10] José M. Cadenas , M. Carmen Garrido, Raquel Martínez. Feature subset selection Filter–Wrapper based on low quality data. J.M. Cadenas et al. / Expert Systems with Applications 40 (2013) 6241–6252

[11] Ladha, L., and T. Deepa. "FEATURE SELECTION METHODS AND ALGORITHMS." International Journal on Computer Science & Engineering 3.5 (2011).

[12] Prasad, V. S. N., Faheema, A. G., & Rakshit, S. (2002). Feature Selection in Example-Based Image Retrieval Systems. In ICVGIP.

[13] Esmat Rashedi, Hossein Nezamabadi-pour*,Saeid Saryazdi. A simultaneous feature adaptation and feature selection method for content- based image retrieval systems. Department of Electrical Engineering, Shahid Bahonar University of Kerman, P.O. Box 76169-133, Kerman, Iran.

[14] Esmat Rashedi, Hossein Nezamabadi-pour *, Saeid Saryazdi, GSA: A Gravitational Search Algorithm. 0020-0255/$ - see front matter _ 2009 Elsevier Inc. All rights reserved.

[15] Disruption: A new operator in gravitational search algorithm S. Sarafrazi, H. Nezamabadi-pour⁎ , S. Saryazdi 1026-3098 © 2011 Sharif University of Technology. Production and hosting by Elsevier B.V. All rights reserved. Peer review under responsibility of Sharif University of Technology.

[16] E. Rashedi, H. Nezamabadi-pour, S. Saryazdi, Filter modeling using gravitational search algorithm, Engineering Applications of Artificial Intelligence 24 (2011) 117–122.

[17] E. Rashedi, H. Nezamabadi-pour, S. Saryazdi, BGSA: binary gravitational search algorithm, Natural Computation 9 (2010) 727–745.

[18] S. Sarafrazi, H. Nezamabadi-pour, Facing the classification of binary problems with a GSA-SVM hybrid system, Mathematical and Computer Modelling 57 (2013) 270–278.

[19] S. Sarafrazi, H. Nezamabadi-pour, Facing the classification of binary problems with a GSA-SVM hybrid system, Mathematical and Computer Modelling 57 (2013) 270–278.

[20] Yanjie Zhu,1 Yongqiang Tan,1 Yanqing Hua,2 Mingpeng Wang,2 Guozhen Zhang,2 and Jianguo Zhang1. Feature Selection and Performance Evaluation of Support Vector Machine (SVM)-Based Classifier for Differentiating Benign and Malignant Pulmonary Nodules by Computed Tomography. Journal of Digital Imaging, Vol 23, No 1 (February), 2010: pp 51Y65

[21] Joachims T: Text categorization with support vector machines. In: Proceedings of European Conference on Machine Learning (ECML), 1998

[22] Brown M, Grundy W, Lin D, Cristianini N, Sugnet C, Furey T, Ares M, Haussler D: Knowledge-based analysis of microarray gene expression data using support vector machines. 1999. http://www.cse.ucsc.edu/research/compbio/genex/genex. Santa Cruz, University of California, Department of Computer Science and Engineering

[23] http://en.wikipedia.org/wiki/Feature_selection

◈◈◈