

# Content Based Image Retrieval By Preprocessing Image Database

Kommineni Jenni

A Thesis Submitted to  
Indian Institute of Technology Hyderabad  
In Partial Fulfillment of the Requirements for  
The Degree of Master of Technology



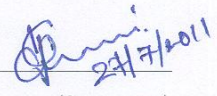
भारतीय प्रौद्योगिकी संस्थान हैदराबाद  
**Indian Institute of Technology**  
Hyderabad

Department of Computer Science and Engineering

July 2011

## Declaration

I declare that this written submission represents my ideas in my own words, and where ideas or words of others have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited, or from whom proper permission has not been taken when needed.

  
24/7/2011

(Signature)

Kommineni Jenni

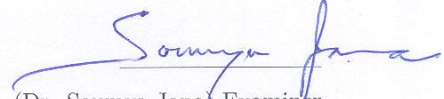
(Kommineni Jenni)

CS09G002

(Roll No.)

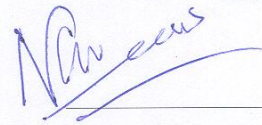
## Approval Sheet

This Thesis entitled Content Based Image Retrieval By Preprocessing Image Database by Kommineni Jenni is approved for the degree of Master of Technology from IIT Hyderabad



(Dr. Soumya Jana) Examiner  
Dept. of Electrical Engineering

IITH



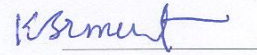
(Dr. Naveen Sivadasan) Examiner  
Computer Science and Engineering

IITH



(Dr. C. Krishna Mohan) Adviser  
Dept. of Computer Science and Engineering

IITH



(Dr. Sri Rama Murty Kodukula) Chairman  
Dept. of Electrical Engineering

IITH

## Acknowledgements

I would like to express my sincere gratitude to Dr. C. Krishna Mohan for providing me with the opportunity to do my research work under his guidance. His emphasis on steady and committed effort has motivated me during the course of the research work. I have immensely benefited from the excellent research environment that he has created and nurtured.

I am profoundly grateful to Dr. C. Sastry for his guidance and encouragement throughout my research work. I sincerely thank Dr. Soumya Jana and Dr. Sri Rama Murty Kodukula for their help and suggestions during the research work. Their suggestions have helped in refining the content and presentation of this thesis.

I am extremely thankful to all faculty members of the Department of Computer Science and Engineering for sharing their views and giving valuable suggestions during the discussion of my work in department reviews.

I thank all my classmates and research scholars for their friendly support who made the stay at this institute enjoyable, we shared joy and knowledge. I thank all my friends at IIT Hyderabad for the same.

I deeply express my loving thanks to my mother and father for encouraging me to do higher studies. I express my heartfelt appreciation and gratitude to my dear sister Sofia and brother-in-law Naresh for their esteemed support.

Finally, I thank everyone who helped me directly or indirectly during my stay at IIT Hyderabad.

# Dedication

*To Lord Balaji*

## Abstract

Increase in communication bandwidth, information content and the size of the multimedia databases have given rise to the concept of Content Based Image Retrieval (CBIR). Content based image retrieval is a technique that enables a user to extract similar images based on a query, from a database containing a large amount of images. A basic issue in designing a content based image retrieval system is to select the image features that best represent image content in a database. Current research in this area focuses on improving image retrieval accuracy. In this work, we have presented an efficient system for content based image retrieval. The system exploits the multiple features such as color, edge density, boolean edge density and histogram information features.

The existing methods are concentrating on the relevance feedback techniques to improve the count of similar images related to a query from the raw image database. In this thesis, we propose a different strategy called preprocessing image database using k means clustering and genetic algorithm so that it will further helps to improve image retrieval accuracy. This can be achieved by taking multiple feature set, clustering algorithm and fitness function for the genetic algorithms.

Preprocessing image database is to cluster the similar images as homogeneous as possible and separate the dissimilar images as heterogeneous as possible. The main aim of this work is to find the images that are most similar to the query image and new method is proposed for preprocessing image database via genetic algorithm for improved content based image retrieval system. The accuracy of our approach is presented by using performance metrics called confusion matrix, precision graph and F-measures. The clustering purity in more than half of the clusters has been above 90 percent purity.

# Contents

Declaration . . . . .	ii
Approval Sheet . . . . .	iii
Acknowledgements . . . . .	iv
Abstract . . . . .	vi
<b>Nomenclature</b>	<b>viii</b>
<b>1 Introduction to Content Based Image Retrieval</b>	<b>3</b>
1.1 Tasks involved in content based image retrieval . . . . .	4
1.2 Computational features of content based image retrieval system . . . . .	5
1.2.1 Color . . . . .	5
1.2.2 Texture . . . . .	6
1.2.3 Shape Retrieval . . . . .	6
1.2.4 Semantics . . . . .	6
1.2.5 Edge density and Boolean edge density . . . . .	6
1.3 Database indexing in content based image retrieval . . . . .	7
1.4 Issues addressed in this thesis . . . . .	7
1.5 Organization of the thesis . . . . .	8
<b>2 Overview of Approaches for Content Based Image Retrieval and Relevance Feed-back</b>	<b>9</b>
2.1 Existing methods for content based image retrieval . . . . .	9
2.1.1 Major content based image retrieval systems . . . . .	10
2.1.2 Applications of content based image retrieval system . . . . .	11
2.2 Measure of similarity . . . . .	12
2.3 Issues addressed in traditional content based image retrieval systems . . . . .	13
2.4 Relevance feedback of content based image retrieval . . . . .	13
2.4.1 Need for relevance feedback . . . . .	15
2.4.2 Feedback strategies . . . . .	15
2.4.3 Automated feedback . . . . .	15
2.5 Summary . . . . .	16
<b>3 Clustering Technique for Content Based Image Retrieval and Genetic algorithm</b>	<b>17</b>
3.1 Clustering technique for content based image retrieval . . . . .	17
3.2 Introduction to genetic algorithms . . . . .	18

3.3	Crossover and Mutation functions . . . . .	19
3.4	Summary . . . . .	19
<b>4</b>	<b>Preprocessing Image Database Using K-Means Clustering and Genetic Algorithms</b>	<b>21</b>
4.1	Results and Discussions . . . . .	24
4.2	Performance Metrics . . . . .	28
4.2.1	Confusion matrix . . . . .	29
4.2.2	Precision Graph . . . . .	30
4.2.3	F-Measures for Previous Approach . . . . .	31
4.2.4	F-Measure for Proposed Approach . . . . .	32
4.3	Summary . . . . .	33
<b>5</b>	<b>Summary And Conclusions</b>	<b>34</b>
5.1	Contributions of the work . . . . .	34
5.2	Directions for further research . . . . .	34
	<b>Bibliography</b>	<b>36</b>



# List of Figures

1.1	Architecture of CBIR . . . . .	5
2.1	Traditional Content Based Image Retrieval system . . . . .	10
2.2	Relevance Feedback block diagram . . . . .	15
3.1	clustering block diagram . . . . .	17
4.1	CBIR with augmented preprocessing stage . . . . .	22
4.2	dinosaurs retrieval . . . . .	24
4.3	dinosaurs retrieval . . . . .	24
4.4	flower(s) retrieval . . . . .	25
4.5	flower(s) retrieval . . . . .	25
4.6	Bus(es) retrieval . . . . .	26
4.7	Bus(es) retrieval . . . . .	26
4.8	Horse(s) retrieval . . . . .	27
4.9	Horse(s) retrieval . . . . .	27
4.10	Precision Graph . . . . .	30
4.11	F-Measure for Previous Approach . . . . .	31
4.12	F-Measure for Proposed Approach . . . . .	32

# List of Tables

4.1	Confusion Matrix For Calculating Clustering Purity . . . . .	29
4.2	Accuracy calculation for previous and proposed approaches . . . . .	30
4.3	F-Measure of Previous approach . . . . .	31
4.4	F-Measure of proposed approach . . . . .	32

# Chapter 1

## Introduction to Content Based Image Retrieval

The problem of searching similar images from large image repositories on basis of their visual contents is called content-based image retrieval [1]. The term content in Content Based Image Retrieval (CBIR) refers to colors, shapes, textures [9] or any other information that can be obtained from the image itself. There are two significant phases in the CBIR:

1) Indexing phase, where in the image information like the color, shape and texture is specified into features that are consequently stored in an index data structure along with a link to the image. Database images are stored in structured manner.

2) Retrieval phase, where in the searching of an image in the CBIR index needs the description of the properties of the image of interest either by supplying a sample image or denoting the image features. Based on the similarity measure between database images and query image the relevant images will be retrieved.

Previously searching an image database was based on human annotation that is each image in a database is given some keywords to denote the semantic meaning of the image. Then all the keywords are used to index images. Thus, searching and retrieving images is based on the keywords of images. This type of image retrieval is called as Text Based Image Retrieval (TBIR)[22]. Now many search engines that claim to do text based image retrieval. Google and AltaVista do text based image retrieval. These search engines search the text around the image such as captions, file names, and paragraphs located close to the image to search for relevant items to the query. This TBIR approach has many limitations namely the size of image collection gets increasingly large and manually giving each image annotation is very difficult. Annotating an image based on human perception is individual. Different people may give different annotations to images with similar visual contents.

In the early 1990's content based image retrieval was proposed to overcome the limitations of text based image retrieval. There are many differences between content-based image retrieval systems and classic information retrieval systems. The major differences are that in CBIR systems images are indexed using features extracted from the content itself and the objective of CBIR systems is to retrieve similar images to the query rather than exact matches. So, retrieval results are not perfect matches of the query image. The similarity in most CBIR systems is quantified and the database

entries are ranked based on their similarity to the query image. Similar images are retrieved as result of a query image. The different users may be interested in different parts of the same image. So, similarity-based retrieval is a more flexible than exact matching, and gives better performance in queries such as finding the images similar to the given image.

Image retrieval is related with techniques for storing and retrieving images both efficiently and effectively. Available image retrieval methods locate the desired images by matching keywords that are assigned to each image manually. These manual annotations are highly dependent on the subjectivity of human perception [23]. That is, for the same image content different people may perceive the visual content of the image differently.

There are many primitive features which denote some general visual characteristic including color, shape, texture, spatial relationships among objects and these features can be used in most CBIR applications. Among various primitive features, the color information has been taken to analyze the images because of its invariance with respect to image scaling and orientation. In the proposed approach, the image database is structured by using techniques called clustering via genetic algorithm. Since clustering will make the association to be strong between members of the same images and weak between members of different images, so similar images will fall into the same cluster and different images will fall into different clusters. This way CBIR system in our approach is more efficient and accurate in achieving the results.

## 1.1 Tasks involved in content based image retrieval

The objective of content based image retrieval is to develop techniques to automatically extract and retrieve relevant similar images from the huge database. In conventional content based image retrieval systems, the query image is given to the CBIR system where the CBIR system will retrieve images from raw (unstructured) image database related to query image. In the next stage, the relevance feedback is used to refine the results such that the retrieved images will be more similar to the query image. In order to get the good result set, the relevance feedback process is repeated several times. The process will be stopped when the satisfactory results are shown or the user quits. In the previous CBIR system, the image database is unstructured. Basically content based image retrieval involves three major tasks is shown in Figure 1.1.

### **The major functions of the CBIR:**

- Analyze the contents of the source information and represent the contents of the analyzed sources in a way that will be suitable for matching user queries. This step is normally time consuming since it has to process all the source information (images) in the database.
- Analyze user queries and represent them in a form that will be suitable for matching with the source database. Which is similar to the source images in the database.
- Define an approach to match the search queries with information in the stored database. Retrieve the images relevant to the query image.

## Architecture of CBIR:

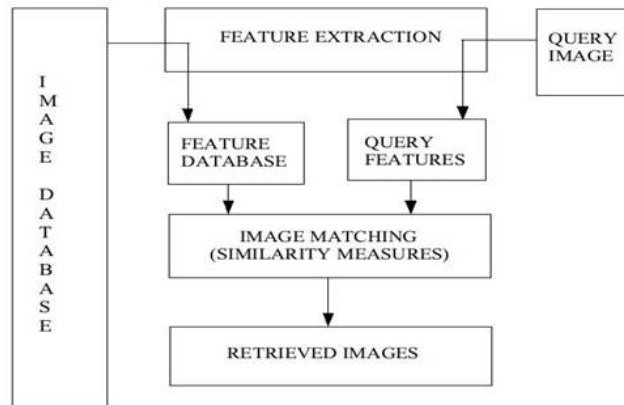


Figure 1.1: Architecture of CBIR

## 1.2 Computational features of content based image retrieval system

Feature extraction is the process of describing the image by considering parameters known as features (color, edge, texture etc) from a given image. A feature is defined as a "descriptive parameter that is extracted from an image" [50]. The effectiveness of image retrieval depends on the effectiveness of features/attributes used for the representation of the content. An important issue is the choice of suitable features for a given task. Effective image retrieval can be achieved by collaboratively using color [8], edge density [10], boolean edge density and histogram bins. These features are discussed in this section.

### 1.2.1 Color

Color has been the most effective feature and almost all systems use colors. Although most of the images are in the RGB (Red, Green, Blue) color space, this space is rarely used for indexing and querying as it does not related well to the human color perception. It only like reasonable to be used for images taken under exactly the same conditions each time such as trademark images. Other spaces such as HSV (Hue, Saturation, Value) or the CIE Lab and Luv spaces are much better with respect to human perception and are more frequently used. This means that differences in the color space are similar to the differences between colors that humans perceive.

There are different types of color spaces available which are appropriate for different purposes. Some of the color spaces that we often come across are RGB, HSV, CIE Lab and Luv [8]. Color feature can be comprised of histogram bins or average, standard deviation or variance in an opted color space.

### 1.2.2 Texture

Texture [6], is another important property of images. Texture features [3] of images refer to the visual patterns that have properties of homogeneity that do not result from the presence of only a single color or intensity. Image texture content provides information of image properties such as smoothness, coarseness, and regularity which is useful in a CBIR system. Basically, texture representation methods can be classified into two categories: structural and statistical. Structural methods including morphological operator and adjacency graph, describe texture by identifying structural primitives and their placement rules. Structural methods tend to be most effective when applied to textures that are very regular. Statistical methods, including Fourier power spectra, co-occurrence matrices, Shift-invariant Principal Component Analysis (SPCA), Tamura feature, World decomposition, Markov random field, fractal model and multi-resolution filtering techniques such as Gabor [11] and wavelet transform, characterize texture by the statistical distribution of the image intensity.

### 1.2.3 Shape Retrieval

Shape features [3], the objects or regions have been used in many content-based image retrieval systems. Compared with color and texture features, shape features are usually described after images have been segmented into regions or objects. Since robust and accurate image segmentation is difficult to achieve, the use of shape features for image retrieval has been limited to special applications. The methods for shape description can be classified into boundary or region-based methods. A good shape representation feature for an object should be invariant to translation, rotation and scaling.

### 1.2.4 Semantics

Most current CBIR systems retrieve images from a collection, on the basis of the low level features of images such as color, texture and shape. Nevertheless, some systems attempt to find images that are semantically similar to a given query. Semantically similar is meant in the sense of human visual similarity perception (or called high level in CBIR).

### 1.2.5 Edge density and Boolean edge density

Edges are identified from each image using sobel operator. To improve the pixels that belong to the edges and boundaries by using a standard edge detector. Sobel operator finds the gradient(change) in intensity at each point in the image. Based on this intensity change towards horizontally or vertically we can move around the image edge. Sobel operator exits for x-order and y-order derivatives and also for mixed partial derivatives. Pixels far from edges will drop to zero and those near to an edge will increase to maximum. Calculated the mean pixel value of the resultant image. From the edge density, the image is represented as edge pixels are white (1) and non-edge pixels are black (0). Count white pixel in the image. The mean of these white pixels are considered as boolean edge density.

### 1.3 Database indexing in content based image retrieval

Practically image content information is represented using a high dimensional format that is (X, Y) coordinates in an image. Most commonly a tree structure is utilized to store image information since it has high dimensional image attributes. R-tree [13], R\*-tree[14], VP-tree structure [15] and Hybrid Tree [16] are some of the widely used tree structures. R-tree is a data structure similar to B-tree used for spatial(or image) access methods (R-stands for rectangular). A database system requires an index mechanism for faster access and retrieval of image data efficiently from image database, as required in image object search applications. These tree based data structures splits image space with hierarchically nested components called MBRs (minimum bounding rectangles or bounding boxes). Often these bounding boxes are overlapped with parents. Each node in an R-tree contains variable number of elements. The number of elements in a node is limited up to some pre-defined maximum size. Each element within a non-leaf node holds two pieces of data: Address of a child node, and the bounding box of all entries within this child node. Each element within a leaf node holds two pieces of data; bounding box of the actual data or image object property and address of actual image property or attribute, and the bounding box of the data element. The operations on R-trees are same as like B-trees.

A variant of R-tree employed in the indexing of spatial information is known as R\*-tree. R\*-trees support point and spatial data at the same time with a slightly higher cost than other R-trees. Since the Indexing tree structure can perform efficiently in dictionary operations. It can't be used for finding similarity among the images in the database. So indexing tree structure was limited to structure the image database so that efficient retrieval is possible. For finding similarity among image database we are using clustering techniques, which are discussed in the next section.

### 1.4 Issues addressed in this thesis

This section deals with the major issues addressed in content based image retrieval. The key issues are the choice of the features for the representation of images, the choice of a similarity/distance metric and an algorithm that is general enough for managing huge amount of image database. We address these issues on the basis of significant changes exhibited by a small subset of color, edge and texture features. A novel approach for preprocessing image database is proposed based on the k-means clustering algorithm with appropriate fitness functions of the genetic algorithm. We also examine the effect of objective functions crossover and mutation on the performance of preprocessing image database.

The problem of image database clustering is addressed in the context of content based image retrieval. An important issue is the presentation of images, so that resultant features adequate capture class-specific information. Another issue is the managing huge amount of database. Our approach to this problem is preprocessing image database with appropriate model called clustering algorithm. In this thesis, preprocessing image database is to cluster the similar images as homogeneous as possible and separate the dissimilar images as heterogeneous as possible. This thesis also proposes new method for preprocessing image database using k means clustering algorithm with the support of genetic algorithm fitness functions. The main aim of this work is to find the images that are most similar to the query image and new method is proposed for preprocessing image database

for better image retrieval.

## 1.5 Organization of the thesis

The thesis is organized as follows. An overview of the existing approaches to content based image retrieval and relevance feedback is presented in Chapter 2. Some research issues are identified from the existing approaches. In Chapter 3, the clustering technique with genetic algorithm are briefly explained for processing image database. K-means clustering algorithm is used for finding cluster centers and genetic algorithm fitness functions are proposed to find the best cluster center positions. In Chapter 4, multiple feature extraction methods are briefly explained. The basis for this method is the significant change exhibited by a few color components over a sequence of images. Preprocessing image database is performed using feature vectors with 136 dimensions and fitness functions, namely crossover and mutation. The similarity measure is performed using euclidean distance measure. The comparison of results with previous and proposed approaches are examined. The results are explained with confusion matrix, precision graph and F-measures are shown in this chapter. In Chapter 5, summarizes the research work carried out as part of this thesis, highlights the contributions of the work and discusses directions for future work.



## Chapter 2

# Overview of Approaches for Content Based Image Retrieval and Relevance Feedback

This chapter reviews some of the existing approaches to content based image retrieval. The problem of image database is briefly described in section 2.1. In section 2.2, the similarity measures are briefly explained. The existing approaches to content based image retrieval are reviewed, with particular focus on the relevance feedback of content based image retrieval. Some research issues arising out of the review of existing methods are identified, which are addressed in this thesis.

### 2.1 Existing methods for content based image retrieval

As the amount of collection of digital images increase, the problem of locating a desired image in an huge collection also becomes very difficult. Therefore the need of an efficient method to retrieve digital images is recognized by the public. There are two approaches to image retrieval, Text Based approach and Content Based approach. The previous solution is a more traditional approach which is keyword based image retrieval. The keyword indexing of digital images is useful but requires a considerable level of effort and often limited for describing image content. The alternate approach, the content based image retrieval indexes images by using the low level features of the digital images and the searching depends on features being automatically extracted from the image.

Content Based Image Retrieval is the term used to describe the process of retrieving images from a database on the basis of the internal features of images. In CBIR, digital images are indexed [2] by summarizing their visual contents through automatically extracted features such as texture, color and shape. CBIR retrieves stored digital images from a collection by comparing features extracted from the images. The most common features used are mathematical measures of color, texture or shape [1]. The CBIR system identifies those stored images whose feature values match those of the query most closely and displays images to the user. In the following section, the traditional content based image retrieval approach will be described.

Initially selecting an appropriate feature set for the image database. The selection of feature

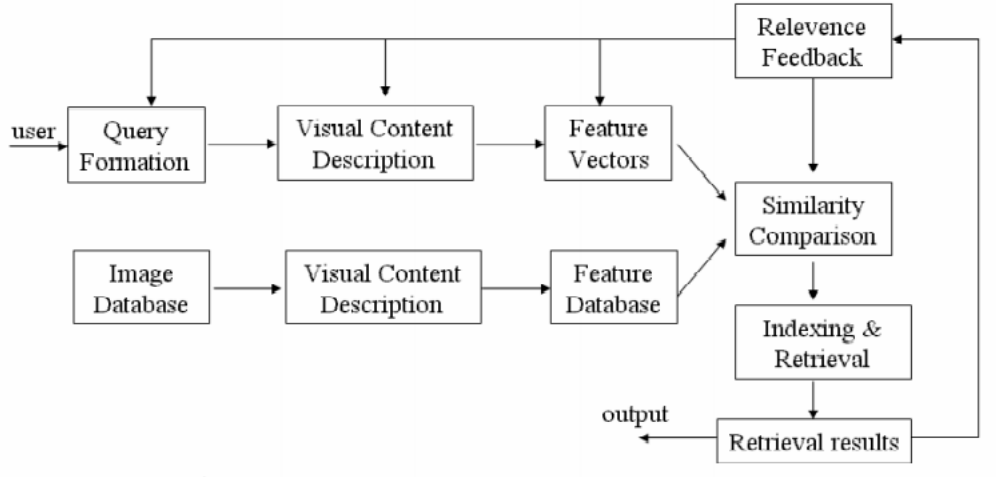


Figure 2.1: Traditional Content Based Image Retrieval system

set should be in a way that it should approximate images as close as possible in a feature space. Preparing a query for the retrieval (i.e. form a query feature vector). Select appropriate distance or similarity measure for the retrieval. In the first iteration, retrieve images from the image database related to the query image (i.e. retrieve images which are closer to the query image using distance or similarity measure). After first iteration we send the retrieved images to the user feedback. The user identify the similar images related to the query by some indication (such as markings ). After getting the feedback from the user we will hand it over to the learning step. There are two types of learning techniques viz. Short term learning and Long term learning. By using relevance feedback learning techniques the results will be refined and the final results are given as output. (NOTE: multiple relevance feedbacks will be conducted before final results) The traditional content based image retrieval is shown in Figure 2.1.

### 2.1.1 Major content based image retrieval systems

A brief overview of the major content based image retrieval systems was presented in this section. Methods like QBIC, Photobook, MARS, IMatch, Blobworld and Netra systems were discussed.

**QBIC** : IBM developed the image retrieval system, Query By Image Content (QBIC) [24]. It extracts simple features from objects or images which are color, texture and shape. Color features computed are; the 3D average color vector of an object or the whole image in RGB, YIQ, Lab, Munsell color space and a 256-dimensional RGB color histogram. The texture features used in QBIC are modified versions of the coarseness, contrast, and directionality features. The shape features consist of shape area, circularity, eccentricity, major axis orientation and a set of algebraic moment invariants. A method of retrieving images based on a rough user sketch was also implemented in QBIC. For this purpose, images in the database are represented by a reduced binary map of edge points. QBIC allows combined type searches where text-based keywords and visual features are used in a single query.

**Photobook** : The Photobook system [25] allows users to retrieve images by color, shape and texture features, and was developed at Massachusetts Institute of Technology. This system

provides a set of matching algorithms, including Euclidean, Mahalanobis, divergence, vector space angle, histogram, Fourier peak, and wavelet tree distances as distance metrics. The method in which users can define their own matching algorithms, was provided in most recent version. The system includes a distinct interactive learning agent (FourEyes), which is a semi-automated tool as well as, can generate query models based on example images provided by users. This adds the advantage for users to directly address their query demands for different domains and, for each domain, an optimal query model.

**MARS :** The advantage of MARS[26] is that, it allows combined features queries, and was developed at UIUC. Moreover, it allows combinations of global or local image features with textual keywords associated with the images. Color is represented by using a 2D histogram over the HS coordinates of the HSV space. Texture is represented by two histograms, one measuring the coarseness while the other one for the directionality of the image, and one scalar defining the contrast. In order to extract the color/texture layout, the image is divided into 5 x 5 sub images. Fourier Descriptors (FD) was used to represent the shape of the boundary of the extracted objects. Mars used relevance feedback techniques from the information retrieval (IR) domain in content-based image retrieval, to permit interactive CBIR.

**IMATCH:** The IMatch [48] system allows users to retrieve images by color, texture, and shape. IMatch supports several query methods to query similar images: Color Similarity, Color and Shape (Quick), Color and Shape (Fuzzy), and Color Distribution. Color Similarity queries for images similar to an example image based on the global color distribution. Color and Shape (Quick) queries similar images for a given image by combining shapes, textures, and colors. Color and Shape (Fuzzy) performs additional steps to identify objects in example images. Color Distribution allows users to draw color distributions, or specify the overall percentage of one color in desired images. IMatch also supports non-CBIR features to identify images: binary identical images, duplicate images that have been resized, cropped, or saved in different file formats, and images that have similar file names to the given images.

**BlobWorld:** Expectation Maximization (EM) algorithm is used in this image retrieval system, to segment the images into regions of uniform color and texture (blobs). UC Berkeley developed this system and named it as Blob World [27]. The color is described by a histogram of 218 bins of the color coordinates in Lab-space and the texture is represented by mean contrast and anisotropy over the region. Shape is represented by approximate area, eccentricity, and orientation. Query-by-example is performed based on a region from one of the images presented to the user. In Blob world, it allows the user to view the internal representation of the submitted image and the query results; facilitating better understanding of the retrieval results.

**Netra:** Netra [28] is a system developed at the University of California, Santa Barbara and is based on regions of homogeneous colors. For image indexing and retrieval it uses color, texture, shape and spatial location information. Images are segmented off-line using an edge flow segmentation technique, and each segment is characterized by its local features.

### 2.1.2 Applications of content based image retrieval system

Various applications of Content Based Image Retrieval System [29] were discussed in this section.

**Crime Prevention** Generally Law enforcement agencies maintain large archives of visual evidence, including past suspects facial photographs, fingerprints, tyre treads and shoe prints. When-

ever a serious crime is committed, for comparing evidence from the scene of the crime for its similarity to records in their archives, CBIR is very helpful.

**Education and Training** It is often difficult to identify good teaching material to illustrate key points in a lecture or self-study module. The availability of searchable collections of video clips providing examples of (say) avalanches for a lecture on mountain safety, or traffic congestion for a course on urban planning, could reduce preparation time and lead to improved teaching quality.

**Fashion and Interior Design** Similarities can also be observed in the design process in other fields, including fashion and interior design. Here again, the designer has to work within externally imposed constraints, such as choice of materials. The ability to search a collection of fabrics to find a particular combination of color or texture is increasingly being recognized as a useful aid to the design process.

**The Military** Some of the examples of Military applications where CBIR can be used are, recognition of enemy aircraft from radar screens, identification of targets from satellite photographs, and provision of guidance systems for cruise missiles.

**Intellectual Property** This has been prime application area of CBIR from long time. Trademark image registration, where a new candidate mark is compared with existing marks to ensure that there is no risk of confusion.

**Medical Diagnosis** Even though the prime requirement for medical imaging systems is to be able to display images relating to a named patient, there is increasing interest in the use of CBIR techniques to aid diagnosis by identifying similar past cases.

**Geographical Information Systems GIS and Remote Sensing** Satellite images are extensively used by Agriculturalists and physical geographers, both in research and for more practical purposes, such as identifying areas where crops are diseased or lacking in nutrients or alerting governments to farmers growing crops on land they have been paid to leave lying fallow.

**Architectural and Engineering Design** The use of stylized 2-D and 3-D models to represent design objects, the need to visualize designs for the benefit of nontechnical clients, and the need to work within externally imposed constraints, often financial; were some of common features shared by Architectural and Engineering design. By keeping such constraints in mind, the designer needs to be aware of previous designs, particularly if these can be adapted to the problem at hand. Hence the ability to search design archives for previous examples which are in some way similar, or meet specified suitability criteria, can be valuable.

**Cultural Heritage** Museums and art galleries also deals with inherently visual objects. The ability to identify objects sharing some aspects of visual similarity can be useful for both researchers trying to trace historical in uences, and art lovers looking for further examples of painting or sculptures appealing to their taste.

## 2.2 Measure of similarity

Similarity measurement[49] is one of the key point in content based image retrieval (CBIR). An important step in most clustering is to select a distance measure, which will determine how the similarity of two elements is calculated. In CBIR, images are represented as features in the database. Once the features are extracted from the indexed images, the retrieval becomes the measurement of similarity between the features. Many similarity measurements exist.

Common distance functions:

- The Euclidean distance (also called distance as the crow flies or 2-norm distance). A review of cluster analysis in health psychology research found that the most common distance measure in published studies in that research area is the Euclidean distance or the squared Euclidean distance.
- The Manhattan distance (aka taxicab norm or 1-norm)
- The maximum norm (aka infinity norm)
- The Mahalanobis distance corrects data for different scales and correlations in the variables
- The angle between two vectors can be used as a distance measure when clustering high dimensional data.
- The Hamming distance measures the minimum number of substitutions required to change one member into another.

Euclidean distance[47] is the most common metric for measuring the distance between two vectors, and is given by the square root of the sum of the squares of the differences between vector components. We used euclidean distance measure in our approach.

## 2.3 Issues addressed in traditional content based image retrieval systems

An observation arising out of the review of the existing approach is that an algorithm with only one type of features and/or similarity metric is not general enough to find relevant images from the database. Moreover, most of the algorithms are sensitive to the threshold used on similarity/distance metric. In this thesis, we attempt to address these issues both at the level of features used and managing large amount of image database using efficient algorithm for preprocessing.

In order to detect only a few color features for the feature vectors, we propose a combination of multiple features and preprocessing image database that exploits the accuracy of better image retrieval.

This is in contrast to existing approaches that compare a image information of previous and proposed approaches. Our objective is to find most relevant images from the huge database so that the proposed algorithm performs well and gives accurate result.

## 2.4 Relevance feedback of content based image retrieval

In previous CBIR systems extraction and revival of more akin image objects to a given image query are obtained using relevance feedback techniques. But for our proposed technique we are concentrating on preprocessing image object inventory to get more refined result set.

Content Based image retrieval is a process to find and extract image objects which are similar in visual content to a given image query from image inventory or database. This image retrieval is mainly depends on a comparison of low level attributes or characteristics, such as color, texture

features with the extracted image objects. At the early stage of CBIR, research primarily focused on expressing various feature representations, hoping to find a best representation for each feature. For example, for texture feature itself almost a several representations have been proposed, including Word decomposition, Fractals, Gabor filter and Wavelets, etc [43]. So the associated system design is to respond to the first best akin representations for the visual features.

The early CBIR system is to first find the best representations for the visual features. Steps in querying and retrieval Process:

- Specification of single or multiple features and their weights the user is interested in.
- Based on these specified features and weights the retrieval system finds the best alike match and then extracts it.

These types of systems are treated as centric systems. While retrieving the best match we need deal with accuracy. Accuracy is the main concern addressing best image retrieval process. As of now the performance of content based image retrieval methods are still limited, much research effort needed to address CBIR issues. The limited retrieval accuracy is because of the big gap between semantic concepts and low level features. The computer centric approach assumes the mapping between high level concepts to low level feature is easy for the user to do. While in some cases it is true, the mapping between the high level concepts represents the actual physical object (fresh mango) and the object attributes color, shape and etc are the low level features of high level concept. In other cases, this may not be true. So, the gap exists between the two. Relevance Feedback is one technique that may bridge the gap. Relevance feedback is a supervised learning technique used to improve the effectiveness of image retrieval systems [44]. The main concept of relevance feedback is to get optimum solutions using positive and negative feeds provided by the user to improve the systems performance. For a given query image, the system retrieves the best possible search of images and give them rank based on positive and negative feeds based on the similarity metric [45]. In the next stage the system refines the query by just including positive ranked feeds and by eliminating negative feeds to get the optimum search.

In the figure 2.2, we can see the relevance feedback process in detail, First user gives a query image, features are extracted from that query image, same features are extracted from all data base images and a similarity measurement is calculated and results are given to user, user selects the relevant images from that and again similarity measure and results are given, this process repeats until user satisfaction or user quits.

Steps in Image retrieval refinement process:

- User gives query image.
- Features are extracted from that query image.
- Similarity feature measurement is calculated from the image database.
- Same featured images get retrieved from the image inventory and results are given to user.
- User selects the relevant images from that and again similarity measure and results are given.
- This process repeats until user satisfaction or user quits.

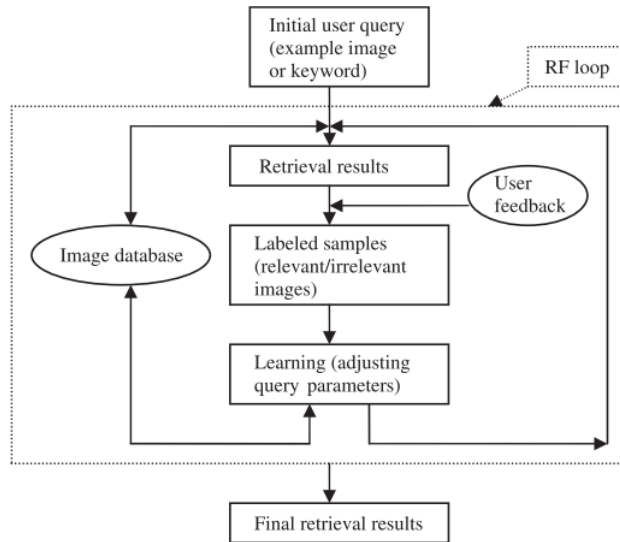


Figure 2.2: Relevance Feedback block diagram

### 2.4.1 Need for relevance feedback

Relevance Feedback (RF) has been defined as the process of adjusting an existing query using information fed back by the user about the relevance of previously retrieved documents.

- It is not always easy for the user to express his needs using an example based query.
- The retrieval system may fail in translating the users needs into image features and similarity measures.

Because of the above reasons, we should include RF techniques in the process of Content Based Image Retrieval(CBIR).

### 2.4.2 Feedback strategies

The two main strategies for RF are either

- Make separate queries for each ranked feedback image and merge the query results.
- Form a pseudo-image from the feedback images and execute a query with this image. this image.

### 2.4.3 Automated feedback

Automated Relevance feedback is only possible once the user judgments exist on the iamge results. Once the user initiates an image query, relevant result set comes as result. Now the user needs to make positive or negative judgment against the individual result image object by comparing the query image features. By feeding back the images the user judged (positive/negative) as relevant we can refine the query retrieval. Thus a reproducible RF for every user can be simulated based upon the judgement and the initial query results of a system. This technique can be used to compare different feedback strategies or to enhance user queries by automatically creating negative feedback.

### **Only positive feedback**

After ranking positive result set for the image query the system weights the features of these images more strongly. As all high ranked returned images have many features in common, the non-relevant images may also be ranked highly in the next step. For this feedback, we select as relevant all the images from the initial query result which the user judged to be relevant.

### **Positive and negative feedback**

Image query result greatly improved by using negative feedback. The user needs to make sure of which images to mark negative, because there is possibility of losing more important positive features. Many systems have problems with too much negative feedback. A query from a user who only uses positive feedback can be improved by automatically supplying non-selected images as negative feedback.

### **Several steps of feedback**

For an image query, In a single step of process we cannot get the optimum result set. So we need to use RF techniques to refine the query to get the best possible results. RF always improves the results. However, too much negative feedback can destroy the query. This can be avoided by using a technique of separately weighting positive and negative features. Using a larger number of images as a source for feedback improves results, but this potential is limited by the number of images a user really inspects. Using a variety of automated RF strategies, we can evaluate the flexibility of a CBIRS. It is important that using several steps of feedback continues to improve the results so, that feature space can be explored thoroughly.

## **2.5 Summary**

In this chapter, some of the existing approach to content based image retrieval and relevance feedback were reviewed. The key component of content based image retrieval are the features used to represent images and the measure of similarity/distance used to find relevant images. The survey suggests that there is a need for robust features and algorithms which are general enough to manage large image database. In this thesis, we propose novel algorithm for preprocessing image databaase for better image retrieval to address this issue, and also examine the effectiveness of efficient features. In image database clustering, most lagorithms are still based on low- level features, since deriving more meaningful information at a higher level is challenging task. We explore, low-level color-based features, edge-based feature and texture feature for preprocessing image database and for image retrieval. We also study the effect of combining evidence obtained from multiple feature and clustering, on the performance of clustering.



## Chapter 3

# Clustering Technique for Content Based Image Retrieval and Genetic algorithm

This chapter deals about the clustering techniques for content based image retrieval. Clustering can be considered as the most important unsupervised learning problem. A cluster is a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters.

### 3.1 Clustering technique for content based image retrieval

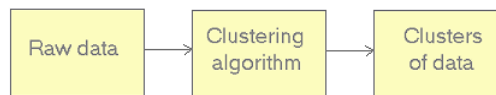


Figure 3.1: clustering block diagram

Clustering [19] is a tool for data analysis, which solves classification problems. Its objective is to distribute classes (people, objects, events etc.) into groups, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. This way each cluster describes in terms of data collected, the class to which its members belongs and forms clusters are shown in Figure 3.1.

There are different types of clustering techniques [17] [18] available in the literature. Many clustering algorithms require the specification of the number of clusters to produce in the input data set, prior to execution of the algorithm. A large number of clustering methods [32-36] have been developed in many fields, with different definitions of clusters and similarity metrics. It is well known that no clustering method can sufficiently handle all sorts of cluster structures and properties (e.g. overlapping, shape, size and density). Clustering is an important technology of the data mining study, which can effectively discovered by analyzing the data and useful information. It groups data objects into several classes or clusters so that in the same cluster of high similarity among objects, and objects are vary widely in the different cluster [37].

Clustering algorithms have been developed and used in many fields. Hierarchical and partitioning methods are two major categories of clustering algorithms. An extensive survey of various clustering techniques are explained in [38], [39]. In this section, we highlight work done on image clustering. Many clustering techniques have been applied to clustering documents. The survey is provided on applying hierarchical clustering algorithms [40] into clustering documents. Adapted various partition-based clustering algorithms to clustering documents. Another popular approach in image clustering is agglomerative hierarchical clustering [41]. Algorithms in this family follow a similar template: Compute the similarity between all pairs of clusters and then merge the most similar pair.

Different agglomerative algorithms may employ different similarity measuring schemes. K-means and its variants [42] represent the category of partitioning clustering algorithms. One of the variants, bisecting k-means [42], performs basic k-means as well as the agglomerative approach in terms of accuracy and efficiency. The bisecting k-means algorithm first selects a cluster to split. Then it utilizes basic k-means to form two sub-clusters and repeats until the desired number of clusters is reached. The K-means algorithm is simple, so it is widely used in image clustering. Due to the randomness of the initial center selection in the K means algorithm, the results of its operation are stable. In our approach K-means clustering is used for finding the cluster centers.

## 3.2 Introduction to genetic algorithms

A genetic algorithm (GA) is a procedure used to find approximate solutions to search problems based on the evolutionary ideas of natural selection and genetic. Genetic algorithms use biologically inspired techniques such as genetic inheritance, natural selection, mutation, and sexual reproduction (recombination, or crossover). Along with Genetic Programming (GP), they are one of the main classes of Genetic and Evolutionary Computation (GEC) methodologies. GAs were first introduced by Charles darwin-1859 (Origin of the species) and John Holland 1975 (Artificial Survival of the Fittest).

Huge techniques and algorithms have been developed to produce optimum development strategies. These GA procedures quickly converge to optimal solutions after examining only a small fraction of the searchspace and have been successfully applied to complex engineering optimisation problems.

Genetic Algorithms are implemented using computer simulations in which the best ways to solve the problem are specified. To the problem specified there may exist different candidate solutions in the solution space called Individuals, and they are represented using abstract representations called chromosomes. These set of individuals collectively known as Population. The GA consists of an iterative process that evolves population toward an objective function, or fitness function. A fitness function is used to evaluate individuals, and reproductive success varies with fitness. Traditionally, solutions are represented using fixed length strings, especially binary strings, but alternative encodings have been developed.

It starts from a population of individuals randomly generated according to some probability distribution, usually uniform and updates this population in steps called generations. Each generation, multiple individuals are randomly selected from the current population based upon some application of fitness, bred using crossover, and modified through mutation to form a new population. The Algorithms

- Generate an initial population  $M(0)$  randomly.
- Determine the fitness of the population by applying fitness function  $u(m)$  to each individual  $m$  in the current population  $M(t)$ .
- Reproduce the population using the fittest parent of the last generation by Define selection probabilities  $p(m)$  for each individual  $m$  in  $M(t)$  so that  $p(m)$  is proportional to  $u(m)$ .
- Determine the crossover point, this can also be random.
- Determine if mutation occurs.
- Generate  $M(t+1)$  by probabilistically selecting individuals from  $M(t)$  to produce offspring via genetic operators.
- Repeat step 2 until satisfying solution is obtained.

The paradigm of GAs described above is usually the one applied to solving most of the problems presented to GAs. Though it might not find the best solution. more often than not, it would come up with a partially optimal solution[30].

### 3.3 Crossover and Mutation functions

Genetic operators such as reproduction, crossover and mutation modify individuals within a population in such a way, so as to produce new individuals to behave more efficiently. Reproduction of new individuals from existing population is based on weighted probability. During reproduction some will survive, some will reproduce and some will die. Whenever you apply a crossover operator between two individuals, a random crossover point is selected and a pair of new solutions is produced. Whenever you apply a mutation operator to the individual of a population, there is a small percentage of the population changes are made.

Crossover probability is a probability measure in which the next generation is produced by applying crossover operation. If crossover is 100%, changes are made to the generation to get the more generalized generation. If crossover is zero, there is no change in the produced generation.

Mutation probability is a probability measure in which some of the random elements from the individual are changed into something else. If mutation is 100%, then there may be a chance of destroying existing behavior of individual. So we choose the mutation probability as much as less to protect the existing behavior and as well as to get the evolution in behavior.

### 3.4 Summary

In this chapter, clustering algorithm types and genetic algorithms are briefly discussed. The basis for this method lies in the significant change occurring in a small number of color features, in the neighbourhood of a clustering algorithm. The technique is robust to preprocessing image database using k-means clustering and filters of genetic algorithm. The objective function used in our approach is crossover and mutation over transition probability 0.95 and 0.01 respectively. Mean square error helps to find the local minimum for optimized solution. Also, modification to the

existing clustering algorithm these modifications are, namely preprocessing image database using k-means clustering algorithm. This chapter also explains the different clustering techniques and use of clustering algorithm. The genetic algorithm objective function are explained briefly. It was also observed that such a combination of multiple features improves the accuracy of image retrieval. The use of genetic algorithms helps to find optimized best cluster centers.

## Chapter 4

# Preprocessing Image Database Using K-Means Clustering and Genetic Algorithms

In previous CBIR systems, retrieving similar images related to a query are obtained or improved entirely by the relevance feedback learning techniques. But in our proposed approach we are concentrating on preprocessing image database so that it helps in obtaining more number of similar images.

The selection of feature set should be in such a way that it should approximate images as close as possible in a feature space. Feature extraction is the main task in content based image retrieval. Feature extraction is the process of describing the image by considering parameters known as features (color, edge, texture etc) from a given image. In our approach, we implemented multiple feature extraction by using combination of three features such as RGB color space, Edge information and Histogram bins. These features collectively form 136 dimensional features vector for each image in the database. Feature vector formation is described as follows.

Database in our approach is having 1000 images that are Wang dataset with 10 classes each class having 100 images. The feature vector for all these images are 1000\*136 dimensional feature vector. RGB mean and variance are the two features selected in RGB color space. Mean is the average of all Red, Green and Blue pixel information in each image, it tells us how the colors are distributed in color space. Variance of red, green and blue pixel is calculated. These two features in color space collectively forms Red-mean, Green-mean, Blue-mean, Red-variance, Green-variance and Blue-variance. So, each image is represented in color space as six features in feature vector. The second feature we used is edge information. In this edge features, the edge density and boolean edge density are calculated. Sobel operator finds the gradient(change) in intensity at each point in the image. Based on this intensity change towards horizontally or vertically we can move around the image edge. Sobel operator exits for x-order and y-order derivatives and also for mixed partial derivatives. Edge density of each image is calculated by using mean of all the edges which are identified using Sobel operator. First, the RGB image is converted into black and white image, and then Sobel operator finds the gradient in intensity between pixels. Where the intensity is maximum

considered it as the edge, find the mean of all the edges which are found by Sobel operator. It gives the edge density of one image; similarly we have to find out edge density for all images in the database. Boolean edge density, once the image is converted into black and white image fixes the white pixel as (1) and black pixel as (0). Then, consider the mean of all white pixel value; it gives the boolean edge density. Similarly for each image in the database we have to find out boolean edge density. These features collectively form two features in feature vector. The third feature we used is histogram information. Each gray scale image is having 0 to 255 bins(colors), from these 256 colors every time two colors information is considered as one level and similarly it forms 128 levels for each image. We will get 128 features in feature vector for each image. Finally, by using these three features we will get six feature values from RGB color space, two feature values from edge information and 128 feature values from histogram information. These feature values will collectively forms 136 dimensional feature vector for every image in database.

Clustering technique (kmeans) is applied on the image database to cluster them so that the degree of association to be strong between members of the same cluster and weak between members of different clusters. Genetic Algorithms(GA) were used to find the best cluster centers. Mutation and crossover functions were taken as the objective functions for the genetic algorithm. Using genetic algorithm we identified 15 best cluster centers by giving 500 iterations for the objective functions. Each image was assigned to the nearest cluster center using euclidean distance as the similarity measure. By doing this 1000 images were segregated into 15 clusters. In the first iteration, similarity between the query image and each cluster center were compared. Based on the matching score , the query image was entered into particular cluster center among 15 cluster centers which is nearest. In the second iteration similarity measure was calculated between the query image and the each image in the selected cluster. Based on the matching score, the most relevant similar images were retrieved.(i.e. retrieve images which are closer to the query image using distance or similarity measure).The proposed approach of preprocessing image database is explained and it is illustrated in Figure 4.1.

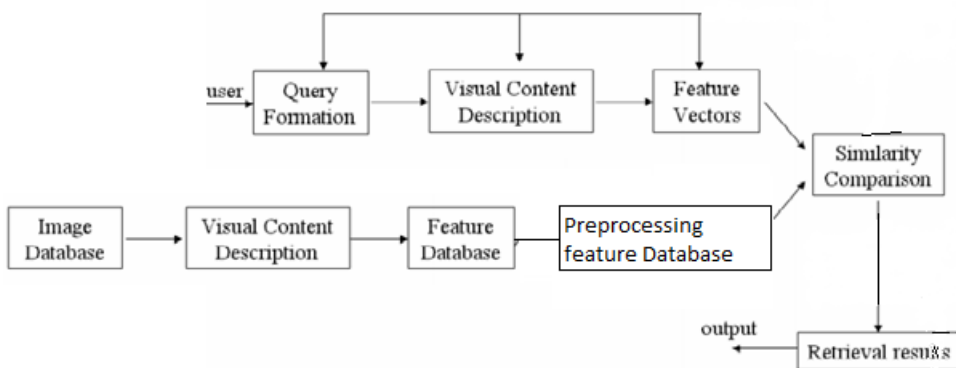


Figure 4.1: CBIR with augmented preprocessing stage

K is the number of clusters. We have taken mutation and cross over probabilities as .01 and

.95 respectively and set the number of generations as 500.

$$\min_{m_k, k=1,2,\dots,K} Cluster\_Dispersion \quad (4.1)$$

- within genetic algorithm
  1. Determine cluster centers.
  2. Partition the labeled data by distance to closest cluster center.
  3. Find non-empty clusters; assign a label to non-empty clusters by majority class vote within them.
  4. Compute dispersion

Mean square error (MSE) is used for evaluating cluster dispersion.

Cluster dispersion is a measurement to find the rate of expansion of a cluster. Mean square error is used as a objective function to know how the data is distributed in every cluster group. If the mean square error is low then that cluster purity is good otherwise cluster purity is low.

$$MSE = \frac{1}{N} \sum_{k=1}^K \sum_{x \in C_k} ||x - m_k||^2, \quad (4.2)$$

N: Total number of images in each cluster group.

K: Total number of cluster centers that is 15.

$x \in c_k$  : x is a element in kth cluster group.

$m_k$ : Cluster center.

Once the random clusters are generated from k-means algorithm, these random cluster centers are given input to the genetic algorithm. Along with these random clusters the number of generations that is 500, crossover and mutation probability, mean square error parameters uses the genetic algorithm and produces best cluster centers. Genetic algorithm helps efficiently in database preprocessing.

## 4.1 Results and Discussions

In this section we present the results obtained for traditional and proposed CBIR systems. In proposed CBIR system, accurate results were obtained by preprocessing the image database with clustering technique and by using multiple features.

### Test 1: Previous approach

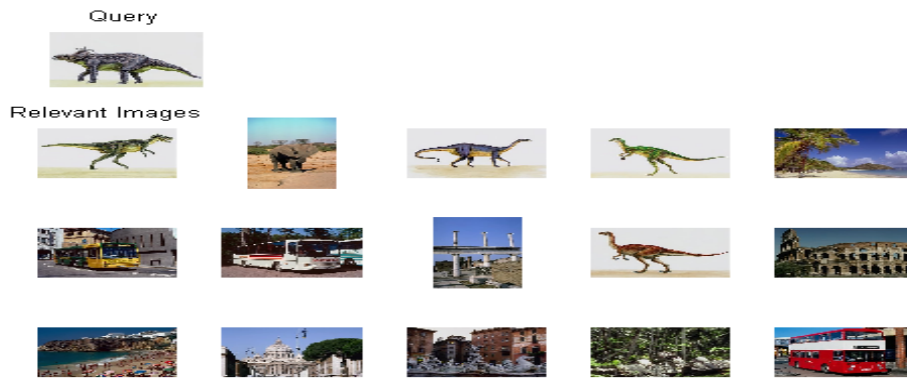


Figure 4.2: dinosaurs retrieval

### Proposed approach :

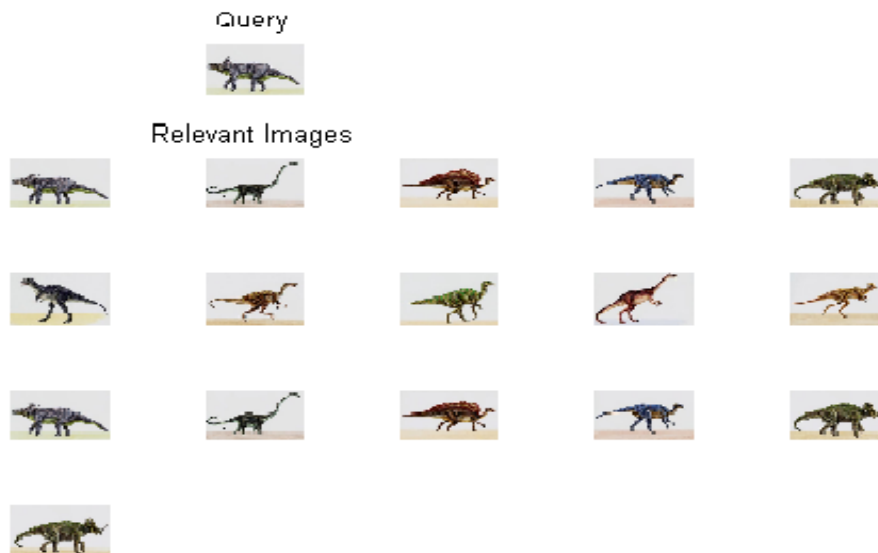


Figure 4.3: dinosaurs retrieval

In Figure 4.2, out of the top 15 images retrieved from the database, only 2 images are relevant. In Figure 4.3, by using proposed approach, number of positive images retrieved was increased from 2 to 5.



## Test 2: Previous approach

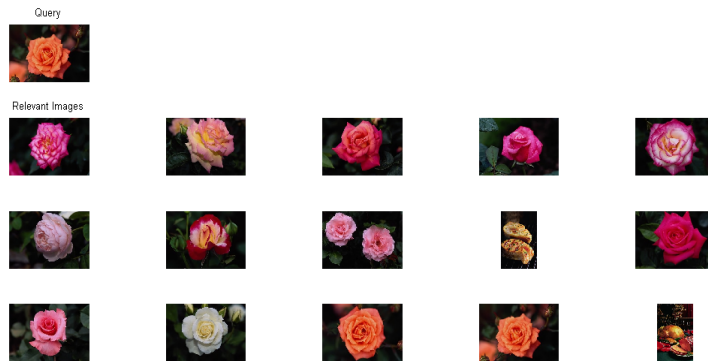


Figure 4.4: flower(s) retrieval

In Figure 4.4, the flower image is query from the database. Out of the top 15 images retrieved from the database, only 4 image are relevant.

## Proposed approach

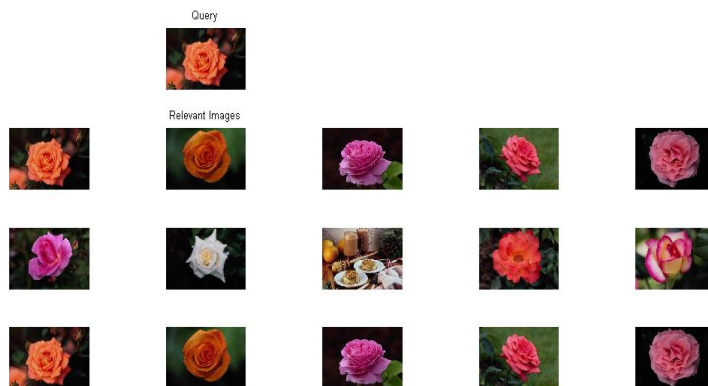


Figure 4.5: flower(s) retrieval

In Figure 4.5, the flower image as query from database. Out of the top 15 images retrieved from the database, only 7 images as relevant. By using proposed approach, number of positive images retrieved was increased from 7 to 10.

### Test 3: Previous approach

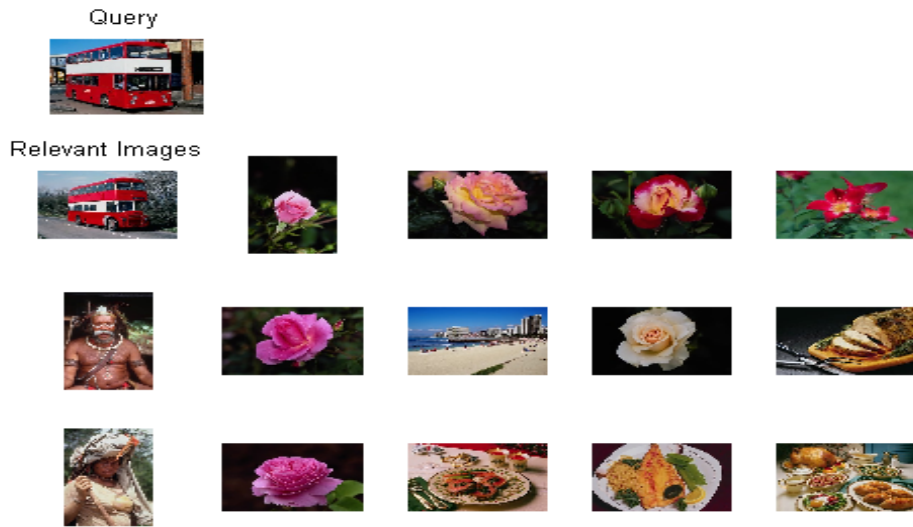


Figure 4.6: Bus(es) retrieval

In Figure 4.6, the bus image as query from the database. Out of the top 15 images retrieved from the database, only 1 image as relevant.

### Proposed approach

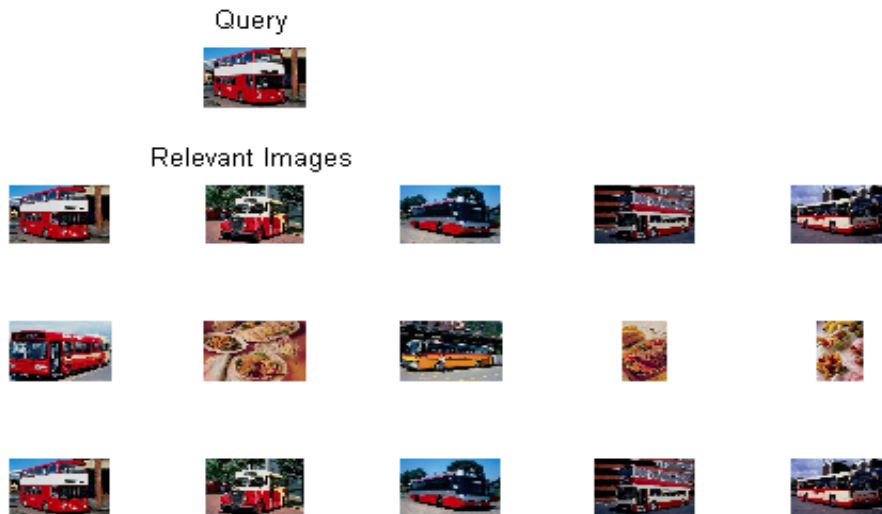


Figure 4.7: Bus(es) retrieval

In Figure 4.7, the bus image as query from the database. Out of the top 15 images retrieved from the database, only 4 images are relevant. By using proposed approach the number of positive images retrieved was increased to 3.

### Test 4: Previous approach

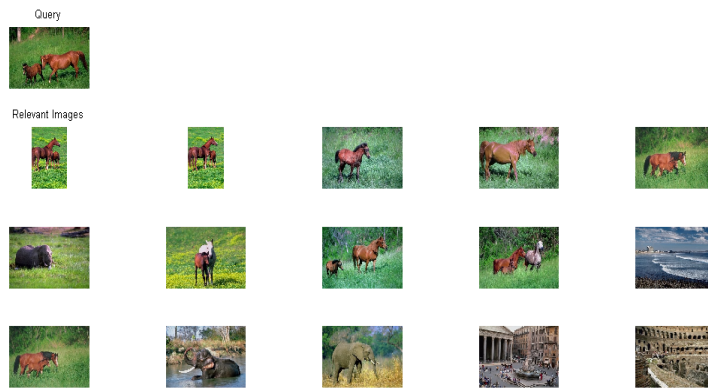


Figure 4.8: Horse(s) retrieval

In Figure 4.8, the horse image as query from the database. Out of the top 15 images retrieved from the database, only 8 image as relevant.

### Proposed approach

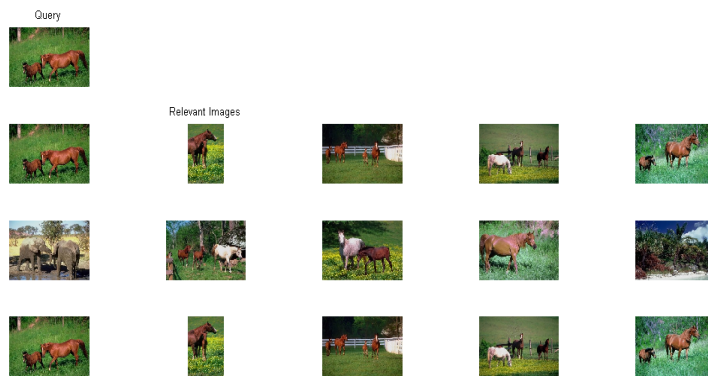


Figure 4.9: Horse(s) retrieval

In Figure 4.9, the horse image as query from the database. Out of the top 15 images retrieved from the database, 12 images are relevant. By using proposed approach, number of positive images retrieved was increased to 4.

## 4.2 Performance Metrics

In this section the proposed system results are analysed by using confusion matrix, precision graph and F-measures. These are defined and described respectively. The clustering purity is presented in confusion matrix by taking 15 observed and 10 actual classes. An image dataset of size 1000 (which is a Wang image database) was taken and it was categorized into 10 different classes, each of 100 images.

The following are the different class labels used in image database.

1. Human(s)
2. Seashore with sky embedded
3. Building(s) with sky
4. Bus(es)
5. Dinosaurs
6. Elephant(s)
7. Flower(s)
8. Horse(s) with greenery
9. Mountains with sky
10. Vegetables of different color.

For each image in the database features were extracted. Different features such as average and variants in RGB color space, Edge density and Boolean edge density and histogram pixel information were taken which collectively form a 136 dimensional feature vector.

### 4.2.1 Confusion matrix

We generated confusion matrix for the 10 actual clusters and 15 observed clusters. Cluster purity is one of the ways of measuring the quality of a clustering solution. The purity is higher it shows that it is the better solution. The purity of each cluster is as shown in Table 4.1.

From the Table 4.1 we observe that humans and horses fall into the same clusters and the elephants and horses are also fall into the same cluster this is because of their color similarity, structural similarity respectively.

Table 4.1: Confusion Matrix For Calculating Clustering Purity

Actual/observed classes	1	2	3	4	5	6	7	8	9	10	purity
1	0	0	0	0	11	0	0	0	0	0	100%
2	0	0	0	0	47	0	0	0	0	0	100%
3	3	47	3	1	0	5	2	2	10	2	62.66%
4	0	0	0	0	24	0	0	0	0	0	100%
5	8	0	4	4	0	0	2	1	0	10	34.48%
6	46	25	13	8	0	4	5	43	14	20	25.8%
7	7	10	24	3	0	56	1	45	15	11	32.55%
8	2	7	4	1	0	0	0	32	9	9	47.05%
9	0	0	0	0	0	0	15	0	0	0	100%
10	0	0	19	0	0	0	0	0	0	0	100%
11	30	7	12	77	0	11	2	0	30	30	38.69%
12	0	0	0	0	0	0	11	0	1	0	91.16%
13	3	4	3	2	1	24	0	1	18	10	36.36%
14	1	0	8	4	0	0	31	0	3	13	51.16%
15	0	0	0	0	16	1	0	0	0	0	94.11%

In the above table each observed clusters of different sizes had different purities. More than half of the clusters has above 90 percent purity.

## 4.2.2 Precision Graph

**PRECISION:** The precision of the image retrieval is calculated as follows:

$$Precision = \frac{\text{Number of relevant images}}{\text{Total number of retrieved images}} \quad (4.3)$$

Table 4.2 shows precision of each class for previous and proposed approaches separately.

Table 4.2: Accuracy calculation for previous and proposed approaches

Class Number	Query Image Number	Previous Approach	Proposed Approach
1	15	0.32	1
2	115	0.31	0.84
3	215	0.1	0.8
4	315	0.4	0.3
5	415	0.4	1
6	515	0.34	0.7
7	615	0.6	0.58
8	715	0.72	0.98
9	815	0.2	0.64
10	915	0.24	0.58
Average	1000	3.63	7.42

Figure 4.10, shows precision of each class separately. *Blue* bars indicate the precision of proposed approach and *red* bars indicate the precision of previous approach.

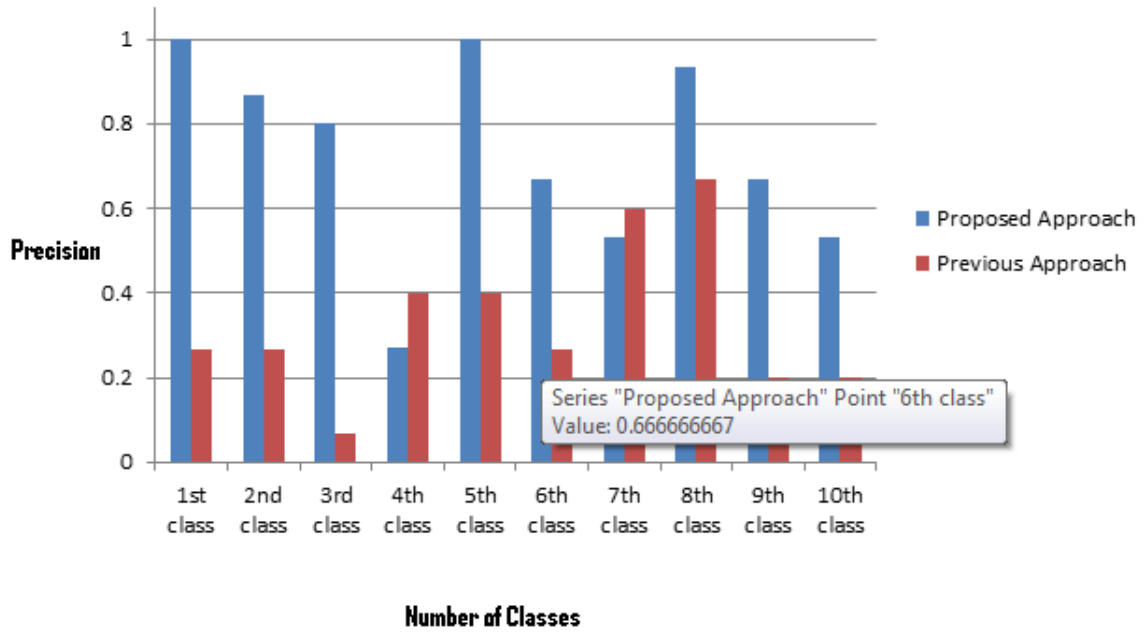


Figure 4.10: Precision Graph

### 4.2.3 F-Measures for Previous Approach

F-measure combines the precision ideas from the image retrieval literature. The higher the overall F-measure, the better the higher accuracy of the resulting image retrieval to the original classes. Table 4.3 shows 39.8% accuracy.

Table 4.3: F-Measure of Previous approach

Class Number	Performance
1	45.01%
2	31.20%
3	20.03%
4	40.03%
5	45.05%
6	34.02%
7	60.01%
8	72.01%
9	32.04%
10	24.02%
Avarage	39.8%

Figure 4.11. shows F-Measure of each class separately. Blue bars indicate the performance of previous approach. The average performance we got in this approach is 39.8%.

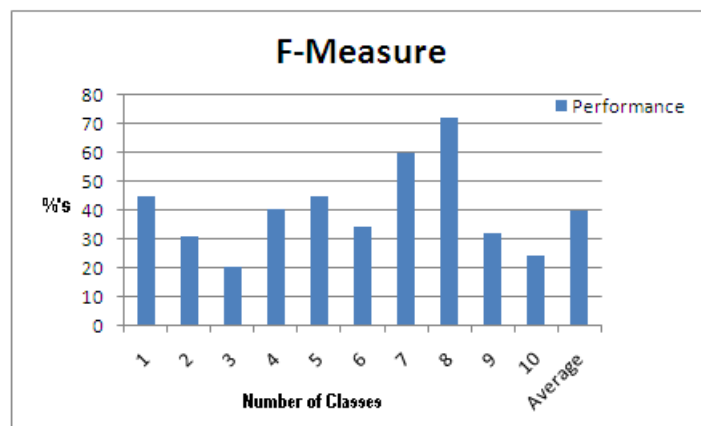


Figure 4.11: F-Measure for Previous Approach

#### 4.2.4 F-Measure for Proposed Approach

F-measure combines the precision ideas from the image retrieval literature. The higher the overall F-measure, the better the higher accuracy of the resulting image retrieval to the original classes. Table 4.4 shows 72.2% accuracy.

Table 4.4: F-Measure of proposed approach

Class Number	Performance
1	100%
2	84.02%
3	80.00%
4	30.00%
5	100%
6	70.00%
7	58.04%
8	98.03%
9	64.04%
10	58.03%
Avarage	74.20%

Figure 4.12 shows F-Measure of each class separately. Blue bars indicate the performance of proposed approach this is with clustering algorithm and multiple features. The avarage performance we got in this approach is 74.2%. These results are encouraging.

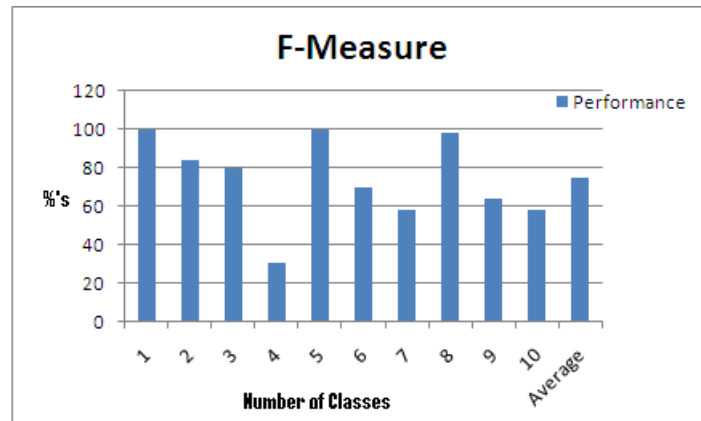


Figure 4.12: F-Measure for Proposed Approach



### 4.3 Summary

In this chapter, the k-means clustering algorithm uses combination of multiple features for preprocessing image database. The different features such as color, edge density and boolean edge density and histogram pixel information were taken which collectively form a 136 dimensional feature vector. Color component consists of average and variance in RGB color space and taken a histogram bins. Edges are identified using sobel edge detector and from that obtain edge density and boolean edge density. These feature vector of 1000 images are given input to the k-means clustering algorithm. The genetic algorithm objective function helps to find the best cluster centers based on the minimum mean square error. The best 15 cluster centers are calculated after 500 iterations. Once the preprocessing image database was finished based on the euclidean measure the most relevant images are shown in results. The performance metric used to represent clustering purity is confusion matrix. This matrix shows the clustering purity is more than half of the clusters has been above 90 percent purity. The comparison of results for both previous and proposed approaches are shown using sample retrieval. In proposed CBIR system, accurate results were obtained by preprocessing the image database with clustering technique and by using combination of multiple features. The precision graph shows the accuracy of each class separately. The analysis of results are presented in this chapter using F-measures. These figures shows the accurate result of content based image retrieval by preprocessing image database.

# Chapter 5

## Summary And Conclusions

### 5.1 Contributions of the work

A novel method for unsupervised learning has been introduced in this work. The basic idea is to take clustering method and simultaneously optimize the mean square error of the resulting clusters. The objective function is a cluster dispersion measure.

After analysing the results obtained we conclude that preprocessing the image database improves the accuracy of CBIR system. In our approach, we preprocessed image database by k-means clustering with the objective functions of genetic algorithms to find the best cluster centers. Here the objective functions are mutation and crossover. Genetic algorithms help us to find an optimized solution. Once the preprocessing stage is completed, all the images in the database are formed into different clusters. Now, based on the euclidean distance between each cluster center and the query image feature vector the most relevant cluster center is decided. Then based on the similarity measure between selected cluster center and images in that selected cluster center the most relevant similar images will be retrieved.

An image database of size 1000 images which is a Wang image database [58] and it was categorized into 10 different classes of each 100 images. For 1000 images 15 best cluster centers were identified. Initially we conducted experiment with 21 cluster centers for finding the cluster purity. Same procedure was repeated by reducing the number of cluster centers till we found the best cluster purity. Finally, cluster center number at 15 we got high cluster purity. By this approach for at least more than half of the cases we were getting more relevant similar images from the database. We compared the results of traditional CBIR system approach with our proposed CBIR system and the results are encouraging.

### 5.2 Directions for further research

In this work we are using only k means clustering to preprocess the image database. We can improve the accuracy of CBIR system using database preprocessing method by considering multiple clustering algorithms like Fuzzy C-means and Hard Fuzzy C-means. The performance of the system can be improved by adding different similarity measures in the preprocessing stage. We can make our approach as a semi-supervised learning by applying labels to the subset of the dataset. Then,

the objective function becomes a linear combination of a measure of cluster dispersion and a measure of cluster impurity. We can also retrieve images from different clusters rather than from a single cluster which will in turn improve retrieval performance.

In our approach, we only used dataset of size 1000 images, we can try our system performance by considering huge dataset of size like 10,000 images. We can improve the accuracy of CBIR using more features like considering Scale Invariant Feature Transforms(SIFT) features and multiple that is combinations of high level and low level features. Similarity measure is a key point for the CBIR. So, we can make our approach more accurate by choosing different similarity measures like Mahalanobis distance and Cosine similarity. We can also make efficient CBIR system by adding relevance feedback approach to this preprocessing image database for better image retrieval. Relevance feedback helps user to improve the results after the number of times of user relevances are used. The accuracy and performance of the CBIR is mainly based on considering the appropriate feature set, efficient algorithm for database preprocessing and efficient similarity measures.

# References

- [1] Remco C. Veltkamp, Mirela Tanase Department of Computing Science, Utrecht University. *Content-Based Image Retrieval Systems: A Survey*, October 28, 2002.
- [2] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger. *The R-tree: An efficient and robust access method for points and rectangles*, In Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 322-331,1990.
- [3] W Niblack, R Barber, W Flickner Equitz, M Glasman, E Petkovic, D Yanker, P Faloutsos, C Taubin, G *The QBIC Project: Querying images by content using color texture and shape*,Proc. SPIE Int. Soc. Opt. Eng., in Storage and Retrieval for Image and Video Databases,vol. 1908,pp 173-187,San Jose, 1993.
- [4] Nallaperumal, K. Sheerin Banu, M. Callins Christiyana, C. M.S. Univ., Tirunelveli *Content Based Image Indexing and Retrieval Using Color Descriptor in Wavelet Domain*,International Conference IEEE Transactions Vol. 3, No.1, pp.185-189, 07 January, 2008.
- [5] Aibing Rao, Rohini K. Srihari, Zhongfei Zhang *Spatial Color Histograms for Content-Based Image Retrieval*, Center of Excellence for Document Analysis and Recognition State University of New York At Buffalo, 1998.
- [6] P. Brodatz, *Textures: A photographic album for artists & designers*, Dover, NY, 1966.
- [7] suematsu, N., Ishida, Y., Hayashi, A., Kanbara, T., *Region based image retrieval using wavelet transform*, In: Proc. 15th International Conf. on Vision Interface May 27- 29, Calgary, Canada, pp. 9-16, 2002.
- [8] Dr. Fuhui Long, Dr. Hongjiang Zhang and Prof. David Dagan Feng *Fundamentals of content based image retrieval: A Survey*, 2002.
- [9] H. Tamura, S. Mori, and T. Yamawaki. *Texture features corresponding to visual perception.*, IEEE Transactions on Systems, Man and Cybernetics, 8(6):460473, 1978.
- [10] Remco C. Veltkamp and Michiel Hagedoorn, *State-of-the-art in shape matching*, In Michael Lew, editor, Principles of Visual Information Retrieval, pages 87119. Springer, 2001. ISBN 1-85233-381-2.
- [11] Etienne Louprias and Nicu Sebe, *Wavelet-based salient points: Applications to image retrieval using color and texture features*, In Advances in Visual Information Systems, Proceedings of

the 4th International Conference, VISUAL 2000, Lecture Notes in Computer Science 1929, pages 223-232. Springer, 2000.

- [12] F. Mokhtarian, S. Abbasi, and J. Kittler *Efficient and robust retrieval by shape content through curvature scale space*, In Smeulders and Jain 73, pages 35-42, 2003.
- [13] Guttman, A *R-tree: A dynamic index structure for spatial searching*, ACM SIGMOD Int. Conf. Management of Data, Boston, MA, pp. 475-484, 1984.
- [14] N. Beckmann, H.-P. Kriegel, R. S. and Seeger, B *The R\*-tree: An efficient and robust access method for points and rectangles*, ACM SIGMOD Intl. Conf. on Management of Data, pp. 322-331, 1990.
- [15] Yianilos, P. N. *Data structures and algorithms for nearest neighbour search in general metric spaces*, SODA: ACM-SIAM Symposium on Discrete Algorithms (A Conference on Theoretical and Experimental Analysis of Discrete Algorithms), 1993
- [16] Chakrabarti, K. and Mehrotra, S *The hybrid tree: An index structure for high dimensional feature spaces*, ICDE, pp. 440-447, 1999.
- [17] <http://mars.csie.ntu.edu.tw/~cychen/olddoc/ClusteringDataMining.html#CURE>.
- [18] [http://en.wikipedia.org/wiki/Cluster\\_analysis](http://en.wikipedia.org/wiki/Cluster_analysis).
- [19] <http://www.bandmservices.com/Clustering/Clustering.htm>.
- [20] Thomas Deselaers<sup>1</sup>, Roberto Paredes<sup>2</sup>, Enrique Vidal<sup>2</sup>, and Hermann Ney<sup>1</sup> *Learning Weighted Distances for Relevance Feedback in Image Retrieval*, Computer Science Department RWTH Aachen University, Spain.
- [21] Asst. Prof. Steven C.H. Hoi, *Relevance Feedback in Interactive Image Retrieval*, Multimedia Search & Mining Group, School of Computer Engineering, Nanyang Technological University Singapore, <Http://www.ntu.edu.sg/home/chhoi> .
- [22] S. Belongie J. M. Hellerstein C. Carson, M. Thomas and J. malik. *Blobworld A system for region-based image indexing and retrieval*, Computer Science Department Proc. Visual Information Systems, 1999.
- [23] N. V. Shirahatti and K. Barnard. *Evaluating image retrieval*. Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2005.
- [24] M. Flickner J. Hafner W. Niblack C. Faloutsos, R. Barber. *Efficient and effective querying by image content*. Proc. IEEE Int. Conf. on Intelligent Information Systems, 1994.
- [25] R. W. Picard A. Pentland and S. Sclaroff. *Photobook: Content-based manipulation for image database*. International Journal on Computer Vision, 1996.
- [26] M. Ortega-Binderberger S. Mahrotra, Y. Rui and T.S. Huang. *Supporting content-based queries over images in mars*. In Proc. IEEE Int'l Conf. Multimedia computing and systems, 1997.
- [27] S. Belongie J. M. Hellerstein C. Carson, M. Thomas and J. malik. *Blobworld: A system for region-based image indexing and retrieval*. Proc. Visual Information Systems, 1999.

- [28] W. Y. Ma and B. S. Manjunath. *Netra: A toolbox for navigating large image databases*. Proc. Eighth ACM Int'l Conf. Multimedia, 2000.
- [29] S. Marchand Maillet P. Clough H. Muller, A. Geissbuhler. *Benchmarking image retrieval applications*. Technical report, Sheffield University funded by the EPSRC, 2004.
- [30] KENNETH A. DE JONG WILLIAM M. SPEARSDIANA F. GORDON. *Using Genetic Algorithms for Concept Learning* Machine Learning, 13, 161-188 (1993) 1993 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.
- [31] V.Kapoor S.Dey A.P.Khurana. *Empirical Analysis and Random Respectful Recombination of Crossover and Mutation in Genetic Algorithms*.IJCA Special Issue on Evolutionary Computation for Optimization Techniques ECOT, 2010
- [32] A. Jain, M. Murty and P. Flynn. *Data Clustering: A Review*.ACM computing surveys,Vol.31,pp:264-323, 1999.
- [33] R. Xu, *Survey of Clustering Algorithms*,IEEE Transactions on Neural Networks, Vol.16, Issue 3, pp: 645- 678, 2005.
- [34] M. Steinbach, G. Karypis and V. Kumar, *A Comparison of Document Clustering Techniques*, Proceeding of the KDD Workshop on Text Mining, pp: 109-110, 2000.
- [35] R. Duda and P. Hart, *Pattern Classification and scene Analysis*. John Wiley and Sons, 1973.
- [36] J. Hartigan and M. Wong. *A k-means Clustering Algorithm*,Applied Statistics, Vol. 28, pp: 100-108, 1979.
- [37] Lan Li, Wan-chun Wu, Qiao-mei Rong, *Research on Hybrid Clustering Based on Density and Ant Colony Algorithm*,.Second International Workshop on Education Technology and Computer Science, 2010.
- [38] A.K.Jain, M.N. Murty and P.J .Flynn, *Data clustering: a reView*", ACM Computing Surveys, vol. 3I( 3), pp 264-323,1999 .
- [39] X. Rui, *Survey of clustering algorithms*", IEEE Transactions on Neural Networks, vol 16(3), pp. 634-678, 2005.
- [40] P.Willett, *Recent trends in hierarchical document clustering: a critical review*, Information processing and management, vol. 24,pp. 577-97, 1988.
- [41] M. Steinbach, G. Karypis, and V. Kumar, *"A comparison of document clustering techniques"*, KDD Workshop on Text Mining'00, 2000.
- [42] R. C. Dubes and A. K. Jain, *"Algorithms for Clustering Data"*,Prentice Hall College Div, Englewood Cliffs, NJ, March 1998.
- [43] M. Crucianu, M. Ferecatu, N. Boujemaa. *Relevance feedback for image retrieval: a short survey*,Report of the DELOS2 European Network of Excellence (6th Framework Programme), October 10, 2004.

- [44] G. Das and S. Ray. "*A comparison of relevance feedback strategies in CBIR*", IEEE, 2007.
- [45] Y. Rui and T.S. Huang. *A novel relevance feedback techniques in image retrieval*, In: Proc. 7th ACMConf. on Multimedia, pp. 67-70, 1999.
- [46] Rui, Y., Huang, T.S., Mehrotra, S. [Sharad], *Retrieval with relevance feedback in MARS*, In Proc of the IEEE Int'l Conf. on Image Processing, New York, pp.815-818, 1997.
- [47] J. R. Smith, F. S. Chang, *Tools and Techniques for Color Image Retrieval* Symposium on electronic Imaging: Science and Technology-Storage and Retrieval for Image and Video Database IV, pp. 426-237, 1996.
- [48] Temenushka Ignatova, Andreas Heuer, *Model-Driven Development of Content-Based Image Retrieval Systems* Journal of Digital Information Management, Volume 6- February 2008.
- [49] Dr. Fuhui Long, Dr. Hongjiang Zhang and Prof. David Dagan Feng, *Fundamentals of content based image retrieval: Second International Workshop on Education Technology and Computer Science* 2008.
- [50] M. A. Smith and T. Chen, *Image and video indexing and retrieval*, Hand Book on Image Processing, pp. 687-704, 2000.
- [51] Chris H. Q. Ding, Xiaofeng He, *Cluster merging and splitting in hierarchical clustering algorithms*, Conference: IEEE International Conference on Data Mining - ICDM , pp. 139-146, 2002.
- [52] Sergio F. da Silva, Marcos A. Batista, Celia A. Z. Barcelos, *Adaptive Image Retrieval through the use of a Genetic Algorithm*, 19th IEEE Conference on Tools with Artificial Intelligence, pp. 557-564, 2007.
- [53] *Wang Image database* from Copcam wiki source, <http://wang.ist.psu.edu/docs/related/>.