Effectiveness of Content-Based Image Clustering Algorithms

BY

MESFIN SILESHI AMBAYE

Advisor: Professor Björn Gambäck
Co-Advisor: Dr. Solomon Atnafu

A thesis submitted to the School of Graduate Studies of Addis Ababa University
in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science

March 2007

Addis Ababa

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF COMPUTER SCIENCE

Effectiveness of Content-Based Image Clustering Algorithms

BY

MESFIN SILESHI AMBAYE

Name and Signature of the Board of members of the examiners board

1. _____
2. _____
3. _____

Acknowledgement

First of all, I am deeply indebted to my supervisor Prof. Björn Gambäck for his continuous, unreserved support and comments he has been giving me, above all for his stimulating suggestions and encouragement that helped me in all the time of research for and writing of this thesis. Prof. Björn Gambäck taught me how to ask questions and express my ideas. He showed me different ways to approach a research problem and the need to be persistent to accomplish any goal, irrespective of the distance that is banned between us.

Special thanks goes to my co-advisor, Dr. Solomon Atnafu, who is most responsible for helping me complete the writing of this thesis as well as the continuous follow up he made to me.

I would like to thank my beloved friend; Ato Kibur Lisanu, currently a PHD student, for his brilliant and brotherly suggestions and comments, without whom, I may not even come into success. I want to thank Ato Samuel Eyassu and Ato Lemma Nigussie for their help, support, interest and valuable hints.

I would like to express my gratitude to all those who gave me the possibility to complete this thesis.

List of Tables

Abbreviations

CBIR: Content-Based Image Retrieval

CBSD: Contour-Based Shape Descriptor

CLD: Color Layout Descriptor

CSD: Color Structure Descriptor

DCD: Dominant Color Descriptor

EHD: Edge Histogram Descriptor

HACM: Hierarchical Agglomerative Clustering Methods

MPEG: Motion Picture Expert Groups

RSD: Region-Based Shape Descriptor

SCD: Scalable Color Descriptor

TBD: Texture Browsing Descriptor

Abstract

Retrieval of a set of similar image documents requires clustering the images based on their similar features. Clustered images are utilized by Content-Based Image Retrieval (CBIR) and querying system that requires effective query matching in large image databases. Content-based image clustering provides a more efficient method of management and retrieval of large number of images documents. The Content-based image clustering facilitates users to browse through only a particular subset of related image documents in an efficient manner.

This study focus in validating the two commonly image clustering algorithms namely: hierarchical and k-means. The validation is based on a set of selected MPEG-7 image feature descriptors. The similarity measure input to these clustering algorithms considers both quantitative and predicate-based similarity measures. We computed two similarity measures total color-based similarity matrix as a weighted sum of the MPEG-7 color descriptors and total similarity matrix as a weighted sum of color, texture and shape features.

The proposed metric to measure the effectiveness of clustering subsets of COREL color photo images is with respect to their semantic meaning. Shannon's information theory is selected in the measuring the image's cluster cohesiveness. The clusters formed are said to be well separated when the distinct clusters formed are associated to a specific image semantic. The separation among clusters becomes better when the semantic association of images to a cluster is predictable. The intra-cluster cohesiveness is also captured by the Shannon's entropy measure in measuring the clusters separation.

The best quality clusters are formed by the hierarchical method that uses the average-linkage method when the same total color similarity matrix is input to all clustering algorithms. Experimental result shows that the quality of clusters formed by k-means clustering is not better than any of the three hierarchical methods. Hierarchical method which uses average-linkage produced quality of clusters three times better as compared to k-means. Even though weighted texture and shape similarity measures were used in addition to total color the average HACM is the best method compared to both the k-means in the formation of both

semantic and cluster cohesive clusters. The other different result obtained is that the addition of texture and shape feature degrades cluster quality for all hierarchical methods.

1. Introduction

1.1. Problem Statement

There is a gap between the high-level concept and the low-level visual features of images, which is called the semantic gap. A lot of research on CBIR systems has been aimed at reducing this semantic gap. The problem of the semantic gap is caused by the difference between visual similarity and semantic similarity. Keywords describe images semantically better than low-level visual features. The visual similarity computed based on low-level image features doesn't coincide with the semantic similarity [21]. The semantics of an image describes the high-level concepts of the image.

An image retrieval system requires the user to provide a query image so that to retrieve similar images from the image database [9]. To compare the similarity, specific features of the query image are extracted to be compared with the pre-computed features of the images in the database. Searching for a similar image is performed by comparing the features of the query image with features of each image in the database. This exhaustive search of similar images makes the retrieval system not to be scalable for larger number of images in the database. The total time required to compare the query image with all images in the database is longer and unacceptable for users even if it takes short time to compare pairs of images.

The conventional CBIR systems return a number of semantically dissimilar images at higher ranks. The rank assigned by such retrieval systems is often not acceptable by users as it doesn't satisfy users' interest. The need of re-ranking the first retrieved set of images is stated by [26, 5, and 33] to minimize the degree of retrieving irrelevant target images being ranked at the top. Relevant images may be ranked at the bottom from the sequence of returned images in decreasing order of their similarity.

CLUE [5], a cluster-based image retrieval system, clusters images by selecting a collection of neighboring target images for a given query image. CLUE proposed the need of further study which enables choosing image clustering that forms better cluster quality. CLUE considers clustering to be a graph partitioning problem and it doesn't select a clustering algorithm among many which may produce high quality clusters. The study also clearly

states that the limitation of CLUE as the lack of cluster centers to semantically represent other members which signifies the poor quality of the clusters. No previous research considers the quality of image clusters formed by different clustering algorithms beyond using different feature selection mechanisms and measure of similarity.

One of the problems that deserve solution is whether the identified cluster results using different methods is an agreement with the prior confirmatory knowledge of the problem [15]. The priori knowledge first refers to the expected relation between specific images so that such items are members of a group, which is the semantic similarity among members. Effective clustering methods is required to confirm this knowledge when its overall result allows to assign semantically similar images to the same cluster rather than considering them as members of two different clusters. The expected compactness of such images that are assigned to a cluster is the other knowledge that enables to validate cluster results of a variety of clustering methods. A clustering method implemented on a large image dataset is required to form clusters in which its density conforms to the expected priori knowledge about total members of image categories.

The other cluster validity analysis in measuring effectiveness of content-based image clustering is if the partitions between the identified image clusters consider their semantic content. In validating or judging the quality of clusters, several validity indices are suggested and compared in clustering data sets [16, 5]. CLUE also suggested Purity and Entropy indices in measuring the quality of image clustering.

The general purpose of image clustering is grouping of semantically similar images in a cluster [5, 33 and 26]. It is desired to categorize semantically similar images in clusters by measuring the distance between images using their corresponding low-level feature descriptors that are directly derived from the image. It is the clustering method used that primarily determines the content or members to be assigned to a cluster in a case where the metric used to measure distance is specifically defined. The image clusters are formed based on the feature descriptors, not based on their semantic annotations. The information used in clustering the images doesn’t include the semantic meaning images. The similarity distance matrix that is input to the clustering algorithms cannot be directly mapped to their semantic

similarity. Content-based image clustering utilizes low-level features with the expectation that semantically similar images to be grouped together.

A good clustering algorithm is supposed to produce quality clusters [5, 16]. Such an algorithm is aimed at achieving maximum density with minimum diversity within a cluster and maximum separation between clusters.

This research attempts to investigate the correlation of the quality of image clusters with the expected outcome of hierarchical and k-means clustering algorithms. These are the commonly used algorithms in clustering images that are expected to group semantically similar images in a cluster. In other ways each of them are required not to group images that are semantically different in the same cluster. The image category information are to be used in testing the proposed metric to measure the effectiveness of the hierarchical and k-means clustering algorithms in terms of the images' semantic meaning in clusters.

A technique is required to verify and validate the correctness of the output of the algorithms used in clustering images based on their content. Application of different clustering methods to image data set and assessing the validity of clusters is an open question that requires cluster analysis study.

1.2. Background

Humans tend to use high-level descriptions of image content. However, the features extracted by the computer vision are mostly low-level. A low-level feature doesn't have a direct link to a high-level concept. CBIR requires the content of the object to be analyzed and related to low-level features. Visual information retrieval and search are broadly categorized into two types: feature-based and text-based. Textual description is a high-level image description that cannot adequately describe the rich contents of an image. The subjectivity of humans in perceiving images makes the text-based image description and retrieval approach difficult and labor intensive. On the other hand, the pixel-level

representation of images by computers is not computationally suitable in measuring similarity of images. The content-based image representation implements indexed techniques in describing images by their visual content.

The basic motive for designing features of the visual media is the computational inefficiency and lack of effectiveness of pixel-based measure of similarity between two or more digitally represented media objects [6]. The low-level features consist of quantitative descriptions of multimedia signals that are captured and processed using appropriate sensors. Color, texture and shape are the three low-level feature descriptors that describe signals from still images. The signal processing includes the implementation of different transformations for encoding the signal. Low-level features are directly derived from the image under consideration and the extracted properties can be expressed as numbers that are represented as points in a vector space. They are the basic components for processing and manipulation of images.

CBIR systems extract low-level features from a query image and compare its similarity with the corresponding pre-computed features of images in the database. Many CBIR systems are developed aiming at improving the retrieval effectiveness and efficiency of the increasingly large amount of video and image data in digital form. They compare the similarity between a query image and target images in the database based on their extracted low-level features. The retrieval time is the sum of the time to calculate the similarity between the query and every image in the database. The cumulative time required to compare the query image with all images in the database is longer even though the time required to compare a pair of images is shorter. The response time by such systems is unacceptable for users and the increasing large amount of image data makes these systems not to be scalable. The set of images that are retrieved and returned to users constitute mostly semantically unrelated to the query image.

A solution to such scalability problems is to avoid the comparison of the query image exhaustively with all images in the database. Instead all images in the database are clustered based on their similarity and a representative image called cluster center represents all members of the cluster. Hence, a query image needs to be compared with all members of a cluster if and only if the query image has the largest similarity to the corresponding cluster

center. This results in a reduction of the number of similarity comparison required and a more scalable system.

The method divides each image into rectangular regions and computes the local color histogram of each in order to incorporate the spatial information of pixels. This results in the image to be represented by a number of histograms which increase the memory requirement. The similarity among pairs of corresponding region histograms is measured from the sum of the histogram intersection measure of the individual histograms of each block. The method results in better retrieval accuracy provided that the increase in computational time in comparing the global histogram from each of the corresponding local histograms.

Efficient retrieval of similar image documents based on their features requires categorizing and therefore clustering them based on some criteria. Image clustering and categorization are means for high-level description of image content. The main aim of clustering as a data mining process is discovering groups and identifying interesting distributions and hidden patterns in the underlying large data sets. The division of large data sets into groups of similar objects is called clustering and the groups formed are called clusters. The clusters are formed by partitioning the given data sets such that objects belonging to a cluster are more similar to each other and dissimilar to objects in different groups. This leads to the formation of semantic groups from randomly distributed large data sets revealing organization of similar but hidden patterns.

There is no predefined template or class for the kind of relationships to be taken as valid among the members of a cluster to be grouped together as in the case of image classification. Clustering is concerned with identifying hidden data concepts among the group. This makes clustering process as an unsupervised learning process.

The first basic step in clustering is the selection of features about the data sets to be grouped [30]. The feature-based approach represents visual properties of images by numerical feature vectors that are extracted directly and expressed as descriptors. A collection of feature descriptors that characterizes the visual information of images are stored in a database. Reducing the low-level information of an image into a manageable amount of relevant

properties is the goal of the feature extraction process. This reduces the complexity of the feature description process and makes the descriptors robust [13].

The feature descriptors of Audio-Visual content used among different applications and devices have to meet a standard to allow effective and interoperable searching, retrieval and browsing operations. MPEG-7 is a visual standard to measure the similarity in images. Content-based image clustering requires selecting feature descriptors that have different discriminative power and assign appropriate weight in measuring distance. The MPEG-7 visual standard for image specifies three basic primary low-level image feature descriptors: color, texture and shape [21].

The second basic step in the clustering process is the selection of proximity measures that quantify the similarity among the selected feature descriptors [30]. The proximity matrix contains the similarity/dissimilarity between many pairs of images. The measurement of similarity between images is quantified as the distance between them. Hence, the clustering result is directly affected by the similarity measure used for each descriptor.

Effective retrieval of similar images based on their low-level features requires clustering using suitable criteria. The image clusters that are formed using a variety of clustering methods require evaluation to validate the final result. This study aims at measuring the effectiveness of the clustering algorithms by measuring the quality of the clusters.

1.3. Scope and Limitations

The development of imaging technology has resulted in producing billions of images. The images considered in the study don't include multivariate images in which each image represents different variables. Multivariate images can use many physical characteristics like wavelength, polarization, etc. Satellites that are sensing the earth surface remotely and the standard tool in medicine Magnetic Resonance Imaging (MRI) generate routinely huge numbers of images. The increased resolutions of such image sensors represent images not only by large number of pixels, but also by larger number of variables. Multivariate images are not selected as raw data sets due to the limitation of the memory capacity and processing speed of the computers used in the experiment. Therefore, the study considers only color photos produced using a specific photo camera.

The type of data that is used as input to many clustering algorithms in different application scenarios is numerical or categorical data, not images. The algorithms used here are required to identify images by their index from the input similarity matrix and provides clusters of the corresponding index. Computation of image similarity requires long hours even for the small data sets considered. Keeping the actual image together with its numerical representation during the overall processing requires a more powerful and high storage capacity-computing machine.

1.4. Objectives

1.4.1. General Objective

The general objective of this study is to measure the effectiveness of image clustering algorithms in grouping semantically similar images together.

1.4.2. Specific Objectives

This research aims at meeting the following specific objectives:

- Extract low-level image features and representing images by feature descriptors.
- Identify and select a set of appropriate image similarity measures to be used among the selected feature descriptors.
- Develop a better technique to cluster image data sets based on the extracted MPEG-7 image feature descriptors.
- Measure the quality of clusters formed by of hierarchical and k-means clustering algorithms using a quality measure index.
- Draw conclusions and recommend future research areas.

1.5. Methodology

The following methods were employed in achieving the above-stated objectives.

1.5.1. Literature Review

An extensive literature review was carried out to get a deeper understanding of low-level image representation, CBIR, image similarity measures and clustering algorithms. The

vector space model was used to interpret the rich image content as vectors in feature space enabling numerical measures of similarity/dissimilarity.

1.5.2. Dataset Selection

The size of each image in each semantic category is 384 x 256. The image descriptors are stored randomly in a database and no information about the image's semantic category is stored together with the corresponding descriptor. It is impossible to know whether two images belong to the same category or not from the descriptor database of all images. A set of images that convey similar semantic meaning when perceived by humans belongs to the same category. An image is not assigned to be member of more than one semantic category. The collection of eight different image semantic categories used in the experiment is the same as the one used in SIMPLIcity [33].

1.5.3. Experimentation

The experiment was conducted in three phases. First, the low-level feature descriptors of each image data set were extracted. The values of each of the selected MPEG-7 descriptors were extracted using M-Ontomat-Annotizer-v0.52. It provides the extracted descriptors in XML format. The feature vectors of the descriptors were parsed from the XML files and stored in a database.

The quantitative and predicate-based distance measures were selected in measuring similarity. The total similarity measure is performed by combining similarity measures at descriptor level. The final result of this phase is a similarity matrix that describes the distance measure of each image with all others. Finally the clustering result of each of the algorithms was analyzed in measuring the cluster quality.

1.6. Justification of the Study

Image retrieval systems are not successful in utilizing high-level semantics in general purpose images due to the difficulty of recognizing and classifying images accordingly. That is why such systems are dependent on the low-level features to classify matching images based on their similarity measure [9]. Identifying information about certain semantic types of

images improves retrieval by simplifying the level of matching schemes to be performed only to a specific semantic category.

The following are some of the reasons that justify the need of measuring the effectiveness of content-based image clustering algorithms.

- Low-level image description requires representation of the image by feature vectors or image descriptors. Using MPEG-7 standard to define the representation of the description allows interoperable searching, clustering and access of image contents.
- Selection of image feature descriptor requires considering the discriminative power of each descriptor.
- The performance of the conventional CBIR needs to be enhanced when the images in a database are clustered.
- Formation of quality clusters requires considering color, texture and shape features by assigning appropriate weight according to their discriminative power.
- Even if the input to the clustering algorithms is a distance/similarity matrix, quality cluster requires considering semantic similarity among images.

1.7. Organization of the Thesis

This thesis is organized into six chapters. The first chapter describes the area of the research. It also lists the statement of the problem, the objectives, the scope and methods employed in the study. Selected works related to this research are discussed in Chapter two. Chapter three is concerned with the brief discussion and evaluation of the three low-level visual descriptors.

The similarity measure that uses the descriptors and properties of clustering algorithms is covered in Chapter four. In Chapter five, the details of the actual experiment are seen and results of the study are discussed. Lastly, the conclusions drawn and recommendations for future research areas are pointed out in Chapter six.

2. Related Work

Efficient retrieval of similar image documents based on their features requires categorizing and therefore clustering them based on some criteria. Image clustering and categorization is a means for high-level description of image content. The key to the retrieval process is similarity among low-level features. Intensive research has focused on content-based image indexing and retrieval in recent years, with the goal of indexing the image data using certain features derived directly from the data. The major challenging problem in the CBIR systems is the difficulty to understand the meaning of an image by computers and the larger size of the general purpose image database.

2.1. Image Classification

Contrary to clustering, the process of assigning a data item to a predefined set of categories is called classification [33]. Images are grouped into semantically more meaningful image categories based on low-level features that are formed using statistical classification methods. The SemQuery [40], SIMPLIcity[33] and ALIP[41] systems are some examples of retrieval systems that use the statistical classification method.

Retrieval systems assume that sets of images that have similar visual characteristics based on their low-level features are also semantically related [33]. Assigning images into different semantic classes is the main goal of image classification. Retrieval systems are not required to understand images the way human being perceive them, but to use image semantic classification techniques for classifying images based on their semantics to facilitate the retrieval process. The image semantic classification assigns images into semantic classes using the similarity result of their corresponding visual characteristics or features.

The problem of semantic classification in region-based systems is the pictorial properties of regions have no direct relationship to the semantics. For example, a red circle can be mapped to a flower, a ball or the sun. Reversely the browsing and retrieval would be easier if the system would understand image semantics and identify the significant features of the constituent objects of images.

SIMPLIcity [33] classifies CBIR systems into three categories depending on the way the low-level features are extracted, namely histogram, color layout and regionbased searches. The color distribution of an image represented by a color histogram is used in finding the similarity among pairs of image. This global histogram image representation doesn't consider shape and texture information.

The color layout search considers a set of local properties by partitioning the image into blocks and uses the average color of each block in finding similarity. The improvement of the color layout over the histogram is that the former retains shape, location and texture information. Images are represented at object level in region-based retrieval system. An ideal segmentation of an image results in the decomposition of the image into objects that are easily understood by the human visual system. The NeTra [42] and the Blobworld [43] systems are two examples of region-based system.

As a textured image consists of similarly shaped objects, the major focus of such images is not shape but rather color and texture. Shape is a critical perceptual visual feature to easily understand non-textured images. Unlike textured images they usually consist of clumped regions that are portioned.

The chi-square statistics that measures the degree of the goodness of fit among the evenly scattered regions of the image is performed for the classification of textured or non-textured images. An image is categorized to its chi-square value with respect to a predefined threshold.

Graph or photograph classification is performed by partitioning each image into blocks. The probability density analysis of the wavelet coefficient algorithm is implemented to classify each blocks of an image. The two feature values that describe the distribution pattern of the wavelet coefficient in high-frequency bands determine if a block is categorized as photograph or not.

2.2. Image Retrieval by Clustering

A retrieval model is developed by G. Park, Y. Back and H.K. Lee [26] in order to lower the highly ranked dissimilar images retrieved to a query image. The model [26] extracts color,

texture and shape features of the query image in the first phase. The HACM is used in analyzing the initially retrieved images so that the post-retrieval clustering enables to re-rank the clusters. Color histograms that use the HSV color space represent the color feature of the image. The texture feature is represented by the Grey Level Co-occurrence Matrix (GLCM) which represents direction and distance between two grey level values. The energy, entropy correlation, inertia and local homogeneity are the five features selected to utilize the information contained in the GLCM. The quantization of the grey levels is in a way that considers the dimension of the feature vectors to be minimal. The edge direction histogram that doesn't need to segment the image represents the shape feature using 73 bins. The normalization method enables to compensate for difference in the images' size.

The similarity of the shape and color features is calculated from the sum of the absolute difference measure from their corresponding histogram of the image under comparison. The Min/Max normalization method is applied to adjust the different range of similarity values from the color, texture and shape features before calculating the total combined similarity through a weighted sum of each feature.

G. Park, Y. Back and H.K. Lee [26] state that non-hierarchical clustering requires parameter choice to be predetermined. For example, the number of clusters is not flexible to users. The hierarchical cluster analysis method is employed in most information retrieval systems. Unlike most CBIR systems that focus on the improvement of retrieval efficiency, the study by G. Park, Y. Back and H.K. Lee was primarily concerned with improving retrieval effectiveness by applying the Hierarchical Agglomerative Clustering Methods (HACM) technique to a small number of images. The implementation of HACM uses the stored matrix approach and the Lance-William formula is applied to update the similarity matrix. Cluster analysis consisting of small numbers of images is not concerned primarily with retrieval efficiency, but with effectiveness.

Different content-based image retrieval techniques have been published that aim at supporting effective searching and browsing of large image databases. One of these techniques, CLUE [5] generates a cluster of images that is tailored to characteristics of a query image. A set of images that are ranked according to their similarity measure is returned. CLUE aims at improving the performance of CBIR systems by considering the

similarity information among the retrieved images by ranking them before returning the result to the user. It clusters the selected neighboring target images by an unsupervised learning method. Instead of displaying all of the highly ranked top matched target images, CLUE enables CBIR systems to show users a collection of representative images of each clusters formed to users first. Then all target images within the specified cluster are displayed to users.

The image is segmented into blocks with 4x4 pixels. Six features are extracted from each of the regions or blocks. Three of them is the average color that uses the LUV color space. Before any compression among regions is done, the extracted feature of each region is associated with a fuzzy feature that will describe its low level features. The region-level similarity is calculated using the fuzzy similarity measure.

Weighted graph-representation of collection of images is used that emphasizes pair-wise relationships is used in measuring the overall similarity measure. The edges that connect pairs of nodes on the graph, representing an image, are labeled by weights that indicate the similarity between the two nodes. When the data items are represented as graphs, clustering the data items is a graph partitioning problem.

The organization of clusters in CLUE employs the order of binary tree traversal that arranges the leaves linearly. This linear organization of clusters keeps both the hierarchical structure of the tree and the distances of the clusters from the query image. The image that is most similar to all images of the cluster is selected as the representative image of the cluster.

The cluster created by CLUE is based on a response to a query image and it is adapted to its characteristics. It uses a graph-theoretic algorithm to generate a cluster. The clustering is dynamic, as the clusters have to be closely adapted to characteristics of the query image. It is a local and dynamic image classification method. An image clustered in a specific class in terms of similarity measures doesn’t necessarily belong to the class. The partitioning algorithm chosen affects the quality of the clusters. The need of testing other graph-theoretic clustering techniques for possible performance improvement is highlighted.

2.3. Performance Improvement by Clustering

M. Abdel-Mottaleb, S Krishnamachari and N.J Mankovich [1] have shown that the effectiveness of clustering images for scalable and efficient retrieval of the query image. The measure of similarity between pairs of images is represented by their corresponding color histograms. The method used to incorporate the local features of the images in the database is to divide them into a fixed number of rectangular regions. The local variation of color information is captured by the local color histogram that is used in computing the similarity between images. Then the similarity measure between the corresponding rectangular regions is combined to calculate a single measure of similarity between the images. Histogram intersection, L1 and L2 norms are the three similarity measures used.

The retrieval results of the query image with and without clustering are compared using quantitative measures. The retrieval accuracy is a function of the number of n top best matches and the number of images that are returned by retrieval. The three similarity measures are not compared against ground truth information for each of the hierarchical and k-means clustering algorithms. For a given number, comparison of the histogram intersection measure offers larger retrieval accuracy for both algorithms than L1 and L2 similarity measures [1].

The result of experiments on the 200 photographed images shows that a larger reduction in the number of comparisons with clustering without scarifying the retrieval accuracy. When the two clustering algorithms implement the histogram intersection similarity measure hierarchical performs better than the k-means clustering algorithm. This is true for both L1 and L2 similarity measures.

The comparison result of the three similarity measures with respect to retrieval performance shows that the histogram intersection measure is slightly better than L1, where as L2 is the lowest. The overall investigations [1] between the three similarity measures suggest that histogram intersection, L1 measures perform very similarly, and result a uniform clusters. Generally it concludes that the hierarchical clustering algorithm performs better in all cases than the k-means.

3. Low Level Visual Features

The vector space model and visual feature transformations are the two basic principles in visual information retrieval. The text information retrieval uses the vector space model to interpret documents as vectors in feature space that enables to measure similarity/dissimilarity of documents as vector distance. A visual object can be represented by a numerical vector as a result of visual feature transformation. The visual feature transformation is required to utilize the vector space model to measure similarities of images. The aim of a number of proposed visual descriptors is basically to imitate human similarity perception properly. MPEG-7 visual descriptors provide a standardized description of visual features that enables applications to identify, categorize and browse images or video [7].

Reducing the low-level information of an image into a manageable amount of relevant properties is the main goal of the feature extraction process. This reduces the complexity of the feature description process and makes the descriptors robust [13]. The low-level features are grouped into local and global features [11, 5]. The global feature of an image represents the overall visual features of an image. The descriptor describes the image by means of histograms (color and texture) and the overall layout of the image. The local visual feature extracts local visual information (color, texture and shape) dividing or partitioning the image into regions or objects.

The high-level features are rather qualitative that requires semantic understanding of images [6]. Image retrieval systems that make use of such features manipulate and process images based on information not directly derived from their content but rather use textual data description about the image [22]. The inability of words in accurately describing the rich image contents and the laborious task in labeling larger volume of images makes text-based image processing more difficult. The different way of perceiving the same image by peoples makes such feature representation more complex.

3.1. Color Descriptors

Previous image retrieval systems like QBIC [44] have used color moments in representing the color content of images. The three color moments that describe the general color distribution of images are the mean, variance and the skew-ness [22, 11]. The color property

of an image is represented with a smaller set of color vectors and only the dominant features of the image color distribution. The technique first identifies regions that contain a predefined set of colors within the image. The different regions are represented by binary vectors. The overall color content of an image is represented using few numerical values that decrease the discrimination power of this color descriptor due to their compactness.

The most common technique used in representing or indexing of images according to their color content is the color histogram [22, 29]. The technique allows the mapping of the constituent colors of an image into a discrete color space containing a predefined number of colors. It then calculates the number of points (pixels) mapped to each point in the corresponding color space. Therefore, a histogram counts and graphs the total number of pixels at each grayscale level. Many image retrieval systems implement color histograms as they are successful in characterizing the global color content of images.

The discrimination power of a color histogram is directly related to the number of bins specified. Different mechanisms [11] were proposed in selecting the number of histogram bins that consider the computational cost affecting the performance of histogram matching. The problem of a general color histogram is that it doesn't consider spatial information of pixels that degrade its discrimination power. Two dissimilar images can be categorized in a cluster if comparison is done using their general color histogram, as the spatial information is not considered [33]. One of the proposed solutions is the segmentation of the image into a predefined number of regions or blocks and calculating their corresponding histograms to incorporate the spatial information. Such a method is called sub-block histogram. The similarity between images is measured by calculating the histogram difference between the corresponding blocks. It needs to consider the number of regions of an image to reduce the memory and the computational time.

The other solution is the color coherence vector [11] that classifies the histogram bins based on whether they belong to a large uniformly colored region or not. The spatial information is incorporated by representing the image as a vector that includes the number of coherent and incoherent number of pixels in each bin.

Different image search and retrieval applications could implement different and independent color space choices, choice of quantization in the selected color space and histogram values of the quantization in using the generic color histogram. These two histogram-based descriptors of the MPEG-7 are designed to limit those varieties of choices by application for interoperability among MPEG-7 systems. The MPEG-7 specifies the number of bins in a color histogram, the quantization scheme (i.e. uniform or non-uniform) together with the color space used [21, 32].

The MPEG-7 visual content descriptors consist of two color histogram-based descriptors for describing the visual content of an image based on its color [32]. These are the Color Structure Descriptor (CSD) and Scalable Color Descriptor (SCD). The others two MPEG-7 color descriptors are the Color Layout Descriptor (CLD) and Dominant Color Descriptor (DCD).

3.1.1. Color Space

A number of color spaces are used for the purpose of manipulating colors. The basic criterion that distinguishes between colors spaces in image retrieval systems is uniformity. Uniformity of pairs of colors refers to the direct and close relationship between the psychological or perceptual similarities and the measured similarity distance between the pairs [21, 11].

RGB is the natural color space that is used by the display computer monitors. It is also used in computer graphics and color television. It is most widely used in image capture and display processes. The RGB color model is based on the three primary colors: red R, green G and blue B. Every color that is displayed on the monitor is the mixture of the shade of these three colors. The 256 different shades for each color component results in the possibility to display 16 million kinds of colors [11].

The basic limitation of the RGB color space is its inability in expressing the color changes in such a way that the change can be sensed by the human eye. It is also a device-dependent color space that lacks uniformity. The MPEG-7 color descriptors used more user-oriented color spaces: HSV, HMMD and YCrCb to solve the limitations of RGB.

Unlike the RGB or CMY color models that define color in terms of the combination of their primaries, the HSV model encapsulates information about color in terms that are more familiar to the way humans tends to perceive and manipulate colors. Hue (H), Saturation (S) and Value (V) are the three constituent components of the HSV color space. The Hue component represents the color in its purest form using a circular region on the cone structure representing the color space. Hue is invariant to changes to illumination that makes it suitable in object retrieval. The 'purity' or richness of the color is represented by the Saturation component that determines the appearance of the color. The Value represents the brightness or intensity of the color. The Saturation is represented by the height and the radius and corresponds to the value component in the cone structure depiction of the HSV color space. HSV is characterized by its perceptually uniform color space. It is commonly used in computer graphics applications [21, 11].

The other color space used in MPEG-7 color descriptors is the Hue-Min-Max-Difference (HMMD). There is no difference in the Hue components of the HSV and HMMD color spaces [1]. The other two basic components that define the HMMD color space are a shade (max) which indicates how close a color is to black and a tint (min) which indicates how close a color is to white. In other words, the min and max components are the corresponding minimum and maximum R, G and B values. The difference between the max and min components is defined as diff that indicates how pure a color is. The intensity or brightness is represented by the sum component that is computed from the average of the min and max [21]. A double cone structure which is basically a modified version of the single cone structure of HSV is used to depict the color model of the HMMD color space.

The two major components of the YCrCb color space are the luminance (Y) and chrominance (CrCb). The luminance component describes the black and white component of a signal or a pixel color. The amount of a light intensity or brightness is indicated by the Y component. Color pixels that need to be visible to human perception require a specified amount of luminance. The color portion is described by the chroma or chrominance component that includes Hue and saturation information together. It includes the color-red component (Cr) and the color-blue component (Cb) of the color. The YCrCb is commonly used in DVD video in compressed color encoding and for transmitting color video images.

The HSV, HMMD and YCrCb color spaces are user-oriented color spaces that make them suitable in content based image application based on colors. These are also more efficient for image search and retrieval. The conversions from RGB color to other color spaces are described in [21]. The transformation from RGB to YCrCb is linear but non-reversible in which the original image cannot be easily reconstructed [13]. Such lossy transformation requires preserving significant image details not to affect its appearance when perceived by humans. The transformation from RGB to HSV is non-linear but reversible.

3.1.2. Color Quantization

Color quantization is a type of vector quantization that enables to choose appropriate set of representative colors based on the selected color space. Displaying of a 24 bit full-color image on devices that support limited number of colors (less than 24 bits, mostly 8 bits) requires mapping colors of the original image to a relatively smaller number of representative colors through quantization process. The complete ranges of values for each of the selected component of the color space are specified and then normalized by the quantizer [24].

The two major categories of color quantization are uniform and non-uniform color quantization. Each of the selected axes of the color space is broken into equally sized regions or segments in uniform quantization. Even though the implementation of uniform quantization is easy, some regions of the color space are wasted during mapping. The non-uniform quantization is performed by segmentation of the color space non-linearly (i.e., logarithmically). Non-uniform quantization provides a better and consistent result at the expense of larger memory requirement due to color mapping complexity.

3.1.3. Color Structure Descriptor

CSD is a color feature descriptor that captures both color content similar to a color histogram and information about the structure of this content. Unlike color histograms it captures some spatial characteristics of the color. The scanning of the image by an 8x8 structuring element enables all locations (pixels) of the image to be visited. The structuring element enables to consider groups of pixels rather than considering each pixel of an image separately [21, 32].

When a specific color is encountered in a structuring element the corresponding CSD bin is incremented by the number of times the color is found. A kind of color structure histogram is used to express such local color structure information of the structuring element. The main difference between such kinds of histograms and the ordinary color histogram is that the histogram that represents local color structure uses the HMMD color space. The histograms are identical in form, but they are semantically different.

Even if the amount of a given color present in two images is identical (i.e., using methods of color histogram) the CSD can distinguish between these images by considering the structure of the groups of pixels having that color. Color values are quantized into 256, 128, 64 and 32 bins. The HMMD color space is divided into five subspaces such that non-uniform quantization is used over these subspaces for the different number of histogram bins. The CSD has the scalability property so that a descriptor with a larger number of bins can be reduced to one with a smaller number of bins keeping the size of the structuring element fixed [31].

The main function of the CSD is image-to-image matching and retrieval of still image [32]. The descriptor provides better similarity-based image retrieval when it is implemented for images that are arbitrarily shaped regions and especially when the regions are disconnected [23]. When it is compared to the ordinary color histogram it provides additional functionality and improved performance for still natural images. The degree of its higher accuracy in similarity-based image retrieval makes the CSD preferable. In the other way both the local spatial distribution and the global color feature are represented by the CSD.

3.1.4. Scalable Color Descriptor

The color space fixed for SCD is the popular HSV. The number of bins is fixed to 256 and quantization is performed uniformly throughout the histogram bins [21]. In the HSV space, the Hue (H) component is quantized to 16 bins and Saturation (S) and Value (V) are quantized to 4 bins each. Each of the 256 bins is represented initially by 4 bits which requires 1024 bits/histogram. The Haar transform is used to encode the histogram and reduce the larger size representation that is to make the descriptor scalable.

The result of Haar transform is a set of high-pass and low-pass coefficients. The high-pass coefficients have relatively low (positive and negative) values due to the redundancy of the originally transformed histogram. The transformed source histogram could be 128, 64, 32 … bins per histogram in the Haar representation after avoiding redundancy in the consecutive adjacent histogram lines. Feature extraction complexity is not added in generating the Haar coefficient from the histogram [21, 31, and 32].

A subset of the Haar coefficient can be used to perform matching for coarse pre-selection of subsets of candidates in a larger image database that will follow using more coefficients to perform a refined matching on the pre-selected items. The use of such coarse-to-fine approach significantly enhance faster similarity search in large image databases.

3.1.5. Dominant Color Descriptor

The DCD identifies a region of interest in an image and clusters all the colors in the region into smaller number of representative colors [21, 32]. One improvement or advantage of the DCD over the above color descriptors is that the colors are directly computed from each image instead of being fixed in the color space which result a more accurate and compact feature representation. This type of color descriptor consists of the representative colors of the region, their percentages, the spatial coherency of the dominant colors and the color variance for each of them.

The maximum number of dominant color representing a region is 8, but varies from image to image. The number typically ranges from 3 to 4 for most images. The smaller numbers of dominant representative colors facilitate the process of searching similar color distributions in the database and combine the result. Therefore the first step in computing the DCD is clustering the colors in a given image or region [21].

The percentage of each of the corresponding dominant colors is computed. The spatial coherency is the measure of the overall homogeneity of all the dominant colors in the image. It enables to distinguish between the degree of concentration and spread of colors all over the image. It is quantized into 5 bits. The color variance is computed from the perceptual weights based on local pixel statistics. Only 3 bits are used to represent the optional color variance.

When the color content of image regions is required to be sufficiently represented by small number of colors the DCD is preferable than other color descriptors [25]. The main application of the descriptor is in similar image retrieval. It is effective in image applications in browsing large image databases based on specific color values.

3.1.6. Color Layout Descriptors

Feature extraction of CLD begins by partitioning the image into 64 blocks (8x8). The main benefit of partitioning the image is to ensure the scale invariance or resolution of the descriptor [21, 32].

The average of all colors in each block is taken as representative colors of the block. This results in the image being represented or expressed in terms of a 2D array of representative colors of the 64 blocks. The applications of Discrete-Cosine Transformation (DCT) on each of the local representative color produce a set of low-frequency and high-frequency coefficients. The non-linear quantization performed on the selected low-frequency coefficients by zigzag scanning forms the descriptor [21, 32].

3.2. Texture Descriptors

Texture refers to the repetition of a basic pattern over a given area. The texture feature is defined in terms of the shape of the basic pattern and its repetition rate. Surfaces are described by their textural properties that express their structural arrangement and their relative relationship to the surrounding environment [22, 29]. A texture feature has spatial extent unlike a color feature which has point-wise (pixel) property. In other words, it characterizes the interrelationship between adjacent pixels of an image. Images that have similar color content, for example sky and sea, are distinguished by their textural similarity.

There are different techniques that describe the texture feature similar to the human perception system [29]. Visual texture properties are recognized by the human eye. The properties that are used by humans to define texture are coarseness, contrast, directionality, regularity and roughness, according to the psychological studies of human perception of texture. The Tamura Feature [11] is one of the computational approximates developed for defining these important visual texture properties.

The wavelet-based texture descriptor is the other descriptor based on the psychological studies in which the descriptor values are related to the signal values. The wavelet transform is utilized in texture analysis that enables the decomposition of a signal into wavelet sub-bands of specific frequency and use the mean and variance extracted from them as a representation of texture [22]. The distribution of each sub-band is used in the construction of the feature vectors. In order to achieve better performance in texture analysis the wavelet transform can be combined with other techniques like co-occurrence matrix and KL expansion [22].

The Gabor wavelet transform also closely models the human vision of texture. The discrimination power of the Gabor filter texture feature extraction is found to be excellent when implemented over textures of broad ranges [30].

The MPEG-7 visual feature descriptor states three texture descriptors in describing the content of an image [21, 32]. These are the Homogeneous Texture Descriptor (HTD), Texture Browsing Descriptor (TBD) and Edge Histogram Descriptor (EHD). Different applications that implement similarity matching in image retrieval require employing the appropriate texture descriptor. In retrieval of similar images from a large image database, the three texture descriptors can be used independently or in combination with the MPEG-7 color and shape descriptors.

3.2.1. Homogeneous Texture Descriptor

The Homogeneous Texture Descriptor is a robust descriptor that provides a quantitative representation and easy to compute. The 2-D frequency space is partitioned into 30 channels and in each of the filtered channels the textured energy is computed. The 2-D Gabor function is used in modeling the individual feature channels [21, 19].

The descriptor is constructed from the feature vector both from values of the mean and energy deviation computed in each of the 30 frequency channels that are the results of the Fourier transform of the image. In addition to these 60 components, the mean intensity and the standard deviation of the image in the pixel domain are calculated and added to the first part of the feature vector. This results in a feature vector of 62 dimensions being extracted for each image [21, 32, and 28].

The texture description technique of the MPEG-7 homogenous texture is in a way that the descriptor corresponds to the Human Visual System (HVS). Psychological experiments on the human perception show that among the widely dispersed spatial frequency domain human responds to only specific bands of frequency [28].

Similar to the decomposition of the spectra into perceptual channels by the HVS, the spatial frequency domain is decomposed and represented by sub-bands. The division of the frequency domain into sub-bands simplifies the computation of quantitative texture values. The sub-bands are also called frequency channels in the frequency domain. The frequency layout of the sub-bands is designed in a way so that the extracted texture information from each band is matching to the human perception system. The basic components of the homogenous texture descriptor, energy and energy deviation are extracted from the 30 sub-bands [28].

3.2.2. Texture Browsing Descriptor

The Texture Browsing Descriptor is a descriptor that provides a qualitative representation of the texture feature of an image similar to the human perception of texture. Texture-based browsing applications require implementing the descriptor in browsing large image databases. The descriptor uses regularity, directionality and coarseness to describe the texture of an image. The three characteristics of the descriptor are compactly represented by only 12 bits.

The regularity of a texture ranges from an irregular or random texture to a highly graded periodic pattern [21]. The quantized four values of regularity requires 2 bits of the descriptor taking values from 0 to 3 with the maximum value indicating periodic pattern. Regularity of an image is described better by a pattern that has well-defined directionality and periodicity.

The six possible values of the directionality component of TBD range from 0 to 1500 in steps of 300. The seven quantized values of the 3 bits x 2 directionality representations begin with value 0 that signifies no dominant directionality followed by the corresponding six directions. The coarseness component ranges from fine grain to a coarse texture in the 2 x 2 bit representation [21].

Retrieval of similar images from large image databases involves several phases for effectiveness and efficiency reasons. Identifying candidate images that are characterized by similar perceptual properties is required before trying to get precisely the required similar images that are to be returned as the final result to a query submitted by users [21]. The texture browsing descriptor can be used in the first phase in finding images characterized by specific features which simplifies the retrieval process. This avoids the exhaustive comparison of the query image with all target images using more than one descriptor type. Implementing the HTD-based similarity retrieval on the list of selected images of the first phase is performed returning the required result.

3.2.3. Edge Histogram Descriptor

The Edge Histogram Descriptor (EHD) is a texture descriptor that represents local edge distribution of an image by dividing the image into non-overlapping 4x4 sub-images [21, 34]. The local edge histogram computed from each sub-image enables to capture the spatial distribution of edges. There are five possible types of edges. The four directional edges are vertical, horizontal, 45 diagonal and 135 diagonal. The last edge type is the nonorientation specific.

Each local edge histogram is represented by five bins corresponding to the five categories. Therefore, for the feature vector corresponding a total of 5x16 = 80 histogram bins are required for the whole image.

The feature extraction process starts by dividing the image into 4x4 sub-images and assigning index according to their location [21, 34]. Each of the sub-images is further divided in to a fixed number (power of 2) of image blocks to perform edge detection inside these sub-images. The edge detector computes the edge strength of each block from the pixel intensity using five different filters. The value of a given histogram bin of a sub-image is incremented when the computed result obtained by the corresponding specific filter type exceeds a certain threshold. The five filters compute five different edge strengths from each image block of a sub-image there by contributing to the corresponding edge histogram bins.

The EHD is especially effective in the representation of natural images with non-uniform edge distribution [21]. The descriptor is targeted to image retrieval systems not for the object-based ones, but for those that require image-to-image matching like QBIC [44].

3.3. Shape Descriptors

The shape of a visual object is defined as a property or characteristics of a set of points in which the object shape is determined by the geometric relationship between the points [14]. The basic assumption in shape analysis is considering shape of an object as characteristics of a binary image region. This allows the description of a shape to be computed from the object's boundary or its interior content.

It is required to identify the constituent objects of an image using techniques like pixel based segmentation, edge detection and skeletonization. The robust and accurate identification of the objects determines the extraction and representation of shape feature descriptors.

The segmentation task of an object results in several regions based on color or texture information. This task is followed by identifying regions of interests. Accordingly; [20] divides shape descriptors into three main categories: namely the area (region) based descriptor, contour (boundary) based descriptor and skeleton-based descriptors. The skeleton-based shape descriptor is formed by mapping the computed skeleton into three structures. Shape similarity among objects is determined by using a tree matching algorithm. The lack of stable skeleton computation that depends on the object's boundary makes the performance of a skeleton-based descriptor weaker.

The contour-based descriptor represents shapes according to the boundary information such as radius, contours and chord length. The global statistical approach for the 2D shape description method includes moments, Fourier shape descriptor and wavelet shape descriptor.

The Fourier transform is a well-known and useful for pattern analysis. Fourier transform of the boundary of an object represents the Fourier shape descriptor. Contour of a 2D object is considered as a closed sequence of successive boundary pixels. Curvature, centroid distance and complex coordinate functions are the three different types of contour representation that

are defined based on the contour. The general and detailed shape properties of objects are described in terms of the low and high frequency coefficient respectively in the frequency domain generated from the three representations.

The MPEG-7 standard for image specifies two shape feature descriptors. These are Region-Based shape descriptors and Contour-Based shape descriptors.

3.3.1. Region-Based Shape Descriptors

The Region-based shape descriptors are derived from the pixel information distribution within a 2D shape region or an object [3]. The 2D Angular Radial Transformation (ART) and moments are the techniques used in region-based shape analysis and description.

The MPEG-7 region-based shape descriptor is an ART-based descriptor that efficiently expresses or describes the pixel distribution of a 2D shape region or object [4]. A number of multiple disconnected or simple regions are described by employing a complex 2D ART. Complex objects with multiple contours and multiple holes inside the object under consideration are well identified by the descriptors.

The region-based descriptor is not adapted to natural color images. The feature extraction of a grey-scale image is performed by size normalization and ART transformation and finally area normalization [4]. It is required to define height and width to normalize the size of the image by means of interpolation. This results in edge maps that are size-invariants in order to apply ART transformation on them. From each image, 35 ART coefficients that are quantized to 4 bits per coefficient are extracted representing the region-based shape descriptor. Region-based shape similarity measure requires normalizing the magnitude of the ART coefficients by dividing each coefficient by the first.

The regional property of the descriptor makes it robust to segmentation noise. The descriptor is invariant to scaling and rotation [3, 4]. The representation of each coefficient using only 4 bits makes the size of the descriptor smaller and compact that facilitate faster extraction and in similarity matching of large image databases.

3.3.2. Contour-Based Shape Descriptor

The Contour-Based Shape Descriptor (CBSD) describes only the closed contour of a single object instead of the entire region with holes or disjoint parts [13]. The descriptor does not capture the shape-interior content and depends only on boundary information. The technique used when there are multiple disjoint regions of a complex object is that the descriptor describes the component contours separately.

The contour-based shape descriptor is based on the Curvature Scale Space (CSS) representation of an object contour that captures its perceptually meaningful shape features [3]. Determining the inflection points, that are points with zero curvature, enables the decomposition of the contour into concave and convex regions. The contour is analyzed at different scales by a smoothing process. The result of the smoothing process continues until the concave parts are flattened and the contour becomes convex as a result of repetitive application of a low pass filter. The CSS image shows how the inflection points or peaks change during the iterative filter operation. The CSS x-axis and y-axis coordinates corresponding to the position of points along the contour and the amount of smoothing iterations needed to remove them, respectively.

M. Bober [3] described that the CBSD comprises three major parts, namely:

- An index indicating the number of peaks
- The magnitude of the largest peak and the position and magnitude of other peaks
- The global and prototype curvature vectors that are built by the eccentricity and circularity values of the original and filtered contour.

The average size of the descriptor per contour is 112 bits, whereas the maximum feature length is 134 bits.

The descriptor emulates the shape similarity perception of the human eye system and provides compact description of objects in the region of interest. It is robust to both the noise present on the contour and distortion in the contours. It is also robust to rigid or non-rigid deformations in the object. The CSBD representation is invariant to rotation, scaling and

mirroring the object contour. The descriptor also discriminates shapes of objects whose regions are similar but have different contour properties.

The contour-based descriptor is suitable for indexing and provides better retrieval performance than the region-based shape descriptor. Shape classification tasks are suitably performed using a region-based descriptor [20]. But CSBD does not capture the shape-interior content and depends only on boundary information.

3.4. Evaluation of MPEG-7 Image Descriptors

The basic descriptors of MPEG-7 are defined and tested mainly based on their performance in the retrieval process. The clustering process can be down-graded as some of the image features can become noise. It indicates the impact of using these different descriptors and the ways of improving their usage in image processing [22].

Eidenberger [10] analyzes the quality of the extracted descriptors by applying the different feature extraction algorithms on different media collections. The analysis selects three media types namely: monochrome texture, color photos and a set of artificial color images with few color gradations. Except for the region-based shape descriptor, all MPEG-7 image descriptors are analyzed on these media.

The result of the analysis suggests the ways of using the extracted features on different media content. The descriptors are tasted such that they satisfy the properties of good descriptors based on statistical indicators, such as variance, robustness and their discrimination power.

3.4.1. Dependency on Media Content

Eidenberger [10] has showed that the only media that performs well with all color descriptors is Corel photos. Except for the DCD other color descriptors do not perform well for both monochrome content and artificial media objects. The DCD performs well on each type of the three media.

Even though the edge histogram performs well on any type of media it doesn’t produce good result for highly textured images (i.e., photos) as a result of the irregular distribution of the

elements due to holes. The Homogenous texture performs poorer on images that that have few color gradation and texture, but well on images with monochrome content. Among the three texture descriptors, Homogenous texture is relatively poorer for texture regions. The region-based shape descriptor performs excellent in all of the three media types [10].

3.4.2. Descriptor Redundancy

The larger number of color descriptors makes the selected MPEG-7 descriptors highly redundant for monochrome media objects. Implementing MPEG-7 color descriptors in processing monochrome images doesn't give reasonably good output. The color information incorporated by each of MPEG-7's color descriptors for each of the media is non-overlapping. The redundancy of the descriptors decreases by half when applied to Corel datasets and the coats-of-arms datasets. The descriptors are less redundant in Corel photos as it contains more details that include colors, significant edges and textures. The redundancy of coats-of-arms is between the monochrome and Corel datasets [10].

The dominant color is the only color descriptor that performs well on monochrome images independent of other color features [10]. The result generated by the two histogram color descriptors is different since different color spaces are used by each of them. The result it generates also has no similarity to the three texture descriptors. Similar to dominant color, Texture browsing is an independent descriptor whereas the Homogenous texture can be fully described by the non-directional edge descriptor.

Statistical evaluation on the MPEG-7 descriptors suggested that the combination of color Layout, Dominant color, Edge Histogram and Texture Browsing provides better result while the two histogram-based MPEG-7 color descriptors perform badly on monochrome media [10].

4. Image Similarity and Clustering

This chapter discusses the major concepts of similarity measures used for MPEG-7 based image descriptors. It also identifies the use of a quantitative model that enables the use of predicate-based similarity measures for images and the need for weighted distance measures when using many descriptors to represent images.

The properties and the steps followed by the two selected clustering algorithms are then discussed. Particularly, some of the widely used linkage methods in agglomerative hierarchical methods are considered, but only the general k-means clustering method. Lastly, the Shannon entropy measure is highlighted as a measure of quality of clusters.

4.1. Image Similarity Measures for MPEG-7 Descriptors

There should exist smaller distance among perceptually similar images but larger distance among those that are not. Among all the distance measures used in image retrieval the ones that reflect human perception are more desirable [37].

Eidenberger [6] listed the most possible similarity structures. The first two similarity structures (or measures) are classified based on the geometric shape of feature spaces. The Euclidean distance structure over a set of pairs of objects assumes the feature space as Euclidean geometry [8]. On the other hand, if no geometric shape of the feature space is assumed, then the similarity measure is said to be a metric over the selected set of objects.

The vector space model for visual similarity measurement is derived from information retrieval theory. The model represents feature vectors as points in feature space and indexes them in a database. The result of a specific distance function that is used in measuring the distance between pairs of feature vectors is represented as a point in n-dimensional distance space. This enables measuring similarity that is metric-based with no assumption on the shape of the vector space. Most Visual Information Retrieval distance measures consider the Euclidean shape that fulfills the following similarity measure properties (metric axioms): Positivity, Identity, Symmetry and the Triangle Inequality.

4.1.1. Color Similarity Measures

The norm-based L1 similarity measure is selected and implemented for better retrieval and accuracy in measuring similarity using SCD, the Scalable Color Descriptor (Section 3.1.4). The application of the L1-norm shows similar results when used in both histogram and Haar transform domains only when identically signed high-pass coefficients are used in the latter case. The complexity is the same in matching coefficients or matching in the histogram bins and when the numbers of coefficients are equal. It is also possible to compare differently sized representation in similarity retrieval [21]. Experimental results show the possibility of achieving reasonable performance in similarity matching using histograms that have smaller number of bit representation per histogram. Good retrieval accuracy is achieved by using the Ll distance measure or the Euclidean measure in measuring similarity among images based on CSD, the Color Structure Descriptor (Section 3.1.3).

If the similarity measure is performed based on the representative colors of each region of the query image and the targets independently then the final computed combined result measures the overall similarity among the pairs of corresponding regions' color distribution in the database. The optional spatial variance and coherency may not be considered in measuring similarity. Unlike the histogram-based descriptors that use L distance measure, the DCD (Dominant Color Descriptor; Section 3.1.5) uses the quadratic distance measure [21].

4.1.2. Texture Similarity

The standardized normatic semantic for the MPEG-7 EHD is described by only 80 bins that represent the local edge. The process of image matching that implements only the local edge histogram doesn't provide sufficient results without considering the global features. A global edge histogram and semi-global edge histograms are computed from the 80 bins local edge histogram [21, 34]. Unlike the local edge histogram that represents only local edges, the edge distribution of the whole image space is represented by the global edge histogram. A 5 bin global edge histogram is obtained from the basic 5 edge types by accumulating and normalizing the corresponding computed local edge histograms of the 16 sub-images.

A third descriptor called semi-global edge histogram is computed from 13 different segments that are defined from the image blocks. The computation of the five edge types from each segment produces a 13x5=65 bin representation of the semi-global edge histogram.

The similarity between two images based on edge histogram needs to consider three histogram bins: local, semi-global and global [21, 34]. MPEG-7 proposes the L1 norm to measure the similarity between the images. Hence a weight factor is required for D_{global}(A, B) since the number of bins of the global edge histogram are relatively small.

The recommended MPEG-7 distance for the Homogeneous Texture Descriptor (Section 3.2.1) is the weighted City-Block distance [21]. The similarity measure is computed from the Homogeneous Texture Descriptor on pairs of images in the database.

In addition to the ordinary matching of images, three different texture similarity matching methods are introduced by Krishnamachari and Adbel-Mottaleb [19]. Removing the first component from the feature vector of the descriptor, i.e., the mean intensity enables applications to perform intensity-invariant matching in measuring similarity.

When the similarity measure is performed between a query image and an image in the database which is scaled differently, the scale-invariant matching mechanism allows the query image to be zoomed in and out appropriately. The minimum distance measure among those measured by zooming-in and zooming-out is taken as the distance measure between the two images [28].

The texture descriptor of a rotated image is an angular shifted version of the original one as sub-band division is performed in polar coordinates. The rotational property of such polar coordinates allows for implementing a rotation-invariant similarity matching method that uses the shifting of the feature vectors of one of the completed images in an angular direction. A maximum of six angular shifts at an interval of 300 has been tested and the minimum distance among these possible distance values is considered to be the distance measure between the two images [21, 28].

4.1.3. Shape Similarity Measures

Fundamental comparison of two specifically shaped objects is the determining of a transformation which casts one of them into another [14]. It assumes the transformation to perform shape matching to be rigid or affine which is not suitable to irregular shape deformations. The non-rigid shape matching method formulates the comparison as the finding of the corresponding feature points among the pair of objects by means of non-rigid transformation. Hence Grigorescu and Petkov [14] defined shape comparison as "either finding pairs of corresponding feature points (i.e. solving a correspondence problem) and/or determining the transformation which maps one point set into the other." The value of the measure of the shape similarity is calculated from the correspondence problem or from the non-rigid transformation.

Moments, Fourier coefficients, CSS, shape context and skeletons are the most commonly used methods in 2-D shape matching and shape-based object retrieval in image databases [18]. Once objects or regions of interest in an image are detected, description of the object requires analysis for simplified representation of the shape. The shape representation needs to preserve the important characteristics of the original shape. Fore example, the silhouettes of 2-D objects that are projections of 3-D objects could change for three different reasons. These are change of a view point, motion of non-rigid object and noise.

4.1.3.1. Partial Shape Matching

The limitations of applications of contour-based shape descriptors in shape similarity measure are reduced when partial matching among shape objects is implemented [20]. The two associated problems to partial shape matching are scale selection and subpart selection. Let a query part Q to be searched as part of target object T, and then there should be appropriate scale selection of the part with respect to the target. In addition to the scale selection, the comparison of Q needs to be performed to all possible subparts of T. The identification of the subpart is by decomposing Q into parts according to a specific criterion or sliding it over all possible positions with respect to T.

A single-directional Hausdroff distance is one of the partial matching approaches that tries to minimize the distance of all points of the query part Q to points in the object T. Enumeration

of all possible scales is the mechanism used to solve the scale selection problem. But the Hausdroff distance results in larger similarity values for two different objects as it doesn't tolerate shape deformation [20].

The preferred scale selection method for contour-based global similarity measure is scaling of the contours of both images to the same length. Such global similarity measures do not face the subpart selection problem.

The similarity measures based on sampling that reduces shape information are not desirable as the measured distance value for two different shapes is larger. Robust similarity measures are characterized by a smaller reflected change in the value of the similarity measure when there is a small change in the shapes The Hausdroff distance is robust to noise on contours but not on region [27]. The overall comparison result shows that the two MPEG-7 standards, CSS and ART are robust to both deformation and noise. In addition to the robustness they perform relatively better than others [27].

4.1.3.2. Evaluation of shape similarity Measures

The similarity measures that are used in image retrieval results in different levels of retrieval accuracy and effectiveness. The higher requirement of computational efficiency of online image retrieval applications needs to choose the appropriate distance measures [37]. Zhang and Lu [37] evaluated the most common similarity measures using two different shape image databases, one region shape database and one contour shape database. Retrieval performance is worst using the quadratic distance measure. The city-block distance and chi-square statistics are better in both retrieval efficiency and effectiveness. Experimental results on shape images show that city-block distance is preferable to the chi-square which is not simple to calculate.

The first part of the MPEG-7 Core Experiment CE Shape-1 specifies some fundamental conditions that should be satisfied by every shape descriptor, namely robustness and invariance to scaling and rotation [27]. The other and main part is designed to measure the performance of the similarity-based retrieval. The test classifies similar objects or distorted versions of a base shape in each category. The result of using 15 different similarity measures on the selected data set shows significant difference in their retrieval rate. Retrieval

of similarly shaped objects using these shape similarity measures on classes of shapes results in objects that are similar to a different class than its own. Any of those shape similarity measures couldn't achieve 100% similarity retrieval due to the shape variation of the objects that belongs to a class.

4.1.4. Combining or Merging Distance Measures

When the number of descriptors used in describing the images to be compared is more than one, the similarity between them is computed from the weighted sum measured distance of each descriptor [16, 8]. The overall dissimilarity between the two images considered is given by the equation:

$$D(i) = \Sigma_j w_j d_j \qquad (1)$$

Where d_j is the measured distance of the j^{th} descriptor of the image and w_j is the corresponding assigned weight for the descriptor. The value of w_j is between [0, 1] such that $\Sigma_j w_j = 1$.

Section 3.4 shows the analysis and the effectiveness of the different descriptors when the feature representation is performed in monochrome and color images. The descriptor that performs badly on a given media need to be given relatively smaller weight than the best descriptor with respect to a specific media type. The value of assigning weight in computing the global distance needs to consider the degree of redundancy of descriptors in addition to the media type.

Greater weight need to be assigned to dominant color descriptors than other descriptors when the comparison is performed between color photos. The weight assigned to the Homogeneous texture descriptor has to be smaller for color images but larger for monochromes content. When more than one texture descriptors are used in texture images, due to its lower discrimination power of textured regions, lower weight has to be assigned to the Homogeneous texture descriptor. The similarity measure based on the computed distance from the feature vectors depends on the media collection of the images and the descriptor type [8].

Visual information retrieval browsing and clustering of still images requires the measuring of similarity/dissimilarity as distance [8]. Comparison of two still images is performed based on their extracted low-level features. The comparison for the similarity among images requires measuring the distance between the corresponding n-dimensional vector representations of the images.

A problem arises due to the complexity of human visual similarity perception as compared to a number of models used in measuring similarity of visual information [8]. The performance comparison results of three similarity models in comparing images based on their low-level features shows that the Vector Space Model (VSM) performs better than two machine learning technique-based similarity models: k-NN (Nearest Neighbor) and Support Vector Machines [17]. The lowest computational requirement of the VSM makes it preferable for the unsupervised clustering of large image database.

The vector space model that implements the Euclidean geometry structure is a commonly used distance measure in most image retrieval systems, but lacks the property to imitate the human similarity perception [8].

The MPEG-7 visual information standard proposed distance measures for each of the descriptors that are derived analytically. The selection of the best distance measure among many for a given feature extraction method is assumed to be by performing quantitative analysis techniques. The quantitative-based distance measure that is applied on continuous data vector elements or quantitative values is representation of low-level information which is used in most visual information retrieval.

On the other hand, the predicate-based distance measure that is employed in human-related sciences such as psychology are suitable to human perception since the predicates represent high-level information. Eidenberger [8] suggested the implementation of a model called the quantization model that enables the use of the predicate-based distance measure on quantitative or continuous data.

Experimental results show that the recommended distance measures for the MPEG-7 descriptors are not the best. The application of the predicate-based distance measure on most

MPEG-7 descriptors shows that they perform better than the MPEG-7 recommendations. Accordingly a Pattern difference, a predicate-based distance measure, is found to be the best to extend its application on continuous media. Pearson correlation coefficient the other a predicate-based distance measure performs best for Texture Browsing Descriptors. The two quantitative distance measures that perform best are the Meehle Index and the Clark's divergent coefficient.

4.2. Clustering Algorithms

The first and basic step in content-based image clustering is the selection of the low-level features that are represented by unlabeled feature vectors [16]. The image clustering process requires the selection among many different kinds of features, so that the images within a cluster are more similar to each other than images belonging to a different cluster. An image that belongs to a cluster should be semantically similar in the selected features according to the weight assigned to the features that varies based on the media type.

The images to be categorized are not assigned to a cluster by measuring their proximity to a predefined specific image class or template. In addition to this, the input data to the clustering algorithm are not labeled, making the clustering process unsupervised. Since members are assigned to a class with no priori information about membership the clustering process is traditionally considered as unsupervised learning [15].

4.2.1. Hierarchical Methods

Hierarchical clustering builds a cluster hierarchy (i.e., a cluster tree), the so-called dendrogram. Agglomerative (bottom-up) clustering algorithms start with one-point clusters and recursively merge two or more most appropriate clusters with regard to a given similarity metric. Divisive (top-down) approaches start with one cluster of all data points and recursively splits the most appropriate clusters with regard to a given similarity metric. Splitting (merging) continues until a given stopping criterion (e.g., number of clusters k) is satisfied [19, 15 and 16].

Hierarchical clustering is aimed at producing a tree-like structure or hierarchy of clusters after the clustering process is completed [15]. The tree-like diagram that is displayed as a

result is called dendogram. The dendogram shows how the clusters are related. The effect of cutting the dendogram at certain desired levels is equivalent to the partitioning of the data items into disjoint groups. There are two types of hierarchical clustering based on the way the technique proceeds.

The Agglomerative hierarchical clustering method iteratively merges small clusters forming larger ones. It selects two groups and combines them into a single group according to a specified clustering criterion, e.g., the measure of similarity in the clusters. The fundamental criterion in hierarchical clustering is the identification of groups that should be linked or combined. The most commonly used group linkage methods of HACM are complete-link, average-link and Ward's method.

The method followed by Divisive hierarchical clustering is in the opposite way to that of Agglomerative method. It first places all objects to be clustered into a single group. The procedure divides the initial single group into two groups based on a predefined threshold distance that results in the objects in one group to be different from the objects of the other group. The Divisive clustering process terminates when the distance between the selected objects is less than a threshold distance.

The complete linkage method measures the maximum inter-group distance among given pairs of groups and links those pairs with the smallest maximum separation. The method forms smaller, tighter and more compact clusters. The average method is a mean method that links groups with the smallest average inter-group distance which is a compromise between the single and complete linkage methods.

The major emphasis of Ward's method is in minimization and quantifying of the information loss in each grouping in terms of the error sum-of-squares criterion. The values of the smallest variance of each pairs of groups to be linked are considered. The centroid of the temporarily merged groups is determined so that the average squared distance to the centeroid or the variance is computed. The merged groups with the smallest value of variance are linked. Therefore the pair of groups that are linked by Ward's method are groups that produce the smallest variance in the merged groups. The method aims at

reducing variations inside groups and results in clusters that are more homogeneous, which makes it an expensive or the most CPU intensive method.

In general the following are the basic steps in hierarchical clustering. Let the similarity matrix S between all pairs of the n images in the database be pre-computed [1].

1. Each of the n images are placed in n distinct clusters indexed as $\{C_1, C_2, \ldots C_n\}$
2. Find two distinct unmerged clusters with closest similarity measures.
3. Merge these two clusters into a new cluster. The new cluster formed by merging two clusters results in reducing the total number of clusters by one in each step.
4. Repeat steps 2 and 3 until the number of unmerged clusters has been reduced to a required number or the largest similarity measure between clusters has dropped to some lower threshold.

4.2.2. Partitioning Relocation Clustering

Clusters are found by iteratively optimizing clusters through relocation of their cluster members (data points). Data points are reassigned between k clusters iteratively until the optimal subsets, with respect to a given similarity metric, are found. Depending on the model/metric to be optimized, algorithms can be grouped in probabilistic clustering (i.e., optimizing a probabilistic model) or k-means/k-medoid methods (i.e., optimizing a dissimilarity or similarity objective function). Partitioning Relocation Clustering algorithms to cluster data points within heterogeneous feature spaces is demonstrated in [2, 15 and 16].

4.2.2.1. k-means clustering

k-means is an unsupervised learning algorithm that divides a set of input data objects into k groups. The procedure implements the least-square partitioning methods in classifying an input data set into a fixed number of clusters fixed a priori. The algorithm starts by initializing k cluster centers, one for each cluster. The method of choosing the initial centroids determines the final cluster result. The input vector that is associated to a nearest centroid is assigned to that cluster. The cluster center is updated every time a new member is assigned to a cluster. The reassigning of new member to a cluster ends only when the cluster center doesn't change [1].

The steps of the k-means algorithm are as follows [1]:

1. Choose k data points to initialize cluster centers. Choice of n_c centers is performed by randomly selecting the indexes of n_c images.
2. Measure similarity between an image and a cluster center and assign the image to the cluster if it has the largest similarity.
3. Compute the centroid of the newly formed cluster.
4. Repeat steps 2 and 3 until no more change in the value of the mean takes place.

The random specification of cluster centers initially determines the final result. To get a better result it is required to try a number of different starting points by running the algorithm many times.

4.3. Cluster Quality Measures

The proposed metric to measure the effectiveness of clustering color photo images is with respect to their semantic meaning. Shannon's information theory is selected in measuring the image cluster cohesiveness. When the distinct clusters formed are associated to a specific image semantic the clusters are said to be well separated. The separation among clusters becomes better when the semantic association to a cluster is predictable. The intra-cluster cohesiveness is also captured by Shannon's entropy measure in measuring the cluster separation [5, 16].

The cohesiveness of a cluster can be defined as the information content of a cluster. Shannon's information theory is implemented to model the semantic cohesiveness within a cluster. Let p_{ic} be the probability of selecting an image of semantic category i within a cluster c. Let n_i be the number of images of semantic category i in cluster c that consists of n_k images. It follows that $p_{ic}=n_i/n_k$ and $\Sigma P_{ic}=1$ for all semantic categories. When most images in a cluster belong to a semantic category, the measured value of the cluster cohesiveness is smaller. The following formulas are applied to the clusters formed in measuring cluster's cohesiveness when m clusters are formed.

$$\textit{Cohesiveness of cluster } c \quad = \quad -\sum_{i=1}^{k} p_{ic} \log_2(p_{ic}) \qquad (2)$$

$$\textit{Total cluster cohesiveness} \quad = -\sum_{c=1}^{m} \sum_{i=1}^{k} p_{ic} \log_2(p_{ic}) \qquad (3)$$

Let S_{ic} be the probability of selecting an image belonging to semantic category i in cluster c among all images belonging to semantic category i. Suppose n_{ic} is the number of images of semantic category i in cluster c and the total number of images in the semantic category i is N_i . It follows that $s_{ic} = n_{ic}/N_i$ and $\Sigma\, s_{ic} = 1$ over all m clusters. The information content of a semantic category i in the entire cluster reflects the cohesiveness of the semantic category. It also indicates the separation of semantics between the clusters formed.

$$\textit{Cohesiveness of semantic category } i \quad = \quad -\sum_{i=1}^{k} s_{ic} \log_2(s_{ic}) \qquad (4)$$

$$\textit{Total semantic category cohesiveness} \quad = -\sum_{c=1}^{m} \sum_{i=1}^{k} s_{ic} \log_2(s_{ic}) \qquad (5)$$

The lower the value of the entropy the better cohesiveness measured and quality of clusters. Categorical cohesiveness is also better for smaller value of the indices.

5. Experiments

5.1. Dataset Selection

The availability of priori knowledge about the image datasets to be clustered is the basic requirement in cluster validity analysis [15, 16]. The priori knowledge about member images in a semantic category enables the measurement of the semantic cohesiveness of each category after implementing a specific clustering method.

The main aim of this study is not to identify semantic similarity among images. We chose existing semantically similar images from the COREL image database [5, 33] that provides rough semantic labels of the images. Even if COREL associates the main subject of an image to the labeled keywords, further semantic classification is required by humans in order to lower the semantic gap. We have chosen eight image semantic categories that were used in SIMPLIcity [33] and in CLUE [5] for cluster-based retrieval of images. The categories are Africa people and villages, Buildings, Buses, Mountains and glaciers, Dinosaurs, Elephants, Flowers and Foods.

Category ID	Category Name
1	Mountains and glaciers
2	Buses
3	Foods
4	Dinosaurs
5	Elephants
6	Flowers
7	Buildings
8	Africa people and villages

Table 1 Semantic image categories and their IDs

Even though the set of images in a category are considered as similar in CLUE and SIMPLIcity, two groups of students were made to select semantically more similar images among the 100 images in each category. The ways humans interpret the same image vary as perception varies among people. The students were neither informed about the category

given by the database, nor able to compare other images to a predefined image template. We took images that were selected by both student groups as our datasets for the experiment.

Most categories include images mainly containing the objects specified in labels. Students have used different criteria of their own in selecting similar images in a way they perceive them. When 100 images of buses are provided, the subject differentiates two or more similar buses by their color. Most of the selected buses are of the same dominant color and all are one-floored, similar-size buses. The direction of motion of the buses and the way the surrounding environment looks like in which the photograph is taken are other parameters used in categorizing buses. Pictures of buses that are taken around city buildings and those around streets with trees on both sides are considered semantically different.

The “Elephant” category consists of single elephant or elephants in groups in forest, grassland, river and lake areas. The primary parameter for selecting the similar “Flower” image category is also their color. Next to their color, the number of flowers in the image was the main criterion of both students group in categorizing “Flowers”. The picture of “Africa people and villages” mainly contains photographs of groups of people dressed and decorated traditionally. The pictures are taken when such people are celebrating variety of festivals. The “Beach” category contains scenery at coasts or river banks.

The images in each semantic category are stored in JPEG format with a size of 384x256 or 256x384 pixels. Among the images that were presented to the students, they selected equal-sized images, taking size as a factor to distinguish between the images’ semantic similarities. A subject categorized semantically similar images of the same category as used in SIMPLIcity [33] as different based on their size. Therefore taking image size as a selection criterion, only images of size 384x256 were used in our experiment.

The minimum number of commonly selected images from category 1, 2, 5 and 8 was 40 by both groups. The commonly selected image numbers for the other four categories ranged from 45 to 65. The students were made to rank the images based on their semantic similarities. We decided to take equal number of member images from each category by taking the highly ranked images from each. Therefore the total numbers of color photos that

were used as ground truth information were 320, where 40 were selected from each of the 8 semantic categories considered.

Measurement of the effectiveness of clustering methods do not necessarily require larger numbers of ground truth image data sets as efficiency is not a major issue. Most previous research has tried to cluster images aiming at improving retrieval performance and accuracy of results for an image query by users. When the image retrieval performance is the major issue it is required to consider large number of image data sets. Better accuracy is obtained as result of the quality of the clusters formed by implementing the appropriate clustering method.

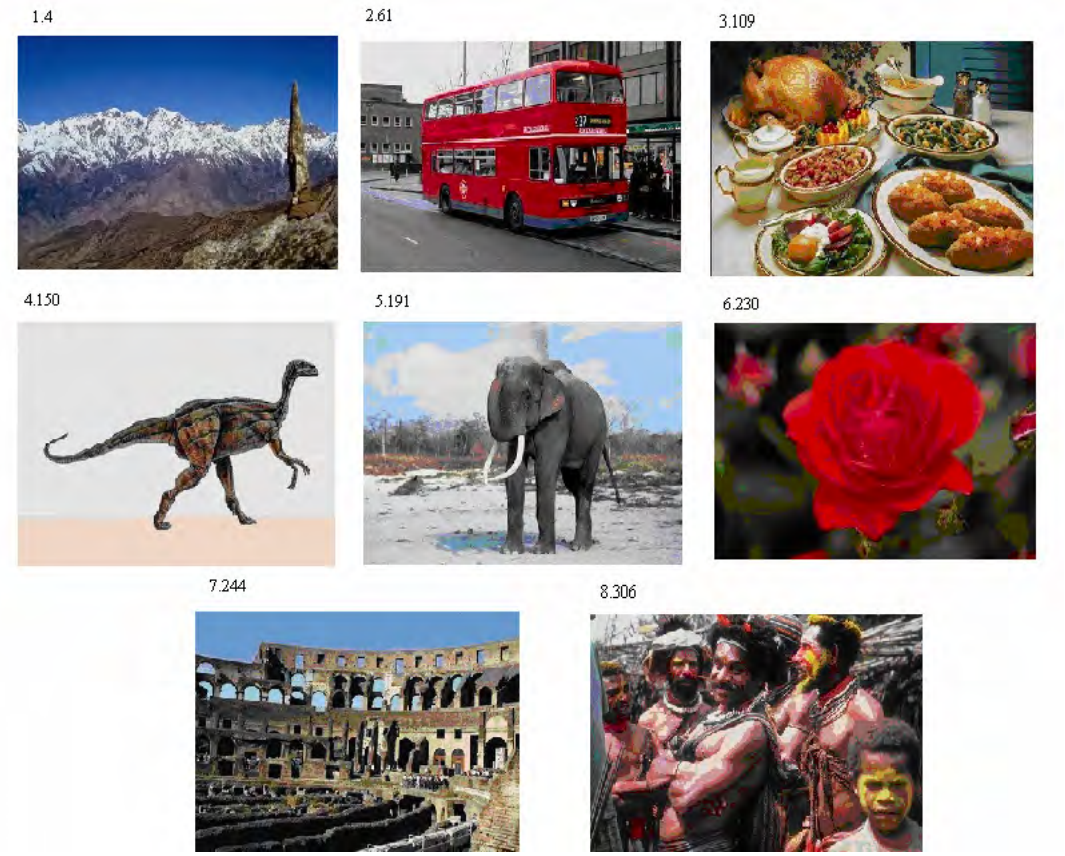

Figure 1 Sample images from each semantic category

Figure 1 shows some sample pictures taken one from each Category as classified in COREL image database. The label at the top of each image identifies the category ID of the image followed by the index assigned to the image to uniquely identify each image in a specific Category.

5.2. Feature Descriptors

Each of the MPEG-7 image descriptors discussed in Chapter 3 extracts some visual properties from the visual media considered. Seven of them were used in representing the selected color image categories. All color descriptors: Color Layout (CLD), Color Structure (CSD), Dominant Color (DCD), and Scalable Color (SCD), two texture descriptors: Edge Histogram (EHD) and Texture Browsing (TBD), and one shape descriptor: Region-based Shape (RSD) is used in comparing similarity. The other basic shape descriptor, Contour-Based Shape, was not used, because it produces almost similar values for all of the semantic categories.

The corresponding bin value of each color descriptor is transformed into a data matrix of 320 rows (color photo images) and 165 columns (bin values of the descriptor element). The two texture descriptors, Edge Histogram and Texture browsing are transformed into a data matrix of 85 columns for the same number of images. Finally the feature vectors the two shape descriptors are transformed into 38 columns. Therefore, each of the images is represented by a total of feature vector of length 288.

Before the beginning of the feature extraction process, each image is indexed randomly independent of its semantic category information. The extraction of each of the four MPEG-7 color descriptors from each images results in the features of the descriptor. Descriptor extraction was performed using the MPEG-7-reference implementation using ACETOOLBOX of M-Ontomat-Annotizer-v0.52 that provides the extracted descriptors in XML format [38, 39]. In the extraction process each descriptor was applied to the whole content of each image of each category and the following extraction parameters were used.

The SCD (Scalable Color Descriptor; Section 3.1.4) was quantized to 64 bins where each bin is represented by 3 bits. The number of coefficients used is 2, but discarding the lowest three bits in the scalable bit representation (i.e., NumberOfBitPlanesDiscarded=”3” [10]. Color in Color Structure was quantized to 32 bins and the HMMD color space was used. Among the five subspaces, the division of the color space used subspace 1 (colorQuant=”1”) at the 32 quantization level.

The YCrCb color space was adopted for the Color Layout Descriptor (CLD, Section 3.1.6). Each of the Cr and Cb coefficients takes 3 bins, but all 6 include the Y coefficient, resulting in a 12-bin representation of the descriptor. Each of these coefficients has one dc value in which each was quantized to 6 bits. The Dominant Color Descriptor (DCD, Section 3.1.5) used eight representative colors that are indexed in a 3-D color space, i.e., YCrCb. It also includes the corresponding color variance value for each index of the color representatives. Including the overall spatial coherency values, Dominant color is quantized to 57 bins.

The description of the Texture Browsing Descriptor (Section 3.2.2) was not suitable for the purpose of this paper. Hence it was transformed to the following form: (regularity, scale: no direction, scale: 0, scale: 30, scale: 60, scale: 90, scale: 120, scale: 150). The possible values for regularity are 0, 1, 2 and 3 corresponds to not regular, slightly regular, regular and highly regular, respectively. Similarly, the values 0, 1, 2, 3 and 4 were used for the corresponding scale bins of no scale, fine, medium, coarse and very coarse respectively. The Edge Histogram Descriptor (Section 3.2.3) was quantized to 80 bins in which each was represented by 3 bits and 35 bins were used for the Region Shape Descriptor (Section 3.3.1).

The number of bits used in representing a bin, that is, an element of a feature vector for each of the eight MPEG-7 descriptors is not the same for all bins. For example, the bin value of the Color Layout Descriptor ranges from 0 to 255 as 8 bits are used in representing each bin of the descriptor. The use of 3 bits in representing a bin of a Region Shape Descriptor allows a maximum value of 7 for its bin. Hence, before measuring the distance between image descriptors, it is necessary to column-wise normalize the resulting values for each vector element, so that it will be in the range [0, 1]. The min-max normalization method used in normalizing the vector element x_{mn} for column n is:

$$X_{mn} = \frac{X_{mn} - X_{\min}}{X_{\max} - X_{\min}} \qquad (6)$$

Where x_{min} and x_{max} are the minimum and maximum values of column n ($x_{min} < x_{max}$). The advantage of such normalization enables preserving the distribution of both rows and columns of the data matrix that is stored in the feature database.

5.3. Similarity Measure

The similarity measures used in measuring similarity among the extracted MPEG-7 feature descriptors for the four color descriptors are from both the quantitative and predicate-based similarity measure domain. A pattern difference which is a predicate-based distance measure is implemented for CSD and SCD. The quantization model [8] is used in computing the distance between each of the image descriptor with all other images. The Meehle Index and the Clark's divergent coefficient are the two quantitative similarity measures used for CLD and DCD.

An empirical evaluation of these four MPEG-7 color descriptors by Ojala et al. [25] in the retrieval of semantic image categories shows that the DCD is worst in retrieving semantic image categories. The combined color similarity measure of all descriptors by assigning more weight to CSD provides the retrieval of the most accurate semantic categories. Based on the recall and precision values, Ojala et al. obtained the weight assigned in the experiment for each of the color descriptors; the appropriate weight is assigned in computing the total color similarity.

The similarity measure used for EHD is the Pattern difference. A total of 80 bins are used in representing the texture feature of an image where 3 bits are used for each bin. Texture Browsing is a compact texture feature descriptor that uses only 12 bits in representing the rich content of images. The Pearson correlation coefficient is used in measuring distance for TBD. The combined texture descriptor similarity needs to assign more weight to EHD [8] due to its higher discriminative power.

The CSD that describes an image using only 3 bins have the same descriptor values for 98% of the images under consideration. Due to its bad discriminative power, this descriptor is not used in measuring the similarity between the color images. The shape feature of the images is represented using only the RSD for which Pattern Difference is used in measuring the shape distance between descriptor values.

The overall similarity among the images is calculated from the summation of the combined color similarity, combined texture similarity and the Region-based shape similarity measures. The relative importance of the color, texture and shape features are not the same

for the color photos used in the experiment. Higher weight is assigned to the color feature descriptor than both texture and shape in measuring the similarity among the color images due to the difference in their number of bins used in their representation.

$$D_{combined} = w_c D_{color} + w_t D_{texture} + w_s D_{shape} \qquad (7)$$

The assigning of weight for each of the three features has considered fundamentally two points. First, the total numbers of bins used in representing each of the features are analyzed. The higher the total numbers of bins of a feature, the larger the weight to be assigned. The other point is the effectiveness of the similarity measure used for each of the feature descriptors [8]. Even though the initial estimate was based on these two reasons, the following optimal results are obtained after several tests. Let n be the total number of bins of a feature and N is the total number of bins of the three features. It follows that the ratio n/N gives the percentage or weight assigned to a given feature. Accordingly; for our experiment in this paper we choose $w_c = 0.45$, $w_t = 0.35$ and $w_s = 0.2$.

5.4. Experimental Results

The implementation of the general k-means and the three HACM (agglomerative hierarchical clustering) algorithms that use average-linkage, complete-linkage and the ward methods for the same input similarity matrix has given different results. When a similarity matrix is input to a clustering algorithm, it assigns a cluster number to each of the image indices as output. The results for each of the clustering algorithms show the corresponding assigned cluster number for each of the indices of the 320 images. This enables the identification of the semantic category of the members of a cluster. For each image cluster formed, the total number of images belonging to the same category is recorded. Therefore the total number of members of a cluster is found from the sum of the total number of each of the constituent categories.

The measure of the cohesiveness of a cluster is best when a cluster consists of member images from a single category type. The smallest possible value 0 corresponds to the best cohesive cluster, that is, equation (2) is evaluated to 0 when all members of a cluster are from one semantic category. The implementation of equation (2) in measuring the degree of

cohesiveness of a cluster always results in the value 0 whether only one or all members of a semantic category belong to a cluster. Table 2 and Table 3 show the result of measuring of the cluster cohesiveness of clusters formed by the corresponding clustering methods used in the experiment when the total color similarity matrix and the total similarity matrix are input, respectively. Hence the value 0 in the total color-based clusters as shown in Table 2 corresponds to a value ranging from 1 to 32 images of the same category in a cluster. This value ranges from 4 to 32 for the cluster cohesiveness measure that input the total similarity matrix as shown in Table 3.

The other interesting result for all clustering methods is when the measured value of cluster cohesiveness is 1.0. This implies a cluster consisting of member images only from two categories with equal number from each.

Cluster	K-means	Average	Complete	Ward
1	1.4355	2.7652	0.0000	0.7532
2	0.7025	0.0000	2.3156	1.7695
3	1.4859	0.4310	0.4537	2.3492
4	2.1713	0.0000	2.2384	0.6582
5	1.7324	0.0000	1.5739	0.4971
6	2.6547	0.2007	0.0000	0.1720
7	1.3792	1.0000	1.2171	0.0000
8	1.5610	0.0000	0.6581	0.0000

Table 2 Cluster cohesiveness measured values for total color similarity measure

The maximum value is obtained for the worst case where each member of all categories is assigned to the same cluster. When equation (2) is implemented for each category of a cluster, it results in a value 0.375, giving the total sum for a cluster 3.0. Therefore, the upper limit value for both Table 2 and Table 3 for cluster cohesiveness is 3.0, which signifies the worst case as a cluster containing all images of each semantic category. The best value 0 on the other hand consists of any number of images, but from a specific category. Therefore the possible value ranges from 0 to 3.0.

Cluster	K-means	Average	Complete	Ward
1	2.6630	0.6889	2.7773	1.2580
2	0.7344	2.7119	0.0000	2.0875
3	1.5510	1.0000	2.1747	2.2738
4	0.4453	1.0000	1.0000	0.6052
5	1.5218	0.0000	1.4825	1.0782
6	2.1426	0.0000	1.0000	1.0000
7	1.1730	1.0000	1.4825	1.0000
8	1.2980	1.0000	2.3220	0.0000

Table 3 Cluster cohesiveness measured values for total similarity measure

The values obtained in the both Table 2 and Table 3 ranges in these intervals, but only values 0 and 1.0000 are calculated for more than one case as shown in both tables. All the other measured values in both tables are different. Table 6 below shows summary information about the total cluster cohesiveness for each of the clustering methods used. It simply integrates all the values of Table 2 and Table 3 in its two columns.

Table 4 and Table 5 show the measured semantic cohesiveness of each Category after implementing the four clustering methods. The values in Table 4 show the results obtained when the total color similarity matrix is input. Table 5 on the other hand shows the results when the input is the total similarity matrix which incorporates the total color similarity. The values in each row show the measured semantic cohesiveness of a Category for the corresponding clustering methods, column-wise after implementing equation (4) for each Category.

Category	K-means	Average	Complete	Ward
1	0.7625	0.4531	1.4890	0.8485
2	1.3762	0.8305	0.6954	0.8539
3	1.9287	1.0368	1.6148	1.2192
4	0.2864	0.0000	0.2864	0.2864
5	1.9584	0.2689	1.8005	1.0638
6	1.3414	0.8822	0.1686	0.9284
7	2.0958	0.1686	0.8538	0.6097
8	1.6128	0.1686	1.0120	1.5229

Table 4 Semantic cohesiveness measured values for total color similarity measure

Category	K-means	Average	Complete	Ward
1	0.4531	0.8286	2.2452	1.1320
2	1.1498	0.4531	2.0144	1.8935
3	1.1935	0.8286	2.2452	1.1320
4	0.9414	0.4531	2.0144	1.8935
5	2.2429	0.2689	1.4001	1.3382
6	1.0295	0.7218	0.8112	0.7691
7	1.9724	0.0000	0.6908	0.5474
8	1.5476	0.0000	1.0869	1.1884

Table 5 Semantic cohesiveness measured values for total similarity measure

Similar to measuring cluster cohesiveness, the value of semantic cohesiveness requires identification of both how many clusters and what amount or number of total images of each Category are distributed. Since the total number of semantically similar images in each Category is 40, we are considering the distribution of each of these images in the different clusters formed.

The semantic cohesiveness of a given Category is best when all of its members are members of only one cluster. The implementation of equation (4) for such Category results in the probability value of $S_{ic}=1$ resulting 0 in the overall result. Unlike the cluster cohesiveness as discussed above, a 0 value of semantic cohesiveness of a Category implies membership includes all members of a Category in a specific cluster.

The cohesiveness of a Category is worst when all image members are distributed in all clusters. The value obtained implementing equation (4) for such a category is 3.0, which is the maximum possible value of cohesiveness of a Category independent of the clustering method used. The upper limit value in measuring the semantic cohesiveness of a Category is 3.0 in which its cohesiveness is worst as its members are distributed equally in all clusters.

The measured values of both semantic cohesiveness tables are in the interval [0.3). The semantic cohesiveness value of two or more image Categories could be the same for different clustering methods. For example the semantic cohesiveness value of Category 1 implementing k-means and that of Category 2 in which average HACM is used is 0.4531. The member images of each of these Categories are distributed in three clusters. The number of images in the three clusters for Category 1 and Category 2 are 2, 1 and 37. The assigned

cluster number of each of the three clusters is not necessarily the same for each Category with respect to the number of images in each cluster. Therefore, similar values of semantic cohesiveness of two or more Categories implies that such Categories are distributed to equal number of clusters and the distribution amount is also the same.

The sum of the semantic cohesiveness of all categories using a specific clustering method gives the total semantic cohesiveness of all categories using that clustering method. This is obtained in column-wise summation of the values of Table 4 and Table 5 and summarized in Table 6 similar to the total cluster cohesiveness. It follows from the description above that each of the values in the total semantic measure and total cluster cohesiveness for each clustering method ranges from 0 to 24.0.

No	Cluster Method	Semantic Cohesiveness		Cluster Cohesiveness		Total Cohesiveness	
		Color Similarity	Total Similarity	Color Similarity	Total Similarity	Color Similarity	Total Similarity
1	K-means	11.3622	10.5302	13.1225	11.5297	24.4847	22.0599
2	Average	3.8087	3.5541	4.3969	7.4008	8.2056	10.9549
3	Complete	7.9205	12.5082	8.4568	13.4302	16.3773	25.9384
4	Ward	7.3328	9.8941	6.1992	9.3027	13.5320	19.1968

Table 6 Overall semantic and cluster cohesiveness measures

5.5. Discussion of Results

The two different similarity matrices that are input to each of the algorithms considered are the total color similarity and the total feature similarity matrices. The clustering algorithms are required to group an input similarity matrix into eight clusters for each of the methods. Semantically similar images are grouped in the same cluster in the best case. The worst

clusters result is when semantically similar images in each cluster are equally distributed among the eight clusters. Therefore, in measuring the effectiveness of the clustering algorithm the number of semantically similar images in each cluster is recorded.

The following part discusses the semantic cohesiveness and cluster cohesiveness measured as a result of the k-means and the three agglomerative hierarchical clustering methods. The discussion analyzes the results obtained when each of total color similarity and total feature similarity matrices are input to the respective clustering algorithms.

5.5.1. Semantic cohesiveness

The distribution of semantic categories for the total color similarity matrix input of the k-means algorithm varies for Category 4 in that its members are distributed only into two clusters. The k-means is worst in partitioning members of Category 7; it partitioned the images of that Category into six different clusters, and also for Category 5 even if the measured value is a little lower.

One of the interesting results in total color-based clustering is that the semantic cohesiveness of Category 4 is best for all clustering methods except for the Complete HACM. Even for the Complete clustering the measured value is the second best next to Category 6. The Average and Complete HACMs are not good in creating cohesive semantics of Category 3. The k-means and Ward produce relatively better cohesive Category 3 semantics, but are lowest in rank with respect to other Categories.

The k-means method partitioned each of the six image Categories with cohesiveness greater than 1.0 into parts ranging from 4 to 6. It is particularly worst for Category 5 and Category 7 where the members are partitioned into six different clusters. The average method is best among all as its worst result partitions each of four of the high-valued categories into four groups each. The complete and Ward methods results in nearly similar semantic cohesiveness values for most categories, but the latter is better. Generally, all the three HACM methods results in better semantically cohesive clusters than k-means for the same total color similarity input matrix.

The total similarity matrix is computed as a weighted sum of the total color, texture and shape descriptors. The total color similarity matrix is one of the constituent parts of the total similarity matrix. It was expected that the addition of texture and shape similarities to the total similarity matrix in a lesser proportion improve both the similarity among images and also the clustering.

The measured value of the semantic cohesiveness of each of the image categories as a result of the implementation for the four respective clustering methods for the total similarity matrix is shown in Table 6.

When each of the eight Categories is considered, the three HACM produce the best cohesive semantics of Category 7. Even though the complete and Ward methods distribute member images of Category 7 in three different clusters, about 80% of the images of each Category belonged to one specific cluster. Similarly, all of the HACM methods also give the next best semantic cohesive values for Category 6 where each method partitioned members of the Category only into two parts.

The semantic cohesiveness of Category 1 and Category 3 are equally worst as a result of implementing average and complete HACM. Especially the complete method distributes the members of each of these Categories on an average equally to six different clusters.

The complete method is not suitable for forming cohesive clusters in the case of the first four categories as shown in Table 5, each with measured value greater than 2.0. This implies that the distribution of each Category into six different clusters, which ranks the complete HACM as the worst among all. The four pairs of similar values measured for the average method also describe the distribution of members of the pairs of Categories that have similar patterns. Unlike all other clustering methods the measured values of the average method are all below 1.0. In the clusters formed by Ward's method, each of four distinct Categories is distributed into three clusters as compared to the k-means which distribute four Categories into more than four different clusters.

Table 6 considers the overall semantic cohesiveness as a measured value of the sum of the semantic cohesiveness of each category with respect to the clustering methods used. The

addition of texture and shape information to the total color similarity matrix is expected to improve the overall cohesiveness of the categories in using the k-means and agglomerative hierarchical clustering methods. But clearly the above results show that there is no significant improvement as the maximum change in the overall measured values is less than 1.0. The texture and shape similarity information degrades the semantic cohesiveness of both complete and Ward methods significantly. The complete HACM is the one that shows the most significant increase in measure of semantic cohesiveness as a result of texture and shape similarity information resulting in higher decrease in quality.

5.5.2. Cluster cohesiveness

The clusters formed by a given clustering method could include images of different Categories as member of a cluster based on the input similarity matrix. A cluster consisting of images of only one semantic Category is the best cohesive cluster as compared to the one that includes images of more than one Category as its members.

The first clusters formed by k-means and complete HACM contain images from all semantic Categories in different proportion. The worst cluster formed by k-means consists of a total of 47 images from the eight categories. The best cohesive clusters with measured values less than 1.0 consist of images from three different Categories. The distribution of image semantic categories is more uniform in the case of average HACM. Two of the clusters formed by average HACM consist of images purely from two Categories, but it forms the worst compact cluster that consists of 226 images from seven different semantic Categories.

The complete HACM is the worst in forming cohesive clusters in which five of the clusters are made up of images from 5 to 8 different image semantics. Two of the other clusters contain equal number of images from two different Categories. This clustering method is the worst as compared to all others in the formation of cohesive clusters.

The four cluster cohesiveness values of the complete HACM each with values grater than 2.0 implies that it is the worst method in forming cohesive clusters as each contains a minimum of seven different image semantic Categories. Ward’s method is relatively better than the complete method in using total feature-based image clustering.

Each of the four clusters formed by the average HACM in the total color-based clustering contains member images from four different categories. The other three clusters also contain a maximum of three different semantic Categories with the last cluster formed from all of the Categories. This makes the average HACM the best method in forming cohesive clusters as compared to all other methods considered.

Most of the clusters formed by k-means are composed of images from more than four Categories. Table 2 shows that only one of its measured values is less than 1.0 as compared to the other clustering methods. This makes k-means the worst method in the formation of cohesive clusters in total color-based clustering.

The overall cohesiveness of the clusters formed by complete and Ward HACMs are comparatively similar. But the cluster cohesiveness by the complete method with values greater than 1.0 describes that each of four of the clusters are formed from images of four and more categories as compared to Ward with only two clusters. This shows that Ward's method is better than complete HACM in the formation of cohesive clusters in total color-based clustering.

The overall measures of cluster cohesiveness of each of the total color-based and total feature-based clustering with respect to the four clustering methods are shown in Table 6. As a result of implementing equation (3); each of the three methods of HACM shows decrease in their respective total cluster cohesiveness due to the addition of texture and shape features information of color photo images to the total color similarity. The increase in cluster cohesiveness in using these features is similar for the Ward and average but larger for complete HACM methods. The k-means is the only method that shows significant improvement in the overall cluster cohesiveness in total feature-based clustering.

6. Conclusions

The thesis addressed the problems of clustering color photo images based on their low-level feature representation. We considered k-means clustering and an agglomerative hierarchical clustering that uses three different cluster linkage methods. CBIR (Content-Based Image Retrieval) systems are required to implement some image clustering algorithm and hence need to select the one which results in the best quality clusters.

We have used the entropy measure of Shannon's information theory to measure the effectiveness of color image clustering. Even if clustering results are affected by the selection of features and proximity measures, the study carefully used the best in the selection process in a way that the algorithms are the only discriminator factor. The homogeneity of a specific image semantic in a cluster measures the effectiveness of color image clustering. The metric we used returns 0 indicating the best semantic cohesiveness.

Among the three HACM used for the total color similarity matrix input, the best quality clusters are formed by the one that uses the average-linkage method in linking pairs of clusters. The quality obtained using this average method is twice that of the complete method for the same hierarchical clustering method. Ward's method produces clusters of similar quality but better than the complete method. The quality of clusters formed by k-means clustering is not better than any of the three hierarchical methods in using total color similarity. The quality of clusters by k-means is three times less than when compared to the best hierarchical method which is the average HACM.

The addition of texture and shape similarity measures (relatively in a lesser weight to the total color similarity) provides a different result. Similar to the total color-based clustering, the average HACM is the best method compared to both the k-means and the other two hierarchical methods in the formation of both semantic and cluster cohesive clusters. Even though the quality of clusters using Ward's method is half of the average method, it results in better quality clusters than the complete method. The k-means clustering method resulted in better cohesive clusters than only the complete hierarchical method when clustering was performed using the three MPEG-7 descriptors of color photo images.

Generally, the hierarchical agglomerative clustering that implements average-linkage in linking clusters gives more than twice as good quality clusters as those of the k-means method regardless of whether total color or total feature similarity is used. Using the complete HACM results k-means to be better than hierarchical method in the formation of quality clusters, Provided that the cluster linkage method was used, the quality of clusters formed by hierarchical clustering is better than k-means.

The other interesting point is that instead of the expected improvement in cluster quality of color photos, the addition of texture and shape feature degrades cluster quality for all hierarchical methods. But using the overall total image similarity matrix resulted in a significant improvement for the k-means method.

We considered validating clustering algorithms that assigns an object exclusively to one cluster. These types of algorithm are not compatible with our everyday life as they do not handle uncertainty of real image's data. Therefore, we propose to validate the fuzzy clustering method that assigns an image as member of several clusters.

References:

[1] M. Abdel-Mottaleb, S. Krishnamachari, and N.Mankovich., “Performance evaluation of clustering algorithms for scalable image retrieval.” ,In Proc.IEEE Workshop on Empirical Evaluation Techniques in Computer Vision, 1998.

[2] P. Berkhin., “Survey of clustering data mining techniques”, Technical report, Accrue Software, San Jose, California, 2002. http://citeseer.nj.nec.com/berkhin02survey.html.

[3] M. Bober, “Mpeg-7 visual shape descriptors”, IEEE Trans. Circuits Syst. Video Technol., vol. 1(6), June 2001

[4] A. Chalechale, G. Naghdy, and A. Mertins, Sketch-Based Image Matching Using Angular Partitioning, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 35, NO. 1, 28-41, JANUARY 2005

[5] Yixin Chen, James Z. Wang, and Robert Krovetz, “CLUE: Cluster-Based Retrieval of Images by Unsupervised Learning”, IEEE TRANSACTIONS ON IMAGE ROCESSING, VOL. 14, NO. 8, pp 1187-1201, AUGUST 2005

[6] Horst Eidenberger, A new perspective on visual information retrieval, Vienna University of Technology,

[7] H. Eidenberger, "A new method for visual descriptor evaluation", Proceedings SPIE Electronic Imaging Symposium, SPIE, San Jose, 2004

[8] H. Eidenberger, “Distance measures for mpeg-7-based retrieval,” in ACM MIR03, 2003.

[9] H. Eidenberger, C. Breiteneder, "Visual similarity measurement with the Feature Contrast Model", Proceedings SPIE Storage and Retrieval for Media Databases Conference, vol. 5021, 64-76, SPIE, Santa Clara, 2003.

[10] H. Eidenberger, "How good are the visual MPEG-7 features?", Proceedings SPIE Visual Communications and Image Processing Conference, vol. 5150, 476-488, SPIE, Lugano, 2003.

[11] David Feng, W C Siu, Hong Jiang Zhang, Multimedia Information Retrieval and Management: Technological Fundamentals and Applications, Springer Published 2003

[12] J. Goldberger, H. Greenspan, and S. Gordon., “Unsupervised image clustering using the information bottleneck method.”, In Proc. DAGM, 2002.

[13] L. Goldmann, M. Karaman, T. Sikora, Human Body Posture Recognition Using MPEG-7 Descriptors, Technical University Berlin, Germany, 2003

[14] C. Grigorescu,, N. Petkov, Distance Sets for Shape Filtersand Shape Recognition, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 12, NO. 10, 1274-1286 OCTOBER 2003

[15] Grira, N., Crucianu, M., Boujemaa, N. (2004) Unsupervised and semi-supervised clustering: a brief survey, July 2004, 11 p., in A Review of Machine Learning Techniques for Processing Multimedia Content, report of the MUSCLE European Network of Excellence (FP6)

[16] M. Halkidi, Y. Batistakis, M. Vazirgiannis. Clustering Algorithms and Validity Measures, Proceedings of the 13th International Conference on Scientific and Statistical Database Management, Pages: 3 – 22, 2001

[17] Daniel Heesch, Alexei Yavlinsky, and Stefan M. Rüger. Performance Comparison of Different Similarity Models for CBIR with Relevance Feedback. In E. Bakker, T. Huang, M. Lew, N. Sebe, and X. Zhou, editors, Proceedings of the International Conference on Image and Video Retrieval (CIVR’03), volume 2728 of Lecture Notes in Computer Science, pages 456–466, Urbana-Champaign, IL, USA, July 2003. Springer.

[18] N. Iyer, S. Jayanti, K. Lou, Y. Kalyanaraman, K. Ramani , Three-dimensional shape searching: state-of-the-art review and future trends, Computer-Aided Design 37 (2005) 509–530

[19] S. Krishnamachari and M. Abdel-Mottaleb. Hierarchical clustering algorithm for fast image retrieval. In Proc. SPIE Conference on Storage and Retrieval for Image and Video databases VII, pages 427–435, 1999.

[20] J. Latecki, R. LakÄamper and D. Wolter, Shape Similarity and Visual Parts, Discrete Geometry for Computer Imagery, Nov. 2003

[21] B. S. Manjunath, Jens-Rainer Ohm, Vinod V. Vasudevan, and Akio Yamada, Color and Texture Descriptors, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 11, NO. 6, 703-715, JUNE 2001

[22] Martin H. C. Law , Mario A. T. Figueiredo , Anil K. Jain, Simultaneous Feature Selection and Clustering Using Mixture Models, IEEE Transactions on Pattern Analysis and Machine Intelligence, v.26 n.9, p.1154-1166, September 2004

[23] Vinay Modi, Color Descriptors from Compressed Images, 2006

[24] JR Ohm, F Bunjamin, W. Liebsch, B. Makai, K Müeller, A. Smolic, D. Zier, Set of visual feature descriptors and their combination in a low-level description scheme, J, Signal Processing: Image Communication, 2000

[25] T Ojala, M Aittola, E Matinmikko - Empirical Evaluation of MPEG-7 XM Color Descriptors in Content-Based Retrieval of Semantic Image Categories, Proc. 16th International Conference on Pattern Recognition, 2002

[26] Gunhan Park, Yunju Baek, Heung-Kyu Lee, “Re-ranking algorithm using post-retrieval clustering for content-based image retrieval” Information Processing and Management, August 2003

[27] Remco C. Veltkamp, Longin Jan Latecki, Properties and Performance of Shape Similarity Measures, 2005

[28] Yong Man Ro, Munchurl Kim, Ho Kyung Kang, B.S. Manjunath, and Jinwoong Kim MPEG-7 Homogeneous Texture Descriptor, ETRI Journal, Volume 23, Number 2, 41- 51, June 2001

[29] Yong Rui and Thomas S. Huang, Image Retrieval: Current Techniques, Promising Directions, and Open Issues, Journal of Visual Communication and Image Representation 10, 39–62 (1999)

[30] Yossi Rubner, Jan Puzicha, Carlo Tomasi, and Joachim M. Buhmann, Empirical Evaluation of Dissimilarity Measures for Color and Texture, Computer Vision and Image Understanding 84, 25–43 (2001)

[31] Philippe Salembier, OVERVIEW OF THE MPEG-7 STANDARD AND OF FUTURE CHALLENGES FOR VISUAL INFORMATION ANALYSIS, Universitat Polit`ecnica de Catalunya, SPAIN, 2003

[32] Thomas Sikora, The MPEG-7 Visual Standard for Content Description—An Overview, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 11, NO. 6, 696-702, JUNE 2001

[33] J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: semantics-sensitive integrated matching for picture libraries," IEEE Trans. attern Anal. Mach. Intell., vol. 23, no. 9, pp. 947–963, Sep. 2001.

[34] Chee Sun Won, Dong Kwon Park, and Soo-Jun Park, Efficient Use of MPEG-7 Edge Histogram Descriptor, ETRI Journal, Volume 24, Number 1, 23-30, February 2002

[35] KM Wong, LM Po, A NEW PALETTE HISTOGRAM SIMILARITY MEASURE FOR MPEG-7 DOMINANT COLOR DESCRIPTOR International Conference on Image Processing (ICIP), 2004

[36] P. Vácha, Texture Similarity Measure, WDS'05 Proceedings of Contributed Papers, Part I, 47–52, 2005.

[37] D. Zhang, G. Lu, EVALUATION OF SIMILARITY MEASUREMENT FOR IMAGE RETRIEVAL, Gippsland School of Computing and Info Tech Monash University, 2003

[38] NE O'Connor, E Cooke, H Le Borgne, M Blighe, T Adamek, THE ACETOOLBOX: LOW-LEVEL AUDIOVISUAL FEATURE EXTRACTION FOR RETRIEVAL AND CLASSIFICATION, 2nd IEE European Workshop on the Integration of Knowledge, 2005

[39] The AceMedia Project, available from http://www.acemedia.com, Last visited September 2006.

[40] [40] G. Sheikholeslami, W. Chang, and A. Zhang, “SemQuery: semantic clustering and querying on heterogeneous features for visual data,” IEEE Trans. Knowl. Data Eng., vol. 14, no. 5, pp. 988–1002, May 2002.

[41] J. Li and J. Z. Wang, “Automatic linguistic indexing of pictures by a statistical modeling approach,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 9, pp. 1075–1088, Sep. 2003.

[42] W.Y. Ma and B. Manjunath, ªNaTra: A Toolbox for Navigating Large Image Databases, Proc. IEEE Int'l Conf. Image Processing, pp. 568-571, 1997.

[43] C. Carson, M. Thomas, S. Belongie, J.M. Hellerstein, and J. Malik, Blobworld: A System for Region-Based Image Indexing and Retrieval, Proc. Visual Information Systems, pp. 509-516, June 1999.

[44] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom et al. Query by Image and Video Content: The QBIC System, IEEE Computer, vol. 28, no. 9, 1995.