CLUSTERING SIMILAR IMAGES USING VARIOUS IMAGE FEATURES

Dr. E. S. Samundeeswari[1], M. Kiruthika[2]

1 Associate Professor, Dept. of Computer Science, Vellalar College for Women, Erode, Tamilnadu, India
2 Research Scholar, Dept. of Computer Science, Vellalar College for Women, Erode, Tamilnadu, India

ABSTRACT: Clustering is an unsupervised classification of patterns into groups (clusters). The images can be clustered into group of visually similar images and can be used in applications like extracting images similar to the query image. The proposed system uses descriptors such as pixel information, foreground/background features, texture features for grouping similar objects. The cluster validation methods applied are internal and stability measures. It includes Connectivity, Silhouette, Dunn index in internal measures and Average Proportion of Non-overlap (APN), Average Distance (AD), Average Distance between Means (ADM), Figure of Merit (FOM) in Stability measures. These methods are used to compare multiple clustering algorithms such as K-Means Clustering, Hierarchical Clustering and PAM clustering for identifying the best clustering approach and the optimal number of clusters. Out of the three Clustering methods, K-Means Clustering provides the best cluster result. The texture feature produces better cluster result when compared to pixel information, foreground/background feature and combination of both texture and foreground/background feature. The system is user friendly and developed using 'R'.

KEYWORDS: Clustering, Descriptors, Hierarchical, K- Means, PAM, Silhouette Cluster Validation, Textures.

I. INTRODUCTION

Content based image retrieval (CBIR) is a technology that in principle helps to organize digital pictures by their optical content. Everything ranging from image similarity function to a robust image annotation engine fall under the Content Based Image Retrieval. It is an application of computer vision to the image retrieval problem and arrangement with the difficulty of searching for digital images in huge databases. CBIR commonly works on the centre of uncertainty by using an image or a part of an image.

Various types of cluster algorithms are extended and these algorithms may vary depending on the application. Content-based image retrieval uses the visual contents of an image such as colour, shape, texture, and spatial layout to represent and index the image.

Features are the information extracted from images in terms of numerical values. Generally, features extracted from an image are of much more lower dimension than the original image. The reduction in dimensionality reduces the overheads of processing the bunch of images. Basically there are two types of features extracted from the images based on the application. They are local and global features. Global features describe the image as a whole to generalize the entire object whereas the local features describe the image patches of an object. Global features include Contour Representations, Shape Descriptors, and Texture features and Local features represent the texture in an image patch. Shape Matrices, Invariant Moments, Histogram Oriented Gradients (HOG) and Co-HOG are some examples of global descriptors. Scale Invariant Feature Transform (SIFT), Speeded-Up Robust Features (SURF), Local Binary Patterns (LBP), Binary Robust Invariant Scalable Key points (BRISK), Maximally Stable External Regions (MSER) and Fast Retina Key points (FREAK) are some examples of local descriptors.

Generally, low level applications such as object detection and classification in global features and higher level applications such as object recognition in Local features are used. Combination of global and local features improves the accuracy of the recognition with the side effect of computational overheads.

Feature detection and extraction are frequently united to resolve computer vision efforts such as object detection and recognition face detection and recognition and texture classification. There are two key approaches to select the features (variables) for the analysis: the minimal-optimal feature selection which recognizes a small (ideally minimal) set of variables that gives the best classification result (for a class of classification models) and the all-relevant feature selection which recognizes all variables that are in some situations relevant for the classification. Automatic feature selection techniques can be used to build many models with different subsets of a dataset and find those attributes that are and are not required to build an accurate model.

II. RELATED WORKS

Chi Zhang et al., (2014) commented that the algorithms of Content-Based Image Retrieval (CBIR) have been well developed along with the explosion of information. These algorithms are mainly differentiated based on features used to describe the image content. The algorithms are based on color feature and texture feature. Color Coherence Vector based image retrieval algorithm is also attempted in the implementation process, but the best result is generated from the algorithms that weights color and texture. 80% satisfying rate is achieved.

Guy Brock et al., (2011) described the R package clValid containing functions for validating the results of a clustering analysis. There are three main types of cluster validation measures available, "internal", "stability", and "biological". The user can choose from nine clustering algorithms in existing R packages, including hierarchical, K-means, Self Organizing Maps (SOM), and model based clustering. In addition, a function to perform the self-organizing tree algorithm (SOTA) method of clustering is provided. Any combination of validation measures and clustering methods can be requested in a single function call. Additionally, the package can automatically make use of the biological information contained in the Gene Ontology (GO) database to calculate the biological validation measures, via the annotation packages available in Bioconductor.

Khalid Imam Rahmani et al., (2014) discussed that clustering makes the job of image retrieval easy by finding the images as similar given in the query image. The images are grouped together in some given number of clusters. They are grouped on the basis of some features such as colour, texture, shape etc. contained in the images. For the purpose of efficiency and better results image data are segmented before applying clustering. The techniques used are K-Means and Fuzzy K-Means.

Lining Zhang et al., (2016) proposed a novel Discriminative Semantic Subspace Analysis (DSSA) method, which can directly learn a semantic subspace from related and unrelated pair-wise constraints without using any explicit class label information. In particular, DSSA can effectively integrate the local geometry of labeled related images, the discriminative information between labeled related and unrelated images, and the local geometry of labeled and unlabeled images together to learn a reliable subspace. Compared with the popular distance metric analysis approaches, this method can also learn a distance metric but perform more effectively when dealing with high-dimensional images. Extensive experiments on both the synthetic data sets and a real-world image database demonstrate the effectiveness of the proposed scheme in improving the performance of the CBIR.

Maria Halkidi et al., (2011) discussed about the fundamental concepts of clustering and surveyed and compared the widely known clustering algorithms. Moreover, the paper addressed an important issue of clustering process regarding the quality assessment of the clustering results. A review of clustering validity measures and approaches available in the literature were also presented.

Syed Hamad Shirazi et al., (2016) discussed that interests to accurately retrieve required images from databases of digital images are growing day by day. Images are represented by certain features to facilitate accurate retrieval of the required images. These features include Texture, Color, Shape and Region. This paper presented a literature survey of the Content Based Image Retrieval (CBIR) techniques based on Texture, Color, Shape and Region. It also reviewed some of the state of the art tools developed for CBIR.

III.METHODOLOGY

3.1 ARCHITECTURE DIAGRAM

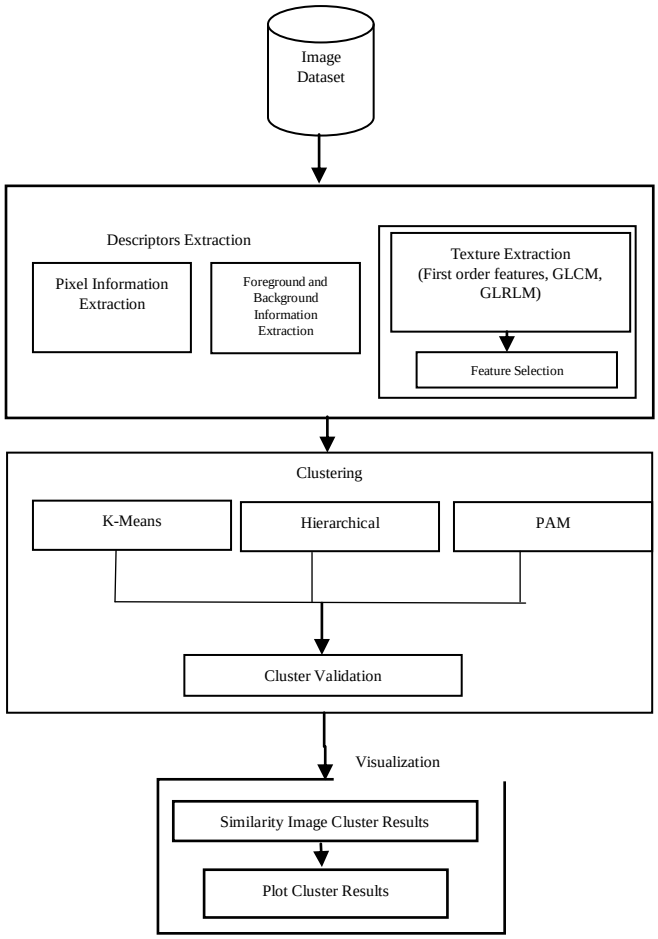

Fig 3.1 System Architecture

3.2 SYSTEM METHODOLOGY

3.2.1 FEATURE EXTRACTION

Feature extraction is the process by which certain features of interest within an image are detected and represented for further processing. The feature is defined as a function of one or more measures, each of which specifies some measureable property of an object, and is computed. This work deals with following features,

- Pixel Information of Image
- Foreground/Background of Image
- Texture of Image

3.2.2 PIXEL INFORMATION FEATURE

The Pixel information feature is to compare the images by intensity of pixel at the same coordinates within the images. In this work 30 images are processed. All pairs of images are compared pixel by pixel and the ratio of similar to total number of pixels is calculated and stored in a matrix. The diagonal elements are 100% showing the similarity of an image with itself. Here intensity refers to the amount of numerical value of a pixel. In grayscale images depicted by the grey level value of each pixel. The intensity of an image could refer to a global measure of that image, such as mean pixel intensity.

3.2.3 FOREGROUND/BACKGROUND FEATURES

The Foreground/Background feature is used to classify the object based on the background model. The classification of background and foreground modelling method which utilizes the background similarity to fuse color and texture feature. The object is extracted from converted greyscale image. Otsu's thresholding method is useful to automatically perform clustering based on image thresholding. The algorithm assumes that the distribution of image pixel intensities follows a bi-model histogram, and separates the pixels into two classes (e.g. Foreground and Background). The optimal threshold value is determined by minimizing the combined intra-class variance. The threshold values of all images are stored in a matrix as the feature matrix.

3.2.4 TEXTURE FEATURES

Texture is a very nebulous concept, often attributed to human perception, as either the feel or the appearance of (woven) fabrics. Image will usually contain samples of more than one texture. Texture descriptions are measurements that characterize a texture and then classification is attributing the correct class label to a set of measurements and then, perhaps to segment an image according to its texture content.

Following Texture Features are considered in this work,

- First Order Features
- Grey Level Co-occurrence Matrix
- Grey Level Run Length Matrix

1. First Order Features

First-order features rely only on the values of specific pixels in the image, and do not express their relationship to other image pixels. For example, the mean/median/minimum/maximum pixel values in the image falls under this category. The first-order features are Energy, Entropy, Kurtosis, Mean, Median, Mean deviation, Minimum, Maximum, Variance, Uniformity, Skewness, Root Mean Square and Standard Deviation.

2. Grey Level Co-occurrence Matrix (GLCM)

The GLCM considers the spatial relationships between two pixels in the image at a time (the reference and the neighbor pixel). The distance between the reference and neighbor pixel can also be chosen. The matrix is built such that each row represents a pixel (reference) in the image and each column represents a pixel (neighbor). The features calculated are Energy, Entropy, Mean, Variance, Contrast, Correlation, Auto Correlation, Cluster Prominence, Cluster Shade, Cluster Tendency, Difference Entropy, Dissimilarity, Homogeneity1, Homogeneity2, Inverse Difference Moment, Inverse Variance, Maximum Probability, Sum Average, Sum Entropy, and Sum Variance.

3. Grey Level Run Length Matrix (GLRLM)

The row of the GLRLM represents again grey levels in the image. However, the columns represent length of the runs, with the entries corresponding to the number of runs of the given length in the image. The features calculated are Grey Level Non-uniformity (GLN), Long Run Emphasis (LRE), High Grey Level Run Emphasis (HGLRE), Long Run Low Grey Level Emphasis (LRLGLE), Low Grey Level Run Emphasis (LGLRE), Run Length Non-uniformity (RLN), Run Percentage, Short Run Low Grey Level Emphasis (SRLGLE), Short Run Emphasis (SRE), Short Run High Grey Level Emphasis (SRHGLE) and Long Run High Grey Level Emphasis (LRHGLE).

3.3 FEATURE SELECTION

Feature selection is the process of selecting a subset of relevant features in model construction. The texture feature set has many features such as First order features, GLCM and GLRLM features. The First order features has nine features, GLCM has twenty one features and the GLRLM has eleven features. A total of 41 features were calculated as texture features. To overcome the curse of dimensionality, the feature set is reduced using correlation.

3.3.1 Correlation Matrix

Correlation is one of a broad class of statistical relationships involving dependence. It

often refers to how close two variables are having a linear relationship with each other represented by correlation coefficients. findCorrelation() method to calculates the correlation matrix of data attributes. This function searches through the correlation matrix and returns a vector of integers that are highly correlated. All the highly correlated features are removed from the feature set.

3.4 CLUSTERING TECHNIQUES

3.4.1 K–Means clustering

K-Means algorithm is the most popular clustering algorithm. It iteratively computes the clusters and their centroids.

ALGORITHM

Input:
D= {$t_1, t_2, \ldots, t_n$} // set of elements
K // no of desired clusters
Output:
K // set of clusters
K-Means Algorithm:
assign initial values for means $m_1, m_2, \ldots m_k$;
repeat
  assign each item t_1 to the cluster which has the closest mean;
until convergence criteria is met;

K-Mean algorithm is a top down approach to clustering. It is used for creating and analyzing the clusters with 'n' number of data points point is divided into 'K' clusters based on the similarity measurement criterion. The results generated using the algorithm mainly depends on initial cluster centroids chosen.

3.4.2 Hierarchical Algorithm

Hierarchical clustering algorithm actually creates a sets of clusters. Hierarchical algorithm differs in how the sets are created. A tree data structure, called a dendrogram, can be used to illustrate the hierarchical clustering technique and the sets of different clusters. The root in a dendrogram tree contains one cluster where all elements are together. The leaves in the dendrogram contain a single element cluster. Internal nodes in the dendrogram represent new clusters formed by merging the clusters that appear as its children in the tree. Each level in the tree is associated with the distance measure that was used to merge the clusters. All clusters created at a particular level were combined because the children clusters had a distance between them less than the distance value associated with the level in the tree.

3.4.3 PAM Algorithm

The PAM (Partitioning Around Medoids) algorithm also called the K-medoids algorithm represents a cluster by medoid. Initially, an arbitrary set of k items is taken to be the set of medoids. Then at each step, all items from the input dataset that are not currently medoids are examined one by one to see if they should be medoids. That is, the algorithm determines whether there is an item that should replace one of the existing medoids.

ALGORITHM

Input:
  D= {$t_1, t_2, \ldots, t_n$} // set of elements
  A // adjacency matrix showing distance between elements
  K // no of desired clusters

Output:
  K // set of clusters
PAM algorithm:
  arbitrarily select k medoids from D;
  repeat
    for each t_h not a medoid do
      for each medoid t_i do
        calculate TC_{ih};
    find i, h where TC_{ih} is the smallest;
    if $TC_{ih} < 0$, then
      replace medoid t_i with t_h;
  until $TC_{ih} \geq 0$;
  for each $t_i \in D$ do
    assign t_i to K_j, where dis(t_i, t_j) is the smallest over all medoids;

3.5 EVALUATION METRICS FOR CLUSTERING

The procedure of evaluating the results of a clustering algorithm is known as cluster validity. Two cluster validation measures are

1. **Internal Measures** use basic information in the data to measure the quality of the clustering.
2. **Stability Measures** assesses the consistency of a clustering result by matching it with the clusters obtained after each column is removed, one at a time**.**

The internal measures include the Connectivity, and Silhouette Width, and Dunn Index. The connectivity indicates the degree of connectedness of the clusters, as determined by the k-nearest neighbors. The connectivity has a value between 0 and infinity and should be minimized. Both the Silhouette width and the Dunn Index combine measures of compactness and separation of the clusters.

The Dunn Index is the ratio between the smallest distances between observations not in the same cluster to the largest intra-cluster distance. It has a value between 0 and infinity and should be maximized. Silhouette validation is validating the result of clustering to find the accuracy of the obtained results from the cluster value. Silhouette cluster validation range from -1 to 1. The Silhouette cluster interpretation result is,

- 0.71 – 1.00 excellent split
- 0.51 – 0.71 reasonable structure has been found

- 0.26 – 0.50 weak structure, could be artificial
- ≤0.25 horrible split

Cluster stability measures include

- The Average Proportion of Non-overlap (APN)
- The Average Distance (AD)
- The Average Distance between Means (ADM)
- The Figure of Merit (FOM)

The APN, AD and ADM are all based on the cross-classification table of the original clustering of the full data with the clustering based on the removal of one column.

- The APN measures the average proportion of observations not placed in the same cluster.
- The AD measures the average distance between observation placed in the same cluster under both the cases(full dataset and removal of one column)
- The ADM measures the average distance between cluster centers for observations placed in the same cluster under both cases
- The FOM measures the average intra-cluster variance of the deleted column, where the clustering is based on the remaining (undeleted) columns.

The values of APN, ADM and FOM ranges from 0 to 1, with smaller value corresponding to highly consistent clustering results. AD has a value between 0 and infinity, and smaller values are also preferred.

IV. RESULT AND DISCUSSION

The proposed system is experimented using R tool. Three descriptors were used and three clustering techniques were used.

4.1 PIXEL INFORMATION FEATURE

The images are compared on Pixel basis and the ratio of similar pixels are calculated (Table 4.1) using the formula,

s=100*sum (img1==img2)/length (img1)

Table 4.1 Similarity Matrix

	./dataset_jpg/1.jpg	./dataset_jpg/10.jpg	./dataset_jpg/11.jpg	./dataset_jpg/12.jpg	./dataset_jpg/13.jpg	./dataset_jpg/14.jpg	./dataset_jpg/15.jpg
1	100.000000	0.00000	0.000000	0.0000000	0.0000000	0.0000000	0.0000000
2	0.000000	100.00000	36.855159	0.0000000	0.0000000	0.0000000	0.0000000
3	0.000000	36.85516	100.000000	0.0000000	0.0000000	0.0000000	0.0000000
4	0.000000	0.00000	0.000000	100.0000000	2.7794732	0.7059733	1.6303168
5	0.000000	0.00000	0.000000	2.7794732	100.0000000	0.3268772	4.1683621
6	0.000000	0.00000	0.000000	0.7059733	0.3268772	100.0000000	0.3997803
7	0.000000	0.00000	0.000000	1.6303168	4.1683621	0.3997803	100.0000000
8	0.000000	0.00000	0.000000	1.9388835	2.4478488	0.3295898	2.6421441
9	0.000000	0.00000	0.000000	1.5316433	1.2481689	1.2356228	1.7022027
10	0.000000	0.00000	0.000000	0.0000000	0.0000000	0.0000000	0.0000000
11	40.000000	0.00000	0.000000	0.0000000	0.0000000	0.0000000	0.0000000
12	58.888889	0.00000	0.000000	0.0000000	0.0000000	0.0000000	0.0000000
13	0.000000	27.12054	7.502480	0.0000000	0.0000000	0.0000000	0.0000000
14	0.000000	53.58383	13.318452	0.0000000	0.0000000	0.0000000	0.0000000
15	0.000000	37.09077	7.762897	0.0000000	0.0000000	0.0000000	0.0000000

4.2 FOREGROUND/BACKGROUND FEATURES

Using OTSU() method, three threshold values are obtained. These values are used to form the matrix for all the images is shown in Table 4.2

Table 4.2 Threshold Value Matrix

	V1	V2	V3
1	0.4160156	0.3886719	0.24023438
2	0.4941406	0.4394531	0.15820312
3	0.3574219	0.2988281	0.17382812
4	0.6542969	0.4863281	0.14257812
5	0.6386719	0.5410156	0.23632812
6	0.5644531	0.5175781	0.07617188
7	0.2285156	0.2089844	0.20507812
8	0.5957031	0.3886719	0.16601562
9	0.3417969	0.2128906	0.10742188
10	0.5917969	0.3144531	0.05664062
11	0.3378906	0.2714844	0.21679688
12	0.4199219	0.1425781	0.11132812
13	0.4980469	0.3261719	0.26367188
14	0.5410156	0.5175781	0.15039062
15	0.5332031	0.5136719	0.42382812

4.3 TEXTURE FEATURES

First Order Features calculated for all the images in the dataset are Energy, entropy, kurtosis, skewness and median features. GLCM calculation includes the value of GLCM mean, GLCM_variance, GLCM_autoCorrleations, GLCM_cProminence, GLCM_CShade, GLCM_cTendency, GLCM_contrast, GLCM_sumAverage and GLCM_sumVariance. The GLRLM features are GLN, HGLRE, LRE, LRHGLE, LRLGLE, RLN and SRHGLE values. The values calculated from these three types of features are combined and Correlation is applied to select the not highly correlated features from the features dataset.

The selected features (not highly correlated) are entropy variation, GLCM variance, GLCM contrast, GLRM_GLN, GLRM_LRLGLE and GLRM_SRHCLE. The snap shot of these values are shown in Table 4.3

Table 4.3 Reduced Texture Feature Set Matrix

	d.calc_entropy	d.glcm_variance	d.glcm_contrast	d.res_glrlm.glrlm_GLN	d.res_glrlm.glrlm_LRLGLE	d.res_glrlm.glrlm_SRHGLE
1	7.413361	59.713424	15.3332403	174.5453	1.4440550	178.48387
2	7.128209	54.422636	4.1699219	140.5967	5.3890254	144.03624
3	6.857108	39.165335	4.6343936	235.3386	6.7957106	87.86648
4	7.468687	34.070943	8.6424851	326.2587	1.1269950	136.42749
5	7.246699	64.737660	17.0214844	236.9017	0.2401186	264.50584
6	6.422768	112.633939	9.0389695	183.4283	5.2983683	204.56350
7	6.841500	36.445124	3.2252604	134.8937	5.2560787	83.88341
8	7.018418	79.421737	10.5478051	158.1593	1.1178991	153.19565
9	6.773528	50.418088	3.9250372	155.8853	25.8087502	67.81614
10	7.111209	32.264928	2.5920759	128.6135	1.9925685	50.56545
11	7.095384	52.672927	8.0185082	201.1742	0.9338877	178.21206
12	6.980880	29.161026	5.3323103	237.1671	5.4561852	63.33458
13	7.478242	51.103217	13.6875930	212.5251	1.4510059	223.78745
14	7.271236	33.536930	4.5817522	175.7510	2.4650318	83.93835
15	7.077696	35.307572	2.9686570	174.9848	1.5600119	93.65270
16	6.915769	45.172142	5.4179688	162.5246	3.0901350	112.29268

Showing 1 to 16 of 100 entries

4.4 COMBINATION OF TEXTURE AND FOREGROUND/BACKGROUND

The texture and foreground/background features were combined. Since the number of features is more, feature selection is carried out using Correlation matrix. The values calculated from foreground/background and texture features are combined and Correlation is applied to select the uncorrelated features from the features dataset. The selected features (not highly correlated) are X1(Threshold), kurtosis, GLCM variance, GLCM contrast, GLRM_GLN, GLRM_LRLGLE and GLRM_RLN GLRM_SRHGLE and is shown in Table 4.4.

Table 4.4 Reduced Foreground/Background and Texture Features set matrix

	d1.X1	d1.calc.kurtosis	d1.glcm_variance	d1.glcm_contrast	d1.res_glrlm.glrlm_GLN	d1.res_glrlm.glrlm_LRLGLE	d1.res_glrlm.glrlm_RLN	d1.res_glrlm.glrlm_SRHGLE
1	column 1: numeric with range 0.1543 - 0.68164	591624721	59.713424	15.3332403	174.5453	1.4440550	1864.6689	178.48387
2		507611592	54.422636	4.1699219	140.5967	5.3890254	1054.1279	144.03624
3	0.3574219	2.793017459	39.165335	4.6343936	235.3386	6.7957106	1714.6323	87.86648
4	0.6543969	0.507269435	34.070943	8.6424851	326.2587	1.1269950	2390.9130	136.42749
5	0.6386719	-1.667094745	64.737660	17.0214844	236.9017	0.2401186	1521.5815	264.50584
6	0.5644531	-1.280972784	112.633939	9.0389695	183.4283	5.2983683	1397.5146	204.56350
7	0.2285156	2.510125599	36.445124	3.2252604	134.8937	5.2560787	741.1070	83.88341
8	0.5957031	-0.711742479	79.421737	10.5478051	158.1593	1.1178991	943.6001	153.19565
9	0.3417969	-0.348676500	50.418088	3.9250372	155.8853	25.8087502	806.4160	67.81614
10	0.5917969	-0.748655798	32.264928	2.5920759	128.6135	1.9925685	477.5096	50.56545
11	0.3378906	0.530384218	52.672927	8.0185082	201.1742	0.9338877	1228.6203	178.21206
12	0.4190219	3.147556183	29.161026	5.3323103	237.1671	5.4561852	1269.0118	63.33458
13	0.4980469	0.438327190	51.103217	13.6875930	212.5251	1.4510059	2003.2662	223.78745
14	0.5410156	2.368039181	33.536930	4.5817522	175.7510	2.4650318	928.0019	83.93835
15	0.5332031	0.489492708	35.307572	2.9686570	174.9848	1.5600119	834.0284	93.65270
16	0.3300781	-0.153817493	45.172142	5.4179688	162.5246	3.0901350	1095.1485	112.29268
17	0.4238281	1.414007094	19.012697	2.4453125	209.2395	7.3512196	687.3712	38.30235
18	0.4550781	1.913567849	32.542000	9.9165737	188.9376	1.5191118	1081.5609	139.54159
19	0.5019531	0.310925741	41.956291	16.1768973	239.2510	0.8025157	1404.2090	235.23006

K- Means, Hierarchical and PAM clustering algorithms were applied and compared. The experimental analysis shows K- means clustering algorithm provide better results when compared to Hierarchical and PAM clustering algorithms is shown in Tabe 4.5.

Table 4.5 Validation Result of clusters

Clustering Methods	Validation Measures	Cluster sizes	
		5	6
Hierarchical	Connectivity	24.0044	28.0544
	Dunn	0.1616	0.1816
	Silhouette	0.4438	0.4185
K-Means	Connectivity	**22.0750**	36.1139
	Dunn	**0.1834**	0.1784
	Silhouette	**0.4537**	0.4170
PAM	Connectivity	26.0433	40.5698
	Dunn	0.1579	0.0998
	Silhouette	0.4347	0.3959

Optimal Scores:			
	Score	Method	Clusters
Connectivity	22.0750	K-Means	5
Dunn	0.1834	K-Means	5
Silhouette	0.4537	K-Means	5

Silhouette plot diagram represents the average value of Silhouette width for pixel comparision. The average value of 0.24 is obtained using this descriptor.

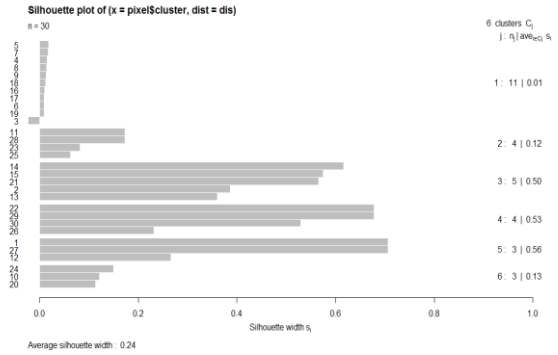

Fig 4.1 Silhouette plot of clusters formed using Pixel Information

The average value of 0.45 is obtained using Foreground/Background descriptor is shown in Fig 4.2

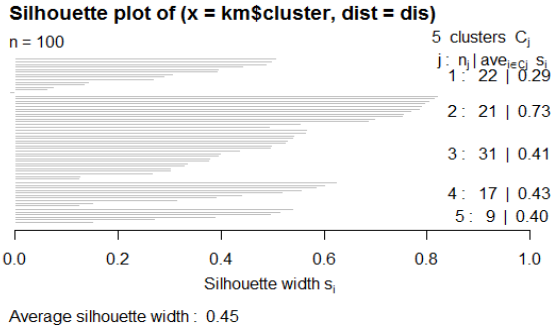

Fig 4.2 Silhouette plot of clusters formed using Foreground/Background Features

The average value of 0.57 is obtained using texture features are shown in Fig 4.3.

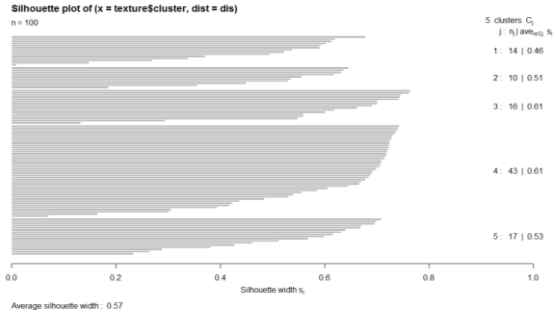

Fig 4.3 Silhouette Plot of clusters formed using TextureFeatures

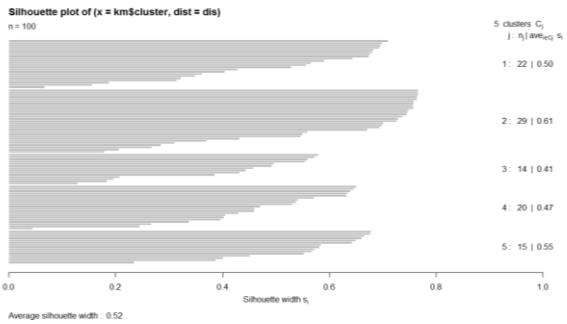

Fig 4.4 Silhouette plot of clusters formed using both Foreground/Background and Texture Features

The average value of 0.52 is obtained using both foreground/background and texture features.

From experiments, it can be concluded that the accuracy of k-means for image dataset is having good evaluation much better than the hierarchical and PAM clustering. A good clustering method produces high-quality clusters to ensure that the inter-cluster similarity is low and the intra-cluster similarity is high.

4.5 COMPARISON RESULT

Table 4.6 shows cluster validation of K-Means clustering for different types of descriptors. Finally Texture feature provides the best cluster result, when compared to other features.

S.No	Features	Silhouette Value (-1 to +1)
1	Pixel Information Feature	0.24
2	Foreground/Background Feature	0.45
3	Texture Feature	0.57
4	Combination of Both Foreground/Background and Texture	0.52

Table 4.6 Comparison Results

Here three different types of descriptors – Pixel Information, Foreground/Background and Texture Features, applications of three clustering methods – K-Means, Hierarchical and PAM clustering for two cluster sizes – 5 and 6. Internal as well as Stability cluster validation methods are used for validating the clusters.

V. CONCLUSION AND FUTUREWORK

Content based image retrieval system is concerned with the representation, storage and retrieval of digital images from a large database. The proposed work mainly focused on grouping similar images based on image descriptors. Three descriptors - Pixel Information, Foreground/Background Feature, Texture Feature and the combination of both Foreground/Background and Texture Feature are used in the proposed system and three clustering algorithms such as K-Means, Hierarchical and PAM clustering algorithms are used for clustering the similar images. All the clustering algorithms are validated using cluster validation techniques. Validation result shows that K-Means clustering produced the best result. The result has been evaluated using Silhouette validation technique. In the experimental result, Pixel Information Feature obtained a Silhouette value of 0.24, Foreground/Background feature obtained a Silhouette value of 0.45, Texture Feature obtained a Silhouette value of 0.57 and combination of both Foreground/Background and Texture Feature obtained a Silhouette value of 0.52, all of the four descriptors the texture feature gives better result. This work can be extended

- To include other descriptors like shape and texture of the objects in images.
- To apply more clustering Techniques and identity the best methods.
- To work with a large dataset.

REFERENCES

1. Chi Zhang and Lei Huang, “Content-Based Image Retrieval Using Multiple Features”, Journal of Computing and Information Technology - CIT 22, 2014, Special Issue on LISS 2013, 1–10 doi:10.2498/cit.1002256.
2. Dong ping Tain, “A Review on Image Feature Extraction and Representation Techniques”, International Journal of Multimedia and Ubiquitous Engineering, Vol. 8, No. 4, July, 2013.
3. Gaurav Mandloi,” A Survey on Feature Extraction Techniques for Color Images”, International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014.
4. Guoyong Duan, Jing Yang and Yilong Yang, “Content-Based Image Retrieval Research”, Published by Elsevier B.V. Selection and/or peer-review under responsibility of Garry Lee, 2011.
5. Guy Brock, Vasyl Pihur, Susmita Datta, and Somnath Datta, “clValid an R package for cluster validation”, Journal of Statistical Software, March 2008.
6. Kannan.A, Dr.V.Mohan, Dr.N.Anbazhagan, “Image Clustering and Retrieval using Image Mining Techniques‖”, IEEE International Conference on Computational Intelligence and Computing Research, 2010.
7. Khalid Imam Rahmani, Naina Pal and Kamiya Arora, “Clustering of Image Data Using K-Means and Fuzzy K-Means”, International Journal of Advanced Computer Science and Applications, Vol. 5, No. 7, 2014.
8. Ladha. L, “Feature selection Methods and Algorithms”, International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No. 5, May 2011.
9. Lining Zhang, Hubert P. H. Shum, and Ling Shao, “Discriminative Semantic Subspace Analysis for Relevance Feedback”, IEEE Transactions on Image Processing, Vol. 25, No. 3, March 2016.
10. Maria Halkidi, Yannis Batistakis and Michalis Vazirgiannis, “On Clustering Validation Techniques”, Journal of Intelligent Information Systems, 17:2/3, 107–145, 2001.