International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.7, No.4 (2014), pp.105-114
http://dx.doi.org/10.14257/ijsip.2014.7.4.10

# An Experimental Study on Content Based Image Retrieval Based On Number of Clusters Using Hierarchical Clustering Algorithm

Monika Jain[1] and Dr. S.K.Singh[2]

*[1]Research scholar, Department of computer science, Mewar university, Rajasthan, India*

*[2]Professor and Head of Department of Information Technology, HRIT Engineering college, Ghaziabad, India*

*[1]monika_smec@yahoo.co.in; [2]singhsks123@gmail.com*

## *Abstract*

*Nowadays the content based image retrieval (CBIR) is becoming a source of exact and fast retrieval. CBIR presents challenges in indexing, accessing of image data and how end systems are evaluated. Data clustering is an unsupervised method for extraction hidden pattern from huge data sets. Many clustering and segmentation algorithms both suffer from the limitation of the number of clusters specified by a human user. It is often impractical to expect a human with sufficient domain knowledge to be available to select the number of clusters (NC) to return. This paper discusses the image retrieval based on NC which is evaluated using hierarchical agglomerative clustering algorithm (HAC). In this paper, we determine the optimal number of clusters using HAC applied on RGB images and validate them using some validity indices. Based on number of clusters, we retrieve set of images. These cluster values can be further used for divide and conquer technology and indexing for large image dataset. An experimental study is presented on real data sets.*

*Key terms: CBIR, number of clusters, hierarchical agglomerative clustering, validity indices, RGB image*

## 1. Introduction

Content-Based Image Retrieval (CBIR) is defined as a process that searches and retrieves images from a large database on the basis of automatically-derived features such as color, texture and shape. The techniques, tools and algorithms that are used in CBIR, originate from many fields such as statistics, pattern recognition, signal processing, and computer vision. It is a field of research that is attracting professionals from different industries like crime prevention, medicine, architecture, fashion and publishing. Clustering algorithms can offer superior organization of multidimensional data for effective retrieval. They allow a nearest neighbour search to be efficiently performed.

Natural scenes are rich in both color and texture and a wide range of natural images can be considered as a mosaic of regions with different colors and textures. Color feature is one of the most widely used features in image retrieval. Colors are defined on a selected color space. Variety of color spaces are available, they often serve for different applications. Description of different color spaces can be found in [5]. Color spaces are shown to be closer to human perception and used widely in RBIR include, RGB, LAB, LUV, HSV (HSL), YCrCb and the hue-min-max-difference (HMMD) [6-10].

Image clustering is the typical unsupervised learning technique for retrieval purpose based on some low level features. It intends to group a set of image data in a way to maximize the similarity within clusters and minimize the similarity between different clusters. Each resulting cluster is associated with a class label and images in same cluster are supposed to be similar to each other. In images, under unsupervised learning process, it is very difficult to predict number of clusters.

The traditional hierarchical clustering and its variations are often used for image clustering. Hierarchical algorithms can be either agglomerative or divisive. The agglomerative (bottom-up) approach repeatedly merges two clusters, while the divisive (top-down) approach repeatedly splits a cluster into two. In divisive, for a cluster with objects, there are $2^{N-1}$-1 possible two-subset divisions, which is very expensive in computation [11]. Therefore, divisive clustering is not commonly used in practice. In recent years, with the requirement for handling large-scale data sets in data mining and other fields, many new HAC techniques have appeared and greatly improved the clustering performance. Typical examples include CURE [14], ROCK [15], Chameleon [16], and BIRCH [17, 2].

To perform agglomerative hierarchical cluster analysis on a data set, we follow this procedure [24]:

1. Find the similarity or dissimilarity between every pair of objects in the data set based on metrics such as ‘euclidean’, ’seuclidean’, cityblock’, ’minkowski’, ’chebychev’, ’mahalanobis’, ’cosine’, ‘correlation’, ’spearman’, ’hamming’, ’jaccard’ [25] .
2. Group the objects into a binary, hierarchical cluster tree using some linkage criterias which are as follows.

### Single Linkage

In single-link (or single linkage) hierarchical clustering, merge the two clusters in each step whose two closest members have the smallest distance (or the two clusters with the smallest **minimum** pairwise distance). Its time complexity is O($n^2$) where n is the number of objects.

### Complete Linkage

In complete-link (or complete linkage) hierarchical clustering, merge the two clusters in every step whose merger has the smallest diameter (or the two clusters with the smallest **maximum** pair wise distance). The worst case time complexity of complete-link clustering is at most O($n^2$ log n).

### Average Linkage

In average-linkage clustering, the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster. The time complexity of average-link clustering is O($n^2$ log n).

### Centroid Method

In the centroid method, the distance between two clusters is defined as the (squared) Euclidean distance between their centroids or means.

### Ward's Method

Ward's method begins with N clusters, each containing one object, it differs in that it does not use cluster distances to group objects.

Instead, the total within-cluster sum of squares (SSE) is computed to determine the next two groups merged at each step of the algorithm. The error sum of squares (SSE) is defined (for multivariate data)

$$\text{SSE}=\sum_{i=0}^{K}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_i})^2 \qquad (1)$$

where $y_{ij}$ is the jth object in the ith cluster and $n_i$ is the number of objects in the ith cluster [13].

As objects are paired into binary clusters, the newly formed clusters are grouped into larger clusters until a hierarchical tree is formed.

3. Determine the cutoff point in the hierarchical tree. Generate clusters by detecting natural groupings in the hierarchical tree or by cutting off the hierarchical tree at an arbitrary point.

### 1.1 Cluster validation

Hierarchical clustering methods report the number and composition of n-1 clusters, n-2 clusters, n-3 clusters, and so on. What then is the optimum number of clusters? In general, a good cluster solution is one in which each cluster is very different from other clusters (between-cluster heterogeneity) and objects in each cluster are as similar as possible (within-cluster homogeneity). Various measures have been proposed to assess the homogeneity and/or heterogeneity of the clustering solution. For a detailed discussion the interested reader is referred to Sharma (1996) and Aldenderfer and Blashfield (1984) [18, 22]. Some of them are discussed here.

### →Cophenet Correlation Coefficient

The cophenetic correlation for a cluster tree is defined as the linear correlation coefficient between the cophenetic distances obtained from the tree and the original distances (or dissimilarities) used to construct the tree. Thus, it is a measure of how accurately the tree represents the dissimilarities among observations.

One way to measure how well the cluster tree generated by the linkage function reflects the data is to compare the cophenetic distances with the original distance data generated by the pdist function. If the clustering is valid, the linking of objects in the cluster tree should have a strong correlation with the distances between objects in the distance vector. The cophenet function compares these two sets of values and computes their correlation, returning a value called the cophenetic correlation coefficient [26]. The more the value of the cophenetic correlation coefficient is closer to 1 means getting more accurate clustering solution. This measure can be used to compare alternative cluster solutions obtained using different algorithms.

### →Root-Mean-Square Standard Deviation

The root-mean-square standard deviation (RMSSTD) measures the homogeneity of the cluster formed at any given step. It measures the compactness or homogeneity of a cluster. Clusters in which objects are very close to the centroid are compact clusters. The smaller the RMSSTD, the more homogeneous or compact is the cluster formed at a given step. A large value of RMSSTD suggests that the cluster obtained at a given step is not homogeneous, and is probably formed by merging of two very heterogeneous clusters.

$$RMSSTD=\left[\frac{\sum_{\substack{i=1\\j=1}}^{n_c,d}\sum_{k=1}^{n_{ij}}\left(x_k-\overline{x_j}\right)^2}{\sum_{\substack{i=1\ldots nc\\j=1\ldots d}}\left(n_{ij}-1\right)}\right]^{\frac{1}{2}} \qquad (2)$$

where $n_c$ is the number of clusters, d is the number of variables(data dimension), $n_j$ is the number of data values of j dimension while $n_{ij}$ corresponds to the number of data values of j dimension that belong to cluster i . Also $\overline{x_j}$ is the mean of data values of *j* dimension.

### →Semi-Partial R-Squared

The Semi-Partial R-squared (SPR) measures the loss of homogeneity due to merging two clusters to form a new cluster at a given step. If the value is small, then it suggests that the cluster solution obtained at a given step is formed by merging two very homogeneous clusters. On the other hand, large values of SPR suggest that the two heterogeneous clusters have been merged to form the new cluster. **SPR** of the new cluster is the difference between the pooled SSw (referring to sum of squares within group ) of the new cluster and the sum of the pooled SSw's values of clusters joined to obtain the new cluster (*loss of homogeneity*), divided by the pooled SSt (referring to the total sum of squares, of the whole data set) for the whole data set. This index measures the loss of homogeneity after merging the two clusters of a single algorithm step. If the index value is zero then the new cluster is obtained by merging two perfectly homogeneous clusters. If its value is high then the new cluster is obtained by merging two heterogeneous clusters.

### →R-Square

R-Square (RS) measures the heterogeneity of the cluster solution formed at a given step. A large value represents that the clusters obtained at a given step are quite different (*i.e.*, heterogeneous) from each other, whereas a small value would signify that the clusters formed at a given step are not very different from each other. Consequently, one would prefer to have a cluster solution with a high RS [19].

$$RS=\frac{\left\{\sum_{j=1\ldots d}\left[\sum_{k=1}^{n_j}(x_k-\overline{x_j})^2\right]\right\}-\left\{\sum_{\substack{i=1\ldots nc\\j=1\ldots d}}\left[\sum_{k=1}^{n_{ij}}(x_k-\bar{x}_j)^2\right]\right\}}{\sum_{j=1\ldots d}\left[\sum_{k=1}^{n_j}(x_k-\overline{x_j})^2\right]} \qquad (3)$$

### →Centroid Distance

The centroid distance (CD) measures the heterogeneity of the clusters merged at any given step to form a new cluster, and is given by the distance of the clusters merged at a given step. If two less heterogeneous clusters are merged the value will be small and if two very heterogeneous clusters are merged the value will be large. The CD index measures the distance between the two clusters that are merged in a given step. This distance is measured each time depending on the selected representatives for the hierarchical clustering we perform. For instance, in case of *Centroid hierarchical clustering* the representatives of the formed clusters are the centers of each cluster, so CD is the distance between the centers of the clusters. In case that we use *single linkage* CD measures the minimum Euclidean distance between all possible pairs of points. In case of *complete linkage* CD is the maximum Euclidean distance between all pairs of data points and so on.

A dilemma arises whether one or all of the measures should be used. It is suggested that all the measures should be used as they relate to various properties of the clusters. RMSSTD, SPR and CD represent homogeneity of the cluster solution and RS represents heterogeneity of the cluster solution. The confidence is obviously high when all of these measures suggest the same number of clusters. On the other hand, if there is no consensus among the measures regarding the number of clusters then it is prudent to examine all the suggested solutions and determine how many clusters are appropriate using other criteria such as interpretability and usefulness of the cluster solutions.

### 1.2 Which Clustering Method Is the Best?

Various hierarchical and non-hierarchical clustering methods are available to achieve same goal. It is obvious that in the hierarchical method a priori knowledge of the number of clusters is not needed. However, for large data sets hierarchical methods require extensive computational resources as large dissimilariy matrices have to be computed and stored. However, this is not an issue for its use as one can draw a random sample and subject it to hierarchical clustering. A commonly used procedure is to use hierarchical method in conjunction with non-hierarchical clustering methods. For example, a hierarchical method could be used initially to determine a number of cluster solutions. Based on this solution, half of the preprocessing is done on the dataset. The examined solution could be used on non hierarchical methods for exact retrieval [18].

## 2. An Experimental Study and Discussion

In this section, we describe the experiment performed to find the optimal number of clusters on real data set. Here we consider 100 real images, a subset of corel image dataset for testing purpose. First we make the dimensions of all images same. We use RGB color space and single linkage to perform hierarachical clustering algorithm. Taking average of R, G and B componenets and similarity at 0.8 or 0.7, we cluster the image. Taking maximum value of clustervector at cutoff point, we find out NC. For finding optimal number of clusters, we use here validity indices cophenet correlation coefficient, RMSSTD [Eq. no. 2] and CD defined in Section 1. Here, we don’t require any previous knowledge of data and it is totally machine dependent.

The materials used in this work consisted of the following

1. Computer: Intel Atom CPU N450 @ 1.66 GHz,1.00 GB of RAM
2. Microsoft windows 7 professional
3. MATLAB version 7.10.0

Test is performed on 9 image categories of corel image database, each containing 100 images. The categories are Beach, Buildings, Buses, Dinosaurs, Elephants, Flowers, Horses, Mountains, glaciers and Food. Randomly, we have chosen 11 images of each category. For elephant category, the retrieved number of clusters proposed by validity indices is shown in Table 1. The optimal values of NC are those for which a significant local change in values of RMSSTD and CD occurs. As regards, cophenet correlation coefficient, an indication of optimal clustering scheme is the point at which its value is nearly approaching to 1. If it is not come in the case of single linkage, we opt other linkages. The essence of clustering is not a totally resolved issue and we may consider different aspects as more significant depending on the application domain. For instance imaged 578 can be considered as having 98 clusters with two of them slightly overlapping or having two well-separated clusters.

We have shown the optimal number of clusters of 11 images of different categories excluding horse category in Table 2. Also we have taken 100 images of only horse catogory. Based on number of clusters of only horse category, some of the retrieved results are shown in Table 3.

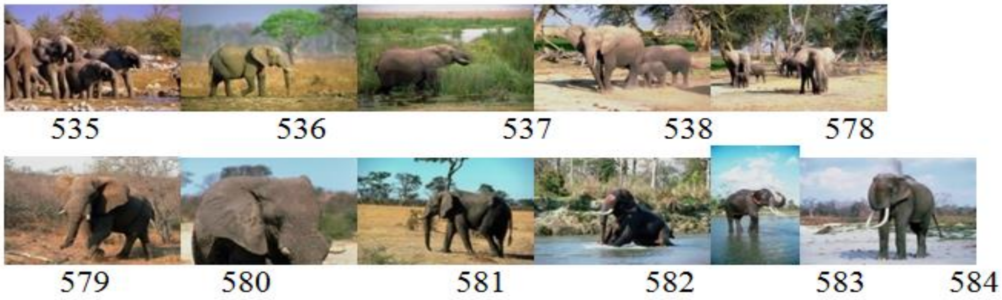

**Figure 1. 11 Images of Elephant Category**

**Table 1. Optimal Number of Clusters Proposed by Validity Indices for 11 Images of Elephant**

| **Image Id's** | 535 | 536 | 537 | 538 | 578 | 579 | 580 | 581 | 582 | 583 | 584 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Cophetic correlation coefficient** | 109 | 123 | 118 | 102 | 98 | 90 | 102 | 77 | 143 | 87 | 81 |
| **RMSSTD** | **110** | 120 | 122 | 105 | 105 | 98 | 111 | 78 | 151 | 88 | 77 |
| **CD** | **100** | 112 | 105 | 102 | 98 | 100 | 92 | 77 | 150 | 85 | 90 |

**Table 2. Optimal Number of Clusters of 11 Images of Different Categories Proposed by Validity Indices**

| **Beach(NC)** | 119 | 123 | 102 | 42 | 80 | 86 | 95 | 94 | 63 | 132 | 81 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **image id's** | 120 | 121 | 125 | 126 | 143 | 158 | 159 | 162 | 163 | 164 | 160 |
| **Monuments(NC)** | 104 | 90 | 149 | 94 | 163 | 114 | 88 | 139 | 110 | 98 | 115 |
| **image id's** | 211 | 212 | 213 | 214 | 217 | 219 | 221 | 280 | 281 | 283 | 284 |
| **Buses(NC)** | 92 | 115 | 90 | 94 | 98 | 110 | 123 | 105 | 60 | 125 | 99 |
| **image id's** | 308 | 309 | 310 | 311 | 360 | 361 | 362 | 363 | 364 | 365 | 366 |
| **Dinosour(NC)** | 95 | 90 | 82 | 110 | 85 | 93 | 99 | 86 | 88 | 90 | 107 |
| **image id's** | 436 | 437 | 438 | 439 | 440 | 441 | 442 | 443 | 444 | 445 | 446 |
| **Flowers(NC)** | 101 | 100 | 106 | 86 | 107 | 105 | 101 | 98 | 86 | 103 | 108 |
| **image id's** | 600 | 601 | 612 | 613 | 614 | 615 | 616 | 617 | 618 | 619 | 620 |
| **Mountains(NC)** | 86 | 193 | 111 | 96 | 99 | 109 | 154 | 119 | 165 | 74 | 103 |
| **image id's** | 800 | 801 | 802 | 803 | 804 | 810 | 813 | 815 | 818 | 820 | 821 |
| **Food(NC)** | 102 | 97 | 95 | 123 | 108 | 97 | 96 | 106 | 95 | 108 | 126 |
| **image id's** | 909 | 910 | 911 | 913 | 920 | 921 | 934 | 937 | 945 | 946 | 947 |

The analysis of Table 2 shows that discrete number of clusters is retrieved while performing HAC on images of same category. The performance analysis performed on horse category is shown in Figure 2. The accuracy depends upon matched images that are retrieved. Also we have selected ten query images of different categories of 1000 corel image dataset. The query is selected in such a way that we have some semantically relevant images. When the experiment is performed on it, the accuracy comes low due to more irrelevant results retrieved rather than only semantically relevant.

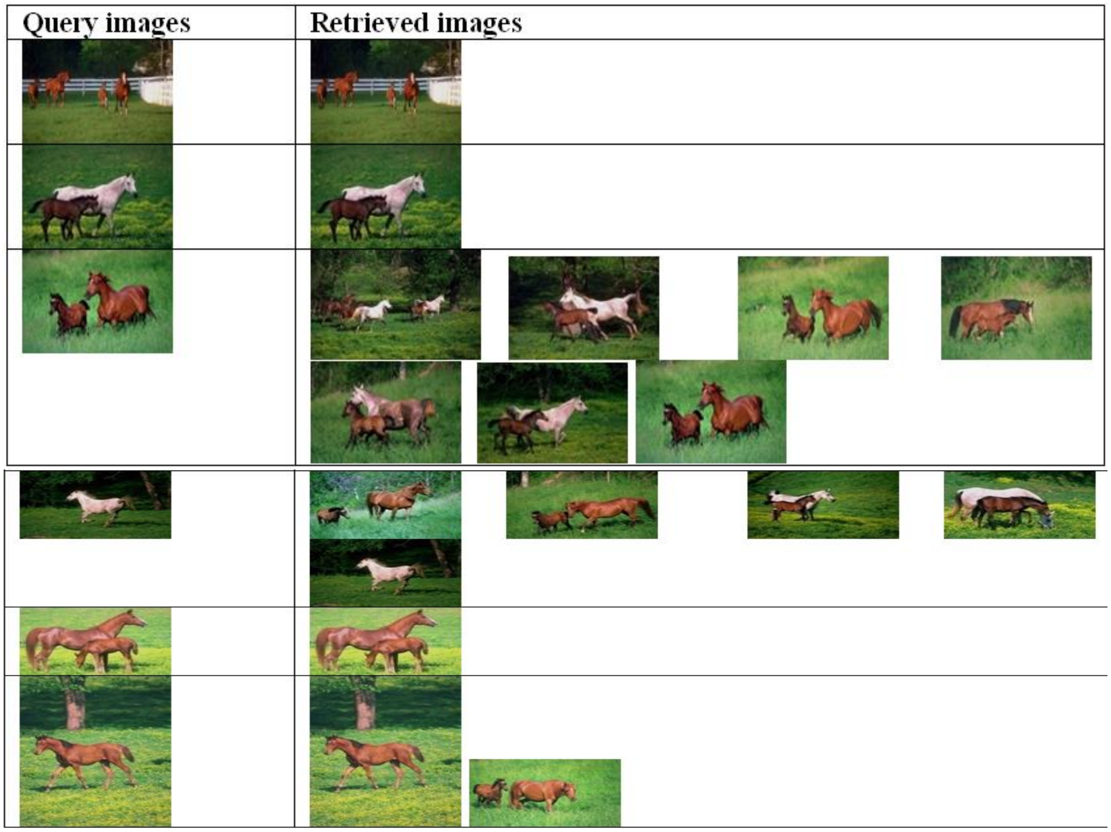

**Figure 2. Some of the Retrieved Results of Horse Category**

**Table 3. Performance Analysis Performed on Horse Category**

| Query image | Accuracy |
|---|---|
| 721 | 1 |
| 731 | 0.99 |
| 744 | 1 |
| 752 | 1 |
| 784 | 0.96 |
| 796 | 0.92 |
| 779 | 0.98 |
| 768 | 1 |
| 788 | 1 |
| 791 | 0.96 |

For image dataset of 100 images of only horse category, the average accuracy is 0.98 which is tremendous. However for 1000 corel image dataset the results are not favorable. The average accuracy is 0.25 which shows it returns more irrelevant results with relevant results. However, the run time complexity of HAC is completely dependent on the number of objects n in the dataset whereas runtime complexity of other partitional algorithms depend upon number of iterations along with n and NC. Thus, poor choice of NC increases the time complexity in partitional algorithms.

## 3. Conclusion

An experimental study on cluster oriented image retrieval with color features is shown in this paper. The selected color features of the image database mainly of same category and the image query are then clustered using HAC algorithm for similarity measurement purpose. The approach is examined in the experimental study with Corel image dataset. The experimental result described in Table 3 showed that the approach reached above 90% accuracy for same category of 100 images by finding out optimal NC using validity indices. However for very large dataset, it can be used as preprocessing and applied on other unsupervised algorithms taking information regarding NC. The results of experiment can be used for indexing and be enchanced for divide and conquer technology. It will become good approach for efficient content based image retrieval system for very large dataset.

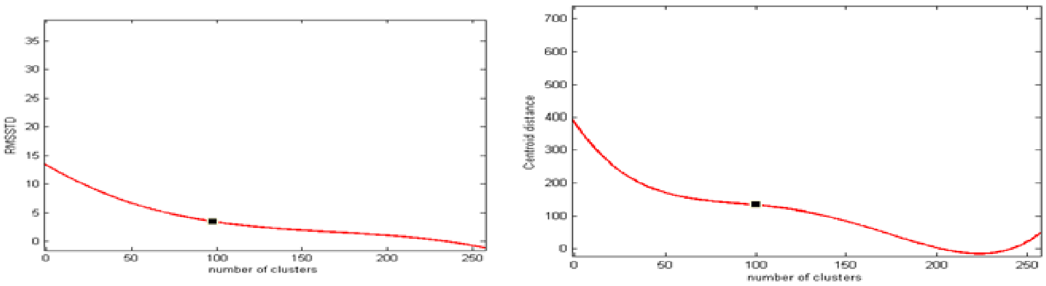

**Figure 3. Plot to Access RMSSTD and Centroid Distance for Image id 579 of Elephant Category**

## References

[1] Y. Liu, D. Zhang, G. Lu and W.-Y. Ma, “A survey of content-based image retrieval with high-level semantics”, Elsvier Pattern Recognition, vol. 40, **(2007),** pp. 262– 282.

[2] M. Jain, Dr. S. K. Singh, “A Survey On: Content Based Image Retrieval Systems Using Clustering Techniques for Large Data sets”, International Journal of Managing Information Technology (IJMIT), vol. 3, no. 4, **(2011),** November.

[3] M. Singdha and K. Hemachandran, “Content Based Image Retrieval using Color and Texture, Signal & Image Processing”, An International Journal (SIPIJ), vol. 3, no. 1, **(2012)** February.

[4] A. R. Barakbah and Y. Kiyoki, “Cluster Oriented Image Retrieval System with Context Based Color Feature Subspace Selection”.

[5] K. N. Plataniotis and A. N. Venetsanopoulos, “Color Image Processing and Applications”, Springer, Berlin, **(2000)**.

[6] P. L. Stanchev, D. Green Jr. and B. Dimitrov, “High level color similarity retrieval”, Int. J. Inf. Theories Appl., vol. 10, no. 3, **(2003),** pp. 363–369.

[7] Y. Liu, D. S. Zhang, G. Lu and W.-Y. Ma, “Region-based image retrieval with perceptual colors”, Proceedings of the Pacific-Rim Multimedia Conference (PCM), **(2004)** December, pp. 931-938.

[8] R. Shi, H. Feng, T.-S. Chua and C.-H. Lee, "An adaptive image content representation and segmentation approach to automatic image annotation", International Conference on Image and Video Retrieval (CIVR), **(2004)**, pp. 545–554.
[9] V. Mezaris, I. Kompatsiaris and M. G. Strintzis, "An ontology approach to object-based image Retrieval", Proceedings of the ICIP, vol. II, **(2003)**, pp. 511–514.
[10] B. S. Manjunath, *et al.,* "Color and texture descriptors", IEEE Trans. CSVT, vol. 11, no. 6, **(2001),** pp. 703–715.
[11] B. Everitt, S. Landau and M. Leese, "Cluster Analysis", London: Arnold, **(2001).**
[12] L. Ferreira and D. B. Hitchcock, "A comparison of hierarchical methods for clustering functional data", Introduction to Clustering Procedures chapter 8.
[13] S. Guha, R. Rastogi and K. Shim, "CURE: An efficient clustering algorithm for large databases," in Proc. ACM SIGMOD Int. Conf. Management of Data, **(1998)**, pp. 73–84.
[14] "ROCK: A robust clustering algorithm for categorical attributes", Inf. Syst., vol. 25, no. 5, **(2000),** pp. 345-366.
[15] G. Karypis, E. Han and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," IEEE Computer, vol. 32, no. 8, **(1999)** August, pp. 68-75.
[16] T. Zhang, R. Ramakrishnan and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in Proc. ACM SIGMOD Conf. Management of Data, **(1996)**, pp. 103–114.
[17] S. Sharma and A. Kumar, "Cluster analysis and factor analysis, **(1996)**.
[18] M. Halkidi, Y. Batistakis and M. Vazirgiannis, "On Clustering Validation Techniques", Journal of Intelligent Information Systems.
[19] S. Ray and R. H. Turi, "Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation," 4th International Conference on Advances in Pattern Recognition and Digital Techniques (I*CAPRDT'99)*, Calcutta INDIA, Narosa Publishing House, **(1999)** December 27-29, pp. 137-143, New Delhi India.
[20] R. Grover and M. Vriens, "Handbook of marketing research: uses, misuses and future advances",
[21] M. S. Aldenderfer and R. K. Blashfield, "Cluster analysis", Thousand Oaks, CA: Sage, **(1984)**.
[22] "Hierarchical clustering – Wikipedia", the free encyclopedia.
[23] http://www.mathswork.in/help/stats/pdist.html.
[24] http://www.mathswork.in/help/stats/linkage.html.
[25] http://www.mathswork.in/help/stats/cophenet.html.
[26] http://www.utdallas.edu/~nkumar/ClusterExample.doc.
[27] SAS/Stat User's Guide, Version 8, 5 Volume Set.
[28] http://wang.ist.psu.edu/docs/related/.

## Authors

**Mrs Monika Jain**, she is B.Sc (H), M.Sc and M.Tech in Computer Science and Engineering. She is Assistant Professor in JSSATE, Noida. She is pursuing Ph.D. in Computer Science & Engineering. She has published research papers in conferences & journals. She has over 10 years of Academic experience. She has handled various B. Tech & MCA Projects. Her area of interests is Computer Graphics, Data Mining and Image processing.

**Dr. S.K.Singh**, he is BE (CSE), M.Tech, and PhD (IT). He is Director, HRIT, Ghaziabad. He has more than 20 years of experience in teaching, research and content writing. He had published more than 40 papers in international refereed Journals and conferences. He is a member of organizing committee of the International conference IEEE' 08, Las Vegas USA. He is also a reviewer of various Journals and conferences of repute. He is currently guiding 7 Ph.D and ten M.Tech scholars.