A Novel Approach for Comparison of Clustering Algorithms on CAD Images

Shelza
CSE Department
SVIET
Banur (Punjab), India
shelzadesires@gmail.com

Balwinder Singh
CE Department
Yadavindra College of Engineering,
Talwandi Sabo,India
puyce@yahoo.com

***Abstract:*-Because of the fast rate of technological progress, the volume of CAD data is increasing in gigabytes day by day ,which are hard to make sense and difficult to analysis when time as resource is scarce and engineers need to take immediate inferences from CAD data. Now the challenge is to represent the data in such a way that it provides insights, trends and tendencies at glance and easy visualization by changing the variations options dynamically. In our research, we shall try to make visualization Framework that will incorporate clustering based on features of CAD images. Various data abstraction techniques are used in visualization systems to facilitate analysis from overview to detail. Also, analysts can compare different abstraction methods using the abstraction quality measures- histogram difference measure to see how well relative data density is maintained and select an abstraction method that meets the requirements of their analytic tasks.**

Keywords:- Clustering, HDM, CAD

I . INTRODUCTION

Very large multivariate datasets are increasingly common in many applications. This proves true for traditional relational databases and complex 2D and 3D multimedia databases that store images, CAD drawings, geographic information, and molecular biology data[2] .We can view relational databases as high-dimensional databases, since the attributes correspond to the dimensions of the data set. The same also holds true for multimedia data. For efficient retrieval, such data must usually be transformed into high-dimensional feature vectors such as color histograms [11] shape descriptors [12] Fourier vectors [13] and text descriptors [14] Many of the mentioned applications rely on very large databases consisting of millions of data objects with several tens to a few hundreds of dimensions.

However, we can apply a number of different data abstraction algorithms to high-dimensional data

1.1.*Data Abstraction*

Data abstraction is the process of reducing a large dataset into one of moderate size dataset, reducing the detail of data while maintaining dominant characteristics of the original dataset[1]. Techniques for data abstraction found within information visualization include

1)Sampling: Sampling is the process of selecting and using subsets of observations to estimate some parameters about a population. Various sampling techniques are simple random sampling, stratified sampling and quota sampling.. Sampling techniques have been well studied in statistics and widely applied in social science. In Computer Science, sampling is used for many tasks, including optimizing queries in databases with approximate information from samples . In recent years, faced with increasingly dense visualizations, researchers have begun to explore combining sampling with visualization. Random sampling can make the visualization of large datasets more perceptually effective. Their Astral Telescope Visualizer employs a 2D zooming interface to show data with different sampling levels.a non-uniform sampling algorithm to select less data in dense areas to reduce clutter, and more data in sparse areas to maintain data characteristics.

2) Clustering: Clustering is the process of partitioning a dataset into groups of objects based on similarity between objects or proximity according to some distance measure . Each group, called a cluster, consists of objects that are similar among themselves and dissimilar to objects in other groups. Clustering is an aggregation method, since a cluster is regarded as a higher level object that represents all objects It is widely used because of two reasons:

a) By visualizing cluster attributes rather than the original data, the number of visual elements displayed can be greatly reduced;

b) Clustering itself is a pattern discovering process. Thus visualizing clusters can explicitly reveal hidden patterns to viewers. Many visualization systems have adopted clustering methods to reduce clutter and analyze datasets.

1.2.*Abstraction Quality Measures*

1) HDM (Histogram Difference Measure) and
2) NNM (Nearest Neighbour Measure)

HDM:-The HDM[1] is derived based on the average relative error of aggregation used in approximate query processing of databases as well as image similarity measures used in image retrieval.

NNM:- The NNM[1] is derived based on the nearest neighbour algorithm used in pattern recognition and an image quality measure used in image compression.

Quality measures are recomputed whenever the above operations are performed. The measures and interactions together form an environment in which analysts can explore multi resolution visualization with abstraction quality information available. Data abstraction is the process of reducing a large dataset into one of moderate size, reducing the detail of data while maintaining dominant characteristics of the original dataset. Some data abstraction methods select a subset of the original dataset as the abstraction, such as sampling and filtering, while other data abstraction methods construct a new, more abstract representation, such as clustering and summarizing. Measurement generally refers to the process of estimating the magnitude of a quantitative property [9]. Measurement is essential for scientific research; with measurement, researchers can compare different objects and evaluate the effectiveness of programs or processes

To facilitate explanation of these measures, we define the DAL & DAQ

DAL:-Data Abstraction Level (DAL) as the ratio between the size of the abstracted dataset and the original dataset, and **DQL**:- Data Abstraction Quality (DAQ) as the degree to which the abstracted dataset represents the original dataset.

II. RELATED WORK

Several researchers have proposed measures for visualization and data abstraction.Kathrin Anne Meier [2] explored data abstractions based on density estimation. The method is to estimate the density **of** scientific data sets is based on the directory **of** a multidimensional data access structure. This data density estimator is called directory estimator. It is based on multidimensional adaptive histograms and is therefore computationally efficient, even for large data sets and many dimensions. They describe the methodology in general and focuses on the estimator's accuracy in particular. The accuracy **of** the directory estimator depends on the parameters **of** the access structures used, such as the bucket capacity. They evaluate the choice **of** bucket capacity theoretically **as** well as empirically with the ISE (Integrated Squared Error) being the measure **of** error and using a gridfile as the data access structure..Qingguang Cui,[1] define two data abstraction quality measures for computing the degree to which the abstraction conveys the original dataset: the Histogram Difference Measure and the Nearest Neighbor Measure. They have been integrated within XmdvTool, a public-domain multiresolution visualization system for multivariate data analysis that supports sampling as well as clustering to simplify data. Several interactive operations are provided, including adjusting the data abstraction level, changing selected regions, and setting the acceptable data abstraction quality level. Conducting these operations, analysts can select an optimal data abstraction level. PAULA FREDERICK [9] use appropriate data structures for storing sets of graphical objects ,it can lead to great performance improvements in the visualization of large figures, such as maps and CAD drawings. They present here a study on the performance of persistent data structures composed by an R-tree and V-trees, for storing and efficiently retrieving 2D graphical objects. Results are shown to demonstrate the efficiency of the proposed solutions when applied to large maps. Alexander Hinneburg, [3] describes an advanced clustering algorithm combined with new visualization methods interactively clusters data more effectively. Experiments show these techniques improve the data mining process.Huy [8] describes new emerging abstractions for parallel data processing, in particular computing clouds, can be leveraged to support large-scale data exploration through visualization. They take a first step using MapReduce framework to implement large-scale visualization techniques.

III. PROPOSED WORK

3.1. METHODOLOGY

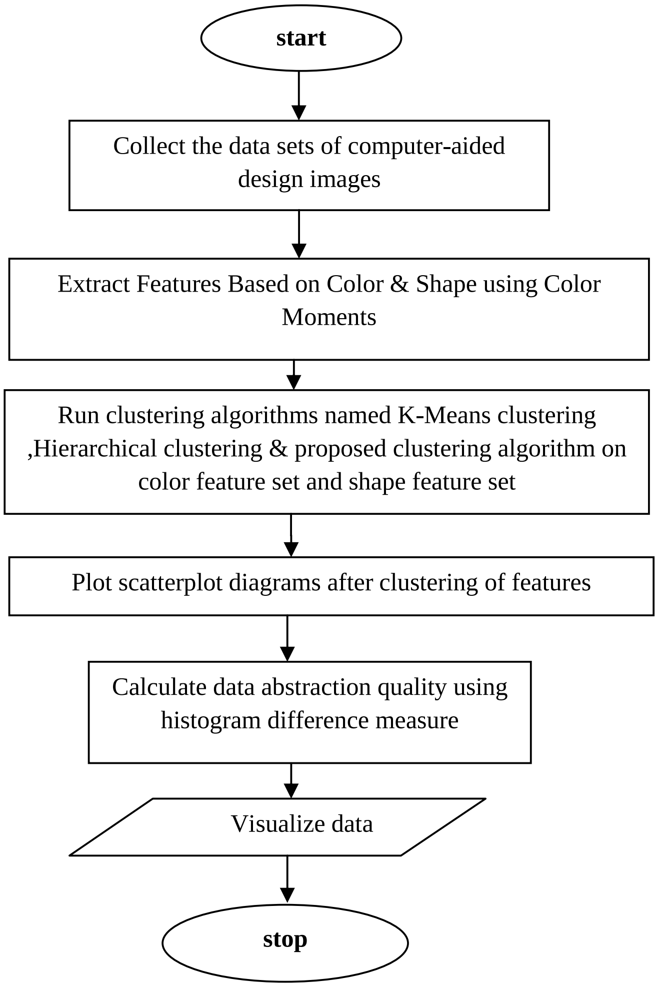
start
Collect the data sets of computer-aided design images
Extract Features Based on Color & Shape using Color Moments
Run clustering algorithms named K-Means clustering ,Hierarchical clustering & proposed clustering algorithm on color feature set and shape feature set
Plot scatterplot diagrams after clustering of features
Calculate data abstraction quality using histogram difference measure
Visualize data
stop

Figure 1. Methodology

In methodology shown in figure 1, First step is to collect data set of CAD images. We collected around 850 CAD images. Second step is to extract color and shape features. Color features are extracted according to color moments

.Three color moments are used: mean, standard deviation & skewness for each color component red ,green & blue. Four shape features are extracted: square ,rectangle , circle & unknown shapes. Third step is to apply clustering algorithms on color and shape features extracted from CAD images. Three clustering algorithms: k-means, hierarchical clustering & proposed clustering are applied, scatterplots of abstracted dataset created by all three clustering algorithms and original dataset are created both on color and shape features. Fifth step is to calculate data abstraction quality 'Histogram Difference Measure' for all three abstracted data sets created by three clustering algorithms .And comparison between clustering algorithms is done on the basis of HDM .Then visualization is done using parallel co-ordinates. Parallel co-ordinates of both original dataset and abstracted dataset are created both on color and shape features.

3.2.Clustering:-According to step 3 of our methodology, Clustering is applied after extracting the color and shape features from image dataset. We have applied three clustering algorithms over feature dataset :K-Means clustering, Hierarchical clustering and proposed clustering approach

3.2.1.K-Means

K-Means is a well-known partitioning algorithm .Objects are classified as belonging to one of the k groups k chosen is priori cluster membership is determined by calculating the centroid for each group .and assigning each object to the group with the closet centroid. This approach minimizes the overall with in- cluster dispersion by iterative reallocation5 of clusters members.

3.2.2.Hierarchical Clustering

Hierarchical clustering groups data over a variety of scales by creating a cluster tree or dendrogram. The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. This allows us to decide the level or scale of clustering that is most appropriate for your application

Algorithm Description

To perform agglomerative hierarchical cluster analysis on a data set, follow this procedure:

a)Find the similarity or dissimilarity between every pair of objects in the data set.

b) Group the objects into a binary, hierarchical cluster tree

c)Determine where to cut the hierarchical tree into clusters.

3.2.3.Proposed Clustering Algorithm

This clustering algorithm partitions the data elements into number of clusters. In clustering, each point has a degree of belonging to clusters, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster, may be in the cluster to a lesser degree than points in the center of cluster.It is based on minimizing the following objective fn:

$$fn = \sum_{i=1}^{n} \sum_{j=1}^{k} d_{ij}^{m} \frac{x_i . c_j}{||x_i|| \; ||c_j||}$$

d_{ij} is the degree of membership of x_i in the cluster j,x_i is the ith of d-dimensional measured data; n is number of data points,c_j is the d-dimension center of the cluster; k is total number of clusters. And objective function is based upon finding the distance between any measured data and the center. In this we have used cosine similarity distance method.

3.2.3.1. Pseudocode

The algorithm works like this

```
▪ Choose a number of clusters
▪ Assign randomly to each point coefficients for being in the clusters.
▪ Repeat until the algorithm has converged (that is, minimum value of objective function is achieved) :
    ▪ Compute the centroid for each cluster.
    ▪ For each point, compute its coefficients of being in the clusters.
```

3.2.3.2.Output of proposed clustering algorithm

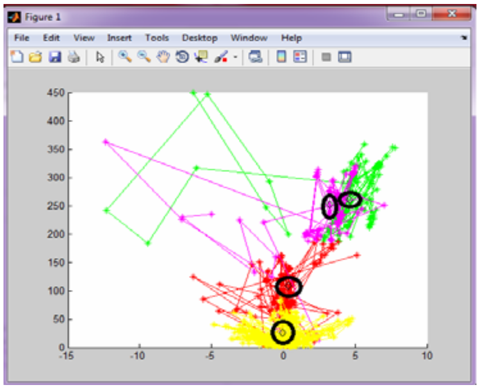

Figure 2.Scatterplot of clustered dataset on color features

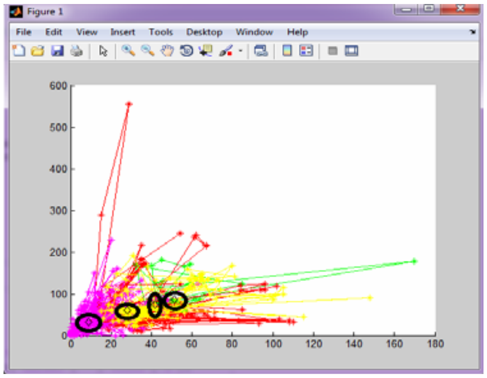

Figure 3.Scatterplot of clustered dataset on shape features

3.3.Data abstraction quality :-According to step 5 of our methodology, abstraction quality is measured for all the clustering algorithms :k-means , hierarchical clustering and proposed clustering algorithm. We have used HDM for calculating abstraction quality.A histogram is an aggregation method that conveys data distribution. To construct a histogram, the data space is partitioned into many small ranges, with each range corresponding to a bin. The height of a histogram bin is determined by the percentage of data points that fall in the corresponding range. It reveals the data density within each subrange.Because a histogram is a common data descriptor and is fast to compute, we propose to use the difference between the normalized histograms of the original dataset and the abstracted dataset as a measure to gauge the DAQ In our paper ,we computed HDM for original dataset and abstracted dataset created by K-means clustering, Hierarchical clustering and proposed clustering ,results of which are shown in next section.

3.4.Parallel Co-Ordinates:- It is a visualization technique in which different classes identified by clustering algorithm based on distance classifier are visualized with different colors. A parallel co-ordinate plot basically gives wholistic view of how data is lying semantically in the complete data set.

IV RESULT & DISCUSSIONS

In our work, three clustering algorithms named K-Means, Hierarchical clustering and proposed clustering are compared using the proposed data abstraction quality measure named HDM .We employ CAD image dataset based on color features and shape features.We use parallel co-ordinates to visualize this dataset. Through this application ,we find out proposed clustering algorithm gives the best measure for data abstraction quality.So we use abstracted dataset created by proposed clustering approach for visualization.

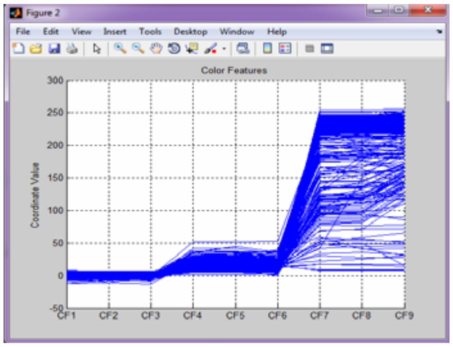

Figure 4. Parallel co-ordinates of original CAD image dataset on color features

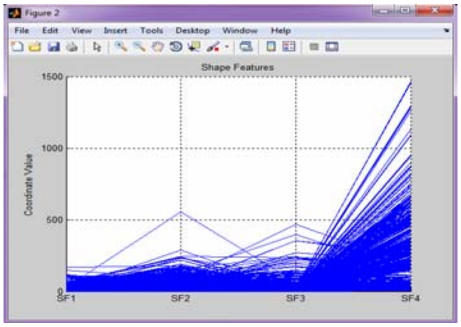

Figure 5. Parallel co-ordinates of original CAD image dataset on shape features

Figure 4 & Figure 5 shows the parallel co-ordinates of original CAD image dataset for color features and shape features respectively. CAD image dataset is of dimension 847*9for color features, it means it has 847 images and 9 color features & 847*4 for shape features ,it means it has 847 images and 9 color features

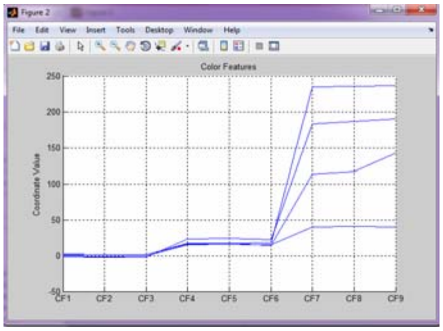

Figure 6. Parallel co-ordinates of abstracted CAD image dataset on color features

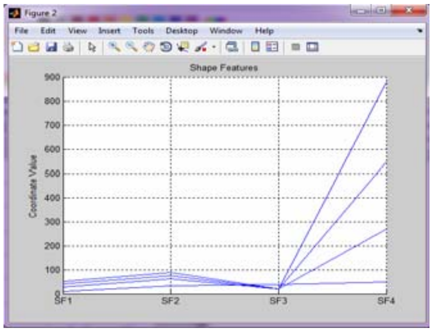

Figure 7. Parallel co-ordinates of abstracted CAD image dataset on shape features

From the figures 4 & 5 it is clear data points are too dense to observe any analytic behavior. We clustered the dataset using proposed clustering approach .Figures 6 &7 shows visualization of abstracted dataset for color features and shape features respectively. From figures 6 & 7, it is clear that visual quality is very good & visual clutter is significantly reduced while maintaining the relative data density.

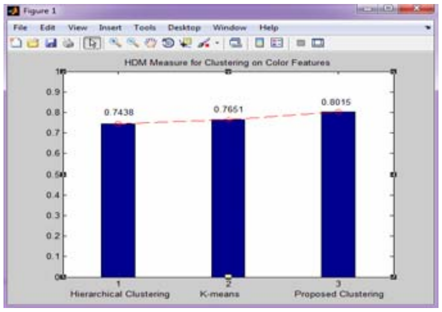

Figure 8. HDM measure for all clustering algorithms on color features

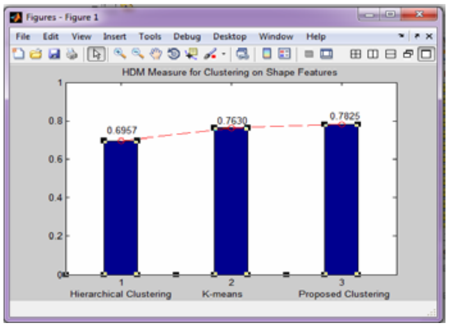

Figure 9. HDM measure for all clustering algorithms on shape features

Why clustered data set created by proposed clustering approach is chosen for visualization, because it gives the best measure for abstraction quality measure HDM.

First we briefly review some characteristics of the HDM . HDM is used to measure the data abstraction quality to know how well abstracted dataset represents the original dataset . The HDM is based on the histogram and minimizes the difference between the distributions of two datasets, so it excels in detecting changes in the relative density of data. HDM measure for all clustering techniques is shown in figure 8 and figure 9 for color features and shape features respectively.

V. CONCLUSION & FUTURE WORK

The methodology presented, opens a novel set of tools and possibilities for data abstraction and visualization. Volume of CAD data is very large, the ability of researchers in the scientific and engineering community to generate or acquire data far outstrips their ability to analyze it.

In this framework, we have identified data abstraction as a common mechanism for dealing with large-scale data visualization. We used data abstraction quality measure 'Histogram Difference Measure (HDM)' to find out how well the abstracted dataset represents the original dataset.. In our case study, we used three clustering algorithms K-Means Clustering, Hierarchical Clustering, and proposed clustering approach for clustering data and applied abstraction quality measures on clustered data produced by these three clusterings and find out the values for HDM for all three clustering algorithms . Among three clustering algorithms, proposed clustering algorithm gives the best results.And then visualization is done for clustered dataset created by proposed clustering approach.

In our future work we can improve abstraction quality more by using new clustering on different distance measures. We will also explore ideas for new data abstraction measures for measuring data abstraction quality based on statistical properties of the data, such as mean value and standard deviation.

REFERENCES

[1] .Qingguang cui .Mathew O. ward ,Elke A . rundensteiner & Jing jang . measuring data abstraction quality in multiresolution visualizations. IEEE transactions on visualization and computer graphics,vol. 12 ,NO. 5 ,pages 709-716 ,2006

[2] Kathrin Anne Meier . data abstraction through density estimation by storage management . proc. IEEE,pages 39-47. 1997

[3] Alexander Hinneburg , Daniel A. Keim and Markus Wawryniuk . hd-eye : visual mining of high – dimensional data. IEEE computer graphics and applications , pages 22-31 , 1999

[4] Jianbing Huang,and Michael B. Carter, Interactive Transparency Rendering for Large CAD Models ,IEEE transactions on visualization and computer graphics, VOL. 11, NO. 5, pages 584-595, 2005

[5] Pierre Georgel, Pierre Schroeder and Nassir Navab , Navigation Tools for Augmented CAD Viewing,IEEE Computer graphics and applications,2009

[6] Kai-Mo Hu, Bin Wang, Bin Yuan and Jun-Hai Yong.,Automatic Generation of Canonical Views for CAD Models,12th International Conference on Computer-Aided Design and Computer Graphics,pages 17-24 ,2011

[7] Jeffrey LeBlanc, Matthew 0. Ward, norman wittels, Exploring N-Dimensional databases,IEEE, pages 230-237 ,1990

[8] Huy T. Vo', Jonathan Bronson ,Brian Summa ,Joao L.D. Comba, Juliana Freire, Parallel Visualization on Large Clusters using Map Reduce,IEEE Symposium on Large Data Analysis and Visualization ,pages 81-88 ,2011

[9] .**P**aula frederick, marcelo gattass, mauricio riguette mediano, Efficient Visualization of Graphical Objects,IEEE,2002

[10] A. Inselberg and B. Dimsdale, “Parallel Coordinates: A Tool for Visualizing Multidimensional Geometry,” *Proc. Visualization 1990*, IEEE CS Press, Los Alamitos, Calif., 1990, pp. 361-370.

[11] H. Shawney and J. Hafner, “Efficient Color Histogram Indexing,” *Proc. Int’l Conf. on Image Processing*, IEEE Press, Piscataway, N.J., 1994, pp. 66-70.

[12] R. Methrotra and J.E. Gray, “Feature-Index-Based Similar Shape Retrieval,” *Proc. 3rd Working Conf. on Visual Database Systems*, Chapman and Hall, London, 1995, pp. 46-65.

[13] T. Wallace and P. Wintz, “An Efficient Three-Dimensional Aircraft Recognition Algorithm Using Normalized Fourier Descriptors,” *Computer Graphics and Image Processing*, Academic Press, Vol. 13, 1980, pp. 99-126.

K. Kukich, “Techniques for Automatically Correcting Words in Text,” *ACM Computing Surveys*, Vol. 24, No. 4, 1992, pp. 377-440e.

AUTHORS PROFILE

Er.Shelza is currently working as Lecturer in Computer Science and Engineering Department at Swami Vivekanand Institute of Engineering and Technology, Banur. She is pursuing M.Tech in Computer Engg. from Yadwindra College of Engineering and Technology, Talwandi Sabo affilated to Punjabi University Patiala .she holds the degree of B.Tech in Computer Science and Technology from Sant longowal institute of Engineering & Technology, Longowal Distt Sangrur,Punjab.She has more than 7 years of teaching experience. Her research interest included programming and data mining.She has authored papers in National Conference & international journals.

Er. Balwinder Singh is working as Assistant Professor (Computer Science) at Yadavindra College of Engineering, Punjabi University Guru Kashi Campus at Talwandi Sabo (India). He received B.E. degree from Guru Nanak Dev Engg. College, Ludhiana (India) and M.Tech. degree from Department of Computer Science & Engineering, Punjabi University, Patiala (India). He has more than 10 years of teaching and research experience. His research interests include Image Processing, Programming and Natural Language Processing