International Journal Of Mechanical Engineering And Information Technology
An Official Publication Of IGM Publication

Image Clustering Using Tree Structured Self Organizing Maps

Authors

R.K.Deebika[1], R. Mahendra Kumar[2]

[1]Assistant Professor, Department of Information Technology, Vivekanandha College of Technology for Women, Elayampalayam, Tiruchengode, 637 205, Tamilnadu, India

[2]Assistant professor, Department of Computer Science and Engineering, Vivekanandha College of Technology for Women, Elayampalayam, Tiruchengode, 637 205, Tamilnadu, India

Email.id: *deebikakuppusamy@gmail.com[1],mahemsd@gmail.com[2]*

ABSTRACT-

The challenges in achieving the efficient image retrieval in the development of look-for mechanisms is to guarantee the deliverance of minimal irrelevant information (high precision) and to make certain that the relevant information is not overlooked (high recall). The unstructured format of images tends to resist the deployment of the standard look-for mechanism and classification techniques. In order to provide efficient organization of images, clustering is the significant feature for the retrieval of images. The ability of the system to retrieve the relevant images based on the look-for criteria could be greatly increased if they were able to provide an accurate clustering method.

An efficient clustering method is desired for the images. The Dynamic Growing Self-Organizing Tree (DGSOT) algorithm outperforms the traditional Hierarchical Agglomerative Clustering algorithm in terms of precision and recall. In Dynamic Growing Self-Organizing Tree algorithm, a hierarchy of images is constructed from top to bottom in a tree-structured format where the similar images are grouped together that provides the most efficient user-friendly manner of image retrieval. A clustering method is proposed based on unsupervised neural networks in addition with the latent semantic indexing method which semantically improves the Dynamic Growing Self-Organizing Tree algorithm.

1. INTRODUCTION

1.1 Image

Images may be two-dimensional, such as photograph, screen display, and as well as a three-dimensional, such as a statue or hologram. They may be captured by optical devices such as cameras,mirrors, lenses, telescopes, microscops etc. and natural objects and phenomena, such as the human eye or water surfaces.

1.2 Image Processing

Image processing is any form of signal processing for which the input is an image, such as a photograph or video frame, the output of image processing may be either an image or, a set of characteristics or parameters related to the image. Most image-processing techniques involve treating the image as a two-dimensional signal and applying standard signal-processing techniques to it. Set of computational techniques for analyzing, enhancing, compressing, and reconstructing images. Its main components are importing, in which an image is captured through scanning or digital photography; analysis and manipulation of the image, accomplished using various specialized software applications and output (e.g., to a printer or monitor). Image processing has extensive applications in many areas, including astronomy, medicine, industrial robotics, and remote sensing by satellites and pattern recognition. Image processing usually refers to digital image processing, but optical and analog image processing also are possible. The acquisition of images (producing the input image in the first place) is referred to as imaging.

1.3 Categorization

Categorization is the process in which ideas and objects are recognized, differentiated and understood. Categorization implies that objects are grouped into categories, usually for some specific purpose. Ideally, a category illuminates relationship between the subjects and objects of knowledge. Categorization is fundamental in language, prediction, inference, decision making and in all kinds of environmental interaction. It is indicated that categorization plays a major role in programming.

There are many categorization theories and techniques. In a broader historical view, however, three general approaches to categorization may be identified:

Classical categorization

Conceptual clustering

Prototype theory

1.4 Image Categorization

The term image categorization refers to the labeling of images into one of a number of predefined categories. Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. The different types of clustering process includes Hierarchical clustering (Agglomerative (bottom-up) and Divisive (top-down)) and Partitional clustering. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with low distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions.

The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the

individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery that involves try and failure. It will often be necessary to modify preprocessing and parameters until the result achieves the desired properties. The general types of categorization algorithm include the following:

Connectivity based clustering (Hierarchical clustering)

Centroid-based clustering (k-means clustering)

Distribution-based clustering

Density-based clustering

The literature review about the papers is given below in chapter 2, the detailed description about the DGSOT algorithm and proposed work is described in chapter 3 and finally chapter 4 deals with the results and conclusion.

2. LITERATURE REVIEW

The resolution of remote sensing images increases every day, raising the level of detail and the heterogeneity of the scenes. Most of the existing geographical information systems classification tools have used the same methods for years. With these new high-resolution images, basic classification methods do not provide satisfactory results. The two different algorithms namely K-means Algorithm and Back Propagation Algorithm of ANN for Segmentation and Classification of Satellite images are implemented. Wide database of images has been used to test both the algorithms. The comparison results obtained by implementing both algorithms shows good accuracy in both the methods [2].

K–Means Algorithm starts with some clusters of pixels in the feature space, each of them defined by its center. The first step consists in allocating each pixel to the nearest cluster. In the second step the new centers are computed with the new clusters. These two steps are repeated until convergence. The basic step of k-means clustering is simple. In the beginning determine number of clusters K and assume the centroid or center of these clusters. Take any random objects as the initial centroids or the first K objects in sequence can also serve as the initial centroids. The K means algorithm will do the three steps below until convergence. Iterate until stable (= no object move group) a) Determine the centroid coordinate. b) Determine the distance of each Object to the centroids. c) Group the object based on minimum distance. After implementation of K –means algorithm, the results of ‘segmentation and classification’ are stored and then same images are given to the neural network classifier. It is found that K- means algorithm gives very high accuracy, but it is useful for single database at a time.

Back propagation algorithm of ANN is a generalization of the least mean square algorithm that modifies network weights to minimize the mean squared error between the desired and actual outputs of the network. Back propagation uses supervised learning in which the network is trained using data for which inputs as well as desired outputs are known. Once trained, the

network weights are frozen and can be used to compute output values for new input samples [2].

The gray scale image clustering is optimized using two traditional methods, these are thresholding technique and genetic algorithm (GA). The clustering optimization is achieved by applying three features (gray value, distance, gray connection) based thresholding technique and genetic algorithm. The clustering optimization includes segmenting the image to find regions that represent objects or meaningful parts of objects depending on the above mentioned three features which base on gray value of image and two standard mathematical theories these are chessboard distance and breshenham's algorithm. These three features make the clustering operation more accurate [10].

With the advancement in image capturing device, the image data been generated at high volume. If images are analyzed properly, they can reveal useful information to the human users. Content based image retrieval address the problem of retrieving images relevant to the user needs from image databases on the basis of low-level visual features that can be derived from the images.

Grouping images into meaningful categories to reveal useful information is a challenging and important problem. Clustering is a data mining technique to group a set of unsupervised data based on the conceptual clustering principal: maximizing the intraclass similarity and minimizing the interclass similarity. By focusing color as feature, Color Moment and Block Truncation Coding (BTC) are used to extract features for image dataset. Experimental study using K-Means clustering algorithm is conducted to group the image dataset into various clusters.

In image retrieval system, the content of an image can be expressed in terms of different features such as color, texture and shape. These low-level features are extracted directly from digital representations of the image and do not necessarily match the human perception of visual semantics. The framework of unsupervised clustering of images is based on the color feature of image. Test has been performed on the feature database of color moments and BTC. K-means clustering algorithm is applied over the extracted dataset. Results are quite acceptable and showing that performance of BTC algorithm is better than color moments. [11]

A novel approach of clustering image datasets is differential evolution (DE) technique. The differential evolution is a parallel direct search population based optimization method. From certain simulations it is found that DE is able to optimize the quality measures of clusters of image datasets. To claim the superiority of DE based clustering the outcomes of DE have been compared with the classical K-means and popular Particle Swarm Optimization (PSO) algorithms for the same datasets. The comparisons results reveal the suitability of DE for image clustering in all image datasets.[12]. It was shown that PSO and DE produced better result compared to K-means with respect to the quantization error, inter- and intra-cluster distances. The local optima problem of K-means was alleviated using PSO

and DE further improved the results. The main advantage of using DE is found to be having almost no parameter tuning. Compared to PSO, DE is able to provide more accurate optimized results for all the investigated dataset. The major merit lies with DE is that it does not require many parameter tuning as compared to PSO.

3. DGSOT

An efficient clustering method is desired for images. First, segment all input images into objects [13] and calculate similarities between each two objects on the basis of color, texture and shape features. Second, cluster similar objects into various groups based on object similarity. On the basis of object groups, deduct weight by calculating term frequency and inverse document frequency. Third, construct a vector for each image and calculate similarities between each two images, using vector model. Image vectors are decided by the objects contained in each image, and the vector length will be the number of object groups. Finally, use some clustering algorithm to cluster images based on image similarity. For this, several existing techniques are available such as hierarchical agglomerative clustering algorithm (HAC)[4,9,5],self-organizing map (SOM) [6,7,8],and self-organizing tree (SOT)[1,4].Existing algorithm named dynamic growing self-organizing tree (DGSOT) is used. The object similarity is determined by the combination of color similarity, texture similarity and shape similarity.

3.1 Color Similarity Measure

To compute color similarity, extract color information from object *i* pixel by pixel and construct a color vector Vi(v1,*i* ,v2,*i* , . . . v*p*,*i* , . . . v*k*,*i*) to express the color histogram. Each item in this vector represents the percentage of pixels whose hue value locates in specific interval. For each object, there is a unique *V*, to evaluate the degree of color similarity between object *i* and object *j* .The value of *k* affects the accuracy of color similarity. Along with increasing of *k*, accuracy will increase also. In this case, consider $k = 12$.

3.2 Shape Similarity Measure

The computation of shape similarity is a little bit more complicated. First, find major axis for each object that is the longest line joining two points on the boundary. Rotate an angle around mass centroid of an object to make its major axis to be parallel to *x*-axis and keep the centroid above the major axis. The reason to do that is to normalize the object to make it invariant to rotation. After doing normalization, create a $q * q$ grid, which is just big enough to cover the entire object, and overlaid on the object. The size of each cell is same. Define a shape vector *Ui* (*u*1,*i* , *u*2,*i* , . . . *u p*,*i* , . . . *uq*2,*i*) sized $q2$ which corresponds to $q2$ cells. Each item in the vector stands for the percentage of pixels in corresponding cell. The higher the *q* value is, the higher the accuracy. Raise the calculation cost as result. If images are simple, it is not necessary to choose a high *q*. In our case, *q* is 4, which is good enough for small

simple images. Calculate the degree of shape similarity between two shape vectors.

3.3 Texture Similarity Measure

The texture retrieval is harder than color and shape, because there is no clear definition of texture". The texture of an image region is decided by gray level distribution. Basically, there are three kinds of approaches to extract texture features: spectral approach, structural (or syntactic) approach, and statistical approach. For statistical approach, according to the order of the utilized statistical function, there exist two categories of descriptors: First-Order Texture Features and Second-Order Texture Features. The difference is that the first-order statistics do not provide information about the relative positions of different gray levels. In this method, intensity value in a region is homogeneous, so the positions of different gray levels are not very important here. To reduce calculation complexity, choose to use first-order texture features to generate texture feature vector.

The DGSOT is a tree structure self-organizing neural network. It is designed to discover the correct hierarchical structure of the underlying data set. The DGSOT grows in two directions: vertical and horizontal. First, in the direction of vertical growth, the DGSOT adds children, and in the direction of horizontal growth, the DGSOT adds more siblings. In vertical growth of a node, only two children are added to the node. In horizontal growth to determine suitable number of children to represent data that are associated with the node is complex. Thus, the DGSOT chooses the right number of sub clusters at each hierarchical level during the tree construction process. During the tree growth, a learning process similar to the self-organizing tree algorithm (SOTA) is adopted.

PROPOSED SYSTEM

A system is developed for image segmentation and clustering. A user can input an image for segmentation, the system will output detected boundary and all segmented regions in thumbnail format. Then from these segmented images, similar regions will be grouped together. Thus, an image will be represented by a vector with a set of objects along with weights. A Latent Semantic Indexing with dynamic growing self-organizing tree (DGSOT) is proposed. This proposed image-clustering algorithm is to be implemented to increase the efficiency and accuracy of clustering of images.

4. RESULTS AND CONCLUSION

The different categories of images are downloaded from the website as an input to the proposed system. The first module includes preprocessing of images. The image preprocessing method converts the color image to gray image then the edge detection process takes place. The next module proceeds with segmentation process followed by Latent Semantic Indexing and clustering using DGSOT.

In order to provide better organization of images, clustering is an important aspect for effective image retrieval. To develop a hierarchy, a

dynamic growing self-organizing tree algorithm (DGSOT) is used. An image-clustering algorithm, Latent Semantic Indexing with DGSOT is to be implemented based on hierarchical tree. The proposed algorithm is to integrate newly coming images into the existing hierarchy structure semantically. The proposed method is to increase the efficiency in time and to provide better accuracy.

REFERENCES

[1] Albers S, Westbrook J. "Self-organizing data structures. In: Online Algorithms: The State of the Art", Springer, Fiat-Woeginger, 1998.

[2] Ashwini T. Sapkal, Chandraprakash Bokhare, Tarapore N. Z. "Satellite Image Classification using the Back Propagation Algorithm of Artificial Neural Network". V.I.T. Pune.

[3] Dopazo J, Carazo JM. "Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree". Journal of Molecular Evolution Vol 44, 226–233, 1997.

[4] Downs GM, Willett P, FisanickW. "Similarity searching and clustering of chemical structure databases using molecular property data". J. Chem. Inf. Comput. Sci. 1994; 34(5):1094–1102.

[5] Ellen M. Voorhees. "Implementing Agglomerative hierarchic clustering algorithms for use in document retrieval". Information Processing & Management, 1986; 22(6):465–476.

[6] Honkela T, Kaski S, Lagus K, Kohonen T. Websom: "Self-organizing maps of document collections". In: Proceedings of Workshop on Self-Organizing Maps 1997 (WSOM'97), Espoo, Finland, June 1997.

[7] Kaski S, Nikkila J, and Kohonen T. "Methods for interpreting a self-organized map in data analysis". In: Proc. 6th European Symposium on Artificial Neural Networks (ESANN98), D-Facto, Brugfes, Belgium, 1998.

[8] Kohonen T. "Self-Organizing Maps", Second Edition, Springer 1997.

[9] Rasmussen EM, Willett P. "Efficiency of hierarchical agglomerative clustering using the icl distributed array oricessor". Journal of Documentation, 1989; 45(1).

[10] Saad K. Majeed, Muna Y. Saghir, "Using Genetic Algorithm in Image Clustering".

[11] Sanjay Silakari, Mahesh Motwani, Manish Maheshwari, "Color Image Clustering using Block Truncation Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 4, No. 2, 2009.

[12] Sudhakar G, Polinati Vinod Babu, Suresh Chandra Satapathy, Gunanidhi Pradhan, "Effective Image Clustering

with Differential Evolution Technique", Int. J. of Computer and Communication Technology, Vol. 2, No. 1, 2010

[13] Wang L, Khan L, Breen C. "Object boundary detection for ontology-based image classification". In: Third International Workshop on Multimedia Data Mining, Edmonton, Alberta, Canada, July 2002.