Indexing of Voluminous Data: Its Needs and Challenges

Minakshi Gogoi[1] **and Jayashree Das**[2]

[1]*Department of Computer Science and Engineering Girijananda Choudhury Institute of management and Technology*
[2]*M. Tech, 3rd Semester Department of Computer Science and Engineering Girijananda Choudhury Institute of management and Technology*
E-mail: [1]*minakshi_cse@gimt-guwahati.ac.in,* [2]*jayashree.kri.das@gmail.com*

Abstract—*In the present era, due to the technological advancement, the rate of growing of internet users are increasing rapidly with the exploration and expansion of huge volumes of databases. Simultaneously, the need for accessing numerous information in terms of various types of data like image, document, video has turned to be a vital part of most people's day to day life. Indexing of image data for fast retrieval and pattern search with higher efficiency and accuracy has turned into a challenging task in the present information retrieval scenario. Image retrieval is enhanced due to the increase in data volume throughout the internet over the decades and also to fulfill the requirement of various applications such as individual authentication, face recognition, biometrics, pattern search, remote sensing etc. Spatial images that covers spectral and non-spectral image data can be indexed based on their content features such as color, texture, shape, spatial layout etc. The techniques used for indexing of spectral data are of particular interest in various application fields such as content based remote sensing, agriculture, astronomy, biomedical imaging etc. Spectral images are nothing but the images of the same object taken in different bands of the electromagnetic spectrum. A spectral image may refer as a hyperspectral or a multispectral image data. They are multidimensional in nature. Other spatial data which are not related to the spectral bands are non spectral data. Currently there are a number of techniques available for indexing and query processing of these types of data (spectral and non spectral) such as pyramid technique, K-d tree, Map Reduce, R-tree, R+ tree, score based method etc. In this paper a brief overview of the current techniques is provided and also their implementation in the indexing of spectral and non-spectral data. Their advantages and limitations are analyzed and their performance efficiency is compared.*

Keywords*: spectral data, indexing, feature extraction, k-d tree, pyramid technique.*

1. INTRODUCTION

With the increase in database volume, fast indexing and retrieval techniques has been considered as demanding for the enhanced network and multimedia technologies. Earlier, managing large spatial databases was used in geosciences and computer aided design (CAD). Later, the newfound applications in computer vision and robotics, computer visualization, geographical information processing, automated mapping and facilities management etc has increased the need for fast and efficient indexing and retrieval of spatial data.

To fulfill the requirement of indexing of images many techniques have evolved in the last few decades. One way is the traditional image database indexing and retrieval approach which is text based. Here the image data is fully converted into an electronic presentation [8]. But with the increase in popularity of the internet and enhancement in multimedia technologies, this approach is disliked by people due to some factors such as lower quality text and higher cost. Difficulties of traditional indexing approach has led to raise the interest level in enquiring and developing the techniques for retrieving images automatically by using content features such as color, shape and texture etc.

The different existing techniques that can be used for indexing of spatial image data are K-d tree, MapReduce, R+-tree, score-based etc. A K-d tree (short for k-dimensional tree) is a space-partitioning data structure for organizing points in a k-dimensional space. K-d trees are a useful data structure for several applications, such as searches involving a multidimensional search key (e.g. range searches and nearest neighbor searches). MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster.

2. NEED OF INDEXING

The best way to express any information is through visual implementation rather than textual. An image or a picture can provide the user with the relevant information he need. Now since the visual content in the World Wide Web (www) along with offices, enterprises, industries having their own servers, is increasing enormously with time, the need for handling this huge volume of resource has become very much important. Every day the internet is loaded with billions of gigabytes of data. And this rate is increasing rapidly day by day. The democratization of images as proper sources of information and education in various fields such as agriculture, economics, astronomy etc resulted in the enhanced importance of providing improved image access. And to make this accessing of voluminous images easier and convenient, we need proper and efficient indexing of images.

Efficient search tools provided by Google, MSN and Yahoo which index trillions of freely available images online are clear evidence that image indexing and retrieval is important and widespread in today's visual world.

3. ISSUES IN INDEXING

There are a number of issues or challenges that are faced while developing an efficient algorithm for image indexing. Some are discussed below.

3.1 Scalability

The concept of scalability is desirable in technology settings. Scalability in this context means the ability that an indexing system should possess by virtue of which it will accept the increased volume of images without impacting the contribution margin. It means that the query result should not be affected if also the voluminous data increases. So an efficient indexing system should also be scalable.

3.2 Robustness

Robustness is the ability of an indexing system to cope with errors during execution or processing. Robustness can also be defined as the ability of an indexing algorithm to continue operating despite abnormalities in input, calculations, etc. In general, building robust systems that encompass every point of possible failure is difficult because of the vast amount of possible inputs and input combinations. But it is feasible to some extent.

3.4 Platform independence

Platform independence in this context means that the indexing system should be independent of the specific technological platform used to implement it. The platform may be hardware or software.

3.5 Miss-categorization

The interest in CBIR has grown because of the limitations inherent in metadata-based systems, as well as the large range of possible uses for efficient image retrieval. Textual information about images can be easily searched using existing technology, but this requires humans to manually describe each image in the database. This can be impractical for voluminous databases or for images that are generated automatically. It is also possible to miss images that use different synonyms in their descriptions. Systems based on categorizing images in semantic classes like "cat" as a subclass of "animal" can avoid the miss-categorization problem, but will require more effort by a user to find images that might be "cats", but are only classified as an "animal". Many standards have been developed to categorize images, but all still face scaling and miss-categorization issues.

4. TYPES OF VOLUMINOUS DATA

When we say voluminous image data, it comprises of all types of images, spatial, spectral, non spectral, big data etc.

A spatial data is a data which is related to pixels. In other words, in a spatial data such as an image, the image characteristics are represented in coordinates in 2 - dimensional, 3 - dimensional or a multi-dimensional space. So a spatial data may be classified as spectral or non-spectral data.

Spectral data or spectral image extends the capabilities of biological and clinical studies to simultaneously study multiple features such as organelles and proteins qualitatively and quantitatively. Spectral imaging combines spectroscopy and imaging. The combination of these two is, however, not trivial, mainly because it requires creating a three-dimensional (3D) data set that contains many images of the same object, where each one of them is measured at a different wavelength. A multispectral image is one that captures image data at specific frequencies across the electromagnetic spectrum. Hyperspectral imaging, like other spectral imaging, also collects and processes information from across the electromagnetic spectrum. The goal of hyperspectral imaging is to obtain the spectrum for each pixel in the image of a scene, with the purpose of finding objects, identifying materials, or detecting processes.

The main difference between a multispectral and a hyperspectral is the number of bands and how narrow the bands are. A multispectral image generally refers to 3 to 10 wider bands that are represented in pixels. Each band is acquired using a remote sensing radiometer. Whereas a hyperspectral image is consists of much narrower bands (10-20 nm). A hyperspectral image could have hundreds of thousands of bands. It uses an imaging spectrometer.

Now the most trending voluminous data is the big data. The concept of big data is like it means really a big data, which is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not merely a data, rather it has become a complete subject, which involves various tools, techniques and frameworks. Big data involves the data produced by different devices and applications. Black box data, social media data, stock exchange data, power grid data, transport data, search engine data, structured data, semi structured data, unstructured data are some of the fields that come under the umbrella of Big Data.

5. LEVEL OF DESIGN OF INDEXING SYSTEM

The design of an indexing and retrieval system mainly focuses on the emerging techniques. K-d tree, R tree, R+ tree, B+ tree, mapreduce, pyramid technique, graph based are some of these which play an important role in the design of indexing system.

Indexing is also a very essential part in the handling of big data. In the traditional approach, an enterprise will have a computer to store and process big data. Here data will be stored in an RDBMS like Oracle Database, MS SQL Server or DB2 and sophisticated software can be written to interact with the database, process the required data and present it to the users for analysis purpose. The limitation of this approach is that it works well only where we have less volume of data that can be accommodated by standard database servers, or up to the limit of the processor which is processing the data. But when it comes to dealing with huge amounts of data, it is really a tedious task to process such data through a traditional database server.

This problem is solved by Google using an algorithm called MapReduce. This algorithm divides the task into small parts and assigns those parts to many computers connected over the network, and collects the results to form the final result dataset.

But in the recent advancement in technologies, the latest technology to handle big data is the use of Hadoop. It is a programming framework that supports the processing of large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation. Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel on different CPU nodes. In short, Hadoop framework is capabale enough to develop applications capable of running on clusters of computers and they could perform complete statistical analysis for a huge amount of data.

6. RELATED WORK

Table 1: Indexing Approaches

Approach / Method used	Description
Hilbert Space-Filling Curve	Indexing multidimensional data can be achieved by the way of mapping multidimensional data into one dimensional data and then using a one-dimensional data indexing method. Hilbert curve is used here as it has superior clustering properties while comparing it with Z-order curve [4].
Pyramid Technique	The Pyramid Technique is an effective high-dimensional data mapping method, by which high-dimensional space point can be mapped to one-dimensional space to exploit one-dimensional indexing structure such as the B+-tree. Here, they presented a new approach for fast indexing the hyper-spectral data based on the Pyramid Technique, and further propose new algorithms to process k Nearest Neighbor queries and spectral partial feature queries [1].
K-d Tree with Feature Level Fusion	Here they first normalized a multi-dimensional feature vector of each trait and then, they projected it to a lower dimensional feature space. Then the reduced dimensional feature vectors are fused at feature level and the fused feature vectors are used to index the database by forming K-d tree. This method reduces the data retrieval time along with possible error rates. Comparing the performance with indexing based on score level fusion they found that this technique based on feature level fusion performs better [2].
Region based for content based retrieval	In their example system images are indexed and accessed based on properties of the individual regions in the image. Regions in each image are indexed by their spectral characteristics, as well as by their shape descriptors and position information. The goal of their system is to reduce the number of images that need to be inspected by a user by quickly excluding substantial parts of the database so that the system avoids exhaustive searching [3].
Graph based approach	Here they concentrated on a graph-theoretic approach to analyze the process of fingerprint comparison to give a feature space representation of minutiae and to produce a lower bound on the number of detectably distinct fingerprints. Their method provided a graph based index generation mechanism of fingerprint biometric data [10].

The image indexing and retrieval techniques deals mainly with applications such as identification of biometric databases, CBIR (content based image retrieval), face recognition, pattern search etc.

Basically there are two ways to index an image for retrieving – the traditional way which is text based and the other one is based on image content. Now-a-days importance is given on different indexing techniques which use image features such as color, texture, shape etc for indexing. But for indexing spectral data, the main thing that we need to deal with is dimesionality. Spectral data such as a hyperspectral image is multidimensional. Li, Jia and Wang, Cheng in their research used the pyramid technique which is an effective high-dimensional data mapping method, to map high-dimensional space point into one-dimensional space to exploit one-dimensional indexing structure such as the B+-tree. Here, they presented a new approach for fast indexing the hyper-spectral data based on the Pyramid Technique, and further proposed new algorithms to process k Nearest Neighbor queries and spectral partial feature queries [1, 9]. The hyperspectral data collected by remote sensors is represented as a high-dimensional spectrum with a spectral reflectance in each band. The hyper-spectrum can be seen as a point in high-dimensional space. So in that case, the Pyramid Technique can used to map the hyper-spectrum to a 1-dimensional value. But there are many invalid bands with negative or zero reflectance.

Instead, the valid value is a number ranging from 0 to 1, and the data space filled by various hyper-spectra is a normal space with the center point (0.5, 0.5, …, 0.5). In the mapping method mentioned above, the pyramid number i of a d-dimensional point is determined by its maximum distance to center point within all dimensions. Then the invalid bands in hyperspectral data will generate an error mapping result. To insure the reflectance of every band within the interval [0, 1] and exclude the impact of invalid bands, we inspect the hyperspectral data band to band, and revise the invalid reflectance to 0.5. They had started their task by building the index based on the Pyramid Technique for hyperspectral data. For a revised hyperspectral vector v, they first determined its pyramid value pv_v and then inserted the vector into a B+ -tree using pv_v as the key. And subsequently stored the hyperspectral data and its pv_v in the according data page of the B+ -tree so that update and delete operation can be done similarly.

Hilbert curve that has superior clustering properties while comparing with Z-order curve is used for Indexing multidimensional data that can be achieved by the way of mapping multidimensional data into one dimensional data and then using a one-dimensional data indexing method [4].

Again, multimodel data means data with multiple features can be indexed first by extracting features and then applying technique such as K d-tree [2]. The technique they used is based on K-d tree with feature level fusion which uses the multi-dimensional feature vector. Here a multi-dimensional feature vector of each trait is first normalized and then, it is projected to a lower dimensional feature space. Then the reduced dimensional feature vectors are fused at feature level and the fused feature vectors are used to index the database by forming K-d tree. If T_i be the feature vector of ith template in the database of d-dimension and is defined as follows:

$$T_i = [\ f_1, ..., f_d\ ]$$

And if the query image Q with feature vector of d-dimension is defined as $Q = [q_1, ...q_d]$, then $\forall j$, q_j may not be same as f_j, where f_j and q_j are the jth feature values of T_i and Q respectively. So for a given query template Q, the problem of identification system is to find the n nearest neighbors in the database consisting of N templates. As they said, this method reduces the data retrieval time along with possible error rates. Comparing the performance with indexing based on score level fusion they found that this technique based on feature level fusion performs better.

T. Kulcsar, G. Sarossy , G. Bereznai, R. Auer, and J. Abonyi in their research presented an idea to combine indexing and visualization techniques to reduce the computational requirement of estimation algorithms by providing a two dimensional indexing that can also be used to visualize the structure of the spectral database. Here the prediction is not to use the high dimension space but can be based on the mapped space too. After analysis they found that their method is able to segment (cluster) spectral databases and detect outliers that are not suitable for instance based learning algorithms [5].

Efficient representation and indexing are the basis for retrieval of images by content as well as associated external information. The region feature of an image is also an important thing based on the property of which images are indexed and accessed. Regions in each image are indexed by their spectral characteristics, as well as by their shape descriptors and position information [3]. The goal is to reduce the number of images that need to be inspected by a user by quickly excluding substantial parts of the database so that exhaustive searching through the image database can be avoided when a query is submitted.

Another approach for indexing is graph based. In this approach indexing is done based on the characteristics of a graph. A minutiae graph is created from each fingerprint image and then an index is generated for each fingerprint which is unique [10]. Here, the index is generated based on four parameters number of vertex, degree of each vertex, highest degree and number of vertices with same degree.

In the mapreduce technique, the voluminous data is mapped into subsets like that in clustering. A MapReduce program is composed of a Map() procedure that performs filtering and sorting, and a Reduce() procedure that performs a summary operation. It operates in three steps. The first step is the map step, the second is the shuffle step and the last step is the reduce step.

7. DESIGN OF AN INDEXING SYSTEM

The following figure is a preferred design of an indexing and retrieval system based on the image feature. Here, first the images in the database are given as input to the feature extraction process. Then the features are stored. Then the features are used to index the corresponding images in the index construction phase. The indexing can be done by one of the techniques like k-d tree, map reduce, pyramid technique, R tree etc. Now the query image is provided as the input to the query feature extraction process. The features of the query image is Extracted and stored in query feature. Then the query feature is matched with all database image features. The best match is given as output.

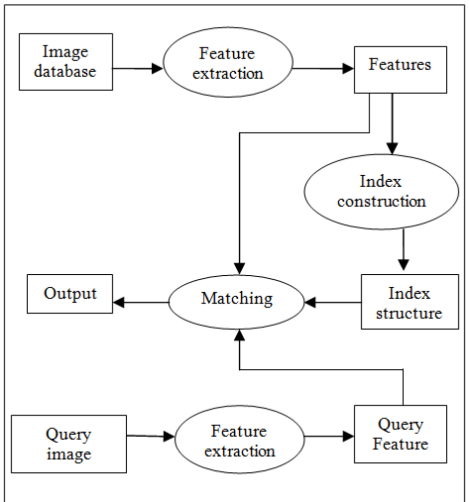

Figure1. Content Based Image Indexing and Retrieval Framework

8. CONCLUSION AND FUTURE SCOPE

Indexing using techniques such as k d-tree, X - tree, R - tree has limitations for large dimensionality. It cannot find the nearest neighbour efficiently. In very high dimensional spaces, the curse of dimensionality causes the algorithm to need to visit many more branches than in lower dimensional spaces. In particular, when the number of points is only slightly higher than the number of dimensions, the algorithm is only slightly better than a linear search of all of the points. But the algorithm can be improved. It can provide the *k*-Nearest Neighbors to a point by maintaining k current bests instead of just one. Branches are only eliminated when they can't have points closer than any of the k current bests. In contrast to all other index structures, the performance of the Pyramid-Technique does not deteriorate when processing range queries on data of higher dimensionality. New indexing techniques have to be developed which will eliminate the limitation of the existing techniques as well as provide more efficiency and accuracy in spatial image retrieval.

REFERENCES

[1] Li, Jia and Wang, Cheng, “Indexing Method for Hyperspectral Data Fast Retrieval by Pyramid Technique,” International Conference on Computer Science and Software Engineering, 2008.

[2] Jayaraman, U., Prakash, S., and Gupta, P., “Indexing Multimodal Biometric Databases Using Kd-Tree with Feature Level Fusion.”

[3] Barros, J., French, J., Martin, W., Kelly, P., and White, J. M., “Indexing multispectral images for content-based retrieval.”

[4] Lawder, J. K., and King, P. J. H., “Querying Multi-dimensional Data Indexed Using the Hilbert Space-Filling Curve.”

[5] Kulcsar, T. G., Sarossy, Bereznai, G., Auer, R., and Abonyi, J., “Visualization and Indexing of Spectral Databases,” World Academy of Science, Engineering and Technology, vol. 6, July 2012.

[6] Sellis, T., Roussopoulos, N., and Faloutsos, C., “The R+-Tree: A Dynamic Index for Multi-Dimensional Objects,” Proceedings of 13th International Conference on Very Large Data Bases, September 1987.

[7] Paliwal, A., Jayaraman, U., and Gupta, P., “A score based indexing scheme for palmprint databases,” Proceedings of 2010 IEEE 17th International Conference on Image Processing, September 26-29, 2010, Hong Kong.

[8] Rahmani, Md. K. I., and Sharma, R., “Image Indexing and Retrieval,” International Journal of Software and Web Sciences (IJSWS), ISSN (Print): 2279-0063, ISSN (Online): 2279-0071.

[9] Berchtold, S., Böhm, C., and Kriegel, H. P., “The Pyramid-Technique: Towards Breaking the Curse of Dimensionality,” Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 142-153, 1998.

[10] Gogoi, M., and Bhattacharya, D. K., "An Effective Fingerprint Verification Technique," Journal of Computer Science and Engineering, Volume 1, Issue 1, May 2010.

[11] Lu, P., Chen, G., Ooi, B. C., Vo, H. T., and Wu, S., “ScalaGiST: Scalable Generalized Search Trees for MapReduce Systems [Innovative Systems Paper].”

[12] Mhatre, A., Chikkerur, S., and Govindaraju, V., “Indexing Biometric Databases using Pyramid Technique.”