ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 7, July 2017

COMPREHENSIVE SURVEY ON CONTENT BASED IMAGE RETRIEVAL FRAMEWORK FOR BIG DATA

Mr.D.Mansoor Hussain[1], Dr.D.Surendran[2], Dr.V.Nandalal[3]
Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology
Coimbatore, India[1,2]
Department of Electronics and Communication Engineering, Sri Krishna College of Engineering and Technology
Coimbatore, India[3]
mansoor.slm@gmail.com[1]

ABSTRACT

Content Based Image Retrieval (CBIR) is one approach to retrieve or search pictures from the smartphone, tablet, desktop or website. There are several types of commonly used features in content based image retrieval including color, texture, shape and object with their relationship. Color, texture and shape features are the common and most popular features in literature. During this comprehensive survey, we found that several researchers used these three features in a better way and proposed multiple methods to provide a better retrieval system. Data set is generally high for this system and due to the explosive growth of multimedia data on the internet creates the need for a best retrieval system on this big data availability. All relevant researches are discussed in detail in this paper.

1. INTRODUCTION

Content-based image retrieval refers to the application of computer vision and data mining methods to the image retrieval problem, that is, the research of identifying similar digital images in large data sets. The word “content” here refers to the details of the image that will be used for retrieval rather than the traditional keyword, tag or description based search mechanisms. Effectiveness of the algorithm is subjective and there are multiple challenges in defining success of the CBIR system. Hence, the limitations in the metadata based systems prompts the need for content-based search and retrieval systems and we will look in to the different methods available for the same in this survey paper. Figure 1 represents the CBIR system:

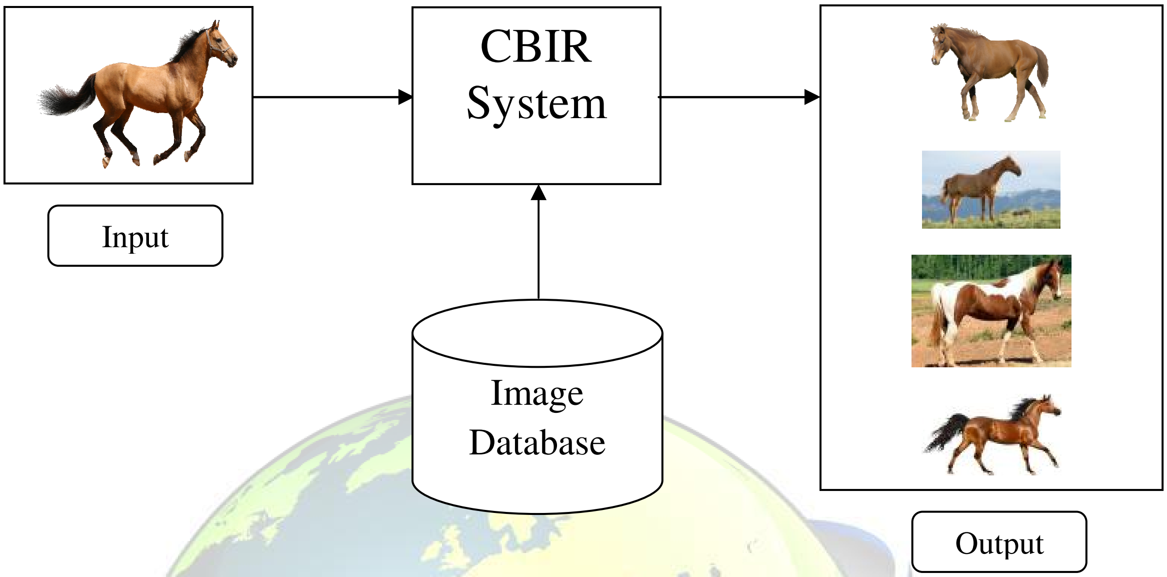

Figure 1. Content Based Image Retrieval

Big data on the other side refers to the data sets that are so large that the usual data mining algorithms may not be sufficient to deal with. Various challenges that are involved with the big data comprise analysis, data capture, exploration, curation, storage, sharing, transmission, imagining, questioning, and bring up-to-date. The human brain along with the visual system is capable of processing millions of multimedia data including audio, video and images from multiple sources simultaneously. With the advancement in technology, smartphones and tablets now record and share multimedia data at an incredibly increasing rate thereby forcing the human brain to process more and more.

Image analytics refers to the algorithmic extraction and analysis of the features from the images using digital image processing techniques. Up to 80 percent of all business and public unstructured multimedia-based data are available for usage. Figure 2 below represents the image analytics:

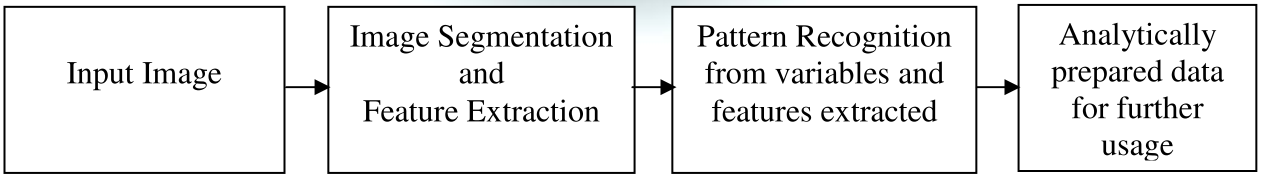
Input Image
Image Segmentation and Feature Extraction
Pattern Recognition from variables and features extracted
Analytically prepared data for further usage

Figure 2.Image Analytics

Segmentation and feature extraction forms the first stage of any content based image retrieval systems. Segments refer to the relevant regions or regions of interest that have a good set of features. It could be based on color, shape or patterns present in the image. As and when the size of data set increases, the process of feature extraction, segmentation and other process has to be made simpler and optimized for quicker retrieval of the results. Figure 3 below represents this big data mining process:

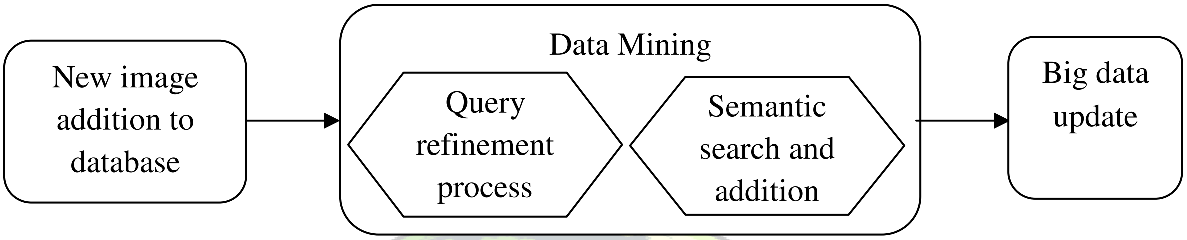

Figure 3. Big data mining

The CBIR system will be evaluated based on its ability to handle very large amount of data and is very important in image analytics and retrieval process. A framework is needed for handling these large datasets and Apache Hadoop is one such open source-programming framework that helps processing and storage of large datasets in a clustered computing environment.

Modules of this framework include libraries needed for running different programs on Hadoop framework, Hadoop Distributed File System (HDFS) for effective storage, Hadoop YARN for scheduling the resources and Hadoop map-reduce for parallel processing of the data. Figure 4 represents the multi node Hadoop cluster:

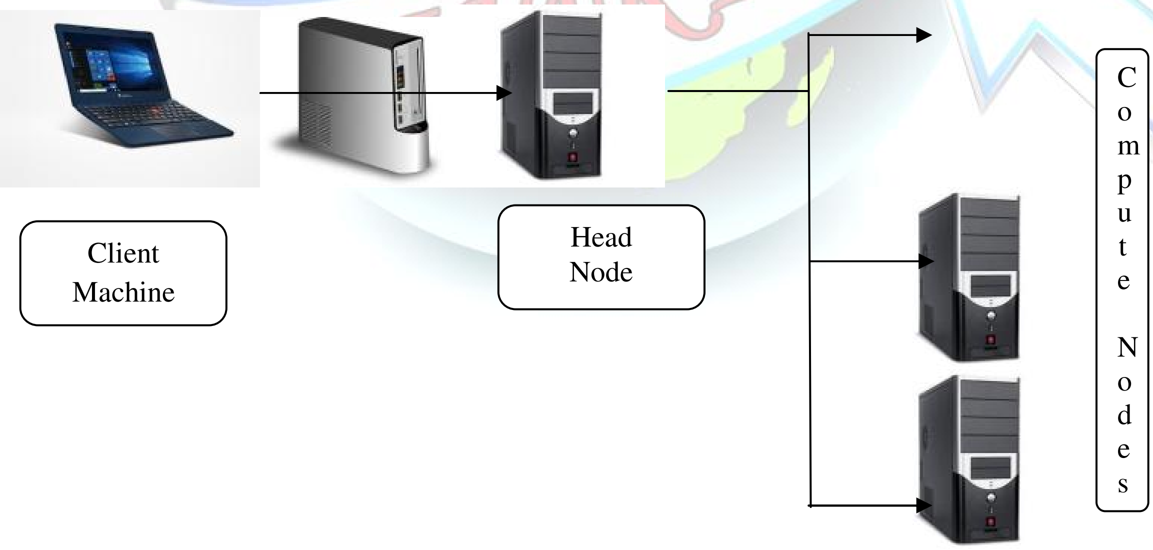

Figure 4. Hadoop cluster Representation

Hadoop cluster will generally contain a master node and a set of worker nodes or compute nodes to do the processing. The role of master node is to assign job, track it and so on. HDFS supports java programs and helps in storing the data while containing portable file system with the scalability, reliability and manageability features. It supports reading in parallel along with data processing like write, read, rename and append operations. It is optimized for streaming reads/writes of large files. Map-reduce program is composed of map function that helps perform sorting and filtering of data sets.

A stand-alone method of content based image retrieval is getting harder due to explosive expansion of information and to fulfill the load of storage and computing, an open source framework like Hadoop is very essential. Such models helps in enabling large datasets processing distributedly across clusters using simple programming techniques.

The major contribution of this paper is threefold:

1) Detailed survey on different feature extraction techniques for effective content based image retrieval.
2) Comprehensive analysis of CBIR framework for big data using open source framework like Hadoop.
3) Scope for future research directions on CBIR and fitting it in to big data framework proficiently.

The rest of the paper is structured as follows. Section 2 details the literature review. Analysis is made on two parts namely the content based image retrieval by means of feature extraction and comparison followed by Hadoop framework. Section 3 presents the analysis including the latest CBIR systems while section four concludes the review paper along with the future directions of the same.

2. LITERATURE REVIEW

In 1992, T.Kato [a] first introduced the terms Content Based Image Retrieval to the external world. He used color and shape features in his experiments and since then CBIR has become more famous pushing for effective and efficient retrieval of data from the database. Multiple techniques, tools and algorithms are used in the process, which are taken from signal processing, probability, statistics and pattern recognition fields [b].

Datasets that are large and complex are called as big data. Data sets grew multifold in the last decade or so due to the advancements in mobile devices, remote sensing, cameras, microphones, RFID’s and wireless sensor networks leading to big data storage and analysis. Though the term big data was coined around 1990’s by John Mashey [c], the actual size grew from few dozen terabytes in 2012 to petabytes of data recently. Apache Hadoop [d], an open source framework is used for processing these larger data sets. The first release of it was made in December 2011 while the latest stable release came in August 2016.

The next two subsection will give the literature review details of content based image retrieval system and the Hadoop framework solutions for big data processing.

2.1 Feature extraction and retrieval process for CBIR systems

Feature extraction refers to the process of building derived values from the original data set and is mostly related to dimensionality reduction. When the input data size is too large and also suspected to be redundant, then feature extraction [e] helps in reducing the data set and extract valuable information from it for further usage. In image processing, features extracted could be edges, blobs, color, texture, shape, lines, circles, ellipses etc.

Color refers to the human visual perception of an object with different names like red, green, blue, orange, purple etc. and the perception of color is derived from the cone cells present in the human eye by EM radiation in the spectrum of light. Though human eye gives more importance to brightness as compared to the color, it is widely used and needed in retrieval systems. There are different ways to extract color information from the image and detailed as follows.

John R. Smith et al [f] proposed a method for color extraction and indexing on image and video databases. They tried to identify the regions in the images that has color information's from predetermined color sets. They have experimented it over 3000 images and found the results to be encouraging as compared with the traditional color approaches.

Jun Yue and team [g] proposed a method to quantify HSV color space followed by color histogram formation. The characteristics including global color histogram, local color histogram are then compared and analyzed for content based image retrieval. They find that the fused features of color and texture brings better visual feeling as compared to the single feature based retrieval mechanism.

Afifi et al [h] analyzed the computational complexity and the retrieval accuracy of the CBIR systems and came up with an approach that uses both color and texture metrics for image retrieval. Color moment is obtained first from HSV space followed by texture feature extraction. They also feel that the combined color and texture feature gives better retrieval results as compared to using them independently.

Reshma Chaudhari and A. M. Patil[i] feel that the quality of CBIR system is heavily dependent on the methods used in feature extraction along with similarity measure techniques. Color coherence vector is used in this work for successive refinement and that helps in improving the accuracy of the retrieval results for them.

Guang-Hai Liu and Jing-Yu Yang [j] have proposed color difference histogram for content based image retrieval where they discuss a novel method for extracting the color feature from the image. In this approach, they calculate the perceptually unvarying color difference among points in diverse backgrounds along with edge orientations in LAB space format.

Manimala Singha and K.Hemachandran [k] used wavelets in their research and found that the color along with texture features help in developing a robust CBIR system in terms of scalability and translation of objects in an image. Wavelet Based Color Histogram Image Retrieval (WBCHIR) is the name of their approach and they have evaluated their solution by relating with the present systems in the literature.

Stehling et al [l] have described the complexity of the image databases that are becoming more and more common in different applications like search engines, medical, remote sensing, criminal investigations etc. They feel color is the most commonly used low level feature in CBIR system because humans easily perceive it when looking at the image, perceptions are easy to realize and implement, widely used in multiple image domains and the results are generally satisfactory. So, they have proposed ways to extract this useful color information along with its usage in content based image retrieval.

Texture means the feel or appearance of a surface. This feature looks for repeated patterns in the image [m] and how they are defined in the spatial environment. They are generally represented by texels. Textures could be classified through co-occurrence matrices, laws texture energy and wavelet transforms.

Image texture refers to the information identification about the color arrangement or intensities present in the whole image or part of it. It could be artificially created or naturally present in the scenes captured by the camera. It helps a lot in image segmentation and classification problems and used widely in content based image retrieval as well. Textures are commonly texel represented where they are placed into a number of sets providing details about the texture and its location in the image. Co-occurrence matrix were used for texture classification while laws texture energy; wavelet and orthogonal transforms can also be used for the same purpose.

Peter Howarth and Stefan Ruger [n] have worked out on a complete assessment of the use of texture features for image retrieval purposes. They used three diverse texture feature types including the statistical approach, psychological and image processing points of view for coming up with a texture based CBIR system. Corel and TRECVID2003 image database were used in their solution evaluation.

Hui Yu et al [o] used Fourier image transformation for representing the texture in an image and also derived eight distinguishing maps for relating the various aspects of co-occurrence associations of image pixels in each color space channel. The maps first and second moments are then calculated which represents the natural color image pixel distribution. They obtain a 48-dimensional metrics and named it as color texture moments for its usage in CBIR systems.

Guoyong Duan and team [p] has combined various features for content based image retrieval. They have focused on various feature extraction process along with their representations for better image matching methods.

Arnold W. M. Smeulders et al [q] have presented a assessment of almost 200 research papers in content based image retrieval. They have discussed in detail about the different patterns, picture types, semantic usages, sensory gaps along with the computational steps for image retrieval systems. Various features like Color, shape, texture etc. are discussed in this paper and are sorted by incremental global features, object and shape features along with signs and structural combinations.

Sundaram R and Satya Sai Prakash [r] have combined novel features for content based image retrieval. Their system combines different approaches to feature based queries. Fuzzy color histogram is used for extracting the color feature, tamura features for texture information and phase congruency for shape information in this research work. The proposed algorithm is tested on the animals and birds data set yielding 96.4% and 92.2% accuracy respectively.

Yossi Rubner and Carlo Tomasi[s] from Stanford University have proposed texture based CBIR system without any segmentation process. Their retrieval approach is based on the earth mover's distance with a suitable ground distance and is proven to deal with both complete as well as fractional multi-textured queries.

R. Bulli Babu, K. Sai Anish and V. Vanitha [t] have proposed a system that not only depend on color, shape and testure but also it traces the underlying image points. Color is the first feature used followed by texture and lastly tracing the underlying graphical structure. Irrelevant images are thus filtered out there by increasing the retrieval accuracy.

Shankar M.Patil [u] refers to the descriptors of texture including mean, standard deviation, angular second moment, inverse difference amount, sum average, contrast, correlations and sum variance. He describes that texture has better information as compared to the color histograms but sensitive to transforms like scaling and view angle.

Sreena et al [v] have used tamura feature for texture information extraction. A fuzzified distance measure technique called as fuzzy hamming distance is used in their approach and their database is ordered based on the similarity measure that is available to the user. The idea is implemented using MATLAB software and is verified against the standard Bordatz texture database.

M Vel Murugan [w] in his research paper have presented a CBIR method based on color and texture features using genetic algorithm and Euclidean distance approach. Gray level co-occurrence matrix is used for texture feature extraction and they find that the color and texture features together provide better retrieval accuracy when tested on google android mobile operating system.

Roshi Choudhary et al [x] defines CBIR as close to human semantics in terms of image retrieval process. They have identified its applications in valorous domains including medical image processing, crime prevention, weather prediction, surveillance and remote sensing. Color moment is used in their research for extracting the color information from the images, local binary pattern is used for texture detail extraction and then they combine both in to a single feature vector for storage and retrieval purposes. They feel that this combined approach provide them with accurate, efficient and less complex system.

Shape is another feature that is commonly used in image retrieval systems. It does not details the whole image shape but to a particular region of interest that is being sought out. Methods like image segmentation or edge detection is generally applied to an image to determine the shape of the region. The shape features that are being extracted should also be invariant to translation, scaling and rotation in order to have an efficient retrieval system.

Aradhana Katare and team [y] have presented a CBIR system for multi object images. They have proposed a GVF active contour method for efficient shape segmentation when there are multiple objects present in an image. Color features were also used in their approach for efficient system.

P.S. Hiremath along with Pujari [z] have detailed a novel framework for conjoining color, shape and texture features to achieve higher retrieval efficiency. They have partitioned the images in to small overlapping tiles of fixed size and the features are then extracted from the tiles. The color moments and gabor filter responses are extracted and stored for color and texture information. Shape information comes through edges of the image and computed using gradient vector flow fields. Thus they form a robust feature set which helps in building an effective and efficient image retrieval system.

Jagadeesh Pujari et al [a1] have used Lab and HSV color spaces to retrieve the edge features of the image. They have also compared their work against the Gray and RGB approaches. Their experimentation results prove that Lab color space gives better performance and accuracy in retrieval when compared with the traditional methods in the literature.

Mary Helta Daisy and others [a2] have used Gabor filters for extracting the texture features from the image. Fourier descriptors and centroid distances are used for shape information extraction and they combined both the features for better accuracy. Euclidean distance is used in finding the closest similarity between query and database images. Precision-recall graph is finally used for performance evaluation.

Deniziak et al [a3] have presented a method to query the database by approximate shape that is representing the given object. A set of geometric primitives and attributes are used to represent shape as per their approach. This method is useful in case of both transformed as well as partially covered objects.

Hwei-Jen Lin et al [a4] have used edge detection in extracting the shape information from the image. They made sure that the feature extracted is invariant to translation, scaling and rotation as well. For matching process, the sustaining deformation contour matching is preferred. Prompt edge detection method is used in detecting the edge points and the proposed method is compared against the Sobel edge detection technique. A novel shape demonstration method called Mountain climbing sequence is proposed in their work.

Amit Jain and others [a5] have presented a new approach for retrieving images with respect to a CAD model database. A linear approximation procedure is used in their method which helps in calculating the depth information of the image along with the 3D data. Similarity measure is then used that combines both shape and depth information for retrieval process.

Though the methods that are discussed above in the literature extracts the metrics in the uncompressed RAW image, most of the digital images in the internet today are JPEG compressed which necessitates us to extract the compressed domain features directly rather than wasting time and other resources in uncompressing the image before extracting the features. For this purpose, there are quite a few approaches available to directly form the feature set from the JPEG image.

Zhe-Ming Lu and others [a6] have discussed about the algorithms that can run directly on the DCT domain rather than uncompressing it for feature extraction. The color, spatial and texture features are mined in the DCT domain in their approach. A 12 dimensional feature vector set is formulated from the image set. They detail that DC component represent the information of energy, AC coefficients represent the frequency information while DCT coefficients in some region will represent the information of the direction in the image. Finally Euclidean distance is used to calculate the distance between the query image and the database images.

Padmashri Suresh and others [a7] have found that the unique usage of the features like kurtosis and skewness helps in improving the retrieval accuracy in compressed domain when used along with the statistical features. Their experimentation results prove that both the speed and the accuracy improves with the proposed algorithms.

Shih-Fu Chang [a8] has discussed the need for effective methods to index, retrieve and search images and videos from large collections stored in the repository. Manual text entry is both erroneous as well time consuming and hence they propose a method to retrieve the multimedia content directly in the stored compressed format without decompressing it. They have used wavelet sub band domain for color and texture information extraction from the still image while using motion vectors and transform coefficients for the video domain. They have also incorporated their work in to the video-on-demand text bed in the image and advanced TV lab at Columbia.

Precision and Recall are the parameters used in image retrieval system evaluation. Though there are many methods considered for the same, these two are widely used across different CBIR systems.

Zhang and others [a9] have described that ranking method is a key element of CBIR system. It can very well affect the retreival performance irrespective of the care taken in feature extraction process. A novel approach to ranking is also discussed which is based on the relative density rather than the traditional ranking approaches. Their method achieve optimal precision and recall values when tested using a database of images.

Huijsmans et al [a10] have detailed that CBIR system gets hampered due to the lack of good evaluation techniques. They have discussed the parameters that define a content based indexing and retreival process. In the evaluation site, all the details including the database, image queries and evaluation scripts are captured.

Veltkamp and Tanase [b1] have detailed a survey on content based image retreival system. They have provided the summary of the temporary image retreival systems in terms of features, queries, matching criteria's, indexing and presentation of results. They have discussed on systems that handles the low level features as well the high level semantics. They have cited the CBIR products from both commercial/production and research systems.

2.2 CBIR based on Hadoop framework – Big data Analytics

Danah Boyd and Crawford [b2] has provided the six provocations for big data including a) automating investigation changes the knowledge definition, b) claims related to objectivity and accuracy are deceptive, c) bigger data does not just represent the better data, d) Not all data are alike, e) it's not always ethical though accessible, f) limited access to Big Data creates new digital divides.

Hadoop is an open source programming framework based on JAVA that supports both the storage as well the processing of large data sets in a distributed environment. It is the platform for big data structuring and solves the problem of organizing it for following analytics purposes. It is created by Doug Cutting and Mike Cafarella in 2005.

Chunhao Gu and Yang Gao [b3] have introduced CBIR system based on Hadoop and Lucene. Performance bottlenecks get overcome with their approach that are brought by computational complexity and big data when building a CBIR engine.

Said Jai-Andaloussi et al [b4] detail that medical images are digitized and stored in large image databases. They apply the MapReduce distributed computing model along with the HDFS storage model in their solution to build a CBIR system. The content of the image is categorized by a) Bi-dimensional Empirical Mode Decomposition with Generalized Gaussian density functions and b) Bi-dimensional Empirical Mode Decomposition with Huang-Hilbert Transform HHT.

Zehra Çamlica and others [b5] have discussed the need for reducing the memory requirements and computational complexity of CBIR systems. They look at the Autoencoding errors of image blocks or tiles to make the decision on the retreival task. Feature dimensionality is hence reduced in their work which in turn speed up the retreival process. Local binary patterns and Support vector machines are used by them in order to validate their proposed scheme. Their experimental results prove that reduced dimensionality helps increase speedup greater than 27% while reducing the accuracy by just 1% using precision and recall parameters.

Juan M. Banda et al [b6] have discussed about big data processing techniques for NASA's solar dynamics observatory mission. They have brought out the importance of moving from traditional data mining and machine learning algorithms to more scalable big data methodologies. They have discussed about multi-label classification in their research work.

Hinge Smita and others [b7] have introduced a new method that can help handle large amount of data along with producing high level of accuracy in retrieval process. They handle large data through parallel processing technique.

Noha A. Sakr et al [b8] proposed an efficient and a fast response CBIR method based on Hadoop Map reduce technology. They have used a chain clustering binary search tree in order to build the pictorial statements for image representation. They have also introduced a methodology for representative's creation for big data high dimensionality solution.

Gautam Muralidhar [b9] detailed that an efficient CBIR system requires different components including the database collection, feature extraction, machine learning algorithms and also mentioned that these components need to run efficiently. They have also demonstrated the best way to realize the concept using pivotal HD with HAWQ and SQL engine for Hadoop.

U.S.N. Raju et al [b10] have presented an approach for implementation of CBIR system on Hadoop Map reduce framework. They discuss in detail about Local tetra patterns (LtrPs) in their paper for the same. Their work can be used for most CBIR techniques that uses distance measures and feature vector computations in it.

3. ANALYSIS OF CBIR SYSTEMS

There are lot of commercial systems [c1][c2] that are developed using different methodologies including the following:

a) Google image search
b) Pixolution
c) Just visual
d) Picalike
e) Elastic vision
f) Yandex image search
g) Baidu image search
h) Imense Image search portal
i) Imprezzeo image search
j) Incogna image search
k) Chic engine
l) Piximilar
m) Empora search engine
n) Galaxy
o) Macroglossa visual search
p) Querbie

The details of these CBIR systems are discussed in table 1.

Table 1: Commercial CBIR systems

CBIR system Name	Description of the tool	Size of the database	Organization type	License type
Google Image Search Engine	Google images is a search service that helps users to search for image data in the web. Primarily introduced in 2001, it accepts both keywords as well images itself as input.	3 peta bytes	Public company	Google Custom Search API access
Pixolution	Pixolution helps in finding similar images, image tagging assistant, find web images in the collection, search images by color and find images with space for text or logos.	32 Million	Private company	Closed
Just visual	The JustVisual application allows creators to access and mix the just	Billions	Start up	API

	visual functionality. The main API method helps in uploading images and to search for visually like images from the database.			
Picalike	Picalike is for mobile and ecommerce and helps show their clients what they like, similarity means relevance in their tool and they also provide individual control as well as easy integration.	No details provided	Private company	Closed
Elastic Vision	Elastic vision is a free software program and image search tool with content based clustering. It is a smart image searcher application.	No details provided	Private company	Closed
Yandex	It is a Russian multinational company working on internet related services and operates the largest search engine in Russia. It searches for images on the web including search by image itself.	10000M	Public Company	Closed
Baidu	It is Chinese web services company and one of the largest internet companies. Their search engine helps in websites, audio files and images.	1000M	Public company	Closed
Imense Image Search	Imense helps in picture search, ID reader, Annotator, Form reader, similar search and auto tagger for its clients.	3 Million	Private company	Closed
Imprezzeo	It is an image-to-image search engine. Their proprietary search software combines content based image retreival with facial recognition technologies.	No details provided	Private company	Closed
Incogna	Incogna is another image search engine that arranges its images based on content and when the user click on any particular image, the tool will retrieve similar looking images.	100 Million	Private company	Closed
Chic	It is a visual fashion search engine.	No	Private	Closed

	They have their application developed where the user can enter a photo of fashion and the tool will find it for sale online immediately.	details provided	company	
Piximilar	This search engine works by color palette uniformity and use a technology called Piximilar visual. The tool analyses image qualities like color, texture, shape, luminosity etc. and retrieves the results.	3 Million	Private company	Closed
Empora	Empora is a fashion search engine which helps its customers to shop from thousands of the most fashionable stores and brands in one place.	0.5 Million	Private company	Closed
Galaxy	Galaxy provides the image recognition platform along with the computer vision consulting. They provide search technology as a cloud based API service.	35 million	Private company	Closed
Macroglossa	It is a visual search engine based on image comparison. To this application, the users can upload photos and the tool retrieves back the results based on specific search categories.	No details provided	Private company	Closed
Querbie	Querbie allows users to search objects and images with image as input. It is a general purpose CBIR search engine.	20 Million	Private company	Closed

QBIC (Query by image content) is the first viable content based image retreival system [c4]. The framework and methods used in it have thoughtful effects on later image retreival systems. QBIC takes queries in the form of example images, drawings or sketches, texture patterns etc. High dimensional feature indexing is taken care in QBIC systems. Beyond the query by example method, Jain and Gupta proposed [c5] a nine-component framework, which provide better results, compared to the original system.

Virage, another CBIR search engine is like QBIC and supports graphic queries based on color, texture, composition and structure. It also give provisions for arbitrary combinations of these four atomic queries. The users also have the provision to adjust the weights themselves based on their interests and emphasis. The virage provides an open framework for developers to 'plug in' primitives in order to explain specific problems in managing images. Few other commercial CBIR systems include Retreival wave, photobook, visualseek, webseek, Netra, MARS and ART MUSEUM. [7] discussed about efficient content-based medical image retrieval, dignified according to the Patterns for Next generation Database systems (PANDA) framework for pattern representation and management. The proposed scheme use 2-D Wavelet Transform that involves block-based low-level feature extraction from images. An expectation–maximization algorithm is used to cluster the feature space to form higher level, semantically meaningful patterns. Then, the 2-component property of PANDA is exploited: the similarity between two clusters is estimated as a function of the similarity of both their structures and the measure components. Experiments were performed on a large set of reference radiographic images, using different kinds of features to encode the low-level image content. Through this experimentation, it is shown that the proposed scheme can be efficiently and effectively applied for medical image retrieval from large databases, providing unsupervised semantic interpretation of the results, which can be further extended by knowledge representation methodologies.

There arealso many other CBIR research projects and open source codebase available including akiwi, ALIPR, Anaktisi, BRISC, digiKam, Caliph & Emir, FIRE, GNU image finding tool, ISSBP, imgseek, IKONA, IOSB, LIRE, Lucignolo, MIFILE, MUVIS, RETIN, SIMBA, VIRAL, Windsurf, PIBE and SHIATSU.

4. EVALUATION STRATEGIES

Statistics, opinions and reviews will help to make a decision when there are different competing products available. Similarly, in CBIR as well, evaluation becomes a critical issue. A benchmark would help select from different discussed ideas and to test new solutions against the older ones. Following aspects are important in any information retreival systems:

a. An appropriate data set for evaluation is essential for testing and measuring the retreival performance
b. Bridging the semantic gap between the users query and details present in the image which is a ground truth for relevance
c. Metrics that helps in determining the content based image retrieval system retrieval accuracy

Two of the most popular metrics that helps in measuring the performance of a CBIR system are the Precision and Recall [c3] parameters.

Precision refers to the percentage of the retrieved digital images that are related to the user query while Recall on the other hand refers to the percentage of all relevant images in the database, which are retrieved. The same is given below:

Precision = Number of relevant pictures retrieved / Total number of pictures retrieved
Recall = Number of relevant pictures retrieved / Total number of relevant pictures in the database

Precision-Recall graphs helps in measuring the image retreival system accuracy. One such precision-recall graph is given in fig 5. We can also interpret precision and recall not just as ratios but as probabilities. Precision is the probability that the retrieved picture is relevant while recall is the probability that the relevant picture is retrieved in a search. Another measure that combines both the precision and recall is called as the F1 score which is given as:

F = 2. Precision . recall / precision + recall

This measure is the average of the two metrics when they both are close and is more commonly the harmonic mean, which, for the case of two numbers, corresponds with the square of the geometric mean separated by the arithmetic mean.

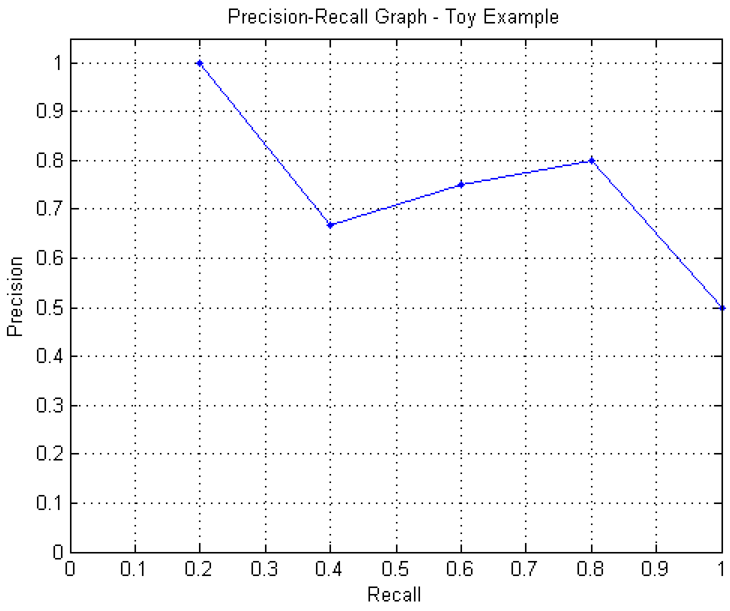

Figure 5. Precision – Recall Graphical Illustration

Though metrics are available for CBIR system evaluation, there are quite a few problems in the same including the following:

a) Defining a common image collection
b) Obtaining relevance judgements

Error rate is a single precision value which corresponds to the number of non-relevant pictures retrieved to the total number of pictures retrieved.

Error Rate = Number of non-relevant pictures retrieved / Total number of pictures retrieved

Hence there is a need for standardized evaluation metrics as different measures are available with slight variations but with the same definition. Also to overcome this problem, a set of standard performance metrics and a standard digital image database is needed. Interactive performance assessments including user relevancy feedback and interaction is essential to build a robust CBIR system.

5. OPEN AREAS

Image similarity search is an important research topic with multiple applications across different domains and industries. Though there are different algorithms and solutions provided in the literature for efficient content based image retreival, there are still several open areas need to be analyzed for the current system to be used practically well.

There is a huge gap between the human and the computer and we need to explore the synergy here. Recent research emphasis is on interactive systems where the user is allowed to give feedback based on the retrieved results and the system will update accordingly. Humans use high level concepts while the imaging algorithms mostly extract low level features. For specific applications like face recognition or finger print identification, this might be useful but for generic applications, these low level features do not have a direct relation to the high level concepts. Further processing through supervised and unsupervised algorithms, neural networks, genetic algorithms etc. help in offline processing to improve the results.

Web based search engines are most needed due to the enormous growth in the multimedia in the World Wide Web. Solutions exist for text based search as of today while multimedia search is still under progress. Technical breakthroughs are still expected to match with the text based search engine retreival results.

High dimensional indexing is the next focus area which is a web expansion by-product. While most of the existing systems could handle only few thousands of images, there is a need for managing millions of images over the internet. Performance evaluation criteria along with identifying a standard test bed is also an open area in developing a CBIR system.

The human perception of image content has to be studied in detail because it’s the humans who are going to use the system or the application for their own purposes. Relevance feedback plays a major role in such systems which has to be studied in detail for developing a robust system. More focus should be on psychological aspects of the human perception in viewing the image. The combination of perception based image features along with metrics helps achieve semantically meaningful outputs.

Building true image databases itself is a major research area associated with CBIR systems. Though lot of efforts are put to develop a good data set, the systems are not that efficient yet. Interdisciplinary research effort is required in building a successful image database system. Integration of multimodal and multimedia content will provide great prospective for indexing and image classification in different domains.

The challenging issues are listed below:

a) Representation of the image
b) Similarity characterization of the images
c) Image annotation through machine learning
d) Formulation of queries
e) Database organization
f) Indexing of the images
g) Result display through ranking mechanisms and assessment
h) Relevance feedback from the user and handling it effectively
i) Updating the database and feedback to improve the results

The above open areas needs to be addressed in order to build a robust and reliable content based image retreival system. Though there are multiple commercial systems available in the market, addressing these issues will help make them popular and also open doors for new systems to enter the world of technology.

6. CONCLUSION AND FUTURE WORK

We have presented a complete survey on content based image retrieval and big data analytics related to it. We have highlighted the current progress in this area, the emerging directions and evaluation methods of content based image retreival systems. There is still lot of scope in this area considering the fact that machine learning, artificial intelligence and data mining techniques gets updated along with the multimedia data size growth in the internet. Robust and dependable image understanding technology will still continue to grow and the future of content based image retrieval depends on collective focus in each aspect of the image retreival.

REFERENCES

[1] T. Kato. 'Database architecture for content-based image retrieval', Proceedings of SPIE Image Storage and Retrieval Systems, volume 1662, pages 112–123, San Jose,CA, USA, February 1992.

[2] Eakins, John, Graham, Margaret. 'Content-based Image Retrieval', University of Northumbria at Newcastle, from the JISC Technology Application Programme. October 1999.

[3] John R. Mashey, 'Big Data and the Next Wave of InfraStress' (PDF) slides from invited talk. Usenix, 25th April 1998.

[4] The Apache Software Foundation, 'Hadoop Distributed File System Architecture Guide', 2008.

[5] Gaurav kumar, Pradeep kumar Bhatia, 'A Detailed Review of Feature Extraction in Image Processing Systems', IEEE Fourth International Conference on Advanced Computing & Communication Technologies, At Rohtak, Haryana, India, February 2014.

[6] John R. Smith and Shih-Fu Chang, 'Single Color Extraction and Image Query', International Conference on Image Processing (ICIP-95), Washington, DC, Oct. 1995.

[7] Christo Ananth, K.Kalaiselvi, C.Kavya, S.Selvakani, P.Sorimuthu Iyan, "Patterns for Next generation Database Systems - A study", International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE), Volume 2, Issue 4, April 2016, pp: 114-119

[8] A. J. Afifi, W. M. Ashour, 'Content-Based Image Retrieval Using Invariant Color and Texture Features', International Conference onDigital Image Computing Techniques and Applications (DICTA), 2012.

[9] Reshma Chaudhari and A. M. Patil, 'Content Based Image Retrieval Using Color and Shape Features', International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 1, Issue 5, November 2012.

[10] Guang-Hai Liu, Jing-Yu Yang, 'content based image retrieval using color difference histogram', Pattern Recognition, Volume 46, Issue 1, January 2013, Pages 188–198.