Distributed Retrieval of Images using Particle Swarm Optimization and Hadoop

Darsana B
Assistant Professor
The Oxford College of Engineering
Bangalore, India

G. Jagajothi,Ph.D
Head, Dept. of IT
Periyar Maniammai University
Tanjore, India

ABSTRACT
Content based Image Retrieval (CBIR) is the problem of searching for digital images in large databases. It is the vital application of computer vision techniques to the image retrieval problem. One inherent problem associated with Content based Image Retrieval is the response time of the system to retrieve relevant result from the image database. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers. The parallel processing of Hadoop can be leveraged to efficiently retrieve images with very less response time. The proposed approach also avoids the semantic gap in image retrieval by utilizing automatic relevance feedback and meta-heuristic optimization. Automatic relevance feedback is implemented using Latent Semantic Analysis, and Particle swarm optimization provides meta-heuristic based development. The goal of proposed approach is to – cluster relevant images using meta-heuristics in less amount of time effectively.

General Terms
CBIR, Distributed File Systems, Image Processing.

Keywords
Content based Image Retrieval, Latent Semantic Indexing, Meta-heuristics, Parallel Processing, Semantic gap.

1. INTRODUCTION
An image retrieval system is a software system for browsing, searching and retrieving images from a large database of digital images. This is of significant importance in the current scenario where in the image archives are growing in rapid speed in every stream of the society, ranging from personal collection of pictures to critical medical images. One of the easier image retrieval techniques is meta-data based approach, where the associated metadata such as keywords, text, etc are used to retrieve images. But, annotating lakhs of images is a practically very difficult, and it is dependent on the subjectivity of human perception. It will bring up evident differences in the retrieved results, since search tags highly varies based on the perception of the human.

The principal challenge of semantic gap is solved by Content based Image Retrieval [1], which filters out images based on the similarities in their contents to the query image. It extracts images' visual signature by analysing various content related data from the image. The performance of image-centric retrieval systems is not satisfactory primarily due to the mismatch between the user's implied concept and the low level visual features. In order to narrow this gap, relevance feedback was introduced as an interactive tool in CBlR [2].
Although relevance feedback can significantly improve the retrieval performance in CBlR systems, the key issue in relevance feedback approaches is how to incorporate positive and negative examples in query and/or the similarity refinement [3].
This paper is organized as follows - Section 2 briefly summarizes the related work on image retrieval, relevance feedback, meta-heuristics and distributed database systems. The proposed approach is described in Section 3. Section 4, presents the experimental setup, results and discussions. Finally, Section 5 draws the conclusions and identifies future research directions.
In essence, we ask you to make your paper look exactly like this document. The easiest way to do this is simply to download the template, and replace the content with your own material.

2. RELATED WORK

2.1 Content Based Image Retrieval
The problems of image retrieval are becoming widely recognized, and the search for solutions an increasingly active area for research and development. CBIR differs from classical information retrieval in that image databases are essentially unstructured, since digitized images consist purely of arrays of pixel intensities, with no inherent meaning. Huge number of images makes it difficult to locate the image searched for, especially when the search is based on the aesthetic value of the image.

CBIR draws many of its methods from the field of image processing and computer vision, and is generally regarded as a subset of that field. It differs from these fields principally through its emphasis on the retrieval of images with desired characteristics from a collection of significant size.
Most of the research has concentrated on feature extraction of an image, e.g., QBIC [4]- which queries images based on their color, texture and shape, VisualSeek [5] - A Fully Automated Content-Based Image Query System, SIMPLicity [6] - Semantics- Sensitive Integrated Matching for Picture Libraries, Blobworld [7] - Image segmentation using expectation-maximization algorithm, Virage [8] - An open framework for image management, [9] specifies how CBIR techniques can most profitably be used and applies the same to image querying.

2.2 Relevance Feedback
One inherent problem with meta-data based approach is that, it is difficult to match the semantic of the image and the subjectivity of human being. The basic Relevance Feedback

mechanism relies on iteratively asking the user to discriminate between relevant and irrelevant images on a given set of results. The resultant feedback drives the next iteration of refined search, based on user input. A binary Relevance Feedback is used to train neural network systems as in PicSOM [10] and in Bordogna and Pasi [11]. Relevance feedback suffers from few vital problems - User interaction for providing feedback is time consuming and it is a tiring process. In order to curb these cons, automatic relevance feedback is used. This eliminates the user interaction totally in the feedback process, thus saving time and energy to a great extent.

2.3 Meta-Heuristics

Biological behaviors drive development of optimization algorithms for various diverse applications. Bio-inspired meta-heuristic optimization approaches provided new ways to achieve nearly-optimal solutions in highly nonlinear, multidimensional solution spaces, with lower complexity and faster convergence than traditional algorithms. Particle Swarm Optimization (PSO) is one such bio-inspired meta-heuristic algorithm, with stochastic nature, inspired by social behavior of bird flocking or fish schooling, introduced in the field of computational intelligence by Kennedy and Eberhart [12] in 1995.

The system is initialized with a population of random solutions and searches for optima by updating generations. The potential solutions, called particles, fly through the problem space by following the current optimum particles. PSO has been successfully applied as an efficient optimization tool in image classification [13]. Particle swarm optimization is used for optimization of local and global feature of weights in Image Retrieval [14].

2.4 Hadoop Distributed File Systems

Hadoop is an open-source software framework that supports data-intensive distributed applications. The Hadoop framework transparently provides both reliability and data motion to applications. It is well suited for distributed storage and distributed processing using commodity hardware. It is fault tolerant, scalable, and extremely simple to expand. Map Reduce, well known for its simplicity and applicability for large set of distributed applications, is an integral part of Hadoop.

HDFS is highly configurable with a default configuration well suited for many installations. Most of the time, configuration needs to be tuned only for very large clusters. In addition, it provides a distributed file system that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. [15] describes an approach for finding image descriptors or tags that are highly reliable and specific using Hadoop. [16] proposes an open source cloud computing deployed with Hadoop which enables content based image retrieval system.

3. PROPOSED APPROACH

This paper proposes a novel approach of, distributed retrieval of images using particle swarm optimization and Hadoop file systems. It addresses the problems of semantic gap and delayed response time in content based image retrieval by coalescing automatic relevance feedback, a stochastic algorithm and distribution of image retrieval.

The image databases will be replaced with the Hadoop file system on a distributed cluster. The name node acts as arbitrator and repository for all HDFS metadata. It executes file system name space operations and determines mapping of blocks to data nodes. The data nodes serve read/write requests from client. It performs block creation, deletion and replication upon instruction from name node. It stores HDFS data in files on local file system and determines optimal file count per directory.

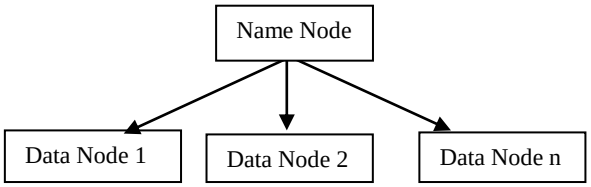

Fig 1: Distribution of Image database in Hadoop File Systems

Each client runs the following sequence of processes to retrieve the relevant results. Input query image is obtained by the system from the user and visual feature is computed for the input image based on the signature composed from Edge detection. The same technique is used to form a visual feature database from the image database using the feature vector calculated from Edge detection. This visual feature database is constructed online for the query vector and offline for the database. This reduces the overhead imposed on feature vector calculation for the images in the database.

Query image is mapped into a feature vector and the distance between the query and image is calculated. The system ranks the whole dataset according to a minimum distance criterion. The distance is the sum of weighted Euclidean distances between pairs of feature vectors – feature vector of the database image and feature vector of the query image.

The Weighted Euclidean Distance WED(X,Y) is calculated as follows:

$$WED(X,Y) = 1/S \cdot \sum_{s=1}^{S} (Xs-Ys)^2 . W_s k \quad \text{----- (1)}$$

where, X,Y are feature vectors corresponding to the query image and the images in the database, S is the dimension of the feature vector, $W_s k$ is the weight associated with feature vector, k is the iteration number. If k=1, $W_s k = 1$ for all values of s.

Based on the computed distance, the nearest images are retrieved from the database and are routed to automatic relevance feedback. Here the images are split into relevant and irrelevant subsets. For the retrieved images, first automatic relevance feedback based on Latent Semantic Indexing (LSI) is generated. Here, LSI is applied in both textual and visual (image key) feature space. The textual feature space is constructed by using the keywords related to the image.

If there is no associated keyword with an image, then textual feature space calculation is automatically ignored by setting the value of α to 0. The visual feature space comprises of a feature vector, which is a combination of Color histogram bins and wavelet texture energy values. The combined similarity of textual and visual feature spaces is evaluated and the images are labeled as relevant or irrelevant, based on the similarity value.

$$TSim(q,i) = \alpha . Tsim + (1- \alpha) . Vsim \quad \text{---- (2)}$$

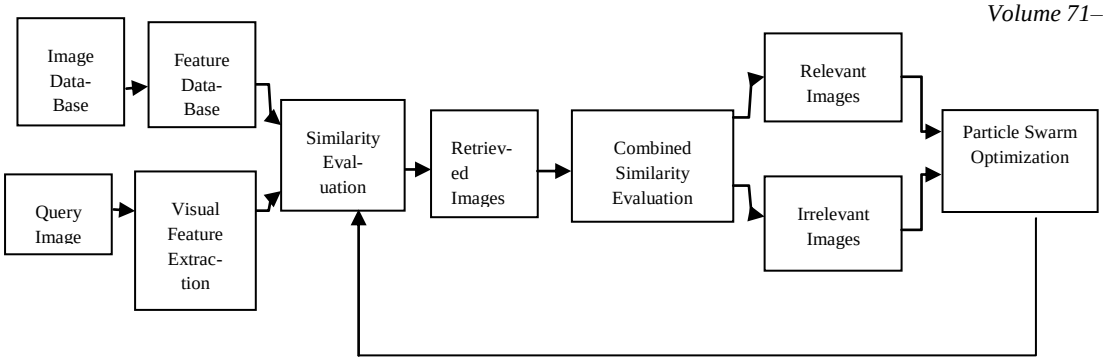

Fig 2: Retrieval of Images using Particle Swarm Optimization

where, α is a constant value. To use only the textual feature space of the image, α value is set to 1, and to use only the visual feature space of the image, when no keywords are associated with the image, α value is set to 0.

Here Tsim represents the textual similarity measure, and Vsim represents visual similarity measure. Accordingly, relevant and irrelevant image subsets are created, which will be progressively populated across iterations, based on the change in weights of individual features, thus changing the distance between the query image and the database images.

The feedback drives a feature re-weighting process and is routed to the particle swarm optimizer. The first iteration of the feature re-weighting process considers equal weight for all the features. From the second iteration on, the weight for feature varies based on the importance of the feature in the current iteration. This serves as the main tool for optimizing the results retrieved from the database.

Particle swarm optimizer provides a powerful optimization tool and an effective space exploration mechanism. A very preliminary version of PSO-CBIR was presented in [17]. After the relevant images are computed through automatic relevance feedback, the swarm is initialized as follows: Initialize each particle of the swarm.

Calculate the fitness value of individual particles. If the fitness value is better than the best fitness value(pBest) in history, then set the current value as the new pBest. After pBest is computed, gBest is computed based on the pBest values. Choose the particle with the best fitness value of all the particles as the gBest.

Calculate the particle velocity for every particle based on the following equation.

v[] = v[] + c1 * rand() * (pbest[] - present[]) + c2 * rand() * (gbest[] - present[]) ---- (3)

where, v[] is the particle velocity, pbest[] and gbest[] are defined as stated before. rand () is a random number between (0,1). c1, c2 are learning factors.

The position of the particles is updated in each iteration, based on the following equation.

present[] = persent[] + v[] ---- (4)

where, persent[] is the current particle (i.e., the solution). This updating is continued till maximum iterations or a minimum error criterion is attained.

4. Experimental Setup and Results

The experimental data comprises of collection of generic images from the Corel image database (http://www.corel.com) [18], SIMPLicity[6], [19] and the web images. We prepared four classes composed of different categories – bag, bat, beer mug and antique.

In this initial implementation, Hadoop is configured to run in pseudo mode. The reduction in response time is inversely proportional to the number of data nodes. The feature vector is computed offline for the database images. This feature vector semantically represents the images in the database. The query image is chosen by the user, and the retrieval processes for the relevant images are triggered. Based on the Euclidean distance between the combined feature vectors, the relevant images are retrieved by the system and the first iteration commences.

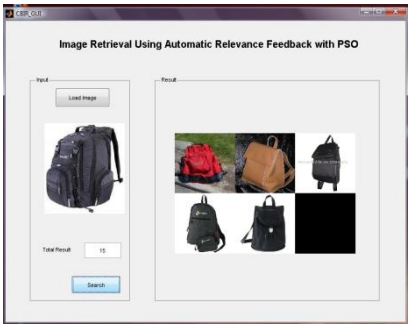

Fig 3: Query Image Selection

The retrieved images are passed for automatic relevance feedback using Latent Semantic Indexing. Here textual feature space and visual feature space are computed for the retrieved images. The combined feature space is used to provide relevant and irrelevant subsets of images, with respect to the query image. Every iteration results with different relevant images, and the top most images of all the iterations are listed as the final result.

In the Particle Swarm Optimizer, the particles are positioned corresponding to the first set of retrieved images and swarm initialization is done. The fitness value of each particle is calculated. The particle with least fitness value represents the best fit image for the query, and that is maintained as global best for that iteration.

From the second iteration, the swarm is split into multiple of two, with respect to the current number of swarms, with each swarm having its own global best value. From then on, the particles start moving towards their corresponding swarm's global best.

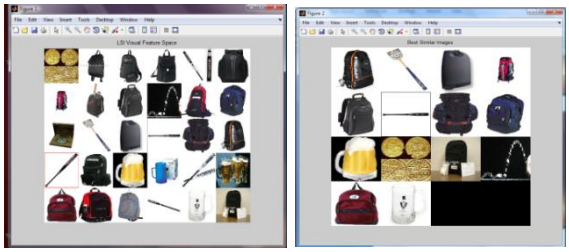

Fig 4: Images retrieved - iteration 1 and iteration 2

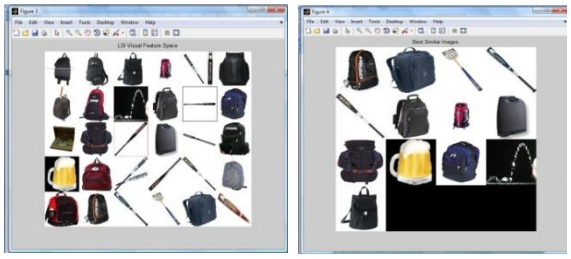

Fig 5: Images retrieved - iteration 3 and iteration 4

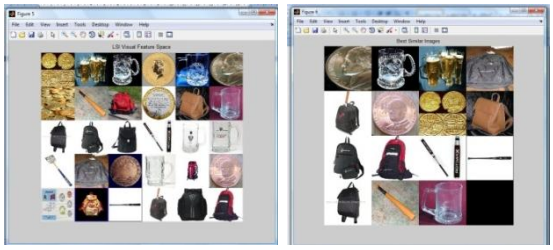

Fig 6: Images retrieved - iteration 5 and iteration 6

The position of the particles is updated in each iteration, based on a random vector value. The random vector value [12] is calculated based on an inertial weight factor which is fixed in the range [0.2,0.7] and two positive constants C1, C2 - called acceleration coefficients, aimed at pulling the particle towards the position related to the personal best and global best (C1=C2=2).

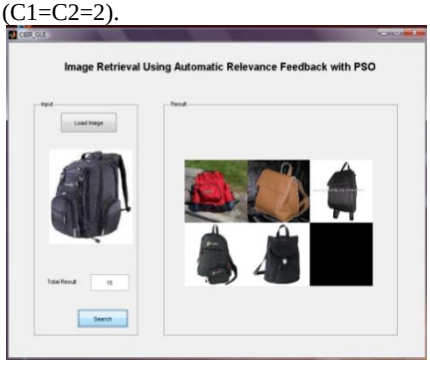

Fig 7: Final Retrieved Relevant Images

The important difference of the proposed approach from typical optimization problems is that the data are not completely available at the beginning, but they are collected from the automatic feedback across iterations in an incremental manner. As iterations progress, a user need input lesser information. This unsupervised version makes the convergence faster, as compared to standard implementations, since the learning procedure is driven automatically without any user intervention.

The maximum average precision turns to be 0.78 and recall is 0.71. For all given classes of images, the proposed approach gives similar and satisfactory results, which gives a clear picture of the effectiveness of the approach for different classes of images. The following graph gives the average precision - recall computation with the given classes.

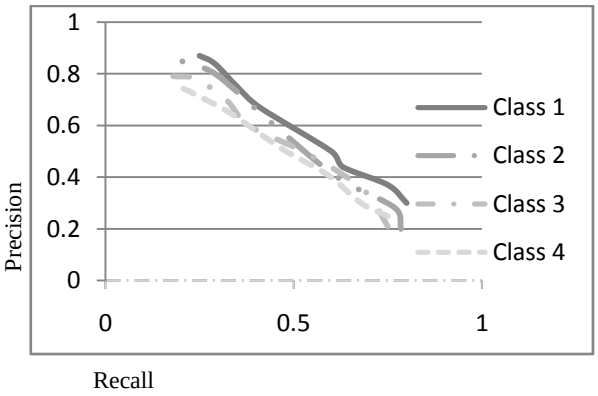

Fig 8. Average Precision Vs Recall

The following table summarizes the comparative study of existing techniques of the proposed scheme – DRIPH. QBIC[4] and VisualSEEk[5] uses purely visual features to discriminate the images. Content-based parallel image retrieval system[20] imposes parallel retrieval of images, but does not use available textual information. DRIPH utilizes maximum available meta-data from images, such as textual information related to images and low-level fine visual features. Most significant challenge of response time is resolved by using HDFS. Also meta-heuristic optimization refines the results for achieving better precision and recall.

Table 1. Comparative Study of existing techniques

	Textual	Visual	Parallel
QBIC	✗	✓	✗
VisualSEEk	✗	✓	✗
CBPIR	✗	✓	✓
DRIPH*	✓	✓	✓

5. CONCLUSION

Distributing the load to data nodes promises significant reduction in response time, which is the major concern in any Content based retrieval system. The initial retrieval is based on a single shape feature which retrieves similarly shaped images faster. The usage of shape feature guarantees efficient retrieval of similar images for the first level of retrieval. The second level uses Automatic relevance feedback to retrieve in-depth low level features of the image using details combination of multiple features. Feature reweighing emphasizes the most discriminating parameters. Optimization and fuzzy results are the drawbacks of Automatic Relevance feedback, and it is addressed by using the stochastic Particle Swarm Optimization algorithm. It takes relevant and irrelevant images as point of attraction and repulsion, and performs effective retrieval. The proposed approach achieves the following goals without any human interaction – retrieving the results in comparatively faster manner, dynamically modifies the feature space by feeding automatic relevance feedback and clustering relevant images using meta-heuristics.

6. REFERENCES

[1] Song Yan, "Research of Image Retrieval Based on color and texture feature", *Computer application and soft computing, 2007*, 9(2),pp.42-50.

[2] Y. Rui, T.S. Huang, and S. Mehrotra, "Relevance feedback a powerful tool in interactive content-based image retrieval", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 644-655, 1998.

[3] H.-J. Zhang, "Relevance feedback in content-based image search", *Conference on New Information Technology NIT'01*. Beijing, China, May 2001.

[4] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Pektovic, P. Yanker, C. Faloutsos, and G Taubin, "The QBIC project: Querying images by content using color, texture and shape," *Proceedings of SPIE Storage and Retrieval for Image and Video Databases*, San Jose, CA, 1994.

[5] J. R. Smith and S.-F. Chang, "VisualSEEk: A Fully Automated Content-Based Image Query System," Proceedings *of the Fourth ACM International Conference on Multimedia '96*, Boston, MA, 1996.

[6] J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics- Sensitive Integrated Matching for Picture LIbraries," *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23*, pp. 947-963, 2001.

[7] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Transaction on Pattern Analysis and Machine Intelligence*.

[8] Gupta. A, "The Virage image search engine: an open framework for image management", *Proceedings in Storage and Retrieval for Image and Video Databases*, Proc SPIE 2670, pp 76-87, 1996

[9] Sutcliffe. A, "Empirical studies in multimedia information retrieval" in *Journal of Intelligent Multimedia Information Retrieval* (Maybury, M T, ed). AAAI Press, Menlo Park, CA, 1997.

[10] M. Koskela, J. Laaksonen, and E. Oja, "Use of image subsets in image retrieval with self-organizing maps," in Proc. *Int. Conf. Image and Video Retrieval (CIVR)*, 2004, pp. 508–516.

[11] G. Bordogna and G. Pasi, "A user-adaptive neural network supporting a rule-based relevance feedback," *Fuzzy Sets System*, vol. 82, no. 9, pp. 201–211, 1996.

[12] J. Kennedy and R. C. Eberhart, "Particle swarm optimization," in *Proc. IEEE Conf. Neural Networks IV*, Piscataway, NJ, 1995.

[13] Kaushik, R.T.; Bhandarkar, M.; Nahrstedt, K., "Evaluation and analysis of Green HDFS: A self-adaptive, energy-conserving variant of the Hadoop distributed file systems", *IEEE Second International Conference on Communication, Networking and Broadcasting*, 2010.

[14] Gonde, Anil Balaji, R. P. Maheshwari, and R. Balasubramanian. "Content-Based Image Retrieval using colour feature and colour bit planes", *International Journal of Signal and Imaging Systems Engineering 3.2 (2010)*: 105-115.

[15] Kennedy, Lyndon, Malcolm Slaney, and Kilian Weinberger. "Reliable tags using image similarity: mining specificity and expertise from large-scale multimedia databases." *Proceedings of the 1st workshop on Web-scale multimedia corpus*, ACM, 2009.

[16] Yang, Zhuo, Sei-ichiro Kamata, and Alireza Ahrary. "NIR: Content based image retrieval on cloud computing.", *IEEE International Conference on Intelligent Computing and Intelligent Systems*, ICIS 2009.

[17] K. Chandramouli, "Particle swarm optimization and self-organizing maps based image classifier," in *Proc. 2nd Int. Workshop Semantic Media Adaptation and Personalization*, 2007, pp. 225–228.

[18] S. C. Hoi, M. R. Lyu, and R. Jin, "A unified log-based relevance feedback scheme for image retrieval," *IEEE Transaction on Knowledge and Data Engineering*, vol. 18, no. 4, pp. 509–524, Apr. 2006.

[19] Jia Li, James Z. Wang, ``Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1075-1088, 2003.

[20] Zhou Bing; Yang Xin-xin, "A content-based parallel image retrieval system," *International Conference on Computer Design and Applications (ICCDA), 2010* , vol.1, no., pp.V1-332,V1-336, 25-27 June 2010