Annotated Image search: Annotated Image Search using Text and Image Features

Prof. Suvarna Nandyal* and Sandhya Koti **

* Assoc. Prof. Dept. of CSE, PDA College of Engineering, Gulbarga, Karnataka, India
**Dept. of CSE, PDA College of Engineering, Gulbarga, Karnataka, India

Abstract- As the diversity and size of digital image collections grow exponentially, efficient image retrieval is becoming increasingly important. In general, current automatic image retrieval systems can be characterized into two categories: text-based and image content-based. For text-based image retrieval, the images are searched using the annotated text. In this framework, manual image annotation is extremely laborious and the visual content of images are difficult to be described precisely by a limited set of text terms. To overcome these difficulties, content-based image retrieval systems index images by their visual content, such as color, shape, texture, etc.

It is a remarkable fact that, neither searching the images based on the content of the image nor searching the images using the annotated text may lead to an accurate result but jointly they tend to produce a perfect result; this is probably because the writers of text descriptions of images tend to leave out what is visually obvious (the color of flowers, etc.) and to mention properties that are very difficult to infer using vision (the species of the flower, say) and the content of the image depicts the description that may be left out by the writer.

An efficient image retrieval system is highly desired. An algorithm which can combine both the retrieval systems i.e. Text Based and Content Based search and then filter out the common images can provide the exact solution for the underlying problem of the retrieval system. Our approach strives to implement the content based search by color and texture features of the objects present in the image using DWT, RGB color filter and color moments and text based search using the simple string match algorithm, and later using both results a similarity comparison is carried out to come up with a final result of the retrieval system. Our image extraction algorithm is based on both the content and the text based retrieval system with high recall rate. The results show that we can improve search accuracy by combining text based search with content based search.

Index Terms- Color Moments; Content Based Image Retrieval (CBIR); Feature extraction; RGB Color Filter; Similarity Comparison; Text Based Image Retrieval (TBIR)

I. Introduction

The Annotated Image search system present a scalable image retrieval system based jointly on text based and visual content based. The solution that we propose is a solution for retrieving images using both their text descriptions and visual content, such as features in color and texture. A query in this system consists of keywords, a sample image and relevant parameters. The retrieving algorithm first selects a subset of images from the whole collection according to a comparison between the keywords and the text descriptions. Visual features extracted from the sample image are then compared with the extracted features of the images in the subset to select the closest.

"Text-based" means the images are searched comparing the annotated text of the image in the database with the user input text using a simple sub-string matching algorithm.

"Content-based" means that the search analyzes the contents of the image rather than the annotated text http://en.wikipedia.org/wiki/Metadata_(computing) such as keywords, tags, or descriptions associated with the image. The term "content" in this context might refer to colors, shapes, textures, or any other information that can be derived from the image itself.

Using both text and image content features, a hybrid image retrieval system is developed in this paper. We first use a text-based image to retrieve images based on the text information on the annotated image to provide an initial image set. An image content based ordering is then performed on the initial image set. Such a design makes it truly practical to use both text and image content for image retrieval. Experimental results confirm the efficiency of the system.

II. Literature Survey

Over the past few years, various techniques have been integrated into CBIR systems to improve the rate of relevant images in the result set. Such techniques include unifying keywords and visual features for indexing and retrieving, using mechanisms of relevance feedback, applying ontology based structures, querying by concept, etc. In the system developed by Zhou and Huang [1], each image was represented by vectors of visual features and text annotations. Keywords were semantically grouped based on user's feedback made during the retrieval process. The system supported joint queries of keywords and example images. Through relevance feedback, retrieving results were further refined.

Zhang and Song [2] implemented a hybrid image retrieval system that was based on keywords and visual contents. Text descriptions of images were stored in a database, on which full-text index catalogues were created. Vectors of visual contents were extracted and saved into a Lucene index. The system was queried jointly by keywords and an example image.

An image retrieval methodology was proposed in [3], where images were divided into regions by a fully unsupervised segmentation algorithm. These regions were indexed by low-level descriptors of colour, position, size and shape, which were associated with appropriate qualitative intermediate-level descriptors that carried high-level concepts. A user could query the system by keywords which carried the high-level concepts. Comparisons were then made with the intermediate-level descriptors and the associated image regions. A relevance feedback mechanism based on support vector machines was employed to rank the obtained image regions that were potentially relevant to produce the results.

A hybrid model of image retrieval was proposed and implemented in [4], where ontology and probabilistic ranking were applied. When the system was queried by a keyword, images annotated by the keyword were selected together with those annotated by keywords conceptually related. The degree of relevance was evaluated by an ontology reasoned whose output were passed to a Bayesian Network to get the rankings.

For large-scale applications of CBIR, linear search over high-dimensional feature vectors must be avoided. Cortina [5], a large-scale image retrieval system for the World Wide Web, was reported to be able to handle over 3 million images. The system had a crawler which collected images and their text descriptions. The text descriptions were stored in a database, where inverted index over the keywords were created. Four MPEG-7 visual features were extracted from the images and stored in another database. To reduce the searching time, the whole dataset was organised in clusters by each descriptor type. When querying the system, a user had to submit a keyword to search through the inverted index to get a set of matching images. The user then had to select one or several images that were visually close to what he/she was looking for. Query vectors from these chosen images were constructed to perform a nearest neighbour search in the spaces of feature descriptors. To avoid a linear search, the visual feature vectors were clustered by the k-means algorithm [6].

The text-based CBIR approaches proposed in [7] were meant to provide quality results within searching times that are acceptable to users who are used to the performance of text search engines. Like Cortina, several MPEG-7 visual descriptors were extracted from the images crawled from the SPIRIT collection [8]. The descriptors were saved as XML documents. An inverted index was created over the terms of the feature vectors. Queries were in the form of example images.

A system architecture for large-scale medical image understanding and retrieval was described in [9], where a hierarchical framework of ontologies was used to form a fusion of low-level visual features and high-level domain knowledge. The implementation was based on the Lucene Image Retrieval Engine (LIRE) and the system supported query by text, by concept and by sample image.

A system for near-duplicate detection and sub-image retrieval was described in [10]. Instead of using global visual features such as colour histograms, the system used a local descriptor, PCA-SIFT [11], to represent distinctive interest points in images. To index the extracted local descriptors they employed locality sensitive hashing [12]. With further optimisation on layout and access to the index data on disk, they could efficiently query indexes containing millions of interest points.

From the above literature survey we come to the conclusion that most of the researches have worked on either on text based or content based retrieval systems. But very few of the researches have addressed the combined model consisting of both the retrieval system. In our work we will address the combined hybrid model which tries to produce highest accuracy result of any retrieval system.

III. The Anno- Search System

With the rapid growth of the numbers of digital images, the need for effectively and efficiently retrieving the images has become demanding. Text based retrieval has been widely used where images are indexed by text terms and retrieved by matching terms in a query with those index. However, text annotations often carry little information about image's visual features. When users wish to retrieve images of similar visual content, a pure text based approach becomes inadequate. Content based image retrieval (CBIR), instead of using text annotations as the basis for indexing and searching, uses visual features extracted from images, such as color, texture, shape and spatial relations of pixels. Unlike text annotations which are subject to human perception, these features make objective representations of images.

The Annotated Image search Systems strives to combine both the searches i.e. text based and content based search to match the result accuracy.

In this paper, we propose a scalable image retrieval system based jointly on text annotations and visual content. Let us discuss the system shown in fig.1 one by one.

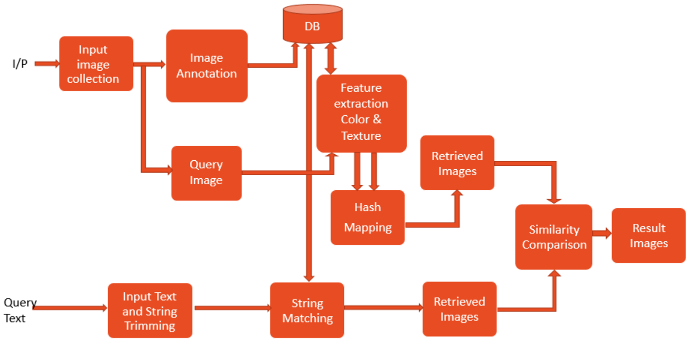


Fig.1 Block Diagram of Annotated Image search System

3.1 Image Collection

Image collection is a process which requires a large collection of images pertaining to different classes. We have collected images that lie in 8 different classes namely Plant, Idol, Lake, Duck, and white flower, Yellow Hibiscus, Pink Hibiscus and Rabbit. Each class has minimum of 10 images with each having different tags. Sample images of each class is as shown in the table below.

Sl.No	Image	Image	Description

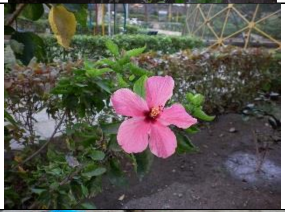
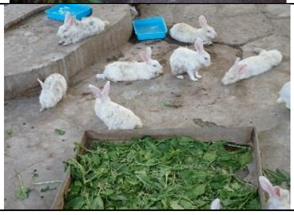
		Class	
1		Plant	Flower Plant in the Garden
2		Idol	Idols with flowers in Front
3		Lake	Building behind the Bush of the Lake
4		Duck	Beautiful White color Duck
5		White Flower	Plant with White Flower and long Green Leaves
6		Yellow Hibiscus	Beautiful Yellow Hibiscus Plant with a Flower
7		Pink Hibiscus	Pink Hibiscus Flower with Green Leaves
8		Rabbit	Rabbit eating Green Leaves

Fig.2 Image Collection

3.2 Image Annotation

Image annotation is a process where each image subjected to the annotation algorithm to extract the image content and form a relevant annotated text for the image. In our work the database is limited to 100 images, so we use manual annotation method to tag the relevant images. Once the database size increases, auto annotation algorithm can be implemented.

3.3 Search by Content

Search by content is a process where the input image is searched against the database images for similar features. The algorithms used in these systems are commonly divided into three tasks:

- Feature Extraction (DWT, Image Segmentation, RGB Color Filter and Color Moments)
- Hash Map Indexing
- Similarity Measures

3.3.1 Feature Extraction

The important task in Feature Extraction is to extract texture features which are most completely describing the information of texture in the image.

Various kinds of texture analysis methods are used to examine textures from different perspectives. Individual method can't be used for all textures multidimensionality of perceived texture. In our approach, a set of parameters which are driven from the variation of pixel elements of texture are used to define an image model. The method used in this work is discrete wavelet transform.

3.3.2 Discrete Wavelet Transform

DWT can be performed by iteratively filtering a signal or image through the low-pass and high-pass filters, and subsequently down sampling the filtered data by two. This process will decompose the input image into a series of sub band images.

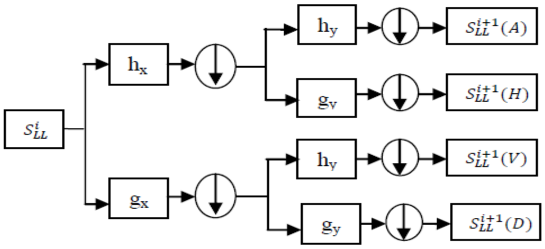


Fig 3. Discrete Wavelet Transform

Figure 3 illustrates an example of DWT, where h and g represent the low-pass and high-pass filter respectively, while the symbol with a down arrow inside a circle represents the down sampling operation. From figure 4, an image S at resolution level i was decomposed into four sub band images after going through one stage of decomposition process. The four sub band images consist of one approximation image and three detail images. The approximation image is actually the low-frequency.

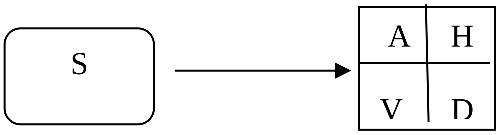


Fig 4. Sub band images for one level of image decomposition using DWT

Detail image contains the information of specific scale and orientation. This means that the spatial information is also retained within the sub band images. Therefore, the detail images are suitable to be used for deriving a set of texture features in the input image. On the other hand, the approximation image can be used for higher levels of decomposition for the input image. Down sampling operation has helps to reduce the useless and redundant samples in the decomposition process. The fig.5 show the DWT of an actual image of an Plant.

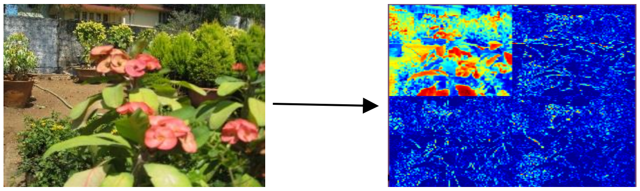


Fig 5. Sub band images for one level of image decomposition using DWT (Actual "Plant" Image as Example)

3.3.3 RGB Color Filter

One of the important features that make possible the recognition of images by humans is color. Color is a property that depends on the reflection of light to the eye and the processing of that information by the brain. We use color to tell the difference between objects, places, and the time of day. Usually colors are defined in three dimensional color spaces. These could either be RGB (Red, Green, and Blue), HSV (Hue, Saturation, and Value) or HSB (Hue, Saturation, and Brightness).

Most image formats such as JPEG, BMP, GIF, use the RGB color space to store information. The RGB color space is defined as a unit cube with red, green, and blue axes. Thus, a vector with three co-ordinates represents the color in this space. When all three coordinates are set to zero the color perceived is black. When all three coordinates are set to 1 the color perceived is white. The other color spaces operate in a similar fashion but with a different perception. In MatLab for example one can get a color histogram of an image in the RGB or HSV color space. Bars in a color histogram are referred to as bins and they represent the x-axis. The number of bins depends on the number of colors there are in an image. Y-axis denotes the number of pixels in each bin. In other words it gives the count of pixels in an image representing a particular color.

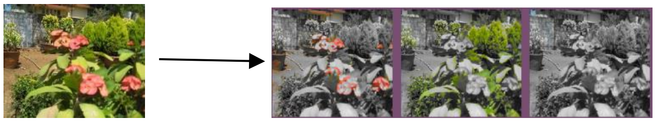


Fig 7. RGB Color Filter application on an Image containing "Plant"

Color Feature Extraction

Color image segmentation is a process of extracting from the image domain one or more connected regions satisfying uniformity (homogeneity) criterion which is based on feature(s) derived from spectral components. These components are defined in a chosen color space model. The segmentation process could be augmented by some additional knowledge about the objects in the scene such as geometric and optical properties.

In our work, the input color image will be coarsely represented using 25 bins. Coarse representation uses the spatial information from a Histogram based windowing process. K-Means is used to cluster the coarse image data. In fig.6 the actual Image of a "Plant" is segmented into 25 color bins.

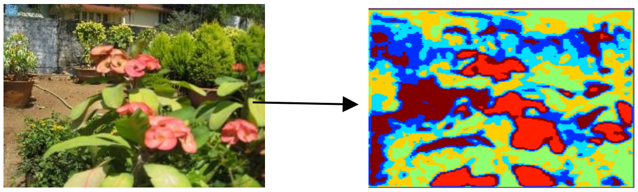


Fig 6. Color Feature Extraction

3.3.4 Color Moments

Color moments are measures that can be used to differentiate images based on their features of color. These moments provide a measurement for color similarity between images. These values of similarity can then be compared to the values of images indexed in a database for tasks like image retrieval.

The basis of color moments lays in the assumption that the distribution of color in an image can be interpreted as a probability distribution. Probability distributions are characterized by a number of unique moments (e.g. Normal distributions are differentiated by their mean and variance). It therefore follows that if the color in an image follows a certain probability distribution, the moments of that distribution can then be used as features to identify that image based on color.

There are three central moments of a image's color distribution. They are Mean, Standard deviation and Skewness. A color can be defined by 3 or more values. Moments are calculated for each of these channels in an image. An image therefore is characterized by 9 moments 3 moments for each 3 color channels.

MOMENT 1 – Mean: Mean can be understood as the average color value in the image.

$$E_i = \sum_{N}^{j=1} \frac{1}{N} P_{ij}$$

MOMENT 2 Standard Deviation: The standard deviation is the square root of the variance of the distribution.

$$\sigma_i = \sqrt{\left(\frac{1}{N} \sum_{N}^{j=1} (P_{ij} - E_i\,)^3\right)}$$

MOMENT 3 – Skewness: Skewness can be understood as a measure of the degree of asymmetry in the distribution.

$$S_i = \sqrt[3]{\left(\frac{1}{N} \sum_{N}^{j=1} (P_{ij} - E_i\,)^3\right)}$$

A function of the similarity between two image distributions is defined as the sum of the weighted differences between the moments of the two distributions. Formally this is:

$$d_{mom}(H,I) = \sum_{i=1}^{r} W_{i1}\left|E_i^1 - E_i^2\right| + W_{i2}\left|\sigma_i^1 - \sigma_i^2\right| + W_{i3}\left|S_i^1 - S_i^2\right|$$

3.4 Hash Map Indexing

Because visual features are generally of high dimensional, similarity-oriented search based on visual features is always a bottleneck for large-scale image database retrieval on search efficiency. To overcome this problem, we adopt a hash encoding algorithm to speed up this procedure.

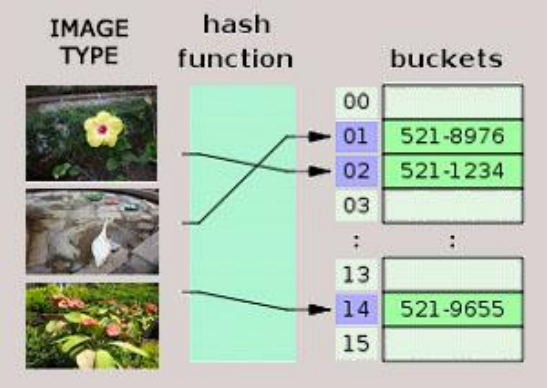


Fig 8. An image feature repository as a hash table

This idea is proposed to encode image visual features to so-call hash codes. Images are divided into even blocks and average luminance of each block is extracted as visual features. These features are transformed by a PCA mapping matrix learned beforehand, and then quantized into hash codes.

3.5 Search By Text

Search by text is a process where the images are searched using the text provided by the user. This is a simple process of comparing the text provided by the user with the annotated/tagged text of the database images. Once the input text matches with the annotated/tagged text of the image, the image will be retrieved, irrespective of the image feature.

3.6 Similarity Comparison

It is difficult to find a unique representation to compare images accurately. In our work we adopt a unique way of comparing the sub-set images of both the searches. In this process the images retrieved by “Image Search” and “Text Search” using similar keyword/image are compared one by one and zeroed upon final result. Similarity comparison involves a simple text based search of the tag of the image retrieved from Image based search. The following flow chart defines the method of comparison:

Step 1: Get the count of number of images from both the searches.
Step 2: Select the images from the search which is having more count (Say Content Based Search)
Step 3: Retrieve the Annotated Text from the Image from Content Based Search.
Step 4: Compare if the same annotated text image is present in Text Based Search.
Step 5: If Present, select for final result else reject the Image.
Step 6: Select the Next Image and go to Step 2.

After completing the similarity comparison the set the images which are retrieved are the final images of this approach.

IV. Result and Discussion

A series of experiments were conducted to evaluate the effectiveness and efficiency of the Anno-Annotation Search system. We created a database of 100 images. They are of high quality and have rich descriptions, such as title, category and comments. Though these descriptions are noisy, they cover to a certain degree the concepts of the corresponding images. These images make-up of the database, from which the relevant images are retrieved to annotate the query image. Two query datasets are used to evaluate the system performance. The first one is 10 images of 5 categories randomly selected. To evaluate the effectiveness of our approach, we deliberately used a few vague query keywords, e.g. we use “Plant” as the query keyword to annotate an image of “Rabbit”. We manually assessed the retrieval results on this dataset.

The second dataset is a content-based image retrieval database. Images in this dataset have about 5 on average manually labeled ground truth annotations. And for many images, not all objects are annotated.

4.1. Experiments on System

It is a remarkable fact that, while text and images are separately ambiguous, jointly they tend not to be; this is probably because the writers of text descriptions of images tend to leave out what is visually obvious (the color of flowers, etc.) and to mention properties that are very difficult to infer using vision (the species of the flower, say). And the content based features like color and the texture of the objects present in the image illustrate the characteristic property like base object classes: plant, flower, duck, rabbit, etc. However, certain characteristics of the image are hard to capture using strictly the visual properties, such as events, people, location etc.

It can be inferred that by combining both Text and Content Based Search the irrelevant images can be pruned out from the search result.

The prefect search result of Annotated Image search System is combination of text based search and content based search by selecting the common images from both searches.

Evaluation Result:

Content Based Search

When a query image (say Flower) is given, we retrieved 7 images that are shown in the fig.9 below. Similarly, when then search is conducted for Plant, Duck, Lake and Idol, we retrieved 7, 7, 7 and 7 images respectively. The histogram graph (Fig.10) is derived based on the search result

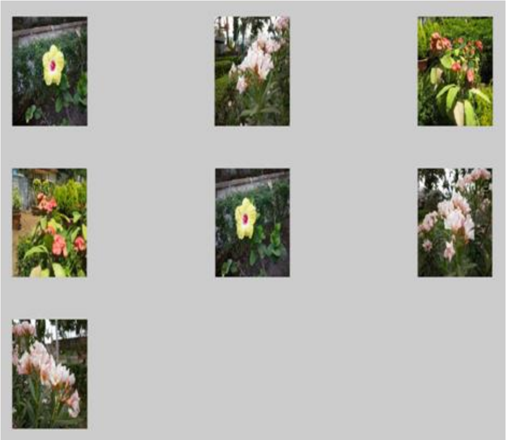


Fig 9 .Search result for Flower

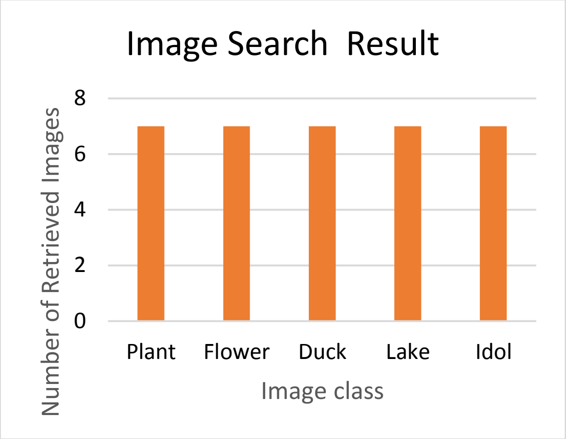


Fig 10. Search Result Graph for Various Image Input

Text Based Search

When a query text (say Flower) is given, we retrieved 9 images that are shown in the fig.11 below. Similarly, when then search is conducted for Plant, Duck, Lake and Idol, we retrieved 9, 9, 6 and 8 images respectively. The histogram graph (Fig.12) is derived based on the search result.

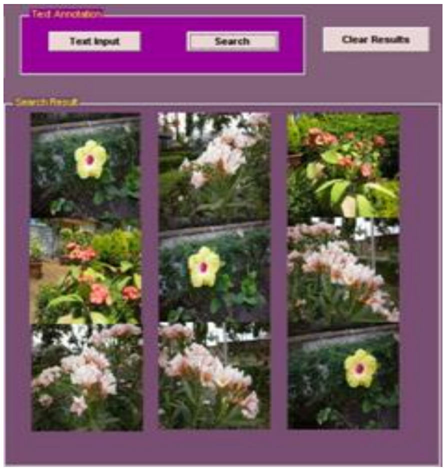


Fig.11 Search result for Flower

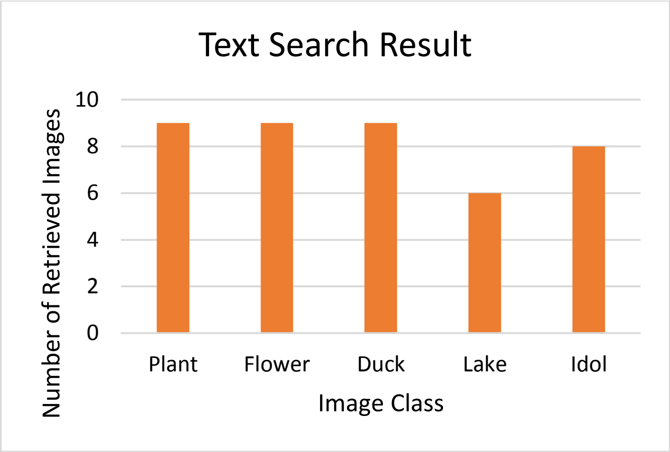


Fig 12. Search Result Graph for Various search keywords

Annotated Image search System Result

By defusing both the system results the final result is inferred that is shown in Fig.13, this result comprise of result for text and content search for similar keywords or images i.e. Plant, Flower, Duck, lake and Idol. The histogram graph (Fig.14) is derived based on the final search result comprising of "Text Search", "Image Content Search" and combination of both.

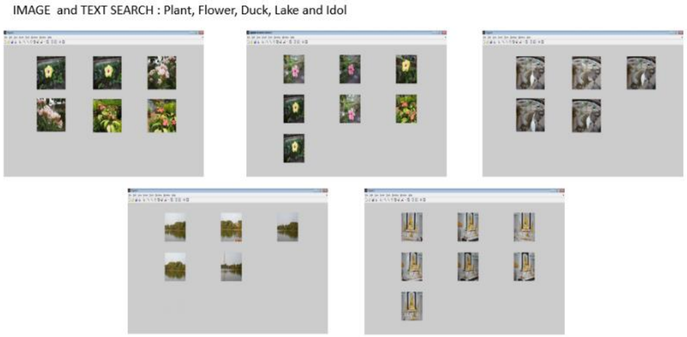


Fig 13. Search result for Annotated Image search System

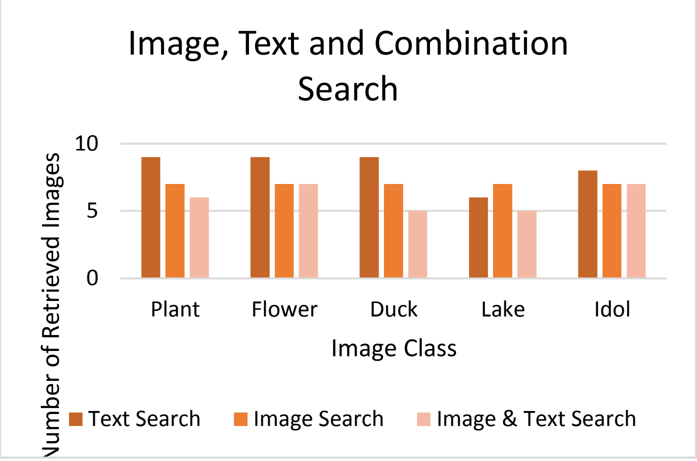


Fig 14. Search Result Graph for Various Searches

The proposed combination model will always result into the best result of any retrieval system

V. Conclusion

We have presented a best image retrieval system. The solution unifies well-established text and content based techniques with the aim of overcoming the semantic gap in image retrieval systems that are solely content based. Content based approaches evaluate similarities in visual domain, which provides more objective representations for images than text annotations. The image selection by the semantic comparison also helps to reduce the size of the data set for the hashing, which further shortens the time spent on the visual comparison. By such a two-staged retrieving strategy, time spent on the content based comparison can be confined within a user-tolerant range. So we believe that the solution has the potential to be scaled up to suit large image collections.

Acknowledgement

I would like to thank our Head of the department **Prof. Rekha Patil** for encouragement and **Prof. Suvarna Nandyal** for the time and guidance provided by her throughout the research, without which this piece of work would not have turned into master piece.

References

[1] X. S. Zhou and T. S. Huang, "Unifying Keywords and Visual Contents in Image Retrieval," IEEE MultiMedia, Vol. 9, No. 2, pp. 23–33, Apr. 2002.

[2] N. Zhang and Y. Song, "An Image Indexing and Searching System Based Both on Keyword and Content," in Proceedings of the 4th International Conference on Intelligent Computing (ICIC2008), ser. LNCS 5226. Springer-Verlag Berlin Heidelberg, Sep. 2008, pp. 1059– 1066.

[3] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, "Region-based Image Retrieval using an Object Ontology and Relevance Feedback," EURASIP Journal on Applied Signal Processing, Vol. 2004, No. 6, pp. 886–901, Jun. 2004.

[4] L. Fan and B. Li, "A Hybrid Model of Image Retrieval Based on Ontology Technology and Probabilistic Ranking," in Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. IEEE Computer Society, 2006, pp. 477–480.

[5] T. Quack, U. Mnich, L. Thiele, and B. S. Manjunath, "Cortina: A System for Large-scale, Content-based Web Image Retrieval," in Proceedings of the 12th Annual ACM International Conference on Multimedia. ACM, 2004, pp. 508–511.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification (2nd Edition). John Wiley and Sons, Inc., 2000.

[7] P. Wilkins, P. Ferguson, A. F. Smeaton, and C. Gurrin, "Text Based Approaches for Content-Based Image Retrieval on Large Image Collections," in Proceedings Of the 2nd European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, Nov. 2005, pp. 281–288.

[8] H. Joho and M. Sanderson, "The SPIRIT Collection: An Overview of A Large Web Collection," ACM SIGIR Forum, Vol. 38, No. 2, pp. 57–61, Dec. 2004.

[9] M. M¨oller and M. Sintek, "A Generic Framework for Semantic Medical Image Retrieval," in Proceedings of the Knowledge Acquisition from Multimedia Content (KAMC) Workshop, the 2nd International Conference on Semantics and Digital Media Technologies (SAMT), Dec. 2007.

[10] Y. Ke, R. Sukthankar, and L. Huston, "An Efficient Parts-based Nearduplicate and Sub-image Retrieval System," in Proceedings of the 12th Annual ACM International Conference on Multimedia. ACM, 2004, pp. 869–876.

[11] Y. Ke and R. Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors," in Proceedings of *the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2004, pp. 506–513.

[12] P. Indyk and R. Motwani, "Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality," in Symposium on Theory of Computing, 1998.

Authors

First Author – Prof. Suvarna Nandyal, Assoc. Prof. Dept. of CSE, PDA College of Engineering, Gulbarga, Karnataka, India.
Suvarna_nandyal@yahoo.com
Second Author – Sandhya Koti, Dept. of CSE, PDA College of Engineering, Gulbarga, Karnataka, India.
Sandhya.chimalgi@gmail.com