INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

AN EFFICIENT CONTENT BASED IMAGE RETRIEVAL USING HIERARCHICAL CLUSTERING APPROACH

[1]Mrs. MALLIKA, [2]Mr. S. ANANTHA KRISHNAN., M.Sc., M.Phil

[1]Research Scholar, Department of Computer Science, Park's College, Tirupur- 5
[2]Assistant Professor, Department of Computer Science, Park's College, Tirupur- 5

Abstract: - The field of image retrieval has been an active research area for several decades and has been paid more and more attention in recent years as a result of the dramatic and fast increase in the volume of digital images. Content-based image retrieval (CBIR) is a new but widely adopted method for finding images from vast and an annotated image databases. CBIR systems index the media documents using salient features extracted from the actual media rather than by textual annotations. Query by content is nowadays a very active research field, with many systems being developed by industrial and academic teams. Results performed by these teams are really promising. Data clustering is an unsupervised method for extraction hidden pattern from huge data sets. With large data sets, there is possibility of high dimensionality. Having both accuracy and efficiency for high dimensional data sets with enormous number of samples is a challenging arena. The proposed CBIR technique uses more than one clustering techniques to improve the performance of CBIR. This optimized method makes use of Hierarchical clustering technique to improve the execution time and performance of image retrieval systems in high dimensional sets. In this similarity measure is totally based on colors. In this paper more focus area is the way of combination of clustering technique in order to get faster output of images.

Keywords: - Query, Image, Retrieval, Content, Clustering, Hierarchical, Performance

1. INTRODUCTION

The use of images in human communication is hardly new – our cave-dwelling ancestors painted pictures on the walls of their caves, and the use of maps and building plans to convey information almost certainly dates back to pre-Roman times. But the twentieth century has witnessed unparalleled growth in the number, availability and importance of images in all walks of life. Images now play a crucial role in fields as diverse as medicine, journalism, advertising, design, education and entertainment.

Problems with traditional methods of image indexing have led to the rise of interest in techniques for retrieving images on the basis of automatically-derived features such as color, texture and shape – a technology now generally referred to as Content-Based Image Retrieval (CBIR). After a decade of intensive research, CBIR technology is now beginning to move out of the laboratory and into the marketplace, in the form of commercial

products like QBIC and Virage. However, the technology still lacks maturity, and is not yet being used on a significant scale. In the absence of hard evidence on the effectiveness of CBIR techniques in practice, opinion is still sharply divided about their usefulness in handling real-life queries in large and diverse image collections. Nor is it yet obvious how and where CBIR techniques can most profitably be used.

CBIR or Content Based Image Retrieval is the retrieval of images based on visual features such as colour, texture and shape. Reasons for its development are that in many large image databases, traditional methods of image indexing have proven to be insufficient, laborious, and extremely time consuming. These old methods of image indexing, ranging from storing an image in the database and associating it with a keyword or number, to associating it with a categorized description, have become obsolete. This is not CBIR. In CBIR, each image that is stored in the database has its features extracted and compared to the features of the query image. It involves two steps:

- **Feature Extraction:** The first step in the process is extracting image features to a distinguishable extent.
- **Matching:** The second step involves matching these features to yield a result that is visually similar.

The Content Based Image Retrieval (CBIR) technique uses image content to search and retrieve digital images. Content-based image retrieval systems were introduced to address the problems associated with text-based image retrieval. Content based image retrieval is a set of techniques for retrieving semantically-relevant images from an image database based on automatically-derived image features. The main goal of CBIR is efficiency during image indexing and retrieval, thereby reducing the need for human intervention in the indexing process.

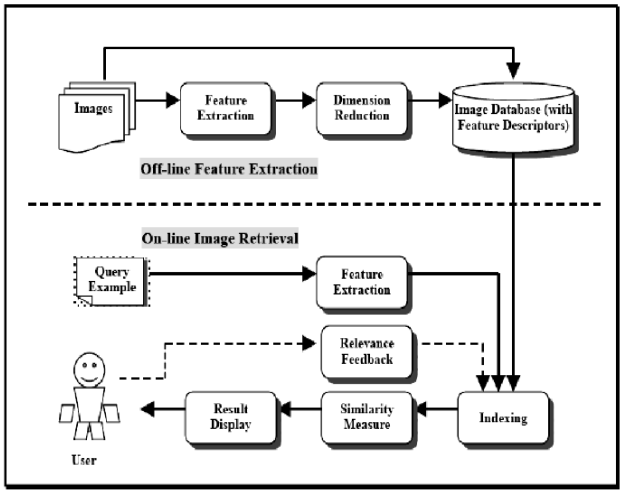

Fig 1.1: - Conceptual View of CBIR

. In typical content-based image retrieval systems (Figure 1-1), the visual contents of the images in the database are extracted and described by multi-dimensional feature vectors. The feature vectors of the images in the database form a feature database. To retrieve images, users provide the retrieval system with example images or sketched figures. The system then changes these examples into its internal representation of feature vectors. The similarities /distances between the feature vectors of the query example or sketch and those of the images in the database are then calculated and retrieval is performed with the aid of an indexing scheme. The indexing scheme provides an efficient way to search for the image database. Recent retrieval systems have incorporated users' relevance feedback to modify the retrieval process in order to generate perceptually and semantically more meaningful retrieval results.

2. Problem Definition

Image databases and collections can be enormous in size, containing hundreds, thousands or even millions of images. The conventional method of image retrieval is searching for a keyword that would match the descriptive keyword assigned to the image by a human categorizer. Currently under development, even though several systems exist, is the retrieval of images based on their content, called Content Based Image Retrieval, CBIR. While computationally expensive, the results are far more accurate than conventional image indexing. Hence, there exist tradeoffs between accuracy and computational cost. This tradeoffs decreases as more efficient algorithms are utilized and increased computational power becomes inexpensive.

The idea behind content-based retrieval is to retrieve, from a database, media items (such as images, video and audio) that are relevant to a given query. Relevancy is judged based on the content of media items. Several steps are needed for this. First, the features from the media items are extracted and their values and indices are saved in the database. Then the index structure is used to ideally filter out all irrelevant items by checking attributes with the user's query. Finally, attributes of the relevant items are compared according to some similarity measure to the attributes of the query and retrieved items are ranked in order of similarity.

The problem involves entering an image as a query into a software application that is designed to employ CBIR techniques in extracting visual properties, and matching them. This is done to retrieve images in the database that are visually similar to the query image.

3. Existing Methodology

Some of the existing CBIR systems extract features from the whole image not from certain regions in it; these features are referred to as Global features. Histogram search algorithms characterize an image by its color distribution or histogram. Many distances have been used to define the similarity of two color histogram representations. Euclidean distance and its variations are the most commonly used. The drawback of a global histogram representation is that information about object location, shape and texture is discarded. Color histogram search is sensitive to intensity variations, color distortions, and cropping. The color layout approach attempts to overcome the drawback of histogram search. In simple color layout indexing, images are partitioned into blocks and the average color of each block is stored. Thus, the color layout is essentially a low resolution representation of the original image.

A relatively recent system, WBIIS, uses significant Daubechies' wavelet coefficients instead of averaging. By adjusting block sizes or the levels of wavelet transforms, the coarseness of a color layout representation can be tuned. Hence, we can view a color layout representation as an opposite extreme of a histogram. At proper resolutions, the color layout representation naturally retains shape, location, and texture information. However, as with pixel representation, although information such as shape is preserved in the color layout representation, the retrieval system cannot perceive it directly. Color layout search is sensitive to shifting, cropping, scaling, and rotation because images are described by a set of local properties

4. Proposed Scheme

The solution initially proposed was to extract the primitive features of a query image and compare them to those of database images. The image features under consideration were color, texture and shape. Thus, using matching and comparison algorithms, the color, texture and shape features of one image are compared and matched to the corresponding features of another image. This comparison is performed using color, texture and shape distance metrics. In the end, these metrics are performed one after another, so as to retrieve database images that are similar to the query. The similarity between features was to be calculated using algorithms used by well known CBIR systems. For each specific feature there was a specific algorithm for extraction and another for matching.

Clustering techniques can be classified into supervised (including semi-supervised) and unsupervised schemes. The former consists of hierarchical approaches that demand human interaction to generate splitting criteria for clustering. In unsupervised classification, called clustering or exploratory data analysis, no labeled data are available. The goal of clustering is to separate a finite unlabeled data set into a finite and discrete set of "natural," hidden data structures, rather than provide an accurate characterization of unobserved samples generated from the same probability distribution.

Images are generated at increasing rate by different sources. Image Retrieval is the processing of searching and retrieving images from huge datasets [1]. Most of the existing image management systems are based on the verbal descriptions to enable their mining. A key-word description of the image content, created by some user on input is used to get the appropriate image. But it is not always a simple task to retrieve the desired images from the large datasets. Retrieving images from large data sets in a less time is a challenge. Content-Based Image Retrieval is defined as a process that searches and retrieves images from large database on the basis of automatically derived features such as color, shape. Retrieval pattern-based learning with fast and refined clustering of the images is the most effective that aim to establish the relationship between the images with similar attributes.

Efficient and effective retrieval of content based images (microscopic, landscape etc) is much essential in Content Based Image Retrieval.

4.1. Steps for Proposed Scheme

Content Based image retrieval system is a widely used method for finding and retrieving images from large databases. The proposed CBIR technique uses more than one clustering techniques to improve the performance of CBIR. This optimized method makes use of Kmeans and Hierarchical clustering technique to improve the execution time and performance of image retrieval systems in high dimensional sets. In this similarity measure is totally based on colors.

Module 1: Input Query Image

In the first module, an image is selected and given as input to the application. The cause behind inputting an image is to find the images that are most similar to it from the database. It is in these large collections that we search for images similar to our query.

Module 2: K-Means Processing

The second module is related to K-Means processing. The resultant set of image groups are processed in K-Means methodology. In this process, primarily the image groups that are most similar are clustered together. Then the images in the groups are sorted in a descending order based on their similarity values.

Module 3: Hierarchical Grouping of Images

This is the process in which the query image is compared with all the images present in the database on the basis of the similarity of their color feature space. The average RGB value of the images is calculated. The RGB value of the query image is compared with each RGB values of the database images. Then Images are hierarchically grouped based on their similarity levels. This results in random groups of images.

Module 4: Resultant Image Display

Finally all the images retrieved through Hybrid Clustering .We have taken an image in a folder as an input query and selected a folder containing several images to be searched.

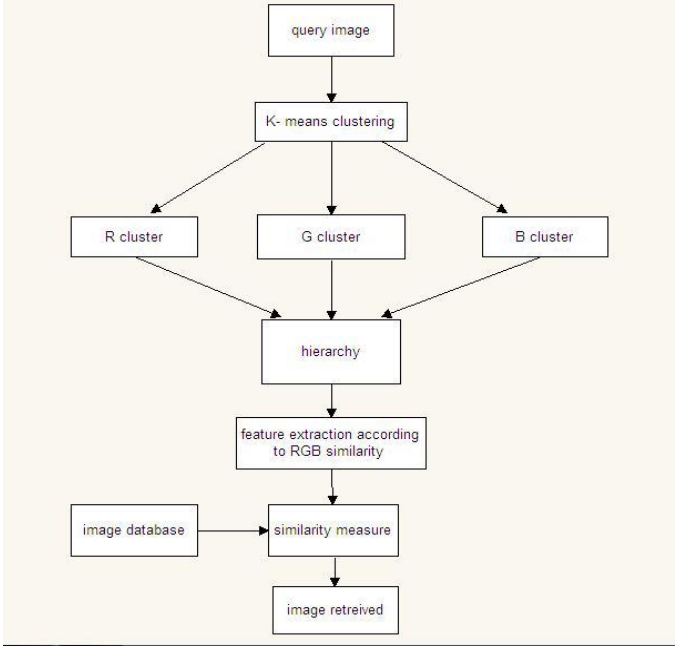

Fig 4.1: - Proposed Scheme Structure

In our first 3 modules includes the segments which is defined on the basis of the color we defined the segment in term of RGB value. In First segment the particular range of the red pixels will be matched with the image in the database and its matching percentage will be shown same way the Green and the blue segments also matched. All these matched images according to the segment are displayed in the folder called red matched, blue matched, green matched. We are performing The K-Means clustering on these on the basics of these clustering

values the images are displayed. Than Hierarchical clustering is performed on the images in that particular folder according to the percentage of the matching image.

In our 4th module includes the searching of the image with the text a pop-up menu is displayed asking the name of the image that we want to search. We just need to write the name of that image and that image will be shown in the new frame. Our 5th module gives us system the more robustness that is the module with the blur image. Whenever we are going to provide any blur image to the system in that case the entire database is scan. To find the exact image from the blur image take place in two steps.

- In first step the entire database is scanned and the most closely image are displayed in the folder.
- Then we are going to display the most closely related image in the frame is the shown out.

Our 6th module includes the matching of the entire image with the images present in the database the matching of the image is based on the Content-Based Image retrieval which means we are taking the RGB value of the complete image that RGB value is matched with the all the image and the result with the source of the image and the percentage of the match of that image with all the other image. Our 7th module includes the mining of desired image from the cluster hierarchy. The main purpose to scan the clustering hierarchies is to reduce the scanning of whole database and it reduces the computational time too as compare to mining from the database.

This Proposed image retrieval system uses the Hybrid algorithm that is the combination of the Kmeans algorithm and the hierarchical algorithm. In our algorithm the k-mean clustering is performed first than we are using the hierarchical clustering. The image is given as the input to the propose system and then some segment of the image is taken and then that segment of the image is matched with the all the images in the database. These segment of the image taken is based on the color value RGB and these RGB value is compared with the images in the database we have taken the three segment based on the color value RGB ad according to the three different segment the image is matched from the database and the Percentage of the matched image is also shown for all the three section.

5. Experimental Results

In improving clustering performance they proposed the use of more than one clustering method. They investigated the use of sequential combination clustering as opposed to simultaneous combination and found that sequential combination is less complex and there are improvements without the overhead cost of simultaneous clustering. In clustering points lying in high dimensional spaces, formulating a desirable measure of “similarity” is more problematic.

Recent research shows that for high dimensional spaces computing the distance by looking at all the dimensions is often useless, as the farthest neighbor of a point is expected to be almost as close as its nearest neighbor. To compute clusters in different lower dimensional subspaces, recent work has focused on projective clustering, defined as follows: given a set P of points in Rd and an integer K, partition into subsets that best classify into lower dimensional subspaces according to some objective function.

Instead of projecting all the points in the same subspace, this allows each cluster to have a different subspace associated with it. Proposed model will be more effective and achieves significant performance improvement over traditional method for most clustering. It will be able to cluster samples of same no of clusters and level and be more efficient and accurate than a single one pass clustering. There is some delimitation for this method. First space should be orthogonal it means there is no correlation among attributes of an object. Second base on application that is used all attributes in an object have the same kind of data types.

In content-based retrieval, precision and recall measures have been frequently used to evaluate the performance of retrieval algorithms. In our case, there is no need to evaluate the subjective quality of the retrieval, but it is only necessary to compare the retrieval with clustering against the exhaustive search.

i) Recall

The following result shows that the proposed approach provides close to the saliency approach and better than other approaches. The saliency retrieval approach was efficient for images having complex spatial structure as in the large database; where query relevant images have large variations and object dissimilarity. The proposed approach provides better recall than the other approaches.

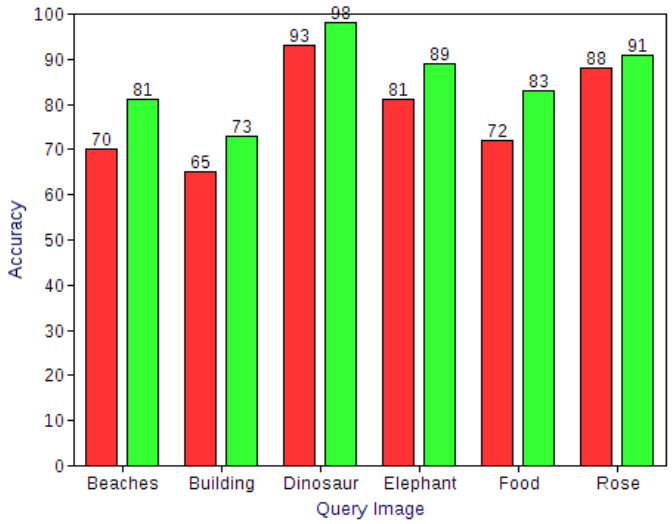

Fig 5.1: - Accuracy of the Image Retrieval

ii) Precision

The proposed retrieval approach works as an extension to the saliency approach. It enhances the precision for images having large smoothed regions and provides better recall than other approaches. In addition, the new approach captures the smooth details of images and provides an approximation between smoothness and complexity for images having large occlusion and complexity in structure.

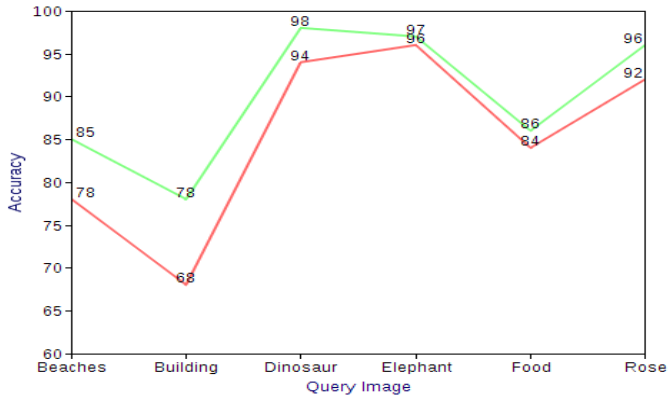

Fig 5.2: - The Precision of the Image Retrieval

iii) Execution Time

The following result shows a performance evaluation of the proposed retrieval approach toward other approaches. The table compares the techniques relative to the number of locally detected feature vectors, the number of globally detected feature vectors, the average segmentation time and the average matching time. On the contrary, the proposed segmentation scheme provides speed up by segmenting images of the Wang database in an average time of 0.318Sec where the entropy rate super pixel segmentation takes 1.036Sec.

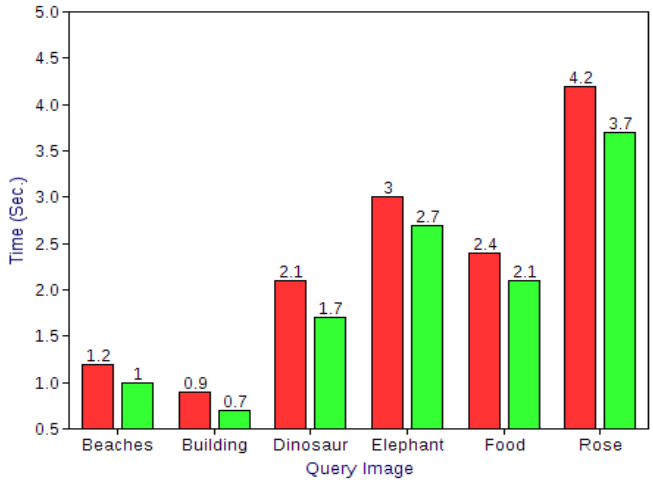

Fig 5.3: - Average execution time for Image Retrieval

iv) Recall & F-Measure

The similarity between images is calculated based on local color properties. Images in the database are clustered into groups having similar color contents. This grouping enables searching only images that are relevant to the query image. For a database containing 6 images, a retrieval accuracy of 90% can be achieved with only 6 similarity comparisons.

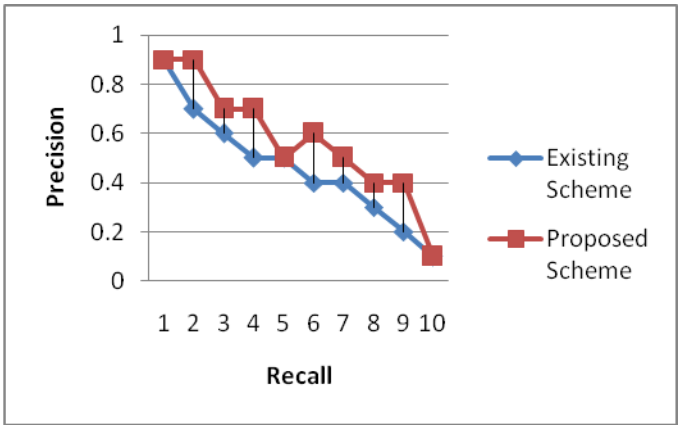

Fig 5.4: - Recall & F-Measure Comparison

CBIR is still a developing science. As image compression, digital image processing, and image feature extraction techniques become more developed, CBIR maintains a steady pace of development in the research field. Furthermore, the development of powerful processing power and faster and cheaper memories contribute heavily to CBIR development. This development promises an immense range of future applications using CBIR.

6. Conclusions

From the very beginning of CBIR research, similarity computation between images used either region based or global based features. Global features extracted from an image are useful in presenting textured images that have no certain specific region of interest with respect to the user. Region based features are more effective to describe images that have distinct regions. Retrieval systems based on region features are computationally expensive because of the need of segmentation process in the beginning of a querying process and the need to consider every image region in similarity computation.

To make the content-based image retrieval truly scalable to large size image collections, efficient multidimensional indexing techniques need to be explored. There are two main challenges in such an exploration for image retrieval.

7. REFERENCES

[1] H. Tamura, and N. Yokoya, "Image Database Systems: A Survey," Pattern Recognition, vol. 17, no 1, pp.29–49, Sep. 1984.

[2] S. Gerard, C. Buckely, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing and Management, vol. 24, no.5, pp. 513-523, Jan. 1988.

[3] Y. Chen, J. Wang, "Image Categorization by Learning and Reasoning with Regions," Journal of Machine Learning Research, vol. 5, pp. 913–939, May 2004.

[4] F. Long, H. Zhang, H. Dagan, and D. Feng, "Fundamentals of content based image retrieval," in D. Feng, W. Siu, H. Zhang (Eds.): "Multimedia Information Retrieval and Management. Technological Fundamentals and Applications," Multimedia Signal Processing Book, Chapter 1, Springer-Verlag, Berlin Heidelberg New York, 2003, pp.1-26.

[5] V. Gudivada and V. Raghavan, "Content-based image retrieval systems," IEEE Computer, vol. 28, no 9, pp18-22, Sep. 1995.

[6] M. Kherfi, D. Ziou, and A. Bernardi, " Image Retrieval From the World Wide Web: Issues, Techniques, and Systems," ACM Computing Surveys, vol. 36, no. 1, pp. 35–67, March 2004.

[7] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, and P. Yanker, "Query by image and video content:The QBIC system," IEEE Computer, vol. 28, no 9, pp.23-32, Sep. 1995.

[8] A. Pentland, R. Picard, and S. Sclaroff , "Photobook: Content based manipulation of image databases," International Journal of Computer Vision, vol.18, no 3, pp.233–254, June 1997.

[9] J. Smith and S. Chang, "Visualseek: A Fully Automated Content-Based Image Query System," Proceedings of the 4th ACM international conference on Multimedia table of contents, Boston, Massachusetts, United States, Nov. 1996, pp. 87-98.

[10] A. Gupta, and R. Jain, "Visual information retrieval," Comm. Assoc. Comp. Mach., vol. 40, no. 5, pp. 70–79, May. 1997.

[11] J. Li, J. Wang, and G. Wiederhold, "Integrated Region Matching for Image Retrieval," In Proceedings of the 2000 ACM Multimedia Conference, Los Angeles, October 2000, pp. 147-156.

[12] Y. Deng, B. Manjunath, "Unsupervised Segmentation of Color -Texture Regions in Images and Video," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 8, pp. 800-810, Aug. 2001.

[13] A. Smeulders, "Content-Based Image Retrieval at the End of the Early Years," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1349- 1380, May. 2000.

[14] J. Caicedo, F. Gonzalez, E. Romero, E. triana, "Design of a Medical Image Database with Content-Based Retrieval Capabilities," In Proceedings of the 2nd Pacific Rim conference on Advances in image and video technology, Santiago, Chile, December 17-19, 2007.

[15] B. Manjunath and W. Ma, "Texture features for Browsing and retrieval of image data," IEEE transactions on pattern analysis and machine intelligence, vol. 18. No. 8, pp. 837-842, August 1996

[16] R. Zhang, and Z. Zhang, "A Clustering Based Approach to Efficient Image Retrieval," Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'02), Washington, DC, Nov. 2002, pp. 339-346.

[17] J. Wang, J. Li, G. Wiederhold, "Simplicity: semantics-sensitive integrated matching for picture libraries," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 23, no. 9, pp. 947–963, Sep. 2001.

[18] J. Smith and C. Li, "Image Classification and Querying Using Composite Region Templates," Int. J. Computer Vision and Image Understanding, vol. 75, no. 2, pp. 165-174, July 1999.

[19] S. Mukherjea, K. Hirata, and Y. Hara, "AMORE: A World Wide Web Image Retrieval Engine," Proc. World Wide Web, vol. 2, no. 3, pp. 115-132, June. 1999.