International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:3, 2007

Grouping and Indexing Color Features for Efficient Image Retrieval

M. V. Sudhamani, and C. R. Venugopal

Abstract**—**Content-based image retrieval (CBIR) aims at searching image databases for specific images that are similar to a given query image based on matching of features derived from the image content. This paper focuses on a low-dimensional color based indexing technique for achieving efficient and effective retrieval performance. In our approach, the color features are extracted using the mean shift algorithm, a robust clustering technique. Then the cluster (region) mode is used as representative of the image in 3-D color space. The feature descriptor consists of the representative color of a region and is indexed using a spatial indexing method that uses R^* -tree thus avoiding the high-dimensional indexing problems associated with the traditional color histogram. Alternatively, the images in the database are clustered based on region feature similarity using Euclidian distance. Only representative (centroids) features of these clusters are indexed using R^* -tree thus improving the efficiency. For similarity retrieval, each representative color in the query image or region is used independently to find regions containing that color. The results of these methods are compared. A JAVA based query engine supporting query-by- example is built to retrieve images by color.

Keywords**—**Content-based, indexing, cluster, Region.

I. Introduction

THE use of low-level visual features to retrieve relevant information from image and video databases has drawn much research attention in recent years. Color is perhaps the most dominant and distinguishing visual feature. Color histogram is the most widely used color descriptor in content based retrieval research. A color histogram captures global color distribution in an image. While color histograms are easy to compute, they result in large feature vectors that are difficult to index and have high search and retrieval cost. In addition, spatial information is not preserved in a color histogram. Thus a large red color blob in a green background will have the same color histogram as an image containing the same number of randomly distributed red and green pixels. Several of the recently proposed color descriptors try to incorporate spatial information to varying degrees. These include the compact color moments [1], [2], binary color sets [3], color coherence vector [4], and color correlogram [5]. The feature vector dimensions of typical color descriptors are quite large. For example, the numbers of bins in a typical color histogram range from few tens to a few hundreds. The high dimensionality of the feature vectors result in high computational cost in distance calculation for similarity retrieval, and inefficiency in indexing and search. Several methods have been proposed to overcome these problems. In [6] the dominant colors in the histogram are used, and a multiresolution color clustering is suggested in [7] to reduce the computational complexity in distance calculation. Singular value decomposition (SVD) [8] and Hilbert curve fitting [9] are used to reduce the dimensionality of the feature vectors. However, these methods have their own drawbacks. In [8] SVD is performed on the quadratic matrix of correlations between the color histogram bins. The resulting eigenvectors are not related to the feature data, and may result in significant errors when lower-dimensional transformed feature vectors are used to approximate the original feature vectors. The results of Hilbert curve fitting depend on the data distributions. Points that are close to each other in the original feature space might be far apart on the Hilbert curve. The distances in the original space might not be preserved well in the curve approximation. The color moments descriptor proposed in [1], [2] has a compact representation. The moment descriptor includes the average, variance, and the third-order moment of the colors in the image. A recent study [10] shows that the color moment descriptor performs slightly worse than a high-dimensional color histogram. One drawback of the moment descriptor is that the average of all the colors might be quite different from any of the original colors. Given a color moment feature description, it is difficult to recover the actual colors in the image.

The proposed descriptor is compact, and is based on the observation that a small number of color/colors are usually sufficient to characterize the color information in an image region. Since the descriptor captures the representative or dominant colors in a given region, we refer to it as the dominant color descriptor. A Euclidian distance measure is used for the color descriptor. However, the difference between the new descriptor and the color histogram descriptor is that the representative colors are computed from each image instead of being fixed in the color space, thus allowing the feature representation to be accurate as well as compact. Unlike the compact color moments descriptor, the dominant color representation allows simple visualization of the color distributions in the image.

Manuscript received March 25, 2007.

M. V. Sudhamani is with the Siddaganga Institute of Technology, Tumkur, Karnataka, India, Pin: 572 103 (phone: 91- 0816-2214064; fax: 91-0816-2282994; e-mail: mvsudha_raj@ hotmail.com).

C. R. Venugopal, is with Sri Jayachamarajendra college of Engineering, Mysore, Karnataka, India, Pin: 570 006 (e-mail: crv@sjce.ac.in).

A. Previous Work in CBIR

Current CBIR systems such as IBM's QBIC, allow automatic retrieval based on simple characteristics and distribution of color, shape and texture. But they do not consider structural and spatial relationships and fail to capture meaningful contents of the image in general. Also the object identification is semi-automatic. The Chabot project integrates a relational database with retrieval by color analysis. Textual meta-data along with color histograms form the main features used. VisualSEEK allows query by color and spatial layout of color regions. Text based tools for annotating images and searching is provided. A new image representation which uses the concept of localized coherent regions in color and texture space is presented [24],[25]. Recently, additional systems have been developed at IBM T.J. Watson [11], VIRAGE [12], NEC C&C Research Labs [13], Bell Laboratory [14], Interpix (Yahoo), Excalibur, and Scour.net. In academia, MIT Photobook [15, 16] is one of the earliest. Berkeley Blobworld [17], CMU Informedia [18], University of Illinois MARS [19], University of California at Santa Barbara NeTra [20], the system developed by University of California at San Diego [21], Stanford WBIIS [22], and Stanford SIMPLIcity [23],[24] are some of the recent systems. Segmentation based on the above features called “Blobworld" is used and query is based on these features. Some of the popular methods to characterize color information in images are color histograms, color moments and color correlograms. Though all these methods provide good characterization of color, they have the problem of high-dimensionality. This leads to more computational time, inefficient indexing and performance. To overcome these problems, use of SVD, dominant color regions approach, and color clustering have been proposed. In this paper, we focus on region-based retrieval of images.

B. Region-Based Retrieval

Region-based approach has recently become a popular research trend. Region-based retrieval systems attempt to overcome the deficiencies of color histogram and color layout search by representing images at the object-level. A region-based retrieval system applies image segmentation to decompose an image into regions, which correspond to objects if the decomposition is ideal. The object-level representation is intended to be close to the perception of the human visual system. Many retrieval systems match images based on individual regions. Such systems include for e.g., the Netra system and the Blobworld system. To query an image, a user is provided with the segmented regions of the image, and is required to select the regions to be matched e.g., color, of the regions to be used for evaluating similarity. Such querying systems provide more control for the users. Blobworld is a CBIR system that fragments an image into regions (blobs), homogeneous with respect to color and texture, by using an Expectation-Maximization clustering algorithm. The Blobworld index-based query resolution algorithm uses an *R*-tree like structure to index color descriptors of blobs. This uses nearest neighbors query on the index. Netra system used the edge-flow algorithm for image segmentation and colors in a region are indexed using hexagonal lattice structure. Our proposed method uses Mean shift robust clustering algorithm for segmentation of images and interested region/regions are indexed using R^*-tree and the algorithm described in section III and uses range query. The remainder of the paper is organized as follows. In Section II, preprocessing and feature extraction is discussed. Section III discusses the indexing methods of the proposed system. In section IV, experimental results are provided and discussed. Section V concludes the paper and future extension of the present work.

II. Preprocessing and Feature Extraction

A. Segmentation

The local color feature extraction starts with color image segmentation. For image segmentation, we use mean shift algorithm [26]**.** Here, color clustering is performed on each image to obtain regions. After segmentation, only small number of color remains. Information like number of regions, time taken to segment an image, boundary points, region points, and region numbers can be extracted.

Large classes of image segmentation algorithms are based on feature space analysis. In this paradigm the pixels are mapped into a color space and clustered, with each cluster delineating a homogeneous region in the image. Pixels were directly associated with the mode to which the path converged. The approximation does not yield a visible change in the filtered image. Recursive application of the mean shift property yields a sample mode detection procedure. The modes are the local maxima of the density. They can be found by moving at each iteration the window by the mean shift vector, until the magnitude of the shifts becomes less than a threshold. The procedure is guaranteed to converge. The number of significant modes detected automatically determines the number of significant clusters present in the feature space. For the color image segmentation algorithm the $L^*u^*v^*$ color space was employed since its metric is a satisfactory approximation to Euclidean, thus allowing the use of spherical windows.

B. Architecture of Proposed System

Fig. 1 shows architecture of a content-based image retrieval system. Two main functionalities are supported: Data insertion and Query processing. The data insertion subsystem is responsible for extracting appropriate features from images and storing them into the image database. This process is performed off-line. The query processing, intern, is organized as follows: the interface allows a user to specify a query by means of a query pattern and to visualize the retrieved similar images. The query-processing module extracts a feature vector from a query pattern and applies a metric as the Euclidean distance to evaluate the similarity between the query image and the database images. Next, it ranks the database images in a decreasing order of similarity to the query image and forwards the most similar images to the interface module. The database images are indexed according to their feature vectors to speed up retrieval and similarity computation. Note that both the data insertion and the query processing functionalities use the feature vector extraction module.

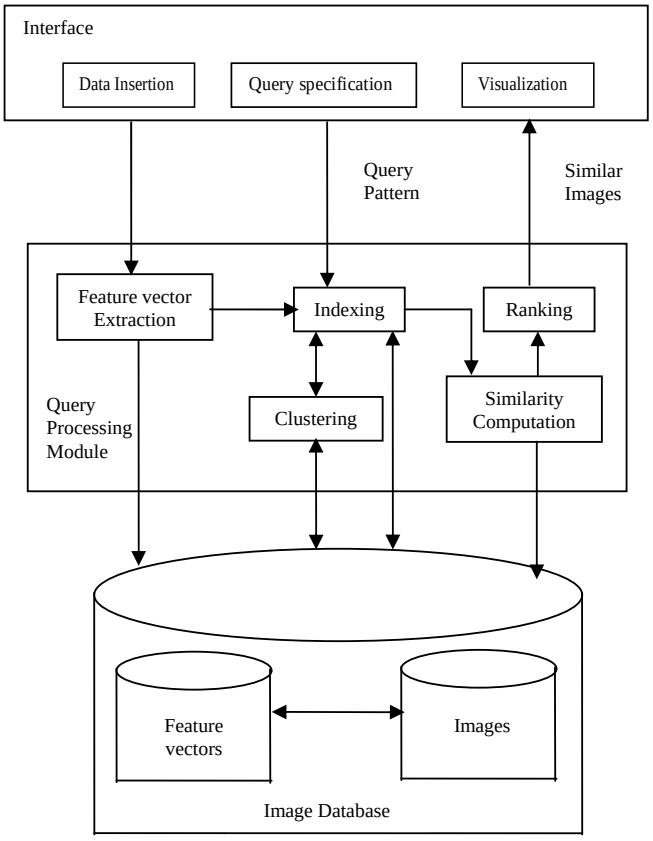

Fig. 1 Architecture of a proposed system

III. Search and Retrieval

Each representative color descriptor F is defined to be F= {c_i }, i=1,., 3, where c_i is 3-D color vector of a region. Each object or region in the database is represented using color descriptor. Given a query image, similarity retrieval involves searching the database for similar color as the input query. Searching for the individual colors can be done very efficiently in a 3-D color space. We consider here only fixed range queries in which the range value limits the search range. We consider here only fixed range queries in which the range value limits the search range.

A. Indexing

In building the database, each image is assigned with unique image ID. The image region ID is a unique integer label that identifies each region in the database. The entries in each index node are sorted by region ID numbers. The proposed indexing scheme allows the database to be dynamic, which means that insertion and deletions of database entries are straight forward and without the need to reconstruct the entire index structure of the database.

Here, we describe cluster-based indexing method aiming to speed-up the evaluation of range queries. This is carried out by reducing the number of candidate images, i.e. images on which the optimal region-matching problem has to be solved.

We chose n-clusters for grouping the features, where n is the number of region features chosen for query. The features in the database are mapped to a corresponding cluster based on Euclidean similarity measure. Each representative color of a cluster is indexed using R^*-tree rather than indexing all the regions thus reducing the time. We only access the images, which are selected for retrieval as candidates.
The procedure is as follows:

1. Given n the number of query regions, for each query region q_j, find the regions belonging to cluster c_j, where $j = 1,..,n$.
2. For each region r_i in the image database
   a. Find the feature vector f_i, for region r_i.
   b. For each query region q_j, in the query set
      i. Find query feature vector f_j, for q_j.
      ii. Find the Euclidean distance between f_i and f_j using: $d_{ij} = \sqrt{\sum_{k=1}^{m} \left(f_{i_k} - f_{j_k}\right)^2}$, where m is the dimension of the feature vector. This score is zero if the regions features are identical, it increases as the match becomes less perfect.
      iii. Measure the similarity between f_i and f_j using $\mu_{ij} = d_{ij} - \tau$, where τ is the search range limit set by user.
      iv. If $\mu_{ij} \le 0$, then f_i belongs to cluster c_j and go to step 2.

After the above procedure is completed, we then index only representatives of c_j, where $j = 1,...,n$ using R^*-tree. Once the query is selected by the user, we apply range search on the tree and the selected regions are retrieved as resultant set. Members of resultant set is ranked according to overall score and return the best matches in decreasing order of similarity along with their relative information.

In case of R^*-tree, all the regions in the database are being indexed. When we pose a query-by-example, based on the range, selected images are displayed as a resultant set according to ranking of similarity in descending order.

In sequential search, all the regions stored in the database are compared for similarity and in turn retrieved for display, making it inefficient.

All three methods yield good performance when the accuracy of resultant set is considered but the proposed method overscores all the above as depicted in table I and II.

IV. Experimental Results

The representative color descriptor is tested on a database of 200 color flag images. After segmentation 440 regions are obtained. Among them, 13 image regions containing a variety

of colors and color combinations are chosen as queries. Table I summarizes experimental data.

Before the evaluation, subjective testing is done to determine the relevant matches in the database to the query image regions. The time complexity associated with proposed method (cluster-index) and other methods are shown in Fig. 2 and listed the values in the Table I. The efficiency of the proposed method is high (less time).

The retrievals from proposed method and other methods are listed in Table II. The retrieval accuracy is measured by precision and recall,

$$\text{Precision}(k) = c_k / k \text{ and } \text{Recall}(k) = c_k / M$$

where k is the number of retrievals, c_k is the number of relevant matches among all the k retrievals, and M is the total number of relevant matches in the database obtained through the subjective testing. The precision and recall values for different queries are listed in Table I. The precision and recall curves are plotted in Fig. 3 and Fig. 4. It can be seen from the tables and Fig. 6 and Fig. 7 that the proposed method achieves good results in terms of the retrieval accuracy. Fig. 5 shows the consistency of precision with respect to different color features of queries. Chosen one query with the red color and another is with yellow as in Fig. 6 and Fig. 7. The retrievals in both examples show good match of colors. Fig. 8 shows the data insertion and display operations from the database.

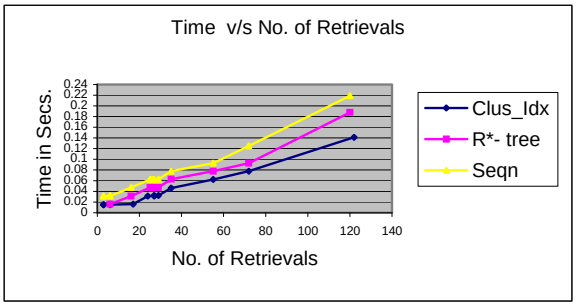

Fig. 2 Time in seconds versus number of retrievals

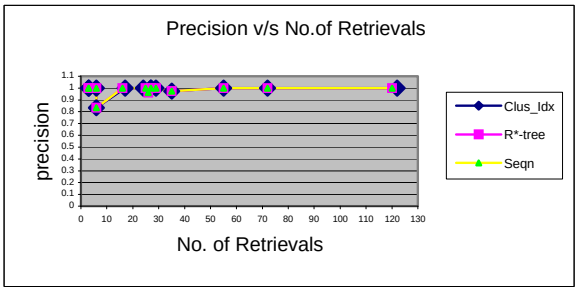

Fig. 3 Precision versus number of retrievals

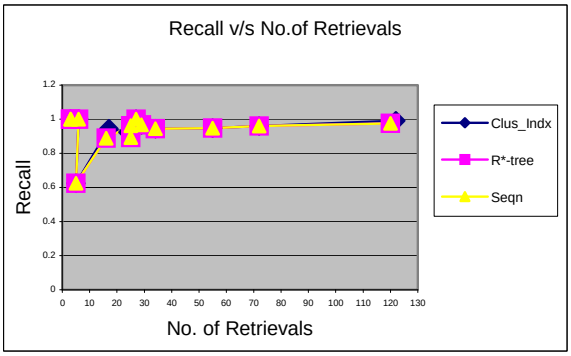

Fig. 4 Recall versus number of retrievals

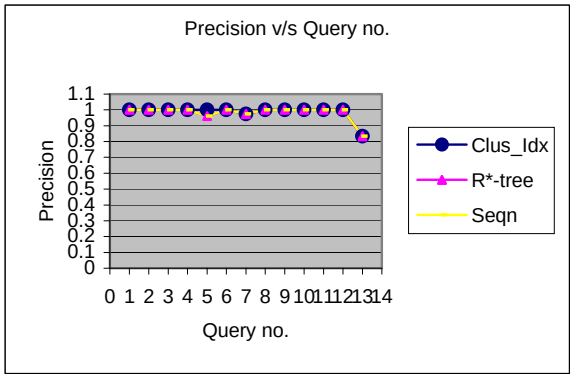

Fig. 5 Precision versus query number

Fig. 6 Example of region-based image search using the representative color descriptor. The query is the yellow color of Columbia flag

Fig. 7 Example of region-based image search using the representative color descriptor. The query is the red color of Bahrain flag

Fig. 8 Example database Insertion and Display operations

TABLE II
RETRIEVALS FOR DIFFERENT COLOR FEATURES WITH THREE METHODS

	Color Features	Cluster-Index (new)			R*-tree index			Sequential Search		
Sl No	Query color	K	C_K	M	K	C_K	M	K	C_K	M
1	Brown	3	3	3	3	3	3	3	3	3
2	Pink Brown	3	3	3	3	3	3	3	3	3
3	Saffron	6	6	6	6	6	6	6	6	6
4	Sky Blue	6	5	8	6	5	8	6	5	8
5	Light green	17	17	18	16	16	18	16	16	18
6	Navy Blue	24	24	26	25	25	26	25	25	26
7	Dark Blue	27	27	28	26	25	28	26	25	28
8	Black	27	27	27	27	27	27	77	27	27
9	Light Blue	29	29	30	29	29	30	29	29	30
10	Yellow	35	34	36	35	34	36	35	34	36
11	Green	55	55	58	55	55	58	55	55	58
12	White	72	72	75	72	72	75	72	72	75
13	Red	122	122	123	120	120	123	120	120	123

TABLE I
TIME , PRECISION, AND RECALL VALUES FOR THREE METHODS

		Cluster-Index (new)			R*-tree Index			Sequential Search		
Sl No	Query color	Time	Precision	Recall	Time	Precision	Recall	Time	Precision	Recall
1	Brown	0.015	1	1	0.016	1	1	0.031	1	1
2	Pink Brown	0.015	1	1	0.016	1	1	0.031	1	1
3	Saffron	0.016	1	1	0.016	1	1	0.032	1	1
4	Sky Blue	0.015	0.83	0.62	0.016	0.83	0.62	0.031	0.83	0.62
5	Light green	0.016	1	0.94	0.031	1	0.88	0.047	1	0.88
6	Navy Blue	0.031	1	0.92	0.047	1	0.96	0.062	1	0.96
7	Dark Blue	0.032	1	0.96	0.046	0.96	0.89	0.62	0.96	0.89
8	Black	0.031	1	1	0.047	1	1	0.062	1	1
9	Light Blue	0.032	1	0.96	0.047	1	0.96	0.063	1	0.96
10	Yellow	0.046	0.97	0.94	0.063	0.97	0.94	0.078	0.97	0.94
11	Green	0.062	1	0.94	0.078	1	0.94	0.093	1	0.94
12	White	0.078	1	0.96	0.093	1	0.96	0.125	1	0.97
13	Red	0.141	1	0.99	0.188	1	0.97	0.219	1	0.97

V. Conclusion

In this work, a representative color descriptor for a region in an image and cluster-based R^*-tree indexing of regions is proposed. Euclidean similarity measure is defined for the proposed method. Experimental results show that the proposed method is fast and accurate over R^*-tree and sequential search methods as we can see from the table values. The proposed method can be used for color-based image retrieval where the fastness and accuracy of the result is the requirement. The same work will be extended for over 5000 natural images to test the efficiency and accuracy of the retrieval.

References

[1] M. A. Stricker and M. Orengo, Similarity of color images, *Proc. SPIE, Storage Retrieval Still Image Video Databases IV, vol. 2420, pp. 381–392, 1996.*

[2] M. Stricker and A. Dimai, Color indexing with weak spatial constraints, Proc. *SPIE Storage Retrieval Still Image Video Databases IV, vol. 2670, pp. 29–40,1996.*

[3] J. Smith and S.-F. Chang, Tools and techniques for color image retrieval, *Proc. SPIE, vol. 2670, pp. 2–7, 1996.*

[4] G. Pass and R. Zabih, Histogram refinement for content based image retrieval, *Proc. IEEE Workshop Applications Computer Vision, pp. 96-102, 1996.*

[5] J.Huang, S R Kumar, M Mithra, W.Zhu, and R. Zabih, Image indexing using color correlograms, *Proc. IEEE conf. Computer vision and pattern Recognition, pp. 762-768,1997.*

[6] H. Zhang, Y. Gong, C. Y. Low, and S.W. Smoliar, Image retrieval based on color features: An evaluation study, *Proc. SPIE Digital Image Storage Archiving Systems, vol. 2606, pp. 212–220,1995.*

[7] X. Wan,C. J. Kuo, A multiresolution color clustering approach to image indexing and retrieval, *Proc. IEEE Int. Conf. Acoustics, Speech, Signals Processing, vol. 6, pp. 3705–3708,1998.*

[8] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack, Efficient color histogram indexing for quadratic form distance functions, *IEEE Trans. Pattern Anal. Machine Intell., vol. 17, pp. 729–736, July 1995.*

[9] G. Cha and C. Chung, Multi-mode indices for effective image retrieval in multimedia systems, *Proc. IEEE Multimedia Computing Systems, pp. 152–159, 1998.*

[10] W. Y. Ma and H. Zhang, Benchmarking of image features for content-based retrieval, *Proc. IEEE 32nd Asilomar Conf. Signals, Systems, Computers, vol. 1, pp. 253–257,1998.*

[11] J. R. Smith, C. S. Li, Image classification and querying using composite region templates, *Journal of Computer Vision and Image Understanding, vol. 23, 2001.*

[12] A. Gupta, R. Jain, Visual information retrieval, *Communications of the ACM*, vol. 40, no. 5, pp. 70-79, May 1997.

[13] S. Mukherjea, K. Hirata, Y. Hara, AMORE: a World Wide Web image retrieval engine, World Wide Web, *vol. 2, no. 3, pp. 115-32, Baltzer, 1999.*

[14] Natsev, R. Rastogi, K. Shim, WALRUS: A similarity retrieval algorithm for image databases, *Proc. ACM SIGMOD, Philadelphia, PA, 1999.*

[15] Pentland, R. W. Picard, S. Sclaro, Photobook: tools for content-based manipulation of image databases, *Proc. SPIE, vol. 2185, pp. 34-47, San Jose, February 7-8, 1994.*

[16] R. W. Picard, T. Kabir, Finding similar patterns in large image databases, *Proc. IEEE ICASSP, Minneapolis, vol. V, pp. 161-64, 1993.*

[17] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, J. Malik, Blobworld: a system for region-based image indexing and retrieval, *Proc. Int. Conf. on Visual Information Systems, D. P. Huijsmans, A. W.M. Smeulders (eds.), Springer, Amsterdam, The Netherlands, June 2-4, 1999.*

[18] S. Stevens, M. Christel, H. Wactlar, Informedia: improving access to digital video, *Interactions, vol. 1, no. 4, pp. 67-71, 1994.*

[19] S. Mehrotra, Y. Rui, M. Ortega-Binderberger, T.S. Huang, Supporting content-based queries over images in MARS, *Proc. IEEE International Conference on Multimedia Computing and Systems, pp. 632-3, Ottawa, Ont., Canada 3-6 June 1997.*

[20] W. Y. Ma, B. Manjunath, NeTra: A toolbox for navigating large image databases," *Proc. IEEE Int. Conf. Image Processing, pp. 568-71, 1997.*

[21] R. Jain, S. N. J. Murthy, P. L.-J. Chen, S. Chatterjee, Similarity measures for image databases, *Proc. SPIE, vol. 2420, pp. 58-65, San Jose, CA, Feb. 9-10, 1995.*

[22] J. Z. Wang, G. Wiederhold, O. Firschein, X. W. Sha, Content-based image indexing and searching using Daubechies' wavelets, *International Journal of Digital Libraries, vol. 1, no. 4, pp. 311-328, 1998.*

[23] J. Z. Wang, *Integrated Region-Based Image Retrieval*, Kluwer Academic Publishers, 190 pp., 2001.

[24] Zaher aghbari, Akifumi makinouchi, Semantic Approach to Image Database Classification and Retrieval, *NII Journal No. 7, 2003.*

[25] Shu-Ching, Chen Stuart H. Rubin, Mei-Ling, A Dynamic User Concept Pattern Learning,Framework for Content-Based Image Retrieval , *IEEE transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 36, no. 6, November 2006.*

[26] M. V. Sudhamani, C.R. Venugopal, Non-parametric classification of image data through clustering: An application for image Retrieval, *Proc. of IEEE Int.Conf. Image and signal processing, Dec 2006.*