System for indexing multi-spectral satellite images for efficient content-based retrieval *

Julio Barros, James French, Worthy Martin

Computer Science Dept.
University of Virginia
Charlottesville, Va 22903

Patrick Kelly

Los Alamos National Laboratory
Los Alamos, NM 87545

ABSTRACT

Current feature-based image databases can typically perform efficient and effective searches on scalar feature information. However, many important features, such as graphs, histograms, and probability density functions, have more complex structure. Mechanisms to manipulate complex feature data are not currently well understood and must be further developed. The work we discuss in this paper explores techniques for the exploitation of spectral distribution information in a feature-based image database. A six band image was segmented into regions and spectral information for each region was maintained. A similarity measure for the spectral information is proposed and experiments are conducted to test its effectiveness. The objective of our current work is to determine if these techniques are effective and efficient at managing this type of image feature data.

Keywords: image databases, content-based retrieval, spectral distributions, similarity measures

1 INTRODUCTION

Digital imagery is an increasingly important and prevalent form of information. As a result, image database management systems are more commonly being used to organize large collections of images. Many of these systems are simple archives while others maintain external information on the images such as imaging parameters and textual descriptions. More advanced systems attempt to provide a content-based retrieval capability by extracting and managing image feature data. The term "content-based

*To appear in the proceedings of IS&T/SPIE: Storage and Retrieval for image and Video Databases III, Feb 1995, San Jose California.

retrieval" occasionally means retrieval of images based on external parameters or textual annotations. However, we use the term here to mean the searching of image databases using the intrinsic properties of the images and not just the external parameters or textual description. Content-based retrieval from image databases is still not well understood and presents a major research challenge.

Current content-based systems, such as QBIC,[7] Photobook[8] and the framework proposed by Yazdani *et al.*,[12] extract a set of features from each image. The feature data is maintained in a data structure that allows efficient access. Queries on the database are processed by searching the managed features and are usually similarity based searches and not for exact values. These powerful feature-based systems show great promise in many applications. However, current systems typically only manage scalar feature data.

Many important features, such as graphs, histograms, and probability density functions (PDFs), have more complex structure. Although there has been much recent interest in this area,[6,10,11,9] general indexing mechanisms for such feature data do not exist or are not well understood. Consequently, if these features are used in an image database system, they are either searched exhaustively or are simplified considerably. Exhaustive search quickly becomes infeasible as the size of the database grows and the effectiveness of a feature is diminished if it is over simplified. Consequently, techniques to directly manipulate complex data need to be developed.

The work we discuss in this paper is an extension of Barros *et al.*[1] and explores techniques for the exploitation of spectral distribution information in feature-based image databases. The objective of our current work is to determine if these techniques are effective and efficient at managing this type of image feature data. An effective technique provides a useful capability to the user and an efficient one avoids exhaustive search. This paper discusses the effectiveness of our approach. We start in Section 2 by introducing an example feature with complex structure and a similarity metric. In Section 3 we describe the experiment background. We continue in Section 4 with a discussion of the test queries and results. We finish with a summary and some directions for future work.

2 SIMILARITY OF SPECTRAL DISTRIBUTIONS

In most feature-based systems, the feature data can be viewed as point data in the feature space. In these systems, similarity based searches are far more common than searches for exact values. Similarity is usually calculated as the (weighted) Euclidean distance between two points in the feature space. To efficiently process queries, any of several point access methods available, such as B-trees[3] R-trees[4] or vantage-point trees,[2] can be used.

However, many important features, such as probability density functions (PDFs), have a more complex structure. PDFs are used wherever there is a need to summarize or approximate underlying data. For example, PDFs can be used to describe both collections of measurements and measurements with known errors. PDFs inherently incorporate the notion of area or distance and are thus different from simple point data and range data. Consequently, the similarity measures and access methods used should take into account particular aspects of PDFs including their size and shape.

We have been working on an approach to organize PDFs such as those describing spectral distribu-

tions. Our system starts with an image that has been segmented into *regions*. Each region is composed of pixels with similar (but possibly identical) values. We model the spectral characteristics of each region with the mean and variance of its component pixels.

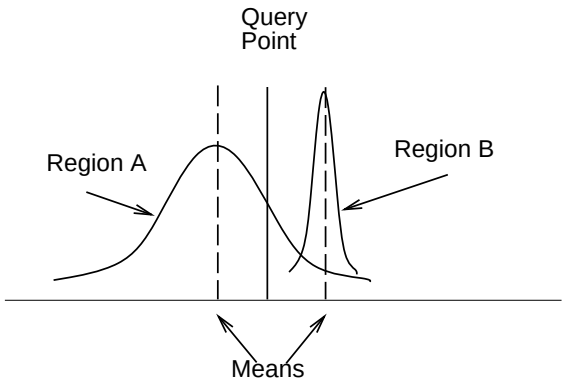

Figure 1: Region variance plays an important role in similarity judgments.

Given this representation, we require a way of measuring the similarity between a region and a query point. Initially, it seems reasonable to use the Euclidean distance between the query point and the region mean as a measure of similarity. The Euclidean distance works well as a similarity measure for many applications. However, the Euclidean distance can be an ambiguous metric. It is not always clear how the distance in feature space relates to the user's concept of similarity or how to choose an appropriate threshold distance. Additionally, a similarity measure would preferably take into account the individual region variances. The variance information can be useful when comparing distances. For example, in Figure 1, the query point is approximately the same distance from the means of the two distributions. However, region A has a larger variance. This suggests that we can consider points at a greater distance to the mean of region A to be similar to region A. Consequently, we consider the query point to be more similar to region A than to region B.

We capture this idea and measure similarity with a normalized distance measure. In this measure the distance between a query point and a distribution mean is normalized by the distribution's standard deviation. The measure we use is:

$$M_i = \left| \frac{\bar{X}_i - q}{s_i} \right| \tag{1}$$

Where q is the query point, X_i is the mean of region i, and s_i is the standard deviation of region i. In other words, M_i is the number of region standard deviations between the query point and the mean of region i.

To use this similarity criterion a user selects a point q as the query point and a threshold value M. Each region i in the database is then examined. If $M_i \leq M$ then region i is judged to be similar to q. The set of all such regions is returned to the user as the *answer* set. This set contains all the regions that are within M standard deviations of the query point.

As presented so far, we must apply the similarity criterion sequentially to each region in the database. However, the test can be modified to be more efficient in many situations, where efficiency is with respect

to the number of comparisons required. The efficiency can be realized through an indexing structure that yields the answer sets without directly calculating M_i for each region i.

This is possible when we notice that the similarity criterion can be written as

$$-M \leq \frac{\bar{X}_i - q}{s_i} \leq M \tag{2}$$

or

$$q - Ms_i \leq \bar{X}_i \leq q + Ms_i \tag{3}$$

In this form we notice that the region mean $\bar{X}_i$ must fall in the range between $q - Ms_i$ and $q + Ms_i$. The query point q specifies the center of the range and M and s_i specify the size of the range. Since q and M are fixed at the beginning of the query the only unknown is the value of s_i. Again, the value s_i changes with each individual region i. When processing the query the largest range is calculated by using the largest s_i, s_{max}, in the database. All *candidate regions*, regions that could possible pass the similarity criterion, lie in this calculated *query range*.

$$q - Ms_{max} \leq \bar{X} \leq q + Ms_{max} \tag{4}$$

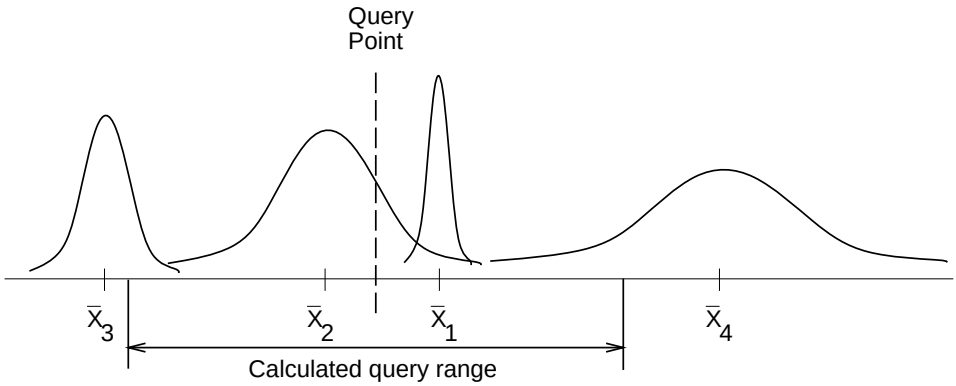

Figure 2: The calculated query range using s_4 as s_{max}

The next step in taking advantage of this information is to create an index of the regions means. The index is created when the database is initialized and is updated each time a new region is added to the database. The index will be used to find regions with means falling within a particular range. Consequently, the index structure used must allow quick updates and efficient range searches. Several acceptable indexing structures, such as the B-tree,[3] are available.

Then, for each query, we calculate the query range, using Equation 4, and use it to perform an efficient range search on the region means. This step retrieves the *candidate* set which contains all the candidate means. The similarity test, Equation 1, can then be performed sequentially on this subset of regions. The set of regions that pass the similarity criterion is then returned to the user as the answer set. It is important to observe that since s_{max} is used for Equation 4, the answer set returned by this two stage process is *guaranteed* to be the same as that returned by the initial method. For example, Figure 2 shows four regions, a query point, and a calculated query range. The regions with means

within the range, i.e., regions 1 and 2, constitute the candidate set returned by the indexing structure, and only these two regions need to receive any further consideration. All other regions can be safely ignored.

The savings realized by this process depends on the size of the candidate set returned by the calculated range search and the degree to which the cost of initially creating the indexing structure can be amortized across multiple queries. If the returned candidate set includes all regions in the database, then we have only incurred additional overhead. However, considerable savings may be realized if this set is significantly smaller than the database. The size of the candidate set, and thus the efficiency of the procedure, is affected by three factors. The first factor is the magnitude of the largest variance. If there is even one region with large variances the size of the query range will be large. The query range may then include a large percentage of the database. In this situation it might be beneficial to use a smaller value than s_{max} in Equation 4. The smaller value would relax the guarantee that the candidate range will contain all regions that should be considered. However, the loss of a few answer regions might be acceptable if the candidate set is significantly smaller than the overall database, yet still contains most of the answer regions. The second factor is the relative positions of the means in the spectral space. If the means are concentrated in a small area, then all queries in this area may retrieve a large percentage of the database. The third factor is the number of queries that can benefit from the indexing structure and thus help amortize the cost of the indexing mechanism. Therefore, the best savings is realized by a database with a relatively small s_{max}, with means that are uniformly distributed throughout the space, and with a high number of queries that can take advantage of the indexing mechanism.

3 EXPERIMENT BACKGROUND

For the experiments described in this paper we used six of the seven spectral bands (the thermal infrared band was ignored) of a 3351×2501 Landsat image. The image pixels were clustered into 241 *clusters* based solely on their spectral values by a modified k-means[5] clustering procedure. The clusters were then manually classified into eleven ground cover *classes* by an analyst. This was done by visually inspecting the image as well as the cluster spectral information. The classes were chosen by analysts independently of this work and are listed in Table 1. The image was then divided into twenty four 512×512 sub-images for purely logistical reasons. The images were median filtered and segmented into *regions* using a connected components algorithm on the class labels. All regions greater than a minimum size of 25 pixels were kept and used in the database. Several features, including area, center, bounding rectangle, and spectral distribution information, were extracted for each region. For this study, only the spectral information is relevant.

The spectral information was calculated for each individual region by pooling the values of the clusters of the pixels that make up that region. In other words, the region mean and variance is the weighted average of the cluster means and variances. Each region weights the cluster information by the number of pixels in that region from that cluster.

Table 1 shows the number and percentage of regions and pixels of each class in the database. For example, there were a total of 22,856 regions accounting for the 5,355,122 pixels in the database. Of these, 2,919 regions accounting for 233,010 pixels were classified as cropland. Additionally, the cropland regions are 12.8% of the regions and 4.4% of the pixels.

Class	Number of regions	Number of of pixels	Percentage of regions	Percentage of pixels
Residential	462	46,398	2.1	0.9
Cropland	2,919	233,010	12.8	4.4
Grassland	4,898	837,881	21.4	15.6
Forest deciduous	4,118	1,552,255	18.0	28.9
Forest evergreen	3,205	1,594,049	14.0	29.8
Scrub/shrub	3,309	202,844	14.5	3.8
Water	186	275,280	0.8	5.1
Wetland	99	5,630	0.4	0.1
Exposed Land	3,024	500,502	13.2	9.3
Artificial surface	460	96,210	2.0	1.8
Exposed surface w/veg	176	11,063	0.8	0.2

Table 1: Number and percentage of regions and pixels of each class in the database.

4 TEST QUERIES AND RESULTS

So far, we have presented a feature and a similarity measure but no method to gauge its effectiveness. Designing such a benchmark is more difficult than it might first appear. We keep in mind that the semantics of the similarity criterion are "find all regions which are spectrally similar to a query point". Unfortunately, there is no way to independently decide which regions actually are "spectrally similar". Consequently, there is no obvious non-circuitous way to evaluate the resultant set of regions.

As an approximation, we settled on a query that could be easily and directly evaluated. The query semantics are "find all regions of the same class as the query point by using only the spectral information". It is not possible to correctly classify all the regions based solely on the spectral information alone. Additionally, there is no reason to believe that all regions of a particular class are more similar to each other than they are to regions of another class. Consequently, the semantics of the similarity criterion and the query are subtly but significantly different. The class labels of the regions are used for evaluation only and are not used during the queries at all.

The test queries are posed against multiple bands by querying each band individually and intersecting their answer sets. We chose to use two bands since two dimensions are easy to visualize and give reasonable performance. Figure 3 is a scatter plot of the regions means in the two dimensions selected. Selecting the number and combinations of bands that give the best results is difficult. We chose to use bands four and five by visually inspecting the projection of the class means on all combinations of two dimensions. This was done to find two dimensions that provide the greatest separation of the classes. We are investigating ways to automate this process.

To measure the effectiveness of the similarity criterion, queries were posed against bands 4 and 5 with varying values of M for each of the classes in the database. The class means were used as the query points during each set of queries. The value of M was varied in steps of .25 from .25 up to at most 10. For each value of M, the query is performed and the results are inspected. If a query achieved 100% recall, the query process was stopped and no further values of M were used.

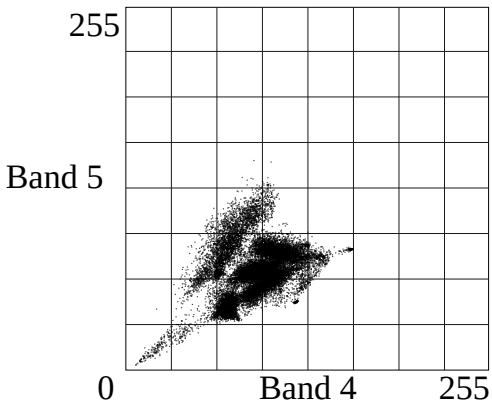

Figure 3: The means of regions projected on bands 4 and 5.

Multiple of M	Water	Forest evergreen	Wetland	Artificial surface	Exposed land	Scrub/ shrub	Forest deciduous	Residential	Cropland	Exposed surf w/veg	Grassland
0.25	2/100	0/100	1/100	5/95	1/95	1/53	0/100	9/100	1/49	2/100	1/89
0.50	5/100	0/100	9/100	13/95	2/92	2/60	1/100	27/95	3/50	8/100	2/85
0.75	6/100	0/100	16/100	26/93	5/92	5/57	2/95	45/91	7/53	14/86	4/83
1.00	8/100	0/80	34/100	37/90	10/90	8/55	3/94	58/83	12/58	31/80	7/82
1.25	11/100	1/76	46/98	48/89	15/92	11/49	6/93	68/75	20/59	44/57	12/84
1.50	16/100	1/76	56/96	58/86	20/93	15/46	8/93	74/68	28/59	55/44	16/84
1.75	18/100	1/81	72/95	70/85	26/93	19/43	12/92	79/58	35/58	64/30	21/84
2.00	20/90	2/82	78/93	78/84	32/93	23/39	16/91	84/51	43/57	69/20	26/85
2.25	25/88	2/85	82/91	84/81	38/93	28/37	21/89	87/44	49/55	77/16	31/85
2.50	30/89	3/84	89/91	89/78	45/93	33/36	25/88	91/40	55/53	84/12	36/85
2.75	35/90	4/87	90/86	93/74	52/92	38/35	30/86	94/35	61/51	92/10	40/84
3.00	42/92	6/88	92/82	96/69	58/92	42/34	35/84	94/31	66/48	96/8	45/84
3.25	48/92	7/85	95/77	97/62	65/91	46/33	40/80	95/27	70/45	98/7	49/83
3.50	54/92	9/82	97/72	98/56	70/91	50/32	45/77	97/25	74/43	99/6	54/81
3.75	60/93	10/78	97/68	99/48	77/91	54/31	50/72	98/23	77/39	99/5	58/80
4.00	66/93	13/76	98/65	99/43	81/90	58/30	54/68	100/21	81/36	99/4	62/78
4.25	72/94	16/75	99/59	99/37	85/89	61/29	59/63	100/19	84/33	100/4	66/76
4.50	78/94	19/74	99/53	99/33	87/88	64/27	62/57		85/30		69/73
4.75	81/93	24/75	99/47	100/30	90/87	68/26	65/52		87/28		72/70
5.00	85/91	28/76	100/42	100/27	92/85	71/25	69/49		89/26		76/67
5.25	89/89	33/76		100/25	93/84	75/24	72/46		90/25		80/64
5.50	94/85	38/76		100/23	94/82	77/23	75/43		91/23		83/62
5.75	95/83	43/77			95/80	80/22	77/41		92/22		86/60
6.00	96/77	49/78			96/78	82/21	79/39		93/21		89/58
6.25	96/73	55/78			97/76	84/21	81/38		93/20		91/56
6.50	98/69	60/78			98/73	87/20	83/37		94/20		94/55
6.75	99/67	65/78			98/71	89/20	85/35		95/19		95/53
7.00	99/62	70/78			99/67	91/20	87/33		96/18		96/51
7.25	99/57	74/77			99/64	92/19	89/32		96/18		97/49
7.50	100/54	78/76			99/61	94/19	90/30		96/17		97/48
7.75		81/76			99/57	95/19	92/28		96/17		98/46
8.00		84/75			99/53	96/18	93/27		96/17		98/45
8.25		87/74			99/50	97/18	94/25		97/16		99/43
8.50		89/72			99/46	98/18	95/24		97/16		99/42
8.75		91/70			100/44	98/18	96/24		97/16		99/41
9.00		93/68			100/41	99/18	97/23		97/15		99/39
9.25		94/66			100/39	99/17	98/22		97/15		99/38
9.50		96/64			100/37	99/17	98/22		97/15		100/37
9.75		97/62			100/35	100/17	99/21		97/15		100/36
10.00		98/59			100/34	100/17	99/21		100/15		100/35

Table 2: Recall and Precision percentages for queries against bands 4 and 5

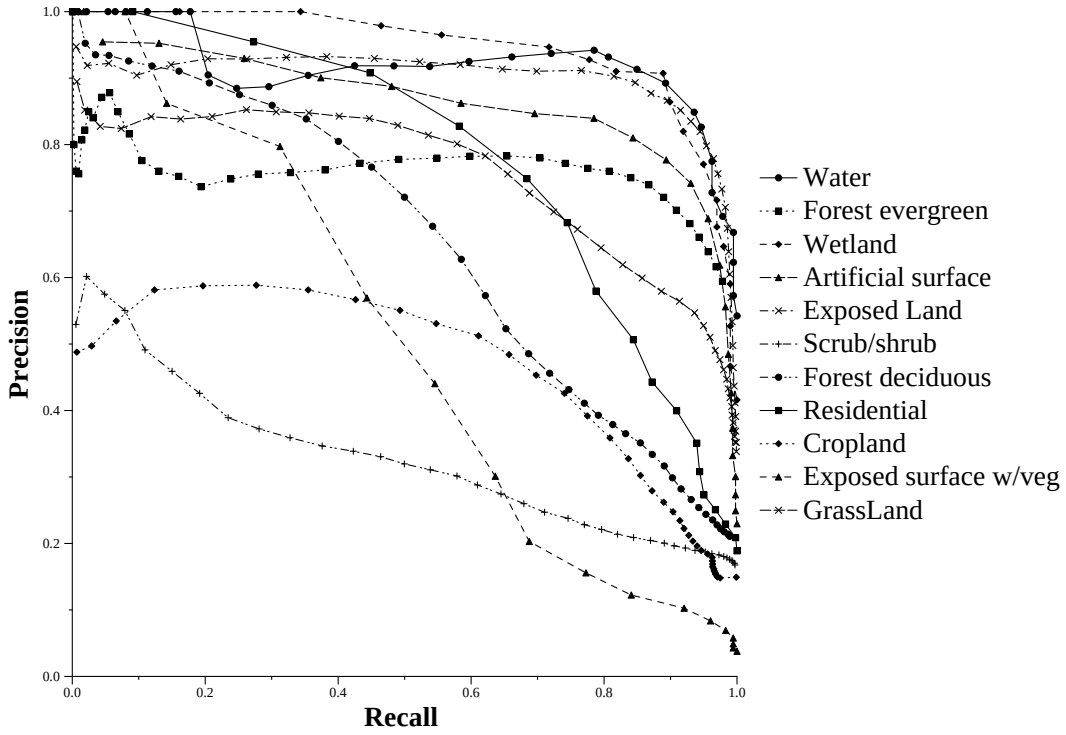

Figure 4: Recall and Precision for queries against bands 4 and 5

Figure 4 and Table 2 show the recall and precision results. Recall is the percentage of desired regions returned by the query. That is, the number of desired regions returned by the query divided by the number of desired regions in the database. For the purposes of this evaluation, a desired region is a region of the same class as the query point. Precision is the percentage of regions returned which were desired regions. That is, the number of desired regions returned by the query divided by the total number of regions returned by the query. The markers of the class lines of Figure 4 indicate the different values of M. Note that since each class will have different recall and precision results the values of M cannot be compared across classes by simply looking along a horizontal or vertical line. To compare results across classes for specific values of M we must refer to Table 2.

Figure 4 shows that the effectiveness varies with the value of M as well as with the class involved. For example, performance for some classes such as water and wetlands is considerably better than others. In these tests it was possible to get 18% recall of the water and 34% of the wetland regions while maintaining about 100% precision. That is, at this setting of M, 18% and 34% of the regions with the desired classification were returned along with few if any regions of an undesired classification. We can see that it is also possible to raise the recall up to about 78% while only sacrificing precision to 94%. However, to get 100% percent of the desired water regions it was necessary to return an answer set of 54% precision.

Other classes, such as scrub/shrub, do not perform as well. For example, the precision of the scrub/shrub answer sets is never better than 60% and to achieve 100% recall the precision drops to about 17%. We are currently investigating possible reasons for the difference in performance. We

believe that it may be due to the strong spectral similarity between the scrub/shrub and cropland classes in these two bands. This again brings up the subtle limitation in the test queries. The purpose of the query mechanism is to find spectrally similar regions; not to make accurate classifications. If in fact the scrub/shrub and cropland classes are spectrally similar we would expect regions of each class to be returned by the query process. However, we are evaluating the answer set based on information not available to the query mechanism. For example the distinction between the two classes may have been made by another spectral band or by examining the physical shape of the regions. Regular shaped regions are likely to be cropland while irregular regions are likely to be scrub/shrub. This information is not available and is not used during the query process. However, it could be included in a more complete image database system.

5 CONCLUSIONS

In this paper we have stated that mechanisms for managing complex feature data are not currently well understood and deserve additional research effort. We have introduced the spectral properties of regions as a type of complex feature. We have introduced a similarity measure and an indexing mechanism to organize and efficiently use the spectral information. The approach discussed would work for varying similarity thresholds and for dynamic environments where new image regions are continuously being added to the system. We have also described an experiment to test its effectiveness. The experiment demonstrates that reasonable performance can be achieved using these non-standard features.

Future work will address issues with selecting appropriate bands with which to resolve queries, modeling the regions spectral information more accurately, resolving multidimensional queries more effectively, using PDFs as well as points as queries, improving efficiency, and exploring other uses of the basic mechanism.

6 ACKNOWLEDGMENTS

The Landsat TM data was provided by the National Biological Survey GAP Analysis Program. The image was interpreted by Susan Benjamin, Bruce Wright (U.S. Geological Survey), and Jesslyn Brown (Hughes STX).

Portions of this work were were performed under a U.S. Government contract (W-7405-ENG-36) by Los Alamos National Laboratory, which is operated by the University of California for the U.S. Department of Energy. This work was also supported by an NSF Graduate Engineering Education fellowship.

7 REFERENCES

[1] Julio Barros, James French, Worthy Martin, Patrick Kelly, and James M. White. "Indexing multi-

spectral images for content-based retrieval", In *Proceedings of the 23rd AIPR Workshop on Image and Information Systems: Applications and Opportunities*, Washington, D.C., October 1994.

[2] Tzi-cker Chiueh. "Content-based image indexing", In *Proceedings of the 20th VLDB Conference*, Santiago, Chile, 1994.

[3] Douglas Comer. "The ubiquitous b-tree", *Computing Surveys*, 11(2), 1979.

[4] Antonin Guttman. "R-tree: A dynamic index structure for spatial searching", *ACM SIG-MOD Proc.*, 1984.

[5] Patrick M. Kelly and James M. White. "Preprocessing remotely-sensed data for efficient analysis and classification", In *SPIE Vol. 1963 Applications of Artificial Intelligence 1993: Knowledge-Based Systems in Aerospace and Industry*, 1993.

[6] Asanobu Kitamoto, Changming Zhou, and Mikio Takagi. "Similarity retrieval of noaa satellite imagery by graph matching", In Wayne Niblack, editor, *Storage and Retrieval for Image and Video Databases*, pages 60–73, Bellingham, Washington, February 1993. IS&T/SPIE, SPIE.

[7] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin. "The QBIC project: Querying images by content using color, texture and shape", In Wayne Niblack, editor, *Storage and Retrieval for Image and Video Databases*, pages 173–187, Bellingham, Washington, February 1993. IS&T/SPIE, SPIE.

[8] A. Pentland, R. W. Picard, and S. Sclaroff. "Photobook: Tools for content-based manipulation of image databases", In Wayne Niblack and Ramesh C. Jain, editors, *Storage and Retrieval for Image and Video Databases II*, pages 34–47, Bellingham, Washington, February 1994. IS&T/SPIE, SPIE.

[9] Markus A. Stricker. "Bounds for the discrimination power of color indexing techniques", In Wayne Niblack and Ramesh C. Jain, editors, *Storage and Retrieval for Image and Video Databases II*, pages 15–24, Bellingham, Washington, February 1994. IS&T/SPIE, SPIE.

[10] Michael J. Swain. "Color indexing", *International Journal of Computer Vision*, 7(1):11–32, 1991.

[11] Michael J. Swain. "Interactive indexing into image databases", In Wayne Niblack, editor, *Storage and Retrieval for Image and Video Databases*, Bellingham, Washington, February 1993. IS&T/SPIE, SPIE.

[12] N. Yazdani, M. Ozsoyoglu, and G. Ozsoyoglu. "A framework for feature-based indexing for spatial databases", In James C. French and Hans Hinterberger, editors, *7th Int. Working Conference on Scientific and Statistical Database Management*. IEEE Computer Society Press, September 1994.