International Journal of Soft Computing 2 (2): 339-345, 2007

Distributed Algorithms for Image Data Base Classification and Retrieval Using Perceptual Features

[1]S. Baulkani and [2]L. Ganesan
[1]Department of CSE, Government College of Engineering, Tirunelveli-627 007, India
[2]Departmen of CSE, A.C.C.E.T. Karaikudi-630004, India

Abstract: The recent growth in the volume of image data being generated and used for a variety of applications have insisted the development of image databases. The richness of content and subjective interpretations of image has rendered text based queries inadequate. Content Based Image Retrieval (CBIR) is a new but widely adopted method for finding images from vast and unannotated image databases. The correctness of retrieval for CBIR depends on efficient and effective indexing and searching schemes. Subjective queries and retrieval demands enormous computation time due to large data sizes of images coupled with large and complex indices required for search. Networks of Workstations (NOW) are a cost effective way of providing the much needed computational power in such applications.This study presents a distributed scheme for the classification and retrieval of images in an image database using NOW system. It uses an initial classification and a heuristic for determining the average feature vectors and distance in that image classes. The results of classification, retrieval, speedup obtained and the correctness of the retrieval are presented. The results indicate the viability and effectiveness of the proposed scheme.

Key words: Image database, distributed algorithms, image classification, CBIR, performance measures.

INTRODUCTION

With the drastic increase in the generation, processing and use of image data in a variety of applications have necessitated the development of image database systems to manage the massive amounts of image data. For such image databases keyword based indexing is inadequate and content based indexing is essential. The reason to go for content based indexing are the inexact nature, subjective interpretations and difficulty in formulating exact keyword based queries. Also it demands for fast and accurate retrievals from user's perspective. Important factors required for achieving fast and accurate retrievals are derivation of prominent features to be used as indices during search, organization of these indices in a suitable data structure and exact measure of similarity corresponding to perceptual similarity. Indices which characterize the image data should be of low dimensionality. Such type of content based retrieval produces a list of objects which are similar to the given query.

Several schemes have been proposed for image retrieval using content based queries (Narasimhalu, 1995, Piamsa and Alexandridis, 1999; Kim and Jung, 2004). Classification of images into groups of similar images facilitates the improvement in speed and accuracy of content based retrievals. Adopting manual classification for such a voluminous database is a difficult task. Automated classification is a suitable alternate. However it places enormous demands on the computation and disk access.

NOW is one of the cost effective way of providing much needed computational power and exploiting the available under utilized resources to meet such high computational needs. Software such as PVM (Parallel Virtual Machine), MPI (Message Passing Interface) provide the user with a unified view of single machine. Generally process and data migration is needed to distribute the load on the machines in a manner transparent to the user.

In this study, we present a distributed scheme for automated classification of images on NOW system based on clustering algorithms (Hartigan, 1975; Jain and Dubes, 1988). The distributed search and retrieval algorithms used in this research are explained. The experimental results are discussed in this study.

SYSTEM MODEL AND ASSUMPTIONS

The distributed systems consist of multiple CPU s with their own way of interconnection and communication (Anderew and Maartun, 2003). In this study we used a

Corresponding Author: S. Baulkani, Departmen of CSE, Government College of Engineering, Tirunelveli-627 007, India

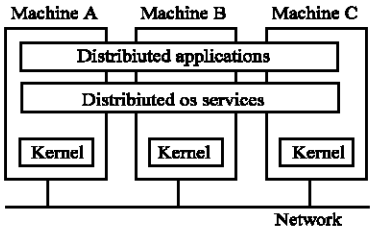

Fig. 1: System architecture of multicomputer operating system

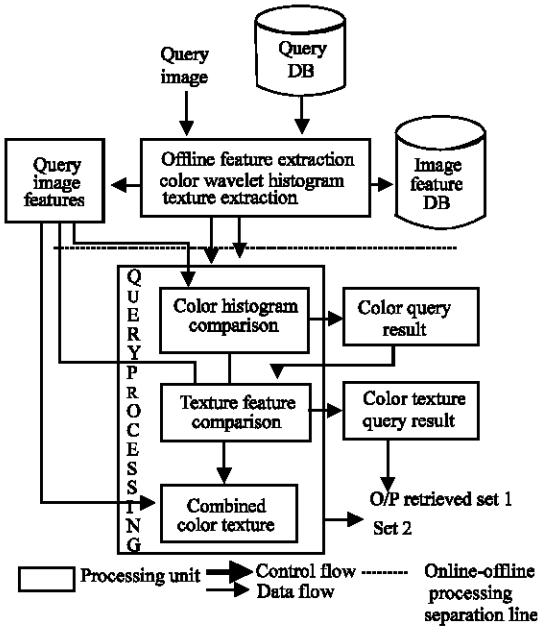

Fig. 2: Architecture of the two-level query feature comparison and retrieval framework

collection of multi computers interconnected through a fast Ethernet and loaded with message passing APIs. The general structure of the model used in our research is given in Fig. 1.

Each node has its own kernel containing modules for managing local resources such as memory, the local CPU, a local disk. Also each has a separate module for handling interprocessor communication that is, sending and receiving messages to and from ether nodes. Above each kernel is a common layer of software that implements the operating system as a virtual machine supporting parallel and concurrent execution of various tasks.

In this model one machine is treated as a master and all other systems are used as slaves. MPI is loaded in all the machines for communication between the machines.

This above proposed architecture is effectively used for our experimentation. To classify such a large database a set of features is computed from the images. Wavelet based statistical features are used as indices during classification and search operations. Images are classified using Haar Wavelet (Liang and Chen, 2004) and the resultant average component is further used to retrieve prominent Haralick features (Chen, 1979). Haralick had proposed 14 statistical features (Fig. 2). Out of which 4 different features are considered in this paper, namely, contrast, entropy, Angular Second Moment (ASM), homogeneity.

This study describes the distributed algorithms for classification and retrieval with few assumptions.

Assumptions:

- One of the machine in the NOW, designated as the console, has all the image data files.
- The data is analyzed and the feature vector is constructed for all the images before classification starts.
- The images are partitioned into clusters using clustering algorithms (Hartigan, 1975; Jain and Dubes, 1988). Then distributed to the machines before the search starts.
- Access of the disk of machine i by j (#i) if necessary, will not generally conflict with access to machine i.
- There is a shared array named balance of size P; balance[i] contains the number of images which remain to be compared to the query at machine i.

DISTRIBUTED CLASSIFICATION AND RETRIEVAL ALGORITHMS

It is assumed that all images have been analyzed and the feature vector has been built. Then the images are partitioned into P (approximately) equal parts and each partition is sent to a machine. The distance between the two images refers to the Euclidean distance between their corresponding feature vector parameters. The steps of the algorithm are described below: Where each of the steps is done by every machine concurrently.

- Apply the Pair wise Nearest Neighbor (PNN) clustering technique to get an initial clustering of the images in the partition at each machine.
- Compute the average value for each of the initial classes. This value is equi-distance from all other images in the class.
- Send this average value at each machine to all others.
- Recluster the images in each machine based on the average value at each machine.ie, each image is compared with all the average values and clustered with the one, with the least distance. Thus, at the end of the step, each machine will have P classes.

- Recompute the average value at each machine and send the new set of average value to all the machines.
- Determine the new average value of the classes which are the averages of the corresponding classes in all the machines.
- Check for convergence when the number of re-classification i.e., the amount of changes in the class size is less than a threshold, then the procedure is considered as having converged. If there is a convergence goes to the next step, otherwise go to step 4.
- Do a final merge and split of the classes across all machines. When the difference between the two average values of the classes is below a threshold the two classes can be combined together into a single cluster. On the other hand, when a cluster is too sparse, then the cluster is split into two classes further.

The PNN clustering technique (Equitz, 1989) is used to derive the initial classes in each machine. The algorithm proceeds iteratively by finding a pair of images which are closest and merging them and determining a new average feature which represents the class. This process is repeated until the number of classes reaches a desired size.

In PNN clustering the Euclidean distance measure is used while merging classes. In our implementations the distance between the two class feature vectors are computed using Eq. 1

$$d(i.j) = \sqrt{\left|x_{i1} - x_{i1}\right|^2 + \left|x_{i2} - x_{j2}\right|^2 + \ldots + \left|x_{ip} - x_{ip}\right|^2} \quad (1)$$

Retrieval algorithm: Feature Vectors are organized as pattern matrices and are used as indices in the searches. The images are classified into classes of similar images using the clustering algorithm (Hartigan, 1975; Jain and Dubes, 1988) and the image classes are distributed to the machines. The query is then broadcast to all machines. The average value and the distances are computed for the image classes at each machine. All the machines then carry out the search in the selected classes and send the result to the console. The sequence of actions at the console and the other machine in the NOW are shown in Fig. 3.

Algorithmic steps for query processing are given below.

DIST1RIMAGE RET (in: I, Q; Out; R) I: set of images, Q: Query, R: Retrieval results

- **Set up :** Basic system initialization

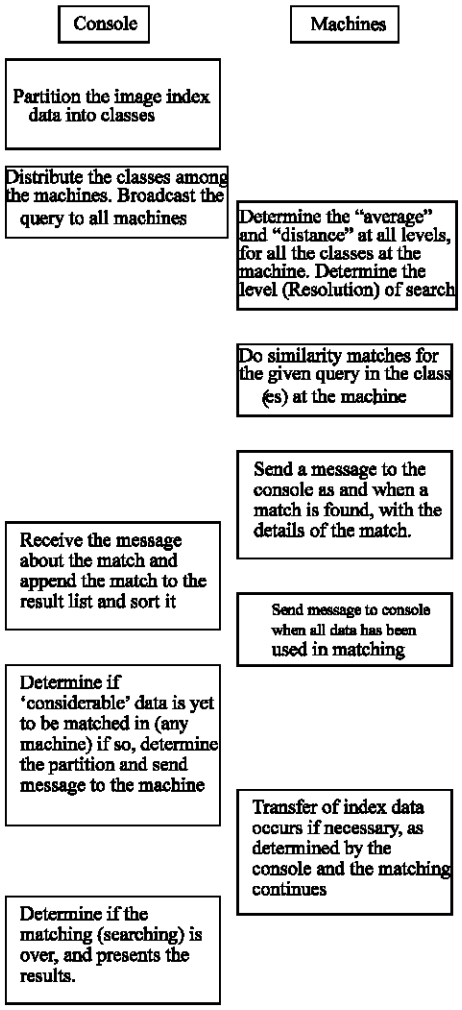

Fig. 3: Sequence of actions at the console and the other machines

Distrimage classify: This scheme automatically classifies the images using NOW as described in this study.

Districlass: This distributes the classes to all machines.

Broadcastquery: Broadcasts the query to all the machines.

- Do concurrently Each machine P does (i) Determine average feature (IP,Avg) (ii) Determine distance (IP,Avg,R) (iii) Select classes(Avg,R,Q)
- Do concurrently console does: (i) Respond to messages from other machines (ii) If balance (iii) = True, For all i then R_best â results, exit End if

Machines:

i) Image Search

7. Until all machines are done.

RESULTS AND DISCUSSION

The proposed algorithms have been tested to classify images in a database of about 5000 images. The database is built using (Brodatz, 1966).The NOW system used for the experiments consisted of five PIV@ 1.7 GHz machines. The machines are connected by a 100Mb/sec Ethernet and running MPI. Four sample classes obtained using the proposed algorithms with minimal set are shown in Fig. 4.

The classification time for different number of images and using different number of machines are shown in Table 1 and Fig. 5a and speedup of the proposed algorithm to a sequential scheme in 5b and c also shown in Table 2 and 3. It is easily seen that the speedup is linear in the number of active machine used in the classification process.

To evaluate the retrieval efficiency of the proposed system, we use the performance measure, *Recall* and *Precision*.

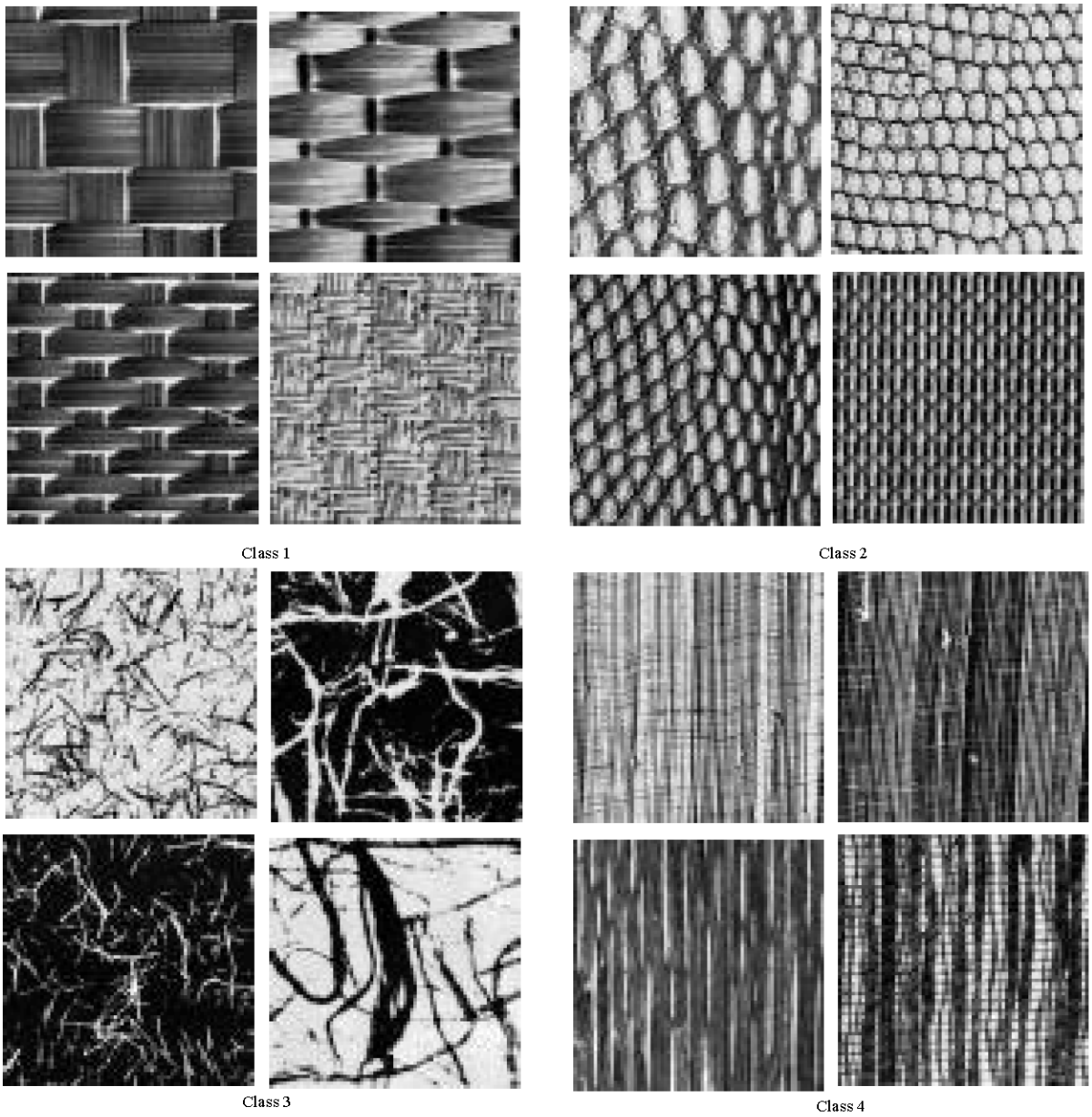

Fig. 4: Sample image classes

Table 1: Classification times

Number of images/ Time in seconds	100	200	300	400
P = 2	5	5	6	4
P = 3	7	8	10	18
P = 4	18	20	24	48
Sequential	22	31	42	75

Table 2: Speedups

Number of machines used in Search/Speedup	With classes	Without classes
1	1	1
2	1.4	1.8
3	2.2	2.8
4	2.6	3.5
5	3.6	4.2

Table 3: Speedups

Number of machines used in Search/Speedup	Computation time	I/O time	Total time
1	1	1	1
2	1.18	1.25	1.24
3	1.19	1.18	1.26
4	1.22	1.24	1.33
5	1.14	1.15	1.35

Table 4: Recall measures

Numberof clusters	H without DWT	H with DWT
5	0.35	0.568
6	0.46	0.515
7	0.37	0.487
8	0.35	0.431
9	0.316	0.431
10	0.305	0.427
11	0.282	0.395
12	0.321	0.376
13	0.316	0.386
14	0.302	0.353
15	0.333	0.349

Table 5: Precision measures

Number of clusters	H with DWT	Haralick without DWT
5	0.31	0.523
6	0.45	0.477
7	0.43	0.454
8	0.48	0.449
9	0.477	0.492
10	0.472	0.490
11	0.474	0.406
12	0.542	0.461
13	0.534	0.488
14	0.524	0.397
15	0.514	0.45

$$\text{Recall} = R_r/T \quad (2)$$

$$\text{Precision} = R_r/T_r \quad (3)$$

where R_r is the number of relevant retrieved images, T is the total number of relevant items in an image database and T_r is the number of all retrieved items.

The first six retrieval results for a given query is presented in Fig. 6. The retrieval operation is performed

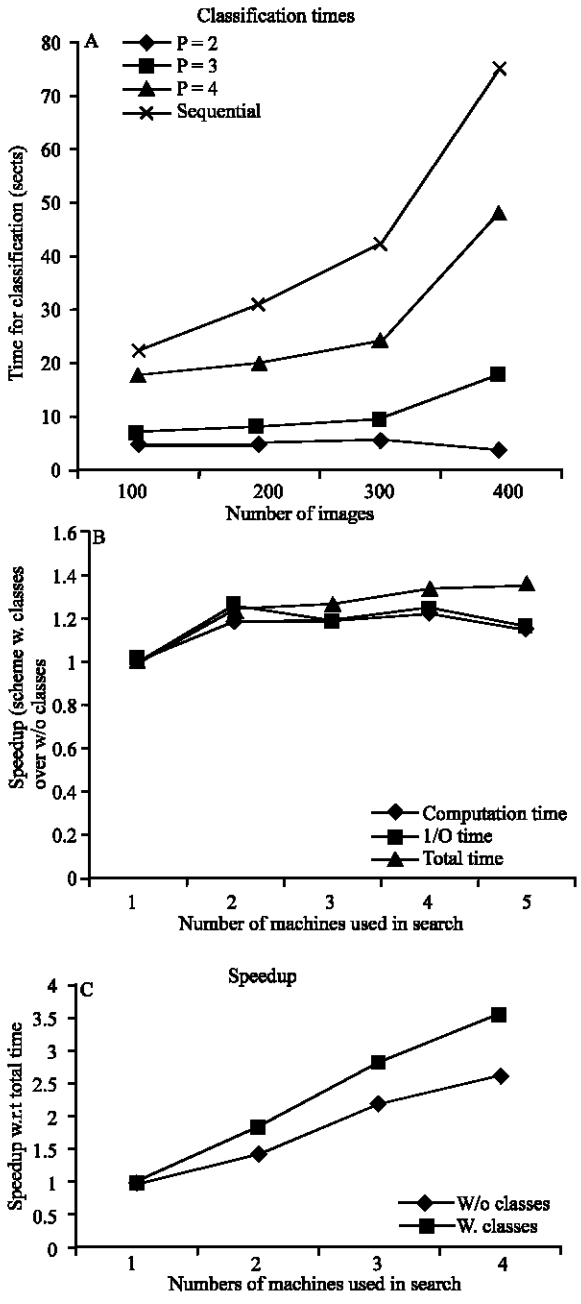

Fig. 5: a) Classification time b) and c) Speedups

in both with and without classification. Time taken for retrieval of the query image with classification is significantly less compared to the retrieval without classification.

The precision and recall calculations for a test set are plotted and the variations are shown in Fig. 7a and b also shown in Table 4 and 5 According to the classification techniques.

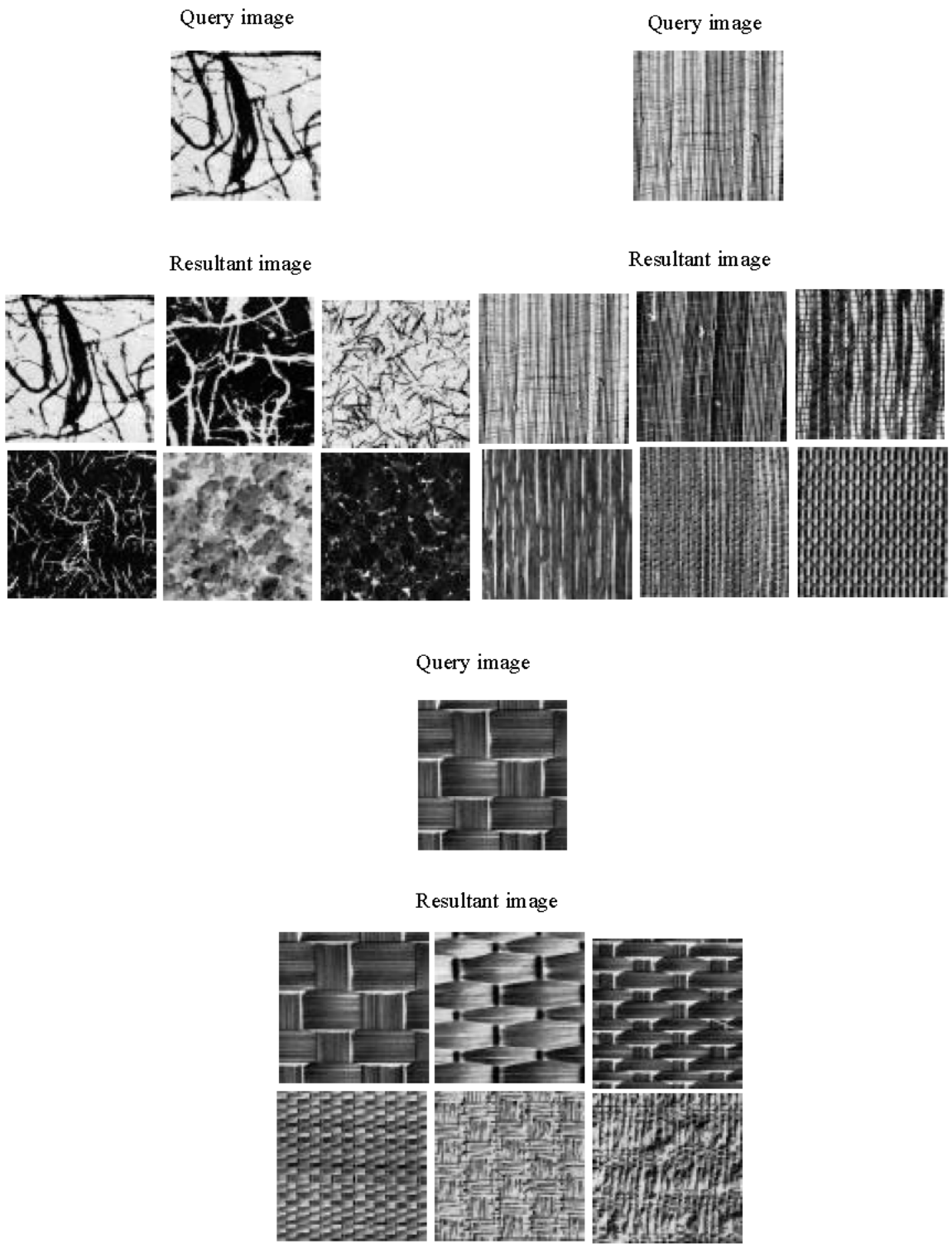

Fig. 6: Representative retrieval results

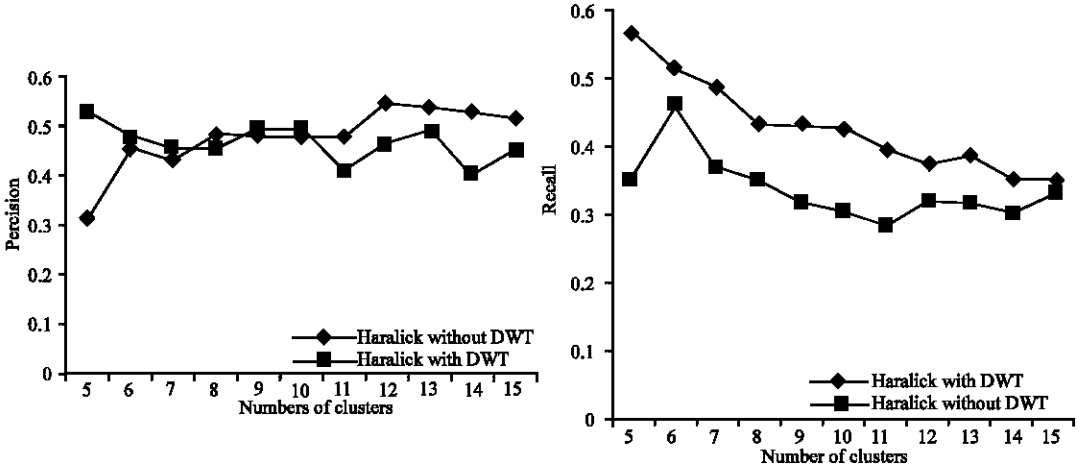

Fig. 7: a) Precision b) recall

CONCLUSION

High volume of images and the requirements for content based retrievals in image databases place enormous demands on storage and computation. NOW are a cost effective way of providing the much needed computation power in such applications. This paper presented distributed schemes for the classification and retrieval of images based on NOW system. The scheme used the feature vector based on wavelet and Haralick features. The average and distance values of feature vector are used to compare the same values of query feature in a class.

The speed up obtained is linear. This increases with the active machines used in the classification process. To validate the results of retrieval the performance measures such as precision and recall are calculated, the values are also promising.

This classification and retrieval operations can be further improved by considering the images as band filter responses, selection of most promising features according to the applications, minimizing the disk access and bus conflicts, application specific data sets (e.g., Medical Imaging, Geosciences processing, Military database, Remote sensing) and query retrieval with relevance feedback.

Authors are currently doing the classification and retrieval for large databases using color and other prominent features present in textured images.

REFERENCES

Andrew S. Tanenbaum, 2003. Maartun Van Steen, Distributed Systems Principles and Paradigms, Prentice Hall Of India Pvt Ltd.

Chung-wei Liang and Po-Yueh Chen, 2004, DWT based Text Localization, Int. J. Sci. Eng., pp: 105-116.

Equitz, W.H., 1989. A New Vector quantization Clustering Algorithm, IEEE. Trans. Acoust, Speech, Signal Proces., 37: 1568-1575.

Hartigan, 1975. Clustering Algorithms, Wiley.

Haralick, R.M., 1979. Statistical and structural approach to texture, In Proceedings IEEE., pp: 67.

Jain, A.K. and R.C. Dubes, 1988. lgorithms for Clustering Data Prentice Hall Of India Pvt Ltd.,

Jin-Ah Kim and Sung-Hwan Jung, 2004. Wavelet-based Texture Feature Extraction for Content-based Image Retrieval, MIPS Lab., Korea, 322.

Piamsa-nga, P.and N.A. Alexandridis, 1999. A Universal K-Tree Model for Content based Multimedia Retireval, International Journal of Computers and their applications Vol. 6.