# Retrieval of Multispectral Satellite Imagery on Cluster Architectures

T. Bretschneider[1] and O. Kao[2]

[1] School of Computer Engineering, Nanyang Technological University, Singapore
[2] Department of Computer Science, Paderborn University, Germany
okao@upb.de

**Abstract.** The retrieval of images in remote sensing databases is based on world-oriented information like the location of the scene, the utilised scanner, and the date of acquisition. However, these descriptions are not meaningful for many users who have a limited knowledge about remote sensing but nevertheless have to work with satellite imagery. Therefore a content-based dynamic retrieval technique using a cluster architecture to fulfil the resulting computational requirements is proposed. Initially the satellite images are distributed evenly over the available computing nodes and the retrieval operations are performed simultaneously. The dynamic strategy creates the need for a workload balancing before the sub-results are joined in a final ranking.

## 1 Introduction

The development and application of remote sensing platforms result in the production of huge amounts of image data. The obtained image data has to be systematically collected, registered, organised, and classified. Furthermore, adequate search procedures and methods to formulate queries have to be provided. The retrieval of images usually starts with a query based on world-oriented descriptions for the images, e.g. location, scanner, date. These characteristics allow a remote sensing professional to retrieve the required data with respect to the specific requirements determined by the application. However, the provided mechanisms are not suitable for users with limited expertise in remote sensing and the acquisition systems' characteristics, respectively. A possible scenario with the corresponding query, which is not supported by the conventional retrieval systems for satellite data, is the following ecological application: *A certain region reveals symptoms of increasing salinity and a satellite image of the area was purchased. It is of interest to find other regions which suffered from the same phenomenon and which were successfully recovered or preserved, respectively. Thus the applied strategies in these regions can help to develop effective counteractions for the specific case under investigation.*

This and other related examples [1] require a search by content rather than by related information, i.e. world-oriented features. Therefore a completely new understanding of the demands for remote sensing databases towards powerful retrieval methods is mandatory which led to the development of the *Retrieval System for Remotely Sensed Imagery* or short $(RS)^2I$.

B. Monien and R. Feldmann (Eds.): Euro-Par 2002, LNCS 2400, pp. 342–345.

## 2 Feature Extraction and Retrieval Strategy

The techniques for content-based retrieval in remote sensing databases are based on spatial content, spectral information, and a combination of both. The published results either require supervision of the feature extraction process or are limited to a specific type of satellite. Although the results might be satisfying for the utilised image sets, the retrieval result in an archive of data from a variety of scanners will lack in quality, i.e. in too many inadequate image answers.

This paper uses a new powerful feature extraction approach solely based on the spectral content by assigning each individual pixel a class membership accordingly to the multispectral radiance values. The successive feature extraction is based on the classes and uses a variety of different criteria for the description, like the class mean, class variance, spatial class neighbours, class compactness in the image etc. – in total the feature vector consists of more than 24 elements. For a detailed description of the proposed method refer to [1].

A satellite image is classificated, i.e. the features are extracted, when it is inserted in the database. Prior sub-division of the whole image is necessary before the processing due to the size of the data and the manifold classes in an entire scene which make computation burdensome. Unfortunately, every a-priori chosen sub-division technique for the images in the database restricts the flexibility of the system. As a conclusion this paper extends the query possibilities by a dynamic component, which has already proven its positive impact on the quality of the retrieval result in general image databases [3]. Instead of calculating the feature vectors solely a-priori they are generated at run-time accordingly to the requirements set by the query image. The combination of static and dynamic features reduces the needed time to obtain the retrieval result. Nevertheless, the dynamic processing is the bottleneck which can be overcome by utilising parallel architectures. Earlier investigations [3] already found that clusters are the most suitable architecture for the considered dynamic image database.

## 3 Parallel Execution of the Retrieval Operation

Dynamic retrieval of satellite imagery requires the analysis of all *image sections* in the remote sensing database and produces an enormous computational load since the requirements grow exponentially by using arbitrary sections instead of pre-defined tiles. Therefore, the utilisation of parallel architectures is necessary for the solution of the performance problem. The developed prototype of a parallel remote sensing database is based on a Beowulf cluster with:

- Master node controls the cluster, receives the query specification, and broadcasts the algorithms to the computing nodes. Furthermore, it unifies the intermediate results of the compute nodes and produces the final ranking.
- Computing nodes perform the image processing and comparisons. Each of these nodes contains a disjunctive subset of the existing images and executes all operations with the data stored on the local devices. The sub-results are sent to the master node.

For the initial distribution of the sections over the available nodes a content independent partitioning strategy is selected, such that the memory size of the sub-images stored on the local devices is approximately equal for all nodes. The resulting set of partitions $P = \{P_1, P_2, \ldots, P_m\}$ over the image set $\mathcal{B}$ for $m$ nodes has following characteristics:

$$\forall P_i, P_j \subset \mathcal{B} : P_i \cap P_j = \emptyset,\ size(P_i) \approx size(P_j), \quad i, j = 1, \ldots, m,\ i \neq j. \quad (1)$$

The advantages of this strategy are the simple implementation and management. All nodes have uniform processing times, if a query needs to analyse all images in the database and a homogenous, dedicated platform is assumed. However, queries combining world-oriented and dynamically extracted features distort the even image distribution and lead to varying node processing times. Therefore, workload balancing is a major issue in the $(\mathrm{RS})^2\mathrm{I}$.

Let $s$ denote the operation sequence performed with world-oriented features and $d$ operation sequence with dynamic feature extraction. The query $q(\mathcal{B}) = d \circ s(\mathcal{B})$ transforms $P_i$ into $\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$ with $A_i := s(P_i)$. The images from $A_i = \{b_{i1}, b_{i2}, \ldots, b_{in_i}\}$ are analysed with dynamically extracted features in the next step. As they do not fulfil necessarily the size condition defined in (1), the even image distribution over the cluster nodes is distorted. A migration of images within the cluster and thus workload balancing is required in order to equalise the processing times of all nodes. Each feature extraction algorithm $p$ implemented in the image database is assigned a measure describing the processing time $t_p$ as a function of the number of pixels. This is usually calculated experimentally and stored in the database. With this information the processing time per image $t_p(b_{ij})$ as well as the system response time $t_r$, minimal processing time $t_{min}$ and the optimal processing time $t_{opt}$ can be estimated [5]. After $t_{min}$ at least one node idles and an image re-distribution is necessary in order to avoid idling compute resources. Note that storing images permanently on the new assigned processing node if the requirements of the succeeding are invariable, e.g. always urban related queries are submitted, can reduce the necessity for a re-distribution. However, experiments under read-world conditions did show that in general a partitioning like given by Equation (1) is favourable.

It can be proven that this problem is NP-complete [4], therefore the heuristic LTF (*Largest Task First*) strategy [5] was developed. Large images with long compute times are mainly processed on the local node and thus the transfer effort is reduced. The processing of a query is divided into three stages. After processing the images on the local storage devices from $0, \ldots, t_{min}$, a temporal transfer of images from overloaded nodes to the other nodes according to the computed schedule follows. Finally, the rest of the images is analysed. The LTF strategy has with $O(n \log n)$ a low computational complexity. Disadvantages result mainly from the image communication between all nodes at nearly the same time, as a network overload is caused [5].

The performance measurements are executed using the implemented prototype of the parallel remote sensing database, which consists of six SMP-nodes (Dual Pentium III, 667 MHz, 256 Mbyte memory) and a 100 Mbit FastEthernet. The retrieval method was applied to multispectral scenes of the satellites Land-

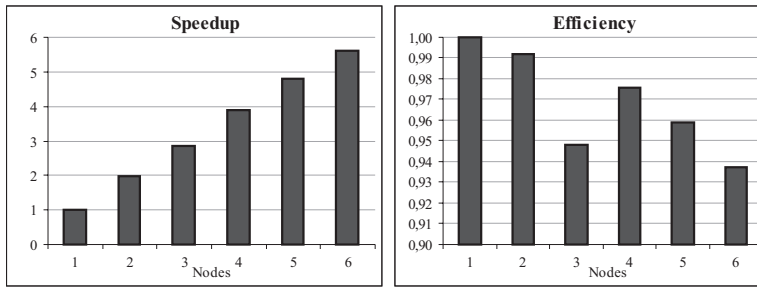


**Fig. 1.** Speedup and efficiency values for the parallel retrieval using the $(RS)^2I$

sat, SPOT, MOMS-02, and IKONOS which add up to an image archive covering an area of more than 204,150 $km^2$ in the 1–30m spatial resolution range. The images show a mixture of urban, agriculture, and forestry areas around the world. An in-depth analysis of the retrieval results can be found in [1] and is omitted here due to the limited space. Figure 1 shows, that a linear speedup with the number of nodes is achieved. Due to the minimised communication between the cluster nodes during query processing the efficiency values remain in a narrow interval between 94% and 98%.

## 4 Conclusions

In this paper an in-sight in the parallel aspects of the $(RS)^2I$ was provided. The system enables dynamic retrieval of remotely sensed data in an image database and shows that the utilisation of a content-based approach enables the retrieval of data even for persons without in-depth remote sensing background knowledge. The developed cluster-based architecture satisfies the computational demands and enables the use of the proposed system in real world applications.

## References

1. T. Bretschneider, R. Cavet, and O. Kao, "Retrieval of remotely sensed imagery using spectral information content", Geoscience and Remote Sensing Symposium, to be published 2002.
2. T. Bretschneider, O. Kao, "Indexing strategies for content-based retrieval in satellite image databases", Conference on Imaging Science, to be published 2002.
3. S. Geisler, O. Kao, T. Bretschneider, "Analysis of cluster topologies for workload balancing strategies in image databases", Conference on Parallel and Distributed Processing and Applications, pp. 874–880, 2001.
4. R. Karp, "Reducibility among combinatorial problems", in: Complexity of Computer Computations, Plenum Press, pp. 85–104, 1972.
5. O. Kao, G. Steinert, F. Drews, "Scheduling aspects for image retrieval in cluster-based image databases", IEEE/ACM Symposium on Cluster Computing, pp. 329–336, 2001.