Cloud Identification Using Genetic Algorithms and Massively Parallel Computation

Bill P. Buckles and Frederick E. Petry
Center for Intelligent and Knowledge-based Systems
Department of Computer Science
Tulane University
New Orleans, Louisiana 70118

Phone: (504)862-3373
Fax: (504)862-8747
email: {buckles, petry}@cs.tulane.edu

Abstract

As a Guest Computational Investigator under the NASA administered component of the High Performance Computing and Communication Program, we implemented a massively parallel genetic algorithm on the MasPar SIMD computer. Experiments were conducted using Earth Science data in the domains of meteorology and oceanography. Results obtained in these domains are competitive with, and in most cases better than, similar problems solved using other methods.

In the meteorological domain, we chose to identify clouds using AVHRR spectral data. Four cloud speciations were used although most researchers settle for three. Results were remarkedly consistent across all tests (91% accuracy). Refinements of this method may lead to more timely and complete information for Global Circulation Models (GCMs) that are prevalent in weather forecasting and global environment studies.

In the oceanographic domain, we chose to identify ocean currents from a spectrometer having similar characteristics to AVHRR. Here the results were mixed (60% to 80% accuracy). Given that one is willing to run the experiment several times (say 10), then it is acceptable to claim the higher accuracy rating. This problem has never been successfully automated. Therefore, these results are encouraging even though less impressive than the cloud experiment. Successful conclusion of an automated ocean current detection system would impact coastal fishing, naval tactics, and the study of microclimates.

Finally we contributed to the basic knowledge of GA behavior in parallel environments. We developed better knowledge of the use of subpopulations in the context of shared breeding pools and the migration of individuals. Rigorous experiments were conducted based on quantifiable performance criteria. While much of the work confirmed current wisdom, for the first time we were able to submit conclusive evidence.

The software developed under this grant was placed in the public domain. An extensive user's manual was written and distributed nationwide to scientists whose work might benefit from its availability. Several papers, including two journal articles, were produced.

1 Introduction

In the Fall of 1992, Tulane University was awarded a grant in the amount of $90,000 for research under a component of the High Performance Computing and Communications Program administered by NASA. The grant category was Guest Computational Investigator and the research activity was Earth and Space Sciences. The overall goal of the HPCC program is the development of software and related technologies that effectively bring to bear the most advanced computer systems in the Nation on the key scientific problems of our day. These problems are termed the "Grand Challenges."

Tulane chose to participate in the Earth Sciences arena by applying new technologies and new problem solving paradigms to remote sensing of the Earth's atmosphere. Within this sphere, we chose the problem of identifying and classifying clouds. A reliable solution to this problem has long-term implications in the collection of data associated with global environmental change. An immediate impact is the ability to quickly collect information needed in GCMs (Global Circulation Models) that are prevalent in the practice of meteorology.

2 Grant Objectives

The original proposal from Tulane stated the objectives of the research as:

> Selection of features as determined by the datasets employed
> and concurrent development of parallel cloud labelling GA will
> be the inititial objective. ... As the study proceeds, we will
> determine the identification granularity for which the method is
> best suited. ... The final phase will be determination of robust-
> ness by examining scenes with "difficult" backgrounds.

These objectives were to be accomplished using the MasPar and genetic algorithms.

3 Accomplishments

Using primarily AVHRR (Advanced Very High Resolution Radiometer) imagery of the Western U.S., significant achievements were obtained in accuracy and speed for cloud classification. The number of cloud speciation categories

used was larger than that prevalent in the literature. Classification was reliable in that approximately the same percentage of accuracy was achieved in each experiment on each image. The MasPar system's capacity was fully utilized in that the algorithmic approach is compute bound and the problem was easily scaled to utilize all processors.

As a cross-check, a secondary experiment was conducted on images depicting mesoscale oceanic features. Here the results were slightly below the accuracy obtained for clouds but still competitive with the best results found in the literature. The speed with which solutions were obtained remained far superior to present techniques.

Our final contribution was a to collect and publish statistics on the algorithm employed. Never before has there been an opportunity to study the behaviour of genetic algorithms in a parallel environment over an extended time period. While much of the information concurred with conventional wisdom, we made available for the first time hard experimental evidence. There were details regarding behavior that had not heretofore been observed.

3.1   Computation Results

The first set of experiments, and the primary focus of the work, was identification of clouds from AVHRR data. Most experiments in the literature utilize a three-level (plus clear) classification. We opted for four:

- Convective Clouds – cumulus and cumulonimbus

- Low Level Clouds – fog, stratus, and stratocumulus

- Middle Level Clouds – altostratus and altocumulus

- High Level Clouds – cirus, cirrostratus, and cirrocumulus

An immediate technical problem was the lack of ground truth data for cloud images. We solved this by obtaining from NRL at Stennis Space Flight Center a workbook/tutorial [2] on sight recognition of cloud types from images. From this, we created our own truth data.

The images employed were similar to that seen on the following page. A rough quadtree region segmentation is first performed. The regions need not be precise because we use only a single pixel from approximately the region's centroid to perform the classification. A schematic is shown in Fig.

2
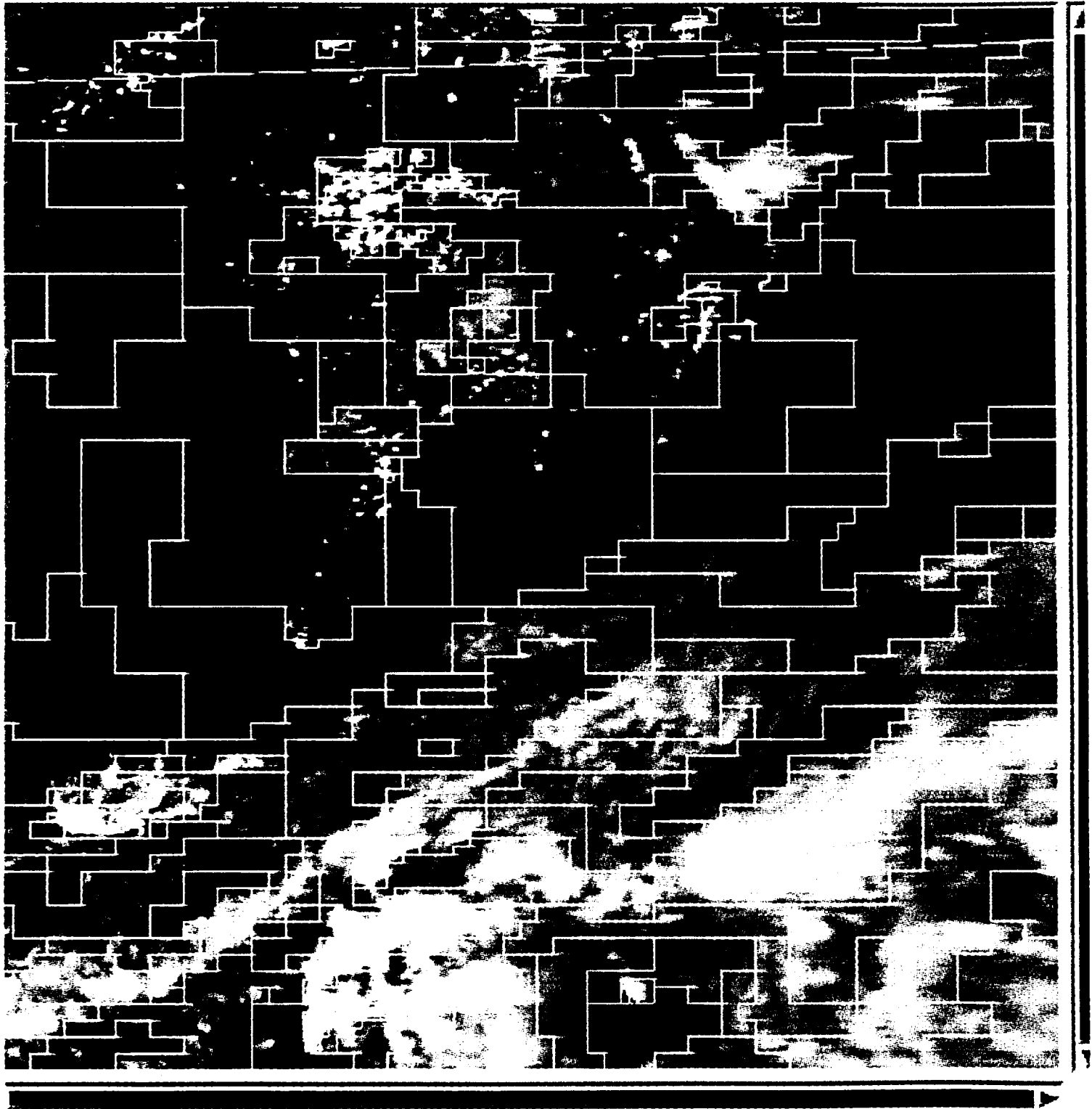
1. Pixels from each region (i.e., their signatures) are given to the genetic algorithm. An individual solution is represented by a string of labels, one for each pixel. Solutions, or *chromosomes*, are collected into small (e.g., 15) subpopulations, one at each processor. Within subpopulations, *recombination* occurs. This involves *selection* of parents, intermixing of the chromosomal material (*crossover*) producing offspring which, after possible *mutation* are grouped to form a new subpopulation at that grid. Subpopulations interact by *migration* – the transfer of individuals, actually a copy of the individuals, from one subpopulation to another.
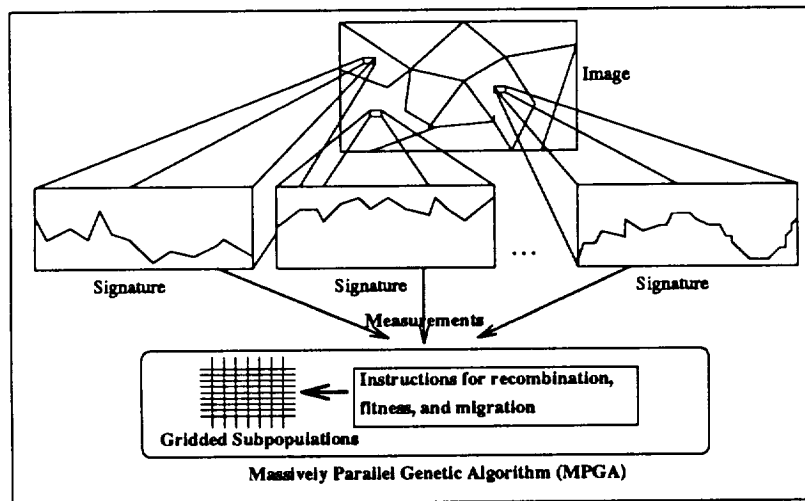

Figure 1: Overview of Massively Parallel Genetic Algorithm

The fitness of individuals is determined by its consistency with prototypical properties of a correct labeling. Prototypical properties are of the nature "channel 4 brightness of nimbocumulus is less than channel 4 brightness of low level clouds." Approximately 30 prototypical properties are assembled in a semantic net structure. The fitness of a particular chromosome is determined by the degree to which it corresponds to the information in the semantic net.

Five images were partitioned to create 11 test cases. The results were remarkably uniform over each case as shown in Table 1. This more than favorably compares with a recently published article [1] for which the researchers used the same data sets and also attempted to detect four cloud

4

Table 1: Accuracy of GA-generated cloud labeling

Run #	Accuracy %	Run #	Accuracy %
1	91	7	91
2	91	8	91
3	91	9	91
4	91	10	91
5	91	11	91
6	91		

classes.

A second series of experiments over a different and more difficult domain was undertaken – that of detecting ocean currents using only the infrared spectrum. Fig. 2 is a representative, cloud-occluded view of the North Atlantic. The objective was to detect the boundaries of the Gulf Stream, warm eddies, and cold eddies.

A different sort of segmentation was employed – edge detection. The edge segmented image for Fig. 2 is shown in Fig. 3. Each line represents a boundary between warm and cold bodies of water. Determining Gulf Stream boundaries and eddies from this image is difficult even for human experts. The accuracy results are shown in Table 2. The range (60% to 80%) was typical to other images. Given several experiments (e.g., 10 as in Table 2) it is possible to determine which renders the best labelling without referring to the original image. Assuming that one is willing to run several consecutive experiments, it is fair to say that the average accuracy of the method is 80% on cloud-occluded images. The accuracy is in the mid-90s for images less obscured by clouds.

3.2 Software Developed and Distributed

A significant addition to the software complement for the MasPar was developed. A user's guide:

D. Prabhu, B.P. Buckles, and F.E. Petry, "MPGA (version 1.1): Users' Guide", Tulane Technical Report, April 1996, 37pp.

Figure 2: Original Infrared Image of the Gulf Stream

The manual acknowledges the support from NASA for the software development.

The software and user's guide is available by anonymous ftp from the site `ftp.cs.tulane.edu` under the directory `/pub/buckles/mpga`. This information has been widely distributed and copies of the manual have been sent directly to researchers whose areas of expertise are most closely correlated.

3.3 Publications

Two keystone publications emanated from this research:

> D. Prabhu, B.P. Buckles, and F.E. Petry, "Genetic Algorithms for Scene Interpretation", submitted to *IEEE Transactions on Systems, Man and Cybernetics*.

> D. Prabhu, B.P. Buckles, and F.E. Petry, "Behavior of Interconnected Subpopulations: Genetic Algorithms in a SIMD Environment", submitted to *Evolutionary Computation*.

6

Table 2: Accuracy of GA-generated oceanic labeling

Run #	Accuracy %	Run #	Accuracy %
1	80	6	63
2	57	7	71
3	66	8	77
4	83	9	71
5	83	10	69

In addition, there have been several minor publications in conferences and workshops. In each publication, there is an acknowledgment of the support from NASA.

4 Benefits

The original proposal from Tulane stated the long-range benefits as:

> The principal [long term] objective is to produce a radiative cloud model that is simple, efficient, and fuses satellite and ground station data. The purpose being to provide a component much abused in present GCMs. This is not a specific objective of [the three-year research plan] but the work proposed is a step in that direction.

This is now, in fact, a possibility.

There are additional benefits. Due to this research, there is now in the literature data that lends a greater understanding as to how genetic algorithms behave in a parallel environment. This might be likened to extending the understanding from micro-genetics to population genetics.

The advances in the use of genetic algorithms for image understanding in general is yet another result. In the near future, we plan to apply the knowledge gained from this work to analyzing hyperspectral images of forestry scenes from the NASA Lewis III satellite.
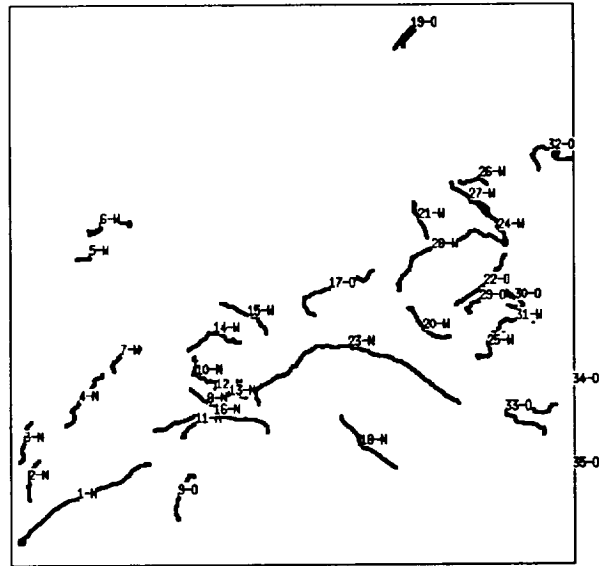
Figure 3: Best Labeling of the Gulf Stream found by the GA

Report Distribution

The following items are being distributed as attachments to this report:

- The dissertation by Deveraya Prabhu listed above

- The article listed above submitted to the *IEEE Trans. on Systems, Man and Cybernetics*

- The article listed above submitted to *Evolutionary Computation*

- The user's guide (also listed above) to the software developed

One set of documents are being sent to:
Mr. James R. Fischer, Code 934
Space Data & Computing Division
NASA/Goddard Space Flight Center
Greenbelt, MD 20771
in the following quantities:

Document	# Copies
Final Report	3
SMC Article	3
EC Article	3
Ph.D. Dissertation	1

A second set is being sent to:

NASA Scientific & Technical Information Facility
ATTN: Accessioning Department
800 Elkridge Landing Road
Linthicum Heights, MD 21090

in the quantity of:

Document	# Copies
Final Report	2
SMC Article	2
EC Article	2

One copy of each is unbound in order to permit miroreproduction.

References

[1] L. A. Bakst and N. N. Fedorova, "Investigation of Cloudiness by Using Multisprectral AVHRR Data from NOAA Satellite for Synoptic Analysis", *Earth Observation and Remote Sensing*, vol. 12 (1995) pp. 475–483.

[2] G. M. Pearson and G. Stogaitis, "Satellite Imagery in Synoptic and Mesoscale Meterorology – A Reference Manual", prepared by Enviroment Canada Atmospheric Environment Service.