Quakefinder: A Scalable Data Mining System for Detecting Earthquakes from Space

Paul Stolorz and **Christopher Dean**
Jet Propulsion Laboratory
California Institute of Technology
(pauls,ctdean)@aig.jpl.nasa.gov

Abstract

We present an application of novel massively parallel datamining techniques to highly precise inference of important physical processes from remote sensing imagery. Specifically, we have developed and applied a system, Quakefinder, that automatically detects and measures tectonic activity in the earth's crust by examination of satellite data. We have used Quakefinder to automatically map the direction and magnitude of ground displacements due to the 1992 Landers earthquake in Southern California, over a spatial region of several hundred square kilometers, at a resolution of 10 meters, to a (sub-pixel) precision of 1 meter. This is the first calculation that has ever been able to extract area-mapped information about 2D tectonic processes at this level of detail. We outline the architecture of the Quakefinder system, based upon a combination of techniques drawn from the fields of statistical inference, massively parallel computing and global optimization. We confirm the overall correctness of the procedure by comparison of our results with known locations of targeted faults obtained by careful and time-consuming field measurements. The system also performs knowledge discovery by indicating novel unexplained tectonic activity away from the primary faults that has never before been observed. We conclude by discussing the future potential of this data mining system in the broad context of studying subtle spatio-temporal processes within massive image streams.

Introduction

Automatic detection and cataloguing of important temporal processes in massive datasets is a problem of overwhelming scale that has so far eluded automation for the great majority of scientific problems. Careful manual inspection of images by highly-trained scientists is still the standard method of extracting important scientific information from even high-resolution images. This is a time-consuming and extremely expensive process, although there has recently been substantial progress in isolated domains (Fayyad, Weir, & Djorgovski 1993). The goal of this paper is to introduce a data mining system that tackles this problem in the context of analyzing the earth's crustal dynamics, by enabling the automatic detection and measurement of earthquake faults from satellite imagery.

The system, Quakefinder, is applied here to the analysis of data collected by the French SPOT satellite. SPOT is a push-broom detector that collects panchromatic data at 10 meter resolution from a satellite in sun-synchronous orbit around the earth. In our application, images of size 2050×2050 pixels are analyzed to detect fault motion to subpixel precision. Although applied initially to the specific problem of earthquake fault analysis, the principals used by Quakefinder are however broadly applicable to a far more general class of calculations involving subtle change-detection in high-resolution image streams. We believe that its success in this domain points the way to a number of such datamining calculations that can directly and automatically measure important temporal processes to high precision from massive datasets. These include problems involving global climate change and natural hazard monitoring, as well as general image understanding problems involving target detection and identification in noisy image streams.

The design and implementation of Quakefinder has been driven by the need to simultaneously address three distinct problems, all of which must be solved in order to enable geophysical analysis of temporal processes from satellite data at the accuracy desired. They are 1) design of a statistical inference engine that can reliably infer the fundamental processes to acceptable precision, 2) development and implementation of scalable algorithms suitable for massive datasets, and 3) construction of an automatic and reasonably seamless system that can be used by domain scientists on a large number of datasets.

The first problem is the design of an inference engine that can infer tectonic ground motion to sub-pixel precision based upon careful comparion of consecutive images. To appreciate the technical challenge involved here, consider briefly the nature of the datasets available in this study, namely SPOT satellite images of the Landers region, taken in June 1991 and June 1992, bracketing the Landers earthquake event. The images consist of panchromatic data taken at 10-meter reso-

lution. Now the ground displacements due to the Landers quake varied in magnitude anywhere up to 7 meters. These ground element motions constitute the fundamental "signal" that we want to mine from the data. Since all the motions are of sub-pixel magnitude, naive pixel-by-pixel change detection methods will fail to extract the events correctly. A careful machine learning technique is required to ensure that reasonable inferences can be drawn about ground displacement to subpixel precision.

A distinct, though related, problem stems from the need to map ground motions at or near single pixel resolution. To be perfectly clear, note the distinction here between the term "resolution", denoting the size of ground element associated with a piece of information, and "precision", denoting the accuracy to which this information (e.g. motion) is known. The relatively sophisticated inference methods needed to infer the motion of a single pixel must be repeated over every pixel of an entire image. For even a modest size image containing 2050 × 2050 pixels, this task represents a huge number of operations. Current workstation technology cannot support these CPU demands, which has led us to implement Quakefinder on massively parallel computing platforms, specifically the 256-node Cray T3D at JPL. The issues of scalable algorithm development and their implementation on scalable platforms that were confronted here are in fact quite general, and are likely to impact the great majority of future datamining efforts geared to the analysis of genuinely massive datasets. They have been investigated previously in the context of data mining and knowledge discovery in (Stolorz et al. 1995).

The third problem is the design and construction of a system architecture that can perform the various computational tasks required automatically. For problems of the scale tackled here, the various components must be linked relatively seamlessly to ensure the rapid delivery of a useful scientific product to geologists. This issue is already important in the context of ground-based analysis, It takes on even greater prominence when considered from the point of view of possible autonomous satellite applications and operations being envisaged by NASA and other agencies and corporations.

The goals of the paper are to describe the basic machine learning techniques used by Quakefinder, to outline the decomposition of these methods onto massively parallel platforms, and to describe their implementation within an overall system usable by domain scientists. We stress the interplay of all three of these components, and argue that all three are essential ingredients for any truly successful datamining system in this domain. We then discuss the results obtained by applying Quakefinder to analysis of the Landers earthquake, outline the overall successes and limitations of the technique, and present future directions that can build upon the work presented here.

Inference Engine

Basic algorithm

The purpose of the basic algorithm is to detect small systematic differences between a pair of images, which we'll call the "before" image I_1 and the "after" image I_2 respectively. This is accomplished at sub-pixel resolution by the following method, dubbed "imageodesy" by its author (Crippen 1992):

1. Match the before and after images by eye as well as possible (i.e. determine the best offsets between the two images in the horizontal and vertical directions).
2. Break the before image up into many non-overlapping templates, each consisting of, let's say, 100 x 100 pixels.
3. For each template, measure the correlation between the before template and the after template at the original position determined in step 1), and at the 24 nearest offset positions.
4. Determine the best template offset Δ_y, Δ_x from the maximum correlation value found in 3).
5. Repeat steps 3) and 4) at successively higher resolution, using bilinear interpolation, or some other interpolation scheme, to generate new templates offset by half a pixel in each direction.

The algorithm relies heavily on the use of sub-pixel image registration for its power. A schematic overview is shown in Figure 1.

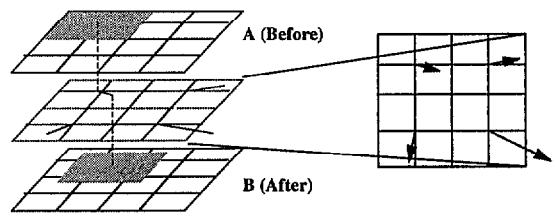

Figure 1: Schematic illustration of the use of subpixel registration to infer displacement maps between image pairs.

This basic idea is a very useful one that has been successfully applied over the years in a number of fields, especially in the context of image enhancement from undersampled image sequences (Crippen 1992; Jacquemod, Odet, & Goutte 1992; Ur & Gross 1992; Wen-Yu-Su & Kim 1994; Kim, Bose, & Valenzuela 1990). Typically, it has been used to automatically account for global effects relating successive images of the same "scene" in an image stream, namely transformations such as translation, rotation and scale changes. We apply the concept here with a highly unusual twist, in that many independent local sub-pixel registrations are performed to uncover the signal of interest, rather than a single global registration.

Figure 2: Architecture of the Quakefinder detection and measurement system

vector field of inferred ground motions from a pair of satellite images. The vector field is then passed to a geometric correction module which automatically corrects for spacecraft artifacts. Upon correction, the resulting displacement map is inspected by geologists for evidence of tectonic activity, with faults being mapped and measured. This information is fed in turn into a further adaptive learning component, described above, in order to refine the fault location and magnitude. This iterative procedure is terminated when sufficient accuracy is obtained. The resulting fault outlines are then registered in a catalog as important events.

Geometric corrections

A correction module is necessary because the original computed displacement field does not in general represent directly the ground motion itself. There are typically systematic effects due to slight differences in spacecraft trajectories. They consist primarily of differences in the yaw, pitch and roll of the spacecraft for each trajectory, as well as small differences in orbital height above the target region. These artifacts can be removed straightforwardly, since they appear as well-parametrized global corrections across the entire displacement vector field. A simplex-based optimization package was used to compute the optimal set of parameters characterizing these corrections. This post-processing component has trivial time complexity compared to the displacement field computation itself, and can easily be implemented automatically on workstations instead of a parallel machine.

There are second order corrections which must in principal be accounted for as well. For example, sun angle and view angle differences can introduce spurious differences between scenes. For the Landers data considered here, these problems are negligible since 1) the images were collected almost exactly one year apart, 2) the SPOT satellite is in sun-synchronous orbit, so time-of-day matches are not a problem, and 3) very similar orbital trajectories were achieved. Radiometric differences are also possible due to vegetation a growth and other events. These are in fact often interesting processes in themselves, although they interfere with the specific task of measuring fault motion. For the desert datasets used here, they are not a major factor in any case. Yaw, pitch and roll can also all vary during a spacecraft overflight of a selected target. Again, this turns out to be a negligible effect for our problem.

Scalable Decomposition on Massively Parallel Processors

The core computation in our approach is that of ground motion inference for a given image pixel. On machines of the workstation class, several hundred such ground motion vectors can be calculated in a matter of days. However, when area-based maps of ground motion are required, workstations are no longer sufficient. For example, for even relatively small images of 2000×2000 pixels, 4 million vectors must now be computed, raising the computational demands by 4 orders of magnitude. The resulting calculation is not accessible to workstations in any reasonable time frame. We have therefore chosen to implement Quakefinder on a 256-node Cray T3D at JPL. The T3D is a massively parallel distributed memory architecture consisting of 256 computing nodes, each based on a DEC Alpha processor running at 150MHz. The nodes are arranged as 3-dimensional tori, allowing each node to communicate

directly with up to 6 neighboring nodes of the machine.

MIMD parallel architectures such as the T3D turn out to be ideal architectures on which to implement many image analysis algorithms, including Quakefinder. The best decomposition consists of simply assigning different spatial portions of each image to different nodes of the machine. The vast majority of the calculation can then proceed independently on each node, with very limited communication. This results in a highly efficient parallel algorithm.

A small amount of communication does of course need to be performed periodically. The storage required for some templates will occasionally overlap boundaries between nodes. A standard technique in parallel decomposition handles this problem by tolerating a small amount of memory redundancy. Each node is simply assigned a slightly larger area of the image at the start of the calculation to allow for these overlaps. The overhead required is very small. Thereafter the calculation can proceed entirely in parallel.

Results for the Landers earthquake

We have obtained the following results from applying Quakefinder to SPOT data bracketing the Landers earthquake of June 22, 1992. The images are 2050×2050 pixels in size covering a 400 square kilometer region of the Southern California desert near the town of Landers. The differences between the two images are extremely subtle and are essentially impossible to detect by eye. Ground motion directions calculated for the Landers quake of June 22, 1992 are shown in Figure 3, superimposed on the 1991 panchromatic SPOT image. The major grey-scale discontinuity along the main diagonal of the map. This is the position of the fault break inferred automatically by Quakefinder with no supervised scientific input, based purely on the two raw before and after SPOT images. The black line is ground truth, the known fault location.

The black line represents ground truth superimposed on the computed image to assess its accuracy. It has been determined by extensive field analysis. Note that the major hue discontinuity corresponds very well to the true fault position, including the bends and steps separating the Emerson fault from the Homestead Valley fault. The general motions are right-lateral, as expected. The motion along the SW block appears to have a north to west trend change as the fault trend itself changes from northerly to more westerly. Thus, the motion tends to parallel the fault, as expected. These observations confirm the value of our approach as an efficient method for automatically detecting and measuring the position of known faults.

The area-mapped nature of the products generated by Quakefinder offer an even more interesting capability, namely discovery of entirely new behavior. In the Landers case, it yields suggestive evidence associated with the NE block of the image. This block seems to have two sub-blocks, with relative left-lateral motions between them, suggesting a surface or perhaps sub-surface fault conjugate to the main break. Efforts are currently underway to refine this prediction and to confirm it via field studies. Note that alternative technological approaches cannot easily supply this type of knowledge, if at all. Interferometric SAR can measure small ground displacements in one dimension, along a line perpendicular to the spacecraft trajectory, but cannot supply a full 2D map of motions. Movable seismic detectors located by GPS technology can measure full 3D motion quite precisely, but only at a limited number of individual locations. For these reasons, much of the information displayed in Figure 3 has never before been obtained.

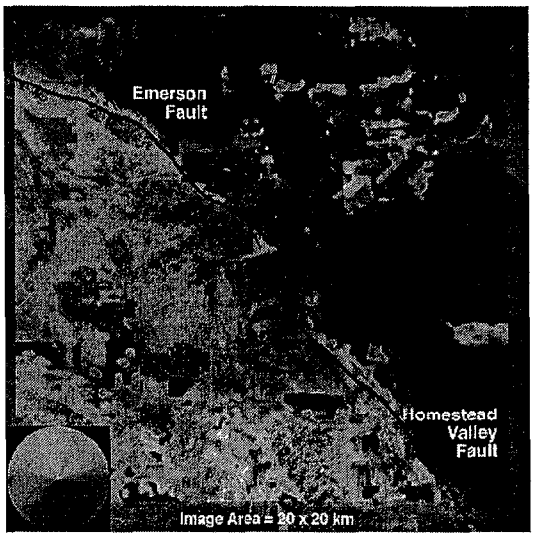

Figure 3: Preliminary results for displacement map of the 1992 Landers earthquake. A grey-scale wheel shown at lower left encodes the direction of motion of each ground element due to the earthquake, inferred by comparison of two SPOT images. The displacement map is superimposed upon the first (1991) SPOT image to show the main topographic features. Note the correspondence of the direction discontinuities to right-lateral main fault breaks (black lines). Blocks northeast of the main breaks indicate left-lateral warpage above a buried conjugate fault, consistent with the seismic pattern. These results give us high confidence in the accuracy of the method as a general area-based mapping technique for seismic activity. A full color image of this map can be found at http://www-aig.jpl.nasa.gov/mls/quakefinder

Why did Quakefinder work?

A number of ingredients contributed to the success of Quakefinder as a data mining tool. To begin with, it was based upon an integrated combination of techniques drawn from statistical inference, massively parallel computing and global optimization. Secondly, sci-

entists were able to provide a concise description of the fundamental signal recovery problem. Thirdly, resulting tasks based upon statistical inference were straightforward to automate and parallelize on scalable platforms, while still ensuring accuracy. The issues of scalable algorithm development and their implementation on scalable platforms that were addressed here are in fact quite general, and are likely to impact the great majority of future datamining efforts Finally, the relatively small portions of the overall task that were not so easily automated, such as careful measurement of fault location based on a computer-generated displacement map, are accomplished very quickly and accurately by humans in an interactive environment and did not pose an enormous bottleneck for the system.

Conclusions and Future Directions

A number of future data mining investigations are suggested by this work. One obvious one is its extension to the continuous domain, measuring very slowly-varying processes instead of abrupt events. This will require the systematic incorporation of scalable I/O resources to allow the rapid ingestion and processing of continuous image streams. The generality of the basic approach indicates that it will also prove scalable as detector and satellite resolutions improve. For example, plans are now underway for the development and deployment of satellites with 1 meter resolution or better. Extensions of Quakefinder will enable physical processes on the scale of centimeters to be straightforwardly detected and measured automatically, opening new avenues of geophysical analysis from satellite images.

The success of the fundamental technique used here has also created another problem, namely the need to register and catalogue the resulting scientific events systematically, rather than simply scattering them amongst flat files. The point here is that the events inferred by Quakefinder can be used as content-based indices to exploration of related remote-sensing datasets, such as Landsat, Synthetic Aperture Radar, and hyperspectral data. The Conquest/Oasis project begin undertaken as a collaboration between JPL and UCLA is an example of a distributed querying and analysis environment that can potentially exploit spatio-temporal events of the type inferred here (Stolorz et al. 1995).

The Quakefinder system addressed a definite scientific need, as there was previously no area-mapped information about 2D tectonic processes available at this level of detail. In addition to automatically measuring known faults, the system also performed knowledge discovery by indicating novel unexplained tectonic activity away from the primary faults that has never before been observed. It shows dramatically the power of a datamining engine that tackles well-posed scientific problems with a coordinated interdisciplinary approach. There are several other areas that can clearly benefit from the application of datamining techniques such as this, for example global climate change and natural hazard monitoring, One particularly intriguing prospect is the idea of performing monitoring tasks completely autonomously from largely self-directed spacecraft. This is a serious possibility for studies such as plate tectonics, because it is clear that almost no external information is needed to perform the most important geometric corrections.

Acknowledgements

The research described in this paper was performed on a Cray T3D supercomputer at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. It is a pleasure to acknowledge our colleauges R. Crippen and R. Blom at JPL, who conceived and implemented the original imageodesy algorithm from which Quakefinder is derived, and who provided the SPOT data, geological domain knowledge and numerous insights during the course of this work.

References

Crippen, R. 1992. Measurement of subresolution terrain displacements using SPOT panchromatic imagery. *Episodes* 15:56–61.

Dempster, A.; Laird, N.; and Rubin, D. 1977. Maximum Likelihood from Incomplete data via the EM Algorithm. *J. Roy. Stat. Soc. B* 39:1–38.

Fayyad, U.; Weir, N.; and Djorgovski, S. 1993. Skicat: A machine learning system for the automated cataloging of large-scale sky surveys. In *Proc. of the Tenth International Conference on Machine Learning*, 112–119.

Jacquemod, G.; Odet, C.; and Goutte, R. 1992. Image resolution enhancement using subpixel camera displacement. *Signal Processing* 26:139–146.

Kim, S. P.; Bose, N. K.; and Valenzuela, H. M. 1990. Recursive Reconstruction of High Resolution Image From Noisy Undersampled Multiframes. *IEEE Trans. Acoust. Speech and Sig. Proc.* 38:1013–1027.

Stolorz et al., P. 1995. Fast Spatio-Temporal Data Mining of Large Geophysical Datasets. In *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining, Montreal, Canada*, 300–305.

Ur, H., and Gross, D. 1992. Improved Resolution from Subpixel Shifted Pictures. *Graphical Models and Image Processing* 54:181–186.

Wen-Yu-Su, and Kim, S. P. 1994. High-Resolution Restoration of Dynamic Image Sequences. *Int. J. Imag. Syst. and Tech.* 5:330–339.

Yuille, A. L.; Stolorz, P. E.; and Utans, J. 1994. Statistical Physics, Mixtures of Distributions and the EM Algorithm. *Neural. Comp.* 6:332–338.