VAST Newsletter 7

September 2011

PI Report

Tara Murphy and Shami Chatterjee

This quarter we are focusing on techniques and algorithms that are being developed as part of the VAST Design Study. We also have a profile of one of our computer scientist VAST members – Kiri Wagstaff from NASA JPL.

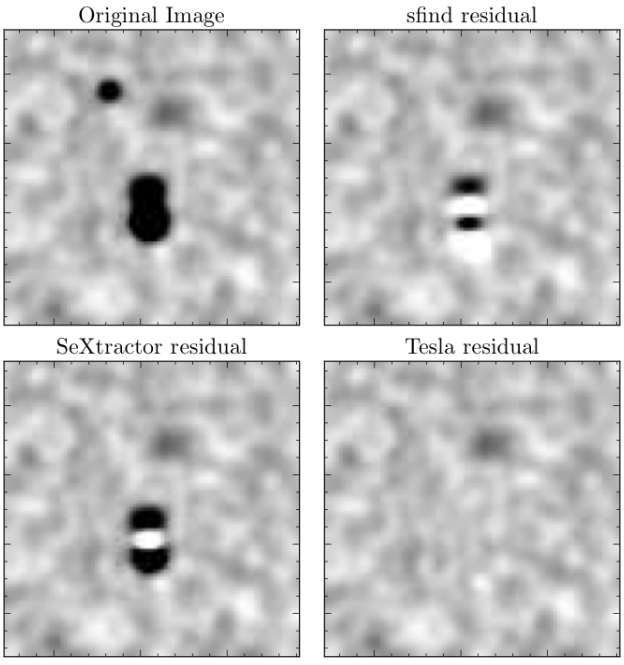
Next month is the internal review, in which all the Survey Science Projects will be reviewed by a committee from CASS. The Working Group chairs are currently preparing a report and any input you have about our achievements over the last year is welcome. We are also planning our BETA commissioning strategy for 2012, when BETA is expected to come online.

You can check the progress of ASKAP antenna deployment on the Murchison site at:

http://www.narrabri.atnf.csiro.au/askap_live

It was great meeting up with many of you at the IAU Symposium in Oxford this month. We had a short but productive discussion about plans for BETA. We look forward to seeing some of you at the ASKAP Working Meeting in November:

http://www.atnf.csiro.au/research/conferences/2011/askap



Science Report

Compact Continuum Source Finding

Paul Hancock, Tara Murphy, Bryan Gaensler, Andrew Hopkins and James Curran

The VAST pipeline will be required to ingest images and export radio light curves in real time. A critical step in this process is the detection and measurement of sources within the ASKAP data cubes. Currently available (image domain) source finding packages report catalogues that are complete and reliable to the 98 to 99% level. However, these packages are often only the first step in the production of a source catalogue, with a significant amount of effort required to remove falsely detected sources and ensure the accuracy of the measured parameters. The 1 to 2% of sources that are falsely detected or missed from one image to the next will contaminate the light curve creation and classification process, generating tens of thousands of false triggers every day.

In order to avoid such a deluge of false triggers VAST will require a source finding algorithm that:

- Has the highest possible completeness and reliability
- Is able to run in real time
- Requires no human interaction

We have been evaluating the performance of some widely used source finding packages: SExtractor, Sfind, Imsad, and Selavy. Each of these packages can produce highly complete and reliable catalogues of sources with source parameters measured with near ideal accuracy. The major stumbling block suffered by these packages is in the way they handle islands of pixels that contain multiple components.

Figure 1: *Top Left*: An example image containing a single and a double source. *Remainder*: The fitting residual for each of the source finding algorithms. Tesla was the only algorithm to fit all three sources, over both islands.

The problem stems from the way that the source characterisation is being handled. Gaussian fitting is sensitive to the starting parameters, and can often converge to a local, though not global, minimum in the difference function. When fitting multiple elliptical Gaussians, it is easy for the fitting routine to converge on non-physical parameters, or to not converge at all. To perform a robust source characterisation for an island of pixels that contains multiple components it is therefore necessary to:

- Be able to estimate the number of components to be fit
- Begin with an initial set of parameters that are close to the true values
- Constrain the parameters to avoid non-physical fits

Due to the difficulty in determining the number of sources, and initial parameters, as well as the difficulty in writing a constrained fitting algorithm, source finding packages have traditionally avoided constrained multiple component Gaussian fits. In order to robustly characterise islands of pixels that contain more than a single component we have developed a new algorithm which is implemented in the source finding package Tesla. This new algorithm uses the curvature (second order spatial derivative) of a continuum image to determine how many peaks are contained within an island of pixels, and to determine a set of starting parameters for the Gaussian fit. By using the freely available mpfit library[1] we are also able to place constraints on the parameters of the fit so as to avoid unphysical solutions. An example of the operation of Tesla is shown in Figure 2.

The difference between the image and a model based on the reported source parameters is called the fitting residual. A function of the fitting residual (usually the sum of the squares) is the metric which a fitting routine aims to minimise. An image of the fitting residual is therefore a useful visual diagnostic for evaluating the effectiveness of a source finding package. In Figure 1 we show an example image as well as the fitting residuals for each of the source finding packages that were tested. It can be seen that Tesla is able to accurately characterise islands of pixels with multiple components. Since such islands make up a large fraction of the sources that were not found by the existing source finders, Tesla is able to achieve a higher completeness and reliability.

A detailed description of our source finding analysis and the operation of Tesla has been submitted to MNRAS for publication. The images used to test the source finding algorithms can be found at

http://www.physics.usyd.edu.au/~hancock/simulations

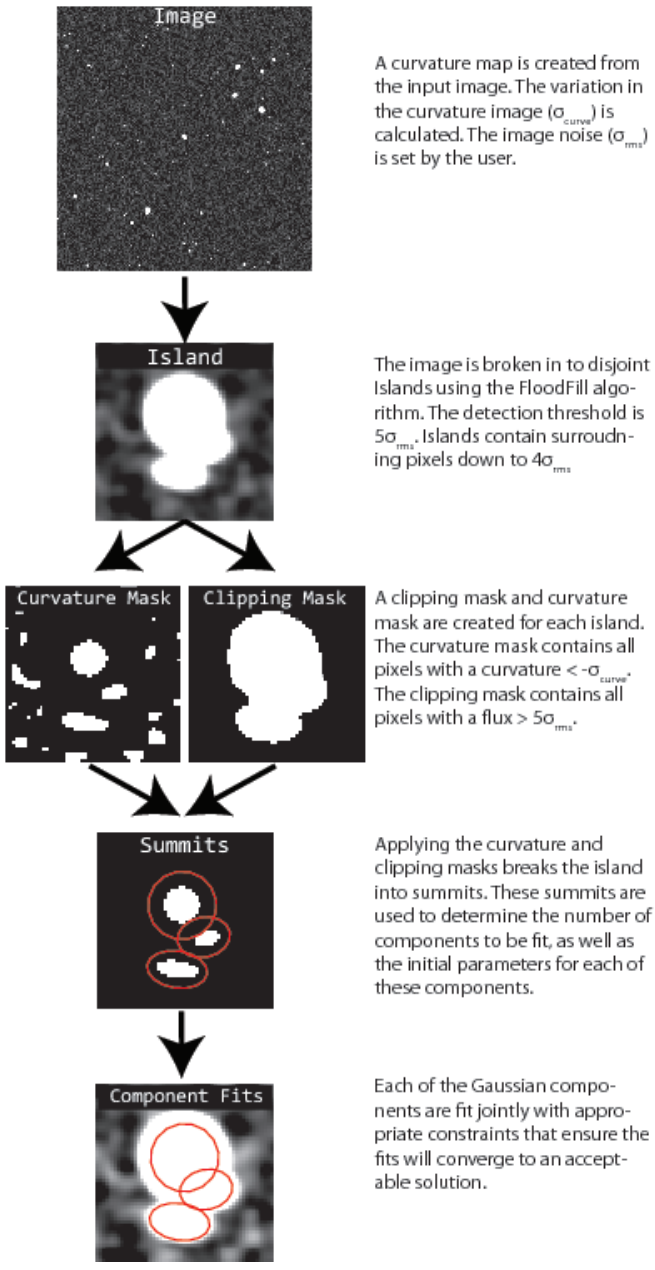


Figure 2: Operation of Tesla source-finding algorithm.

[1] A python version of mpfit can be found at code.google.com/p/agpy/source/browse/trunk/mpfit/mpfit.py

Profile – Kiri Wagstaff

Kiri is a researcher in machine learning at the NASA Jet Propulsion Laboratory in Pasadena. She is currently enjoying a sabbatical at Oregon State University.

What are your main research interests?

I am interested in developing new machine learning methods that can benefit online or onboard data analysis for instruments on spacecraft or on the ground. I believe that increasing the autonomous capabilities of our data collection instruments can advance our understanding of natural phenomena by both reducing the data volume to a manageable level and enabling intelligent action on the basis of local, autonomous analysis (like issuing alerts, adjusting data collection rates, etc.).

What papers are you working on at the moment?

I'm writing up a description of collaborative clustering methods, which are a kind of machine learning that, in the radio array context, is designed to enable individual telescopes or stations to communicate with each other to improve their data analysis capabilities. We've demonstrated benefits of this approach with data from seismic sensors from a volcano sensor network, and have just completed some experiments with VLBA data.

What excites you about ASKAP?

I think ASKAP has excellent potential to not only forge a path for future SKA and other array projects, but also to uncover new scientific knowledge and discoveries of its own. Seeing the project making progress towards an ambitious, functional, physical instrument is very exciting.

What is the main challenge for lightcurve classification?

Right now I'd say that the main challenge is to find the most useful representation of light curve data to provide a solid basis for accurate classification. In a larger sense, this is the fundamental question of all of machine learning: what is the best way to describe the data so that learning is optimized? What matters and what doesn't? Answering that question takes a mixture of theoretical understanding and empirical tests to determine the best approach.

What do you enjoy outside computer science?

I've been delighted to have the chance to learn about radio astronomy, remote astrophysical phenomena, and a bit of cosmology. I'm also fascinated by geology and the process of decoding the rock record to understand the past. My other hobbies include knitting, video game development, and creative writing.

News and Updates

Characterising the Fornax A field for BETA

Jamie Stevens, Simon Johnston, Keith Bannister

As part of ATCA project C2478 we surveyed the same 30 square degree Fornax field as C2479 (PIs: Feain & Johnston) looking for variability. To do this we used the ATCA's on-the-fly mosaicing mode, which allowed us to cover the entire field multiple times in a single 12 hour session, giving us both decent sensitivity and good uv-coverage. Data was taken in three epochs – April, July and August 2011 – allowing us to probe variability on 1 – 6 month timescales. Because the ATCA can be accurately calibrated, we expect that we should be able to reliably detect any variable flux densities within the field. Data reduction is a work in progress: although a reasonable image can be obtained using the standard Miriad individual mosiacing approach, bright sources within the fields limit the dynamic range of the images. Alternative imaging techniques using Miriad and other packages such as CASA are being investigated. When completed, we expect that the results, when combined with the measurements of C2479, will give us a good idea of the magnitude and timescale of variability within these BETA test fields.

VAST Lightcurve Classification

Umaa Rebbapragada, Kitty Lo, Colorado Reed, Kiri Wagstaff, David Thompson, Tara Murphy

We are continuing our study of both online and archival classification of simulated VAST light curves to establish upper bounds on transient classification performance. We presented this work in several posters at IAUS 285.

Our simulation of VAST light curves has expanded to include the following:

Source types: ESEs, IDVs, Novae, Supernovae, Flare Stars RSCVn and dMe, Xray Binaries, and background sources (see Figure 3).

Survey strategies: Wide, Deep, Patches, Logarithmic, Monthly, and Galactic Plain

Feature representations: time domain, wavelets, Lomb-Scargle Periodogram (LSP), statistical features from Richards et al. 2011

Signal to Noise Ratio: ranges from 3.0 to 10.0

Offline classification is performed with Support Vector Machines and Random Forest classifiers. Results demonstrate that classification accuracy degrades with decreasing numbers of observations per light-curve. Classifiers built with statistical feature representations using a Random Forest perform best (see Figure 4).

We are investigating classification performance in an online setting, in which decisions are made as new observations arrive, enabling alerts and fast follow-up observations by other instruments. For VAST wide, we compare performance between a system trained on full (archival) versus partially-observed light curves, and then tested on incoming light curves. We find that training on partially-observed light curves performs best. We are also looking at a cost-sensitive cascade ensemble method that seeks to minimize the number of observations required before a source can be classified, at both training and test time. In general, waiting for more observations to arrive leads to higher performance.

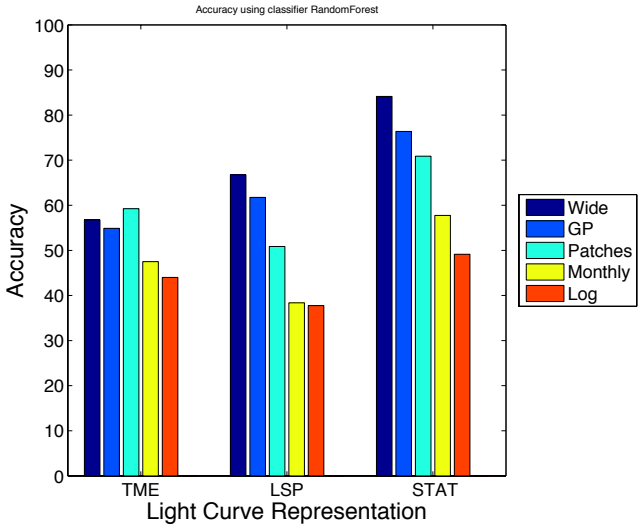


Figure 4: 10-fold cross validation performance using the time domain (TME), LSP and statistical feature (STAT) representations. See wiki for more.

We are currently investigating performance on a per-source-type basis, since different source types benefit from different observing schedules.

Upcoming Meetings

ASKAP Survey Science Working Meeting

CASS, Sydney, November 9 to 11, 2011

http://www.atnf.csiro.au/research/conferences/2011/askap

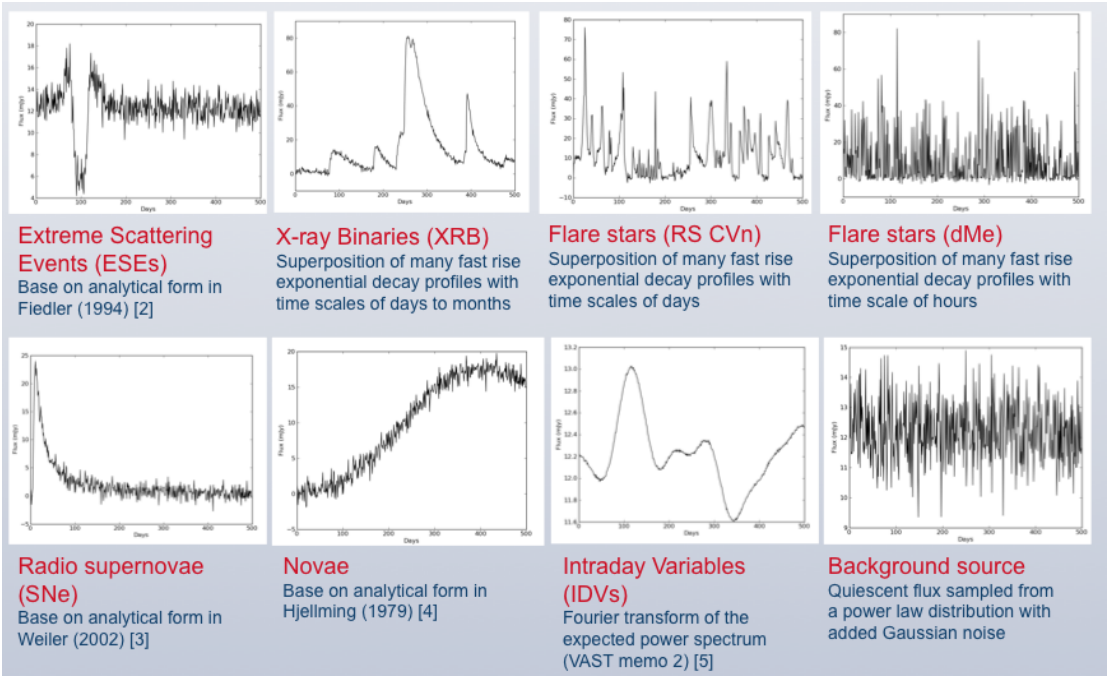


Figure 3: Examples of simulated source types (Kitty Lo).