

Progress Report and Thesis plan

Richard Ashley

July 29, 2014

Student: Richard Ashley
Supervisor: Tom Marsh

Abstract

Since May 2002, the Ultracam high-speed CCD photometric camera (ULTRACAM) has been taking observations of a variety of astronomical objects at the William Herschel Telescope (WHT), Very Large Telescope (VLT) and New Technology Telescope (NTT). Over this period it has generated approximately 10 Terabytes of raw data (CCD frames plus meta-data) taken on 390 observing nights and including approximately 566 target objects. In this data there are objects that may hold scientific interest and have not been investigated. Objects in the data need to be identified, listed and have reductions performed to determine their light-curves. The goal of this 1 year MSc project is to build an automated software pipeline that will process the Ultracam data archive and produce a browse-able version of the light-curves and meta-data for all of the observations.

1 Introduction to the Ultracam

The Ultracam high speed photometry camera (hereafter, ULTRACAM) had its ‘first-light’ at the William Herschel Telescope (WHT) on the 16 May 2002. Since then it has been used on many occasions at three telescopes, namely the William Herschel Telescope (WHT), La Palma, Islas Canarias, the Very Large Telescope (VLT), Cerro Paranal, Chile and New Technology Telescope (NTT), La Silla, Chile. The camera is designed to be portable and can travel between these telescopes.

1.1 About the camera

The camera has three CCD detectors enabling it to capture data in three colour bands simultaneously. Two dichroic beamsplitters divide the light from the collimator into three different beams, which shall hereafter be referred to as the ‘red’, ‘green’ and ‘blue’ channels. The three CCD detectors are mounted at right angles to each other on the camera. Therefore, each detector is at the end of a slightly different optical path. While the exposure timing is synchronised across the detectors, it is, however, possible to have the detector in the blue channel remain exposed and not ‘read-out’ while the other two are going through multiple exposures and ‘read-outs’. This is to allow for longer exposures where there might be less flux in blue, because the transmission of the optics and atmosphere and the intrinsic flux of most objects are lower in this channel.

A key aspect of the design of the camera is its ability to perform at high cadence. It is possible to have the camera read-out at up to 500Hz (frames per channel per second) [1]. This makes the camera useful for observations of rapid transient events and accurate timing.

The filters for each channel can be modified. In usual configurations the SDSS filter (u, g, r, i, z) set are used, but there are a selection of narrow-band filters that can be substituted.

Each CCD has a total pixel area of 2057x1024 pixels. Half of these pixels are masked and never exposed to light. They are used as a temporary buffer for reading out the chip. CCD detectors are read out serially, but in order to decrease the time between exposures, the full image can be moved to the blanked out area of the chip and this can then be read out while the un-masked area of the chip is once again exposed to light.

ULTRACAM gives the observer the ability to reduce the amount of the detector that is used for the exposure. This, again, speeds up the readout of the chip and enables the high cadence. Reducing the number of pixels exposed also decreases the amount of data storage needed for the run. The observer can define pairs of *windows* that are centered on their objects of interest. By making the windows suitably small, the observer can have the camera in extremely high cadence mode.

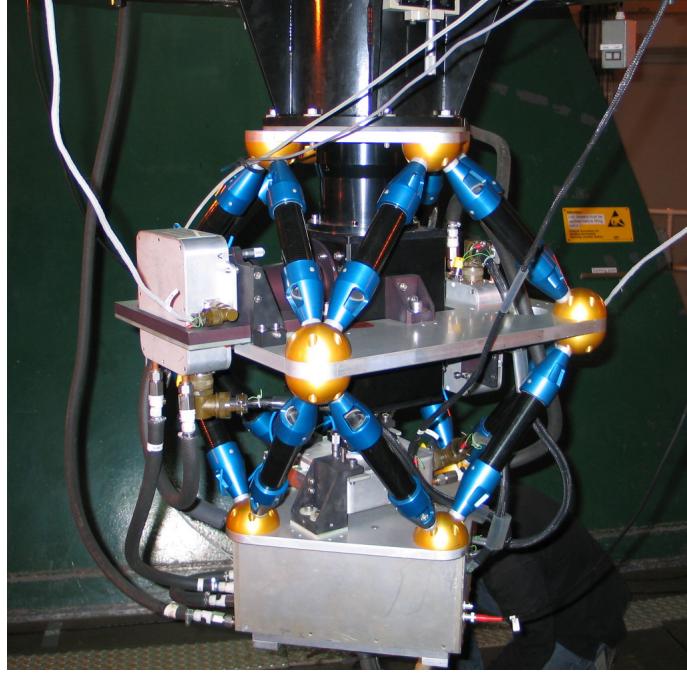


Figure 1: The Ultracam being commissioned in May 2002.



Figure 2: One of the three CCD detectors. The masked-off area is visible in the lower half of the chip surface.

The highest cadence mode is called *Drift mode*. This mode uses the masked area of the CCD chip to store several exposures simultaneously. Only the portion of the CCD that is exposed is shifted into the masked area of the CCD meaning that the unmaksed area is ready to be re-exposed more quickly. This mode requires that only the lower portion of the detector, close to the boundary of the masked and un-masked areas, is exposed. In order to operate in Drift mode, it is likely that the observer will have to rotate the camera in order to place the target object(s) in the correct place on the CCD. This means that we can have any orientation ($0 - 180^\circ$) of the sky coordinates for any particular run of ULTRACAM.

More details on the camera design and operation can be found in *Dhillon et al. (2007)* [1].

1.2 Observation and recording of data

Usually the camera remains installed on the telescope for a week or so and is used for observations on consecutive nights. During the night there are several *runs* made. A *run* can be defined as a period when the camera is active and gathering data. Not all runs are used for gathering *scientific* data. Some runs

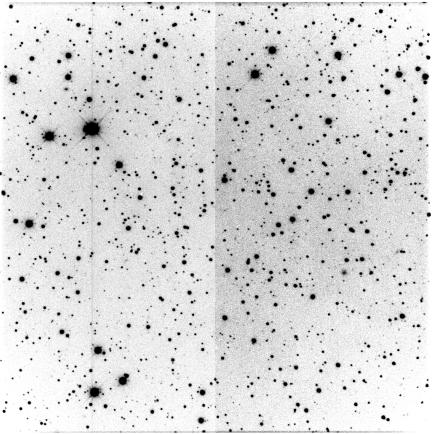


Figure 3: A 'fully' exposed CCD with 1 pair of windows (512x1024 pixels each)

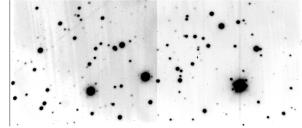
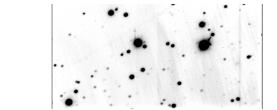


Figure 4: A 'masked' exposure with 2 pairs of windows (350x300 and 250x250 pixels)

are used for target acquisition and camera calibration purposes.

The types of runs are:

- *Science run*: These are the runs that contain the valuable scientific data. They usually comprise the longest portions of the observations during the night, unless the camera is having difficulties or adverse conditions are preventing useful astronomical observations.
- *Acquisition run*: These are runs, usually of short duration (ie a few minutes) during which the telescope is being moved in order to place the candidate object(s) in the field of view. The camera may also be rotated in order to align the CCD such that the targets avoid 'bad' pixels or are near to the lower boundary of the detector (eg for high speed readout in Drift mode).
- *Sky-flats*: At the start and the end of the night (usually during twilight) the observer will take a few runs to create *sky-flats* that will be used later for calibrating the variations in pixel sensitivity across each of the detectors. Sky-flats are generated by exposing the camera to patches of sky during the twilight.
- *Bias*: A short run to build calibration readings for measuring the bias of the detector. This bias will be subtracted from the data during the reduction.
- *Timing calibration run*: One way to check the timing measurements of the camera is to take frames of a well-known rapidly oscillating source. For example, the Crab Pulsar (PSR B0531+21). The timing of the pulses as measured by the camera can be compared to the expected times for the pulsar which serves as a standard clock for calibration.

1.3 The data archive

At the time of writing, the ULTRACAM data archive comprises of about 10 terabytes of saved data. This can be broken down as:

- 390 nights on which ULTRACAM was operational at a telescope.
- 12 256 runs, including Science, Acquisition runs, etc as listed above.
- 119 817 742 frames in total. This is total includes all the frames for each channel; red, green and blue.
- 10 181 269 485 298 bytes of raw image data.

The data set is relatively large and is housed on storage internal to the university computer network. This means that it access to this data from remote locations (for example, by research collaborators) is difficult and relies on manual intervention. If a researcher needs to see some data, then they need to contact a member of department at Warwick or Sheffield and request a data reduction. There is no means to explore this data from their remote location.

1.4 Data reduction

Tom Marsh has developed a set of software tools that allow the observer to reduce the data in ‘real-time’ while the observations are taking place. This serves as a ‘preview’ for the observer and allows adjustments to be made during the run. After the run, the raw data is copied to the archive and this can be used for reductions later. This can happen the following day, or much later when the observer has returned from the telescope site. This data archive can be ‘re-reduced’ at any time as all of the raw data is stored in the archive.

The current data reduction process for ULTRACAM is designed to produce three colour light curves from the raw image data. The pipeline consists of the following stages:

1. Producing *bias* frames that are used to calibrate the CCD detector’s thermal noise characteristics.
2. Producing *flat-fields* to calibrate the pixel sensitivity of each of the 3 CCD detectors. These flat-fields are subtracted from the image data during the reduction process.
3. Defining *apertures* for the objects of interest in the run. This step requires manually choosing the objects of interest in the frames and defining the aperture sizes for each object. Apertures are independently set for each channel (r, g, b).
4. Running the *reduction* software. This code uses the apertures defined in the previous step and measures the flux of each object in each colour. The software is able to track changes in the object’s size and shape due to changes in the point spread function (PSF) by scaling the virtual aperture. It is also able to track small deviations in the positions of the objects.

Although this process is not particularly cumbersome, it *does* include some manual steps and it does not scale well when there are a large number of runs to be processed. The example in Fig. 3 contains more than 1000 objects. Manually defining apertures for each of these objects in each channel is not really practical. An automated method would enable data reduction for all of the objects captured in each run without manual intervention.

2 Automating the Ultracam pipeline

2.1 Project goals

The outcome of this MSc project will be a software pipeline that provides an automated way of reducing the raw data from ULTRACAM. The key outputs will be field identification through astrometry, deep image generation and light-curve reduction for all of the science runs in the data.

The final product should allow researchers to quickly browse the reduced data and examine light-curves of each object in all of the runs. The light-curves and runs should also be graded in terms of quality (eg seeing, atmospheric transmission, etc)

The *minimum* goals for this project are:

- An *automated pipeline* which can process ULTRACAM data to obtain light-curves for all objects within the field of any science run. This data will include light-curves for all objects and all channels in each run. An indicator of quality of the run will also be provided.
- A *light-curve browser* allowing a researcher to query and browse the Ultracam archive by date, object and position.
- *Deep images* of the observed field for all three channels for each run, based on stacking of frames in the run. The quality of the image will be ensured by discarding poor frames and taking into account movement of the frames due to tracking- and/or atmospheric effects.

In addition to these minimum requirements, it would be *desirable* to have results for:

- *Full photometric reduction*. If the quality of the automated reduction is not high enough for science research, then the pipeline could include a stage that prepares and runs the current Ultracam pipeline reduction process for more accurate reductions.
- *Identification* of moving objects seen in the run. For example, these might be Kuiper belt objects (moving at a few arcseconds per hour).

- *Automated alerts* triggered when an interesting (variable or transient) light curve is detected by the pipeline.
- *Movies* (in .mp4 format) of the run displayed in the light-curve browser.

2.2 Approach

The key step in automating the Ultracam pipeline is finding a reliable way of identifying all of the objects in a run and tracking them across all of the frames. It is also necessary to identify the same objects in each of the three colour channels.

2.2.1 Source extraction

A popular software tool used for source extraction is SExtractor [2]. SExtractor is able to process an image field and produce a catalog of sources in that field along with a measurement of the flux count of each object. The flux count is calculated after making efforts to account for the background in the field and then fitting a profile to the object. *Comment: Will need a full chapter on this in the thesis, including tweaks to the configuration parameters of SExtractor and how it affects the results.*

2.2.2 Algorithm

The stages of the reduction process are as follows:

1. Stage 1: Create a list of objects found in the run.
 - (a) Read the raw image file, containing all frames for a particular run.
 - (b) Initialise an empty list of objects
 - (c) For each frame in the run
 - i. For each colour channel in the frame
 - ii. Extract each window from the frame
 - iii. Send the image to the SExtractor software
 - iv. Read the results of the source extraction process, including pixel position and flux measurements for each object
 - v. For each object returned
 - A. Try to match this object with one already in the list, based on nearest distance.
 - B. If the object is not already in a list, add this object to the list as a *new* object.
 - (d) Store the list of objects for each of the three channels.
2. Stage 2: Filter this list, removing objects that are likely to be artifacts. This is done by looking for objects that do not persist across more than a pre-defined percentage of frames; and objects that have a size approximately equal to one pixel.
3. Stage 3: Produce catalogs of the objects ordered by 'brightness' as measured by the flux estimate from SExtractor. Pass these catalogs to the Astrometry.net library to resolve the WCS solution for the fields. Since the three channels have a very slightly different view and different distortions of the image, their WCS solutions will differ by a small amount.
4. Stage 4: Using the WCS solutions, merge the three catalogs to 'cross-identify' each object across the three channels. This may seem to be a trivial step for many ULTRACAM runs because the differences in the fields from channel to channel are minor, however in crowded fields such as the one shown in Fig. 3 simply matching objects based on their pixel coordinates is not enough to disambiguate them.
5. Stage 5: Produce deep images for each channel and export to a web-viewable format, such as PNG.
6. Stage 6: Create .json data files and HTML files to enable viewing the output of the run through a web browser.
7. Stage 7: Publish this 'web-enabled' data to a web site that is accessible outside of the university network.

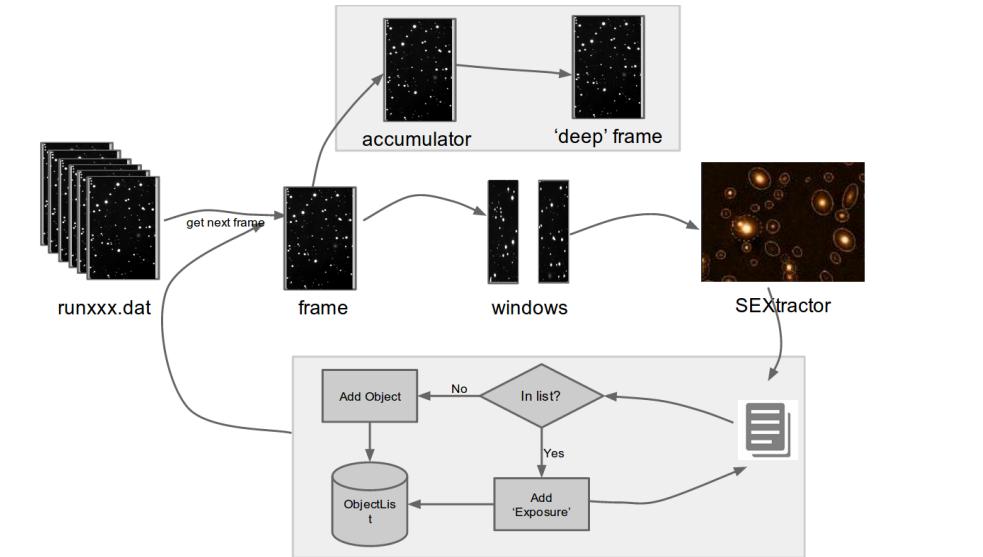


Figure 5: Schematic of Stage 1 of the pipeline.

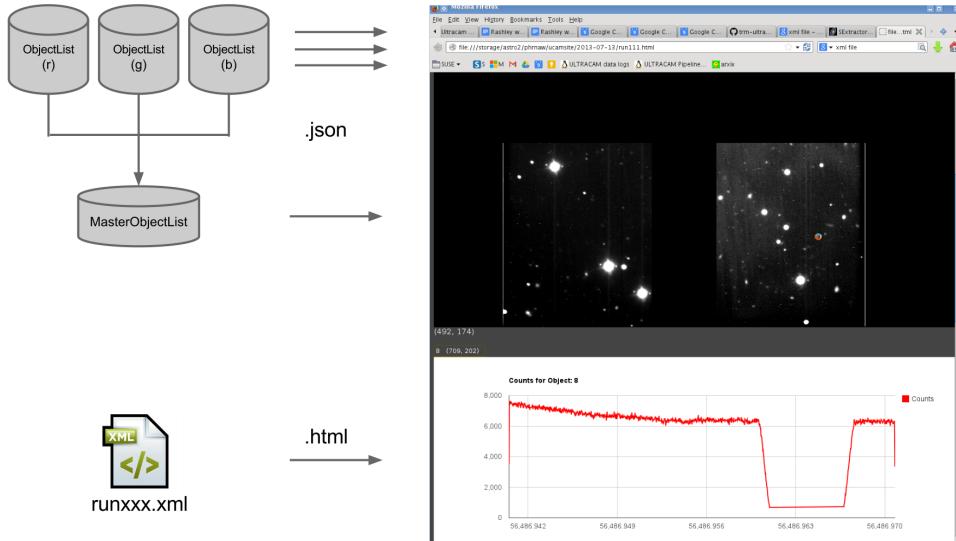


Figure 6: Schematic of Stage 6 and 7 of the pipeline.

2.3 User Interface

The core ‘visible product’ of the project will be a website that allows a user to browse all of the data in the ULTRACAM archive. The key features of this website will be:

- A *search* interface to allow the user to query the archive by object identifier and/or world coordinates (right ascension, declination).
- A catalog of runs organised by calendar date, containing *thumbnail images* of the fields.
- For each run, a web page that shows the user:
 - the *deep images* of the field in individual colours and also a combined (r,g,b) image.
 - the *light-curves* of each object as the user clicks on the object with the mouse.
 - *previews* of the light curves as the user ‘hovers-over’ the object with the mouse.
- light-curves should be plotted as absolute measured flux or a *relative flux* compared to some other objects in the field.

- the web page should also allow the user to *export* the data to a standard format, such as FITs.

See Fig. 7 for an example of the web-page.

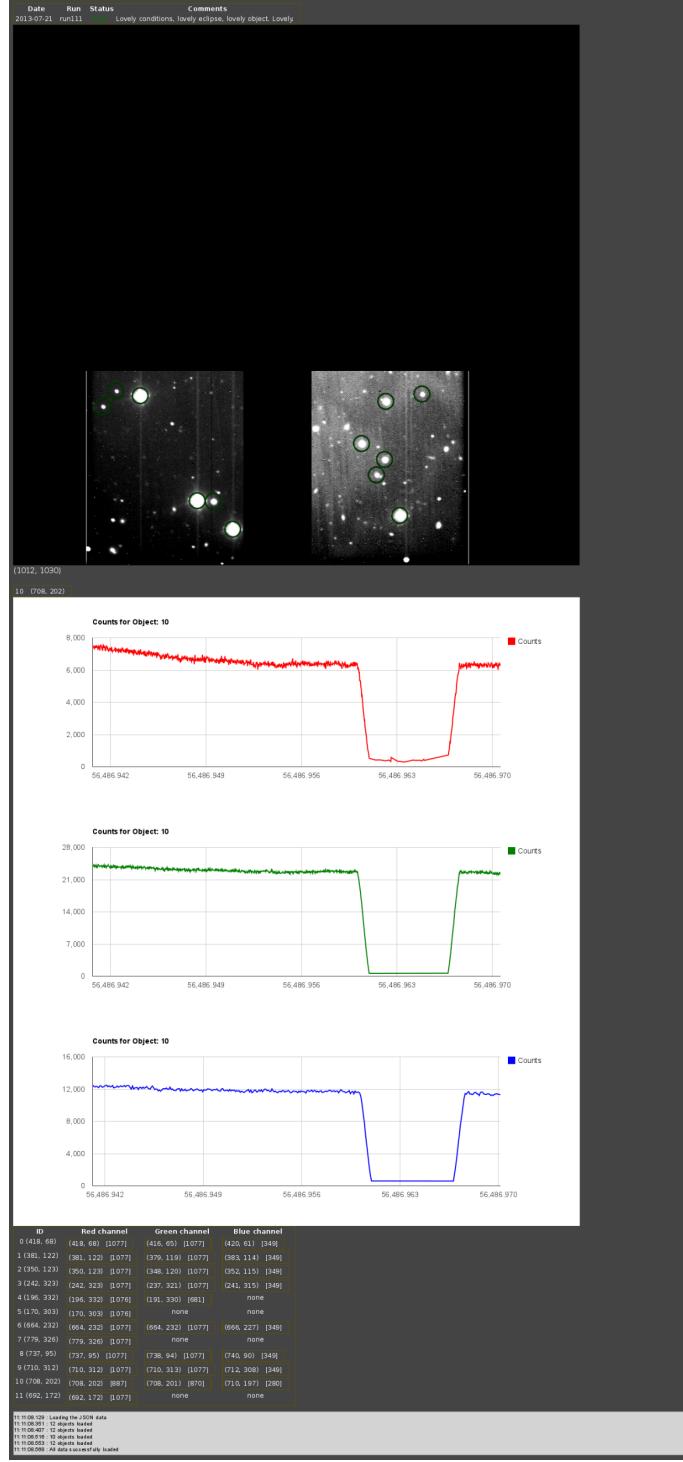


Figure 7: Preview of the webpage.

3 Progress so far

At the moment, the software for the pipeline is at a stage where it is able to produce light curves for many of the runs in the data archive. This means that, given a run, it is possible to produce a web-page

(as described above) that has deep images for the field and light curves for all of the objects in the field. Therefore, at its most basic, the pipeline ‘works’, however there is still significant work to be done.

3.1 WCS solution

After some ad-hoc tests using *SCAMP* [3] and *Astrometry.net* [4] it seemed that the Astrometry.net software was more reliable at finding good WCS solutions to the fields. The software was downloaded to a local machine (including the extensive index files) and compiled. At the moment, this is now part of the pipeline but it does not consistently find WCS solutions for all of the fields.

There are several difficulties in reliably finding a WCS solution for the fields. These are:

- *Lack of pointing data*: The Ultracam system does not integrate with the pointing software of any of the telescopes and does not get pointing information automatically. We rely on the observer to enter a name of the candidate object for each run and then, when the data is archived, a *SIMBAD* lookup is used. This gives us a world coordinate that is somewhere in the field, but it is not known which object (or pixel location) this applies to.
- *Field rotation*: Since the Ultracam can be rotated about the optical axis to allow for optimal alignment of the objects, this means that the field of view can be at any arbitrary rotation giving an extra degree of freedom to the matching task.
- *Windows*: Many ULTRACAM runs are configured to use only portions of the CCD area. You can see an example of this in Fig. 4. This means that there is an incomplete view of the sky for that field. When trying to match to existing indexes, there could be important, bright objects that are in the index file, but do not appear in the Ultracam field due to the masking caused by the windows.
- *Sparse fields*: On uncrowded fields, we might only have 4-5 objects that can be used for field identification.
- *Very small windows*: Some runs, particularly ones in high cadence mode, use very small windows (eg 172x156 pixels) in order to decrease readout time. This means that the Ultracam images might only contain two objects or so. This makes matching to a reference catalog impossible.
- *Choice of reference index by colour*: The Astrometry.net software uses 2MASS and Tycho-2 reference catalogs by default. These are based on infra-red and V magnitudes. This means that the blue channel (which is often using the SDSS u filter) might not match the reference indexes. Indeed, current tests often end with a match in red, a match in green and no match in blue.

The current line of investigation in the project is to try to use positive matches in the red and/or green channel to ‘seed’ the search for matches in the blue channel. If this doesn’t work, it may be worthwhile to seek alternative (bluer) reference catalogs.

For sparse fields, it might be worth looking for deeper reference catalogs.

For the very sparse fields, with 2 to 3 objects on them, it is unlikely that a WCS solution can be found. This is probably not a problem as the number of objects is so low, there is unlikely to be any ambiguity in cross-matching the object across the channels and combining the light-curves can be performed simply by matching based on the object’s pixel coordinates alone.

3.2 Quality of the photometry

The photometry of the objects is simply the flux estimate produced by the SExtractor software and has not been evaluated in terms of its accuracy relative to the existing Ultracam reduction pipeline. Also, there has been no effort to look at the errors in flux measurements (as estimated by SExtractor). This work still needs to be done.

A preliminary check of photometry is shown in Fig. 8.

3.3 Run classification

Due to the diversity of data in the Ultracam archive, not all of the runs can be processed in the same way. Crowded fields with full-frame exposure need to be treated in a different way to high cadence runs of only a few objects. SExtractor parameters need to be tweaked in order to get the best results depending on the nature of the run. At the moment, this step is done manually and with some ‘trial and error’. If

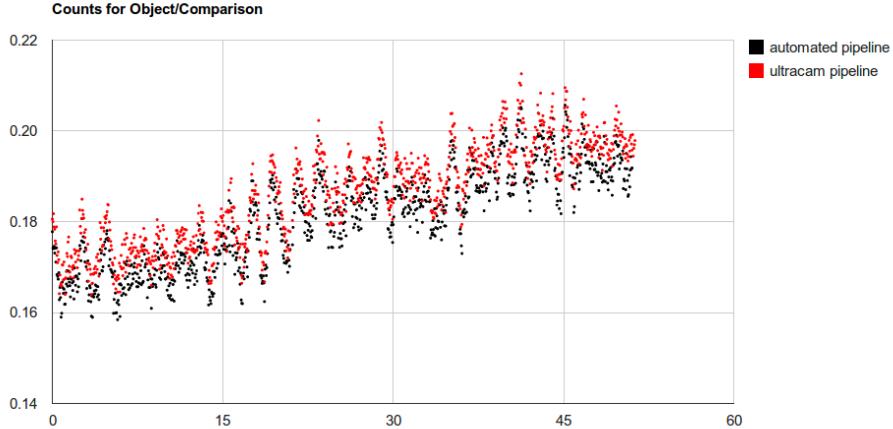


Figure 8: Comparison of the photometry of the Ultracam pipeline with the new automated pipeline (this project). The data shows the 'r' band photometry for 60 seconds in the light-curve of the polar V834 Cen.

the pipeline is to run in a truly automated way, then it needs to be able to adapt to this diverse data and perform a ‘run classification’. This classification step should also attempt to identify non-science runs, such as flat-fields, biases and acquisition runs and deal with them appropriately. Some acquisition runs often become real science runs and contain valuable data.

3.4 Search interface

At the moment, since there is no robust solution for solving WCS coordinates (yet), the search interface has not been built. It should be possible to query the archive using world coordinates. A WCS solution should be available for every science (and potentially also acquisition) run in order to build the search index (within the limitations mentioned above).

3.5 Cover the entire data archive

The browseable archive (the product of this project) should cover the entire ULTRACAM archive, from May 2002 to the present. Once some of the remaining issues have been resolved (mainly WCS solutions and Run classifications) then the pipeline should be set to run across the entire data set.

4 Thesis outline

The outline of the thesis is as follows:

1. *Abstract*
2. *Introduction*
 - (a) *The ULTRACAM instrument*: Discussion about the ULTRACAM instrument and how it works operationally.
 - (b) *Data archive*: Review of the data that is in the archive. How much and how well processed is the data?
 - (c) *Current pipeline*: How the current pipeline works. What are its limitations and strengths?
 - (d) *Automated pipeline*: Motivation for building an automated pipeline and the benefits of having the archive more accessible.
3. *Automating the reduction pipeline*
 - (a) *Algorithm of the pipeline*: Description of the key steps in reducing the data in the archive.
 - (b) *Source extraction*: Discussion of the mechanism to invoke SExtractor and the how it processes the fields

- (c) *Object matching*: Analysis of the techniques used by the pipeline to keep track of the objects from frame to frame and across the channels.
- (d) *WCS solutions*: Analysis of the various approaches to finding WCS solutions for the fields.
- (e) *Web-publishing*: Description of the architectures and technologies used to publish on the web. Including a discussion on choices of ways of encapsulating and displaying the data (JSON, HTML5, charts, etc).

4. Results

- (a) *The archive*: A description of where to find and how to browse the data archive.
- (b) *Data quality*: Analysis of the data quality (eg accuracy of photometry, WCS solutions, tracking of objects, distortion of the fields). Include a discussion on the quality of the photometry compared to the current Ultracam pipeline. How does 'tweaking' of the SExtractor parameters effect the quality of the data?
- (c) *Highlights*: Highlight some 'interesting' objects in the archive.

5. User manual

- (a) *Using the archive*: A detailed description of how to use the browser interface.
- (b) *Running the pipeline*: Detailed instructions for installing and running the automated pipeline.

6. Conclusion and acknowledgements

5 Timetable for work

5.1 One year plan

The high level view of the timeline for the 1 year duration of the MSc project is as follows:

- *October - January*
 - Prototyping and building of automatic ULTRACAM reduction pipeline.
- *February*
 - Demonstration to the department at Warwick through internal seminar.
 - Code refactoring.
 - Checking code into GitHub.
- *March*
 - Code revisions based on feedback from demos, training and presentation.
 - Progress report and thesis plan due, 24th March
- *April*
 - New demonstrations to the department at Monday morning meeting.
 - Identification of objects worth study.
 - Production of data metrics (colour-colour diagrams, magnitude-error plots, etc)
 - Viva with Director of Graduates 14th April 2014
 - Publishing to wider Ultracam audience (eg Sheffield) via an online demo/web interface.
 - Training of three department members (from Boris, Tom, Madelon, Elme, JJ) on how to operate the software.
- *May*
 - Further investigation of 'interesting' identified objects.
- *June-July*
 - Thesis iterations
- *August*
 - Thesis submission and viva

References

- [1] Dhillon, V. S., Marsh, T.R. et al, 2007. *ULTRACAM: an ultrafast, triple-beam CCD camera for high speed astrophysics*, MNRAS, 378.
- [2] Bertin, E. and Arnouts, S. *SExtractor: Software for source extraction*, Astronomy & Astrophysics Supplement 317, 393
- [3] Bertin 2006 *Automatic Astrometric and Photometric Calibration with SCAMP*, ASP Conference Series, Vol. 351, 2006, C. Gabriel, C. Arviset, D. Ponz, and E. Solano, eds., p. 112
- [4] Hogg, D. W. and Lang, D. *Astrometry.net*, <http://www.astrometry.net>