Highlights of Spanish Astrophysics VI, Proceedings of the IX Scientific Meeting of the Spanish Astronomical Society held on September 13 - 17, 2010, in Madrid, Spain. M. R. Zapatero Osorio et al. (eds.)

LEMON: an (almost) completely automated differential-photometry pipeline

Víctor Terrón[1] **and Matilde Fernández**[1]

[1] Institute of Astrophysics of Andalusia, IAA-CSIC, Glorieta de la Astronomía 3. Granada, E-18004 (Spain), {vterron, matilde}@iaa.es

Abstract

We present LEMON, the CCD differential-photometry pipeline, written in Python, developed at the Institute of Astrophysics of Andalusia (CSIC) and originally designed for its use at the 1.23 m CAHA telescope for automated variable stars detection and analysis. The aim of our tool is to make it possible to completely reduce thousands of images of time series in a matter of hours and with minimal user interaction, if not none at all, automatically detecting variable stars and presenting the results to the astronomer.

1 Introduction

The development of the LEMON pipeline emerged from the need to reduce and analyze the data from a photometric monitoring of thousands of stars during a period of eight months distributed over four years. These observations form part of a project led at the Institute of Astrophysics of Andalusia (CSIC) with the goal of studying the variability of pre-main sequence, low mass stars. Isolated peaks in the graphs of light intensity may reveal the occurrence of accretion events, while periodic changes are reasonable indicators of the presence of starspots, binary stars or even, although not a primary objective in our research, exoplanet transits.

This classification of the observed stars, a rather straight-forward task for a trained eye, and which would have traditionally been done after generating by hand the light curve of each star, became an unviable option at the outset of the project, as thousands of stars were to be monitored during each campaign. Furthermore, with two thirty-night observation periods allocated per year at the Calar Alto Observatory's 1.23 meter telescope, it was of special importance to be able to reduce and analyze the data before the next campaign, hence the necessity of developing a pipeline in order to automate the reduction of the data. Its name, LEMON, is a loose acronym which expands to Long-term Photometric Monitoring.

Although it was developed with the 2k × 2k pixels, 17′ FOV, SITE#2b optical CCD installed at the 1.23 m CAHA telescope in mind, LEMON is designed in order to work with FITS images taken with any CCD. This is possible thanks to the information stored in the FITS header, which is presupposed to contain all the information that the pipeline needs in order to correctly handle the images. As this information is not optional but mandatory, the preliminary stage of the pipeline checks for the existence of the required keywords and allows the user to manually specify them in case the necessary values are not where they are expected. After this, the pipeline can proceed with the data reduction with nearly no user interaction.

LEMON is coded in Python, with the strong emphasis on simplicity and code readability that is expected from any program written in this high-level, object-oriented programming language. The most computationally intensive tasks, where high performance is critical, are implemented as Python ANSI C and C++ extensions. Even more important is the Unix tools philosophy ("write programs that do one thing and do it well") that the pipeline follows. Thus, the different stages of the data reduction process are implemented as independent modules that, although most of the time run on a predictable, sequential order, may be combined as needed. If, for example, the astronomer simply needs calculating the offset between two images, then only the *offsets.py* module has to be run. In this sense, LEMON can be viewed as a set of tasks that *may* be used as a pipeline.

2 Data calibration

Being the SITE#2b optical CCD refrigerated with liquid nitrogen, the dark current, directly proportional to the temperature of the detector, is insignificant. At fewer than two electrons accumulated per hour [6], there is no need in our campaign to take dark frames. Thus, only the bias noise, the electronic signal independent of the exposure level and time, has to be removed. LEMON starts by subtracting from each image the mean value of the extra overscan region added to the normal image and which provides an estimate of the pure bias level when the image was taken. As some residual patter may remain, after this first step all the bias frames are combined, resulting in a master bias frame which is subtracted from all the images.

Multiplicative noise, caused by the varying sensitivity of the CCD and the uneven illumination across it, is reduced using the sky and dome flat-field frames. Ideally they would be, analogously to the bias calibration, combined into a single frame by which all the images taken with the same filter would be divided. However, it is here where probably the most arduous difficulty encountered during our automatization efforts, namely the dust, arises. Although the observing conditions at the Calar Alto Observatory could hardly be better, the motes of dust glaringly obvious in the flat-field images are absolutely not static. Additional motes irregularly fall from the filter wheel onto the instrument window, both of which are also periodically cleaned by the CAHA staff. In practice, from the point of view of the astronomer, the motes of dust seem to occasionally hop, so to speak, sometimes several times during the same night, while at times completely disappearing, as illustrated in Fig. 1. As a result of this, it becomes impossible to simply combine all the flat-field frames into a master

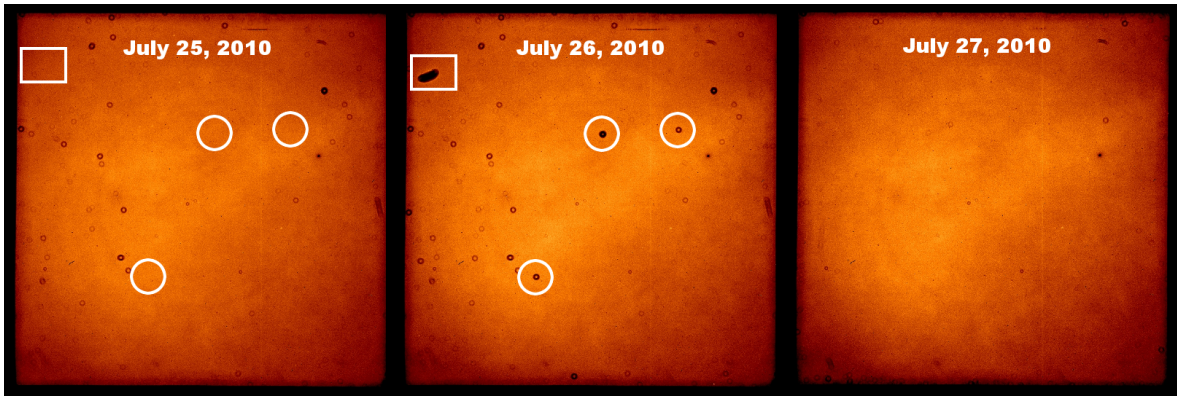

Figure 1: Motes of dust noticeable in the flat-fields taken in the Johnson B band on three consecutive nights of July 2010. It can be seen how some of the motes change their position between the first and second nights, whereas, on the third one, most of them disappear after the periodic cleaning done by the CAHA staff.

one.

The existence of this apparently random motes of dust forces us to identify intervals, that is, series of consecutive flat-fields in which *most* of the motes are in the same position. The flat-fields belonging to each interval are combined into a master frame by which all the images belonging to it are divided. This task, with which the data calibration stage concludes, is the only one that we have not yet been able to automatize in a satisfactory manner. Until a method for evaluating the similarity between two flat-fields is eventually implemented in the pipeline, the identification of these intervals must be done by the astronomer.

3 Offsets calculation

Upon calibration, all the scientific images of the field taken with the same filter are combined into a single one. This image, referred to as the *reference frame* from now on, maximizes the signal-to-noise ratio and is the one on which sources are later identified and astrometry is done.

However, the images cannot be directly combined as, in spite of the telescope tracking and the autoguider, they rarely happen to be exactly aligned. Thus, in order to move them back to their "correct" positions, we need to calculate the translation offsets between all the images and one of them—that with the best astronomical seeing, from which all the other images are considered to have dislodged— so that they can be adequately shifted.

The identification of the image with the best seeing is done by using SExtractor [1], two of whose parameters in the output catalog are $FWHM$ and $FLUX\_MAX$, which for each star give its full width at half maximum and the peak flux above background, respectively. By calculating the arithmetic mean of these values for all the stars we can evaluate the astronomical seeing and transparency of the image as a whole. Another interesting parameter

is *ELONG_RATIO*, defined as A/B, where A and B are the semi-major and semi-minor axis lengths of the star, and that is used in order to automatically discard the rare images which are excessively elongated because of complications with the tracking or the autoguider.

The routine that calculates the offsets is implemented as a Python extension developed upon part of the code of IRDR [8], an ANSI C library of fast image processing routines for automated reduction of infrared observations. Once obtained, the offsets are used to align all the images with the IRAF [10] *imshift* task —although it is not directly run but instead invoked through the PyRAF [4] interface. The same package is also employed in order to combine the images, using the IRAF *imcombine* procedure, averaging the input pixels after discarding a fraction of the top and bottom values.

4 Photometry and astrometry

Stars are detected on the reference image using, once again, SExtractor, with settings that maximize the number of detected sources. False positives are not an issue as these detections will result in random light curves that will be simply ignored when variable stars are detected. Note that sources are detected only once, on the reference frame; their position in the remaining images is calculated with the already measured translation offsets.

Aperture photometry is done on each star using IRAF's *qphot*, albeit as a short-term goal we plan to shift this task to SExtractor. Different apertures and the size of the corresponding sky annulus are used, their suitability being evaluated by the statistical dispersion of the constant stars —as the optimal aperture will minimize it. Note, however, that this evaluation involves generating the light curves. Therefore, LEMON needs to switch back and forth between this and the next stage, testing different apertures until the best one is found. Then, and only then, the data reduction can proceed until the end.

Astrometry is also done at this stage, matching the reference frame against the 2MASS catalog [9] using SCAMP[1], which reads SExtractor catalogs and computes their astrometic solution. Although within the pipeline each star is uniquely identified by its number in the SExtractor catalog for the reference frame, celestial coordinates are indispensable for the comparison of the results with other works.

5 Light curves generation

The light curve of each star in each filter is generated by comparing its brightness (instrumental magnitude) to that of a reference star which results from averaging, with a weight inversely proportional to their statistical dispersion, the most constant stars in the field (differential photometry). The identification of these constant stars is done by evaluating the statistical dispersion of their light curves, as proposed by [2]. Beginning with all the detected sources, the light curve of each one of them is generated by comparing its brightness to that which results from combining the brightness of all the others. In each iteration, those stars

[1] http://www.astromatic.net/software/scamp

with the maximum statistical dispersion are discarded, a process that is repeated until only a small fraction of the stars remains. These are the sources identified as the "most constant" and used as the reference star against which the brightness of the star is compared in order to produce its light curve. This method, of course, is based on the premise that at least a small fraction of the observed stars is constant at the working precision.

It should be remarked that each star is compared against its own reference, artificial star. The reason for this is that the constant stars are identified among those stars satisfactorily detected in all the images that contain the star whose light curve is being generated. In this manner, no matter the observing conditions or how many different exposure times were used in each band, only those images in which the star is measurable with an acceptable signal-to-noise ratio are taken into account in order to generate its light curve.

6 Data analysis

The first tool developed to help the astronomers analyze the results was a web-based user interface, written in PHP and which stored the light curves in a SQLite database. Although this approach offered some advantages, such as facilitating the exchange of information between different users thanks to the use of persistent links, it made it necessary to use a PHP server, which unfortunately complicates the portability of the software.

Due to this issue, we have recently started to develop a wxPython-based GUI intended to replace the web-based user interface shortly. This application parses the result of the previous stage, a XML file that contains all the computed astrometric and photometric information, and presents all the data to the astronomer. Light curves are dynamically generated with the matplotlib library[2], along with additional information which yields valuable hints about the variability of each star:

1. The period of each star is calculated using the string-length method [5], which evaluates several candidate periods and selects that which minimizes the sum of the distances between the consecutive points in the corresponding phase diagram. However, this method, by definition, always returns a "best period", for any observed star, including those that are not actually periodic. This intrinsic limitation of the method requires further analysis of the results. Thus, the periods of each star, separately calculated in each observed filter, are compared. The more of them that are found to be similar by means of evaluating their relative percent difference, the more likely it is that the variable star is actually periodic.

2. For pre-main sequence stars —the objects of our study— periodic variability results in light curves which, in different bands, have different amplitudes but the same period, mainly due to the presence of cold spots. These cause the light curves in the visible filters to be synchronized, although of different amplitudes. Because of this, the Pearson correlation coefficient between the variability observed in the different filters —e.g., the

[2] http://matplotlib.sourceforge.net/

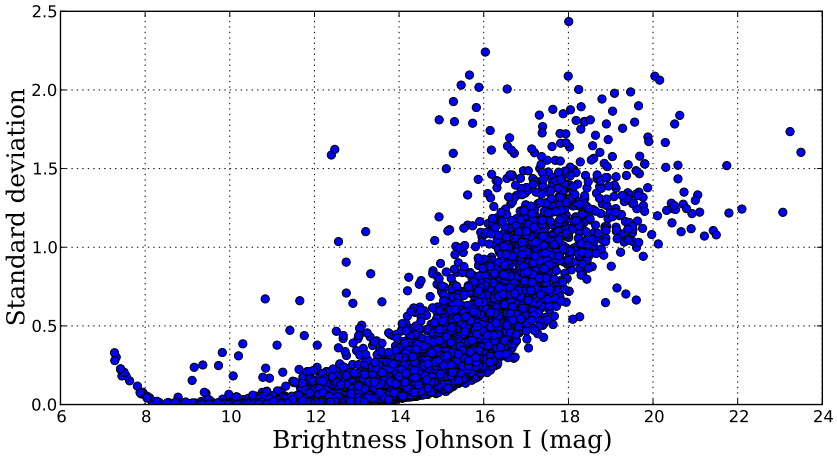

Figure 2: Standard deviation versus brightness of the stars observed in Trumpler 37 on August 17, 2009. Note how the statistical dispersion of some stars is considerably above the average values for those with a similar brightness. The higher standard deviation of the brightest stars is explained by their saturation in some of the images.

Johnson B and I bands— is evaluated. A high correlation is a reliable indicator that the variability of the star is real and not caused by the noise.

3. In each observed field there is a strong correlation between the brightness of a constant star and its statistical dispersion, as the signal-to-noise ratio becomes worse towards fainter stars, as shown in Fig. 2. This fact can be used in order to identify potentially variable stars, by graphing the statistical dispersion of the light curve of each star as a function of its brightness. The more that a star is above the general trend among those stars with a similar brightness, the more possible it is that the dispersion is actually caused by its variability instead of by the error due to a poor signal-to-noise ratio. The identification of these stars is performed by doing a loose boundary fitting so that a fraction of the points with extreme values (i.e., excessive dispersion) fall outside of the upper boundary [3].

The data analysis application does not only allow to filter and sort the stars by the three aforementioned criteria, but, more interestingly, also automatically detects the best candidates for variable stars, which therefore need a more careful examination.

7 Future work

The most immediate priority, on which we are already working, is to develop an algorithm to detect patterns of motes of dust in the flat-field images and automatically define the intervals in which they should be divided for data calibration, as explained in Section 2. After this is achieved, we would like to make use of a Bayesian network classifier in order to identify

the different variability classes among the observed stars, similar to that developed by [7]. This improvement will almost unavoidably be built upon the Fourier Transform of the light curves, information which shall complement the existing methods for detecting variable stars and their periods.

As a long-term goal we plan to move away from the IRAF tasks that LEMON currently depends on. Although a robust, excellent software, some parts of the code are released under a non-free license. This severely conflicts with our goal of releasing LEMON as free software —as a matter of principle; even more critical as the project is publicly funded— under the GNU General Public License and making it fully usable in a free environment.

Acknowledgments

The development of LEMON has been made possible by the financial support of the Spanish Ministry of Science and Innovation (MICINN) through grant AYA2007-64052, the Regional Ministry of Education and Science of the Regional Government of Andalusia (Junta de Andalucía) through grants TIC-101 and TIC-4075, and the project Intramural 2008-50-I-043 of the Spanish National Research Council (CSIC). The thoughtful feedback received from José Miguel Ibáñez, Juan Pedro Cobos and César Husillos during our intermittent conversations in front of the water cooler has repeatedly proven to be of decisive help throughout these months of hectic software development. Although never aware that we intended to use it here, the credit for Fig. 1 goes to Nuria Huélamo. The authors are also deeply indebted to Zac Dettwyler and Livy Siegel for having dutifully assured the linguistic correctness of this manuscript.

References

[1] Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393

[2] Broeg, C., Fernández, M., & Neuhäuser, R. 2005, Astronomische Nachrichten, 326, 134

[3] Cardiel, N. 2009, MNRAS, 326, 680

[4] de La Peña, M. D., White, R. L., & Greenfield, P. 2001, in ASP Conf. Ser. Vol 238, *Astronomical Data Analysis Software and Systems X*, ed. F. R. Harnden Jr., F. A. Primini, & H. E. Payne, 59

[5] Dworetsky, M. M. 1983, MNRAS, 203, 917

[6] Gorosabel, J., Kubánek, P., Jelínek, M., et al. 2010, Advances in Astronomy, 2010

[7] López, M., Bielza, C., & Sarro, L. M. 2006, in ASP Conf. Ser. Vol 351, *Astronomical Data Analysis Software and Systems XV*, ed. C. Gabriel, C. Arviset, D. Ponz, & S. Enrique, 161

[8] Sabbey, C. N., McMahon, R. G., Lewis, J. R., & Irwin, M. J. 2001, in ASP Conf. Ser. Vol 238, *Astronomical Data Analysis Software and Systems X*, ed. F. R. Harnden Jr., F. A. Primini, & H. E. Payne, 317

[9] Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163

[10] Tody, D. 1986, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 627, ed. D. L. Crawford, 733