Astronomical Data Analysis Software and Systems XVII **O2.1**
ASP Conference Series, Vol. 394, © 2008
R. W. Argyle, P. S. Bunclark, and J. R. Lewis, eds.

Automated Astrometry

David. W. Hogg,[1] Michael Blanton

Center for Cosmology and Particle Physics, Department of Physics, New York University, 4 Washington Place #424, New York, NY, 10003, USA

Dustin Lang, Keir Mierle, Sam Roweis

Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

Abstract. We have built a reliable and robust system that takes as input an astronomical image, and returns as output the pointing, scale, and orientation of that image (the astrometric calibration or WCS information). The system requires no first guess, and works with the information in the image pixels alone. The success rate is very high (~ 99.9 percent for shallow UV and optical imaging survey data), with essentially no false positives. We are using this system to generate consistent and standards-compliant meta-data for all digital and digitized astronomical imaging, no matter what its archival state, including imaging from plate repositories, individual scientific investigators, and amateurs. This is the first step in a program of making all of the world's heterogeneous astronomical data searchable and interoperable.

1. Non-text queries

When a user searches the web in 2007, he or she types text into an input box in a search engine; the engine finds web pages that contain that text (or related text). Even if the user is looking for an image or other media, the user must find it using text, and the search engines locate it using either bundled meta-data or else text associated with it in its context on the web. This is a severe limitation (consider cases where the user knows what the object of her or his search *looks like* or *sounds like* but not what it is *called*) and it also requires that the media on the web be correctly tagged with text meta-data. The present generation of search engines are easily confused, by, for example, web pages about dogs containing pictures of cats.

For making non-text data reliably searchable and for making it possible to query large data repositories with queries that are not based on text at all, we require systems that can recognize the content of complex data records and organize them. This problem, stated so generally, is very far from a solution at the present day. Even systems restricted to small domains like pop music and fingerprints are not extremely reliable and specific. This is true despite enormous investment; non-text data search is a multi-billion-dollar opportunity.

[1] david.hogg@nyu.edu

In this contribution, we will describe an extremely successful enterprise in this area. We can take any image of the sky, and return for you its pointing, scale, and orientation, with a very high success probability and essentially no false positives.

Our system is essentially a “pattern recognition” system, but it has the great advantage that the patterns of interest—asterisms—have been catalogued and understood by generations of astronomers. A successful pattern match therefore brings with it a great deal of meta-data, including the precise astrometric calibration of the image, the objects contained in the image, and—in principle—the bandpass through which and the time at which the image was taken.

Our problem is easy. There are many obstacles to systems that recognize objects in images. In general, objects move around relative to their backgrounds and are randomly occluded by complex objects in the foreground, they are often articulated or deformable (think limbs), they can be viewed from any angle, there are large changes possible in the same object or object class (Matt Damon can grow a beard), specific feature matching and identification is non-trivial (find the nose), and they are subject to dramatic changes in lighting. We have none of these problems when we consider astronomical images.

The stars are extremely distant, so our viewpoint is effectively fixed. Of course we are moving as are the stars in our Galaxy, so there are significant (and extremely important) motions on human timescales. However, these configurational and viewpoint changes create individual-star angular movements that are very small relative to the mean inter-star separations in our catalogs (this is the relevant comparison, since it determines the changes to the apparent shapes of asterisms), and smaller than or comparable to the angular resolution of typical imaging. The stars create their own light, there are no differences of illumination, only differences of relative brightness in different observational bandpasses. Stars appear to modern instruments as unresolved points, much brighter than the background level (in most cases). This makes them very easy to detect and measure; we don’t have the usual problems of feature identification.

These properties makes our problem much easier than typical image data interpretation tasks. We also benefit from the fact that large, uniform catalogs exist, containing $> 10^8$ stars observed in multiple bandpasses and over time periods measured in decades. This means that when we *do* find a matching asterism, we can say quite a bit about it.

Our problem is hard. On the other hand, we face a few challenges that are significant. The first is that the sky is big: A typical astronomical image we consider includes substantially less than 10^{-6} of the sky, and contains thousands of stars, only a fraction of which have been previously catalogued. Furthermore, the stars have been catalogued using data in a bandpass which, in general, can be very different from the bandpass in which the input image was taken. There can be arbitrary variation in the sensitivity to particular stars or the probability that a bright star in the image is represented in the catalog.

For many images we see, on the web, on bit-rotting magnetic tapes, and in the basements of amateurs, we do not know very much about the images. We don’t know when they were taken, what telescope was used, what bandpass they are in, how big they are (in an angular sense), where they are centered on the sky, the exposure time, or what mode the telescope or image rotator was in. This leads to an enormous heterogeneity in what we hope to see, and sets a high bar if we want to create a complete census of the world's astronomical data.

The need is great. There is a sense in which astrometric calibration is trivial: Observers know what they are observing, and telescopes know where they are pointing! However, astrometric calibration is also non-trivial: The vast majority of astronomical images in existence have no—or worse, wrong—astrometric meta-data. This can be blamed on many things: The late appearance of the WCS standard for astrometric meta-data (Greisen & Calabretta 2002, Calabretta & Greisen 2002); many astronomers use their own personal conventions for measuring and recording the information; some projects do not require precise global coordinates for any image; many telescopes have control systems that drift relative to the sky as observers tweak the pointing; and many telescope–instrument combinations output images with meta-data that are not standards-compliant or simply wrong.

A system to automatically and blindly calibrate all astronomical imaging coming off of instruments at each of the world's observatories could catch telescope drifts and faults in real time, potentially saving significant amounts of observing time and operations costs. Similarly, an observer coming down the mountain with already-calibrated data saves grant money and time in a non-illuminating part of her or his data analysis.

Amateur observers, many of whom take professional-grade data and some of whom publish IAU circulars and discoveries, do not usually have the tools they need to produce accurate and standards-compliant astrometric calibrations for their images. A system that provides these could bring a whole treasure trove of data to light, backed by a community that has always had enthusiasm for supporting basic research.

Historical plate repositories, filled with photographic plates, irreproducible by any standard, find it easier to scan the plates to digital files than record in a reliable way the plates' meta-data. This is because the meta-data exist in hand-written logs, not all of which survive, and not all of which are legible and organized. Furthermore, these logs contain many errors, and more errors are introduced in data entry. The barriers to making these meta-data accurate and accessible obscure these data—which provide our only record of the sky's time history—from scientific view. Imagine looking for precursor activity for some future Galactic supernova. It may have appeared in dozens of plates over the last hundred years. There is no way to find those data without reliable and standards-compliant astrometric meta-data.

We have only one sky; this makes our project possible. We have only a brief period of time in which we have been observing it, heterogeneously and chaotically; this makes our project necessary.

2. How it works

Most of what is described here is described elsewhere in more detail (Lang et al, 2007).

Our input is an image of the sky (see Figure 1). This can be an image in JPEG, GIF, or PNG format, or a data file in FITS format. Although all of

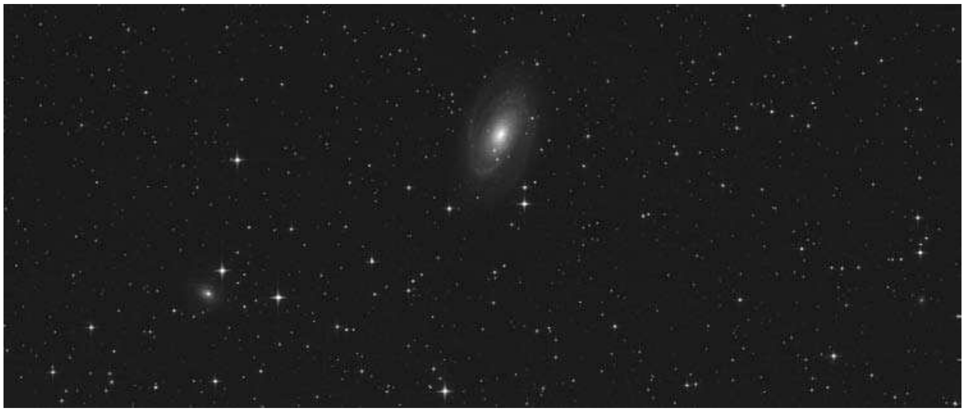

Figure 1. An example input image.

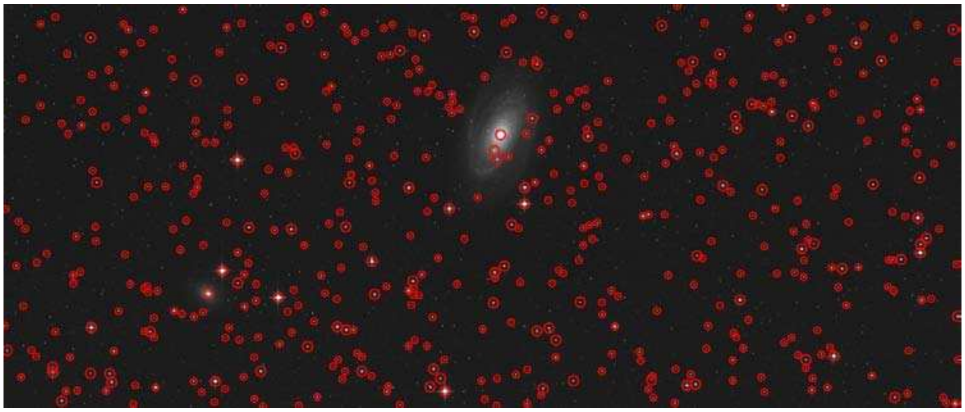

Figure 2. The result of running source detection on the input image.

these formats can carry some meta-data, we ignore all these meta-data. We use a simple but relatively robust heuristic system for identifying the stars in the image, measuring their centers, and ordering them in brightness (see Figure 2). We pass the brightest 200 stars through to our matching system.

Our matching has access to a set of pre-built indices. These indices contain information about sets of four stars ("quads" hereafter) selected from the billion-source USNO-B Catalog (Monet et al, 2003; Barron et al, 2007). The quads we choose to index are selected to cover a wide range of angular scales, to be easily found in a wide range of possible imaging bandpasses, to cover the whole sky uniformly, with plenty of overlap, and to not over-use individual stars (any one of which might be dropped from the input image for reasons of occlusion, error, noise, or defect). Construction of these indices takes hundreds of CPU hours, and the rules by which they are made cost many human months. We typically index $\sim 10^9$ quads.

Because we do not know the pointing or scale or orientation of the input image, we construct a geometric description of each indexed quad that is independent of location on the sky, angular size, and rotation. This geometric description or "hash" is a set of four floating-point numbers, which, in a coordi-

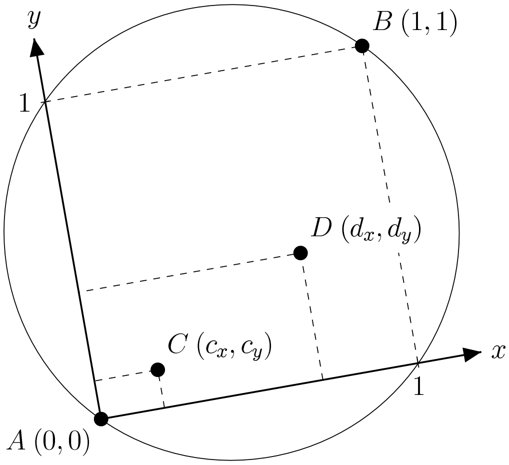

Figure 3. The definition of the quad hash code. The four stars are labeled A, B, C, and D, and the hash code is the position (c_x, c_y) of C and the position (d_x, d_y) of D in the coordinate system defined by A and B as shown. In detail, this operation is performed in the tangent plane with tangent point at the center of the AB line. To be a valid quad for our purposes, both C and D must lie within the circle with diameter AB and we remove degeneracies by requiring $c_x < d_x$ and $(c_x + d_x) < 1$.

nate system defined by the location of the two most widely separated stars in the quad, are the x, y positions of the other two stars (see Figure 3). We store all of these four-dimensional hash codes in a four-dimensional KD-tree data structure so we can do rapid searches given hash codes derived from the input image. Our KD-tree is a balanced tree designed so that it can be read into memory in a fully functional state with a single `mmap` operation with no building of pointers or other complicated operations (Mierle et al, in preparation).

Our hash of the quads has the great advantages that *(a)* if stars are Poisson distributed on the sky, hash codes are uniformly distributed in code space, and *(b)* small movements of the stars in a quad lead to small movements of the code in code space. It has the disadvantage that there is a parity degeneracy: flipping the input image changes the hash codes.

We loop over all 200-choose-4 quads (subject to cuts given in the caption to Figure 3) in the star list derived from our input image, beginning with the quad composed of the four brightest stars, proceeding to the additional four quads possible with the brightest five stars, and so on. Each quad from the input image can be used to generate a hash code (or all eight symmetry-equivalent hash codes, given the degeneracies), and that hash code can be used to look up indexed hash codes in the KD tree of indexed quads, using some neighborhood in code space to allow for jitter or noise in the image and in the USNO-B Catalog. Stellar magnitudes or brightnesses are not used in the matching or anything that follows; it is used only to *order* the quads for consideration, and, similarly, to order the stars at index-construction time.

Any index quad that matches an input-image quad creates a "hypothesis" about the pointing, scale, and orientation (and parity) of the input image on the sky. We "verify" this hypothesis by looking at the explanatory power of

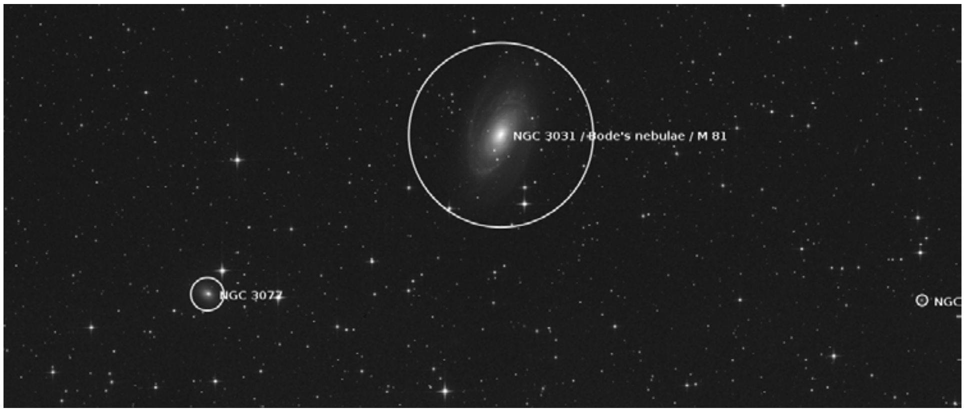

Figure 4. Example output from the system.

the hypothesis for the other stars we have found in the input image. We cast this "verify" step as a well-posed statistics problem about the probability that indexed stars and input-image stars would correlate as well as they do purely by chance. We avoid false positives by requiring a very low "match-by-chance" probability for a hypothesis to be considered verified.

Typically we look up thousands to millions of input-image quads in our index and generate a similar number of hypotheses, the vast majority of which fail to verify. A hypothesis can fail verification because it is spurious—by chance the image quad has a similar shape to some quad in our index—or because the match is correct but is not sufficiently convincing to pass our threshold. As soon as any quad-inspired hypothesis verifies, we terminate the search and declare the input image solved. We return to the user the basic meta data, the WCS, the FITS header containing the WCS, KML, matched object lists, names of known objects in the field, and annotated figures with known objects labeled (see Figure 4). The time to solve depends a lot on luck (that is, on the number of quads that generated unverifiable hypotheses prior to the first verifiable hypothesis), but for many images the location and measurement of the stars in the input image takes longer than the quad search and verification.

3. Successes and limits

Successfully solved images come from a huge range of sources,[2] including professionals with research-grade digital imaging, digitized plates from plate archives, amateur astrophographers, and casual photographers with ordinary digital cameras. In general, it is easier to solve an image that has a larger field of view (because the sky doesn't contain so many quads at the relevant angular scale) and it is easier to solve an image taken in a bandpass spanned by the USNO-B bandpasses (which are essentially B, R, I).

We have performed two statistical tests on the system, one with $9 \times 13\,\mathrm{arcmin}^2$ Sloan Digital Sky Survey imaging "fields", and another with 1 deg circular

[2] See `http://astrometry.net/` for examples.

GALEX near-ultraviolet images from the shallow All-sky Imaging Survey. Of 336,554 r-band fields, we solve all but 451 (99.9 percent success rate) with no false positives. Many of the failures can be attributed to issues with the SDSS fields themselves (focus and telescope motion issues), or the USNO-B catalog (small holes and regions dominated by spurious sources; see also Barron et al, 2007), and therefore the algorithm's success rate is in fact higher. Our success rate is good for the g, i, and z data also, but declines to about 97 percent when we move to u-band images, which detect fewer of the USNO-B stars. With 7077 *GALEX* AIS NUV fields, we successfully solve all but 8 (99.9 percent success rate) with no false positives; again those 8 are not all "our fault".

There is a fundamental limit to the size of a solveable field, which is related to the mean inter-star separation in the USNO-B Catalog (on the order of arcmin). In order for a hypothesis to verify, there must be stars outside the matching quad about which we can ask statistical questions. This requires that any solveable image contain significantly more than four stars. In detail the minimum size is a function of sky position, because the USNO-B Catalog density is a function of Galactic coordinates. It is also a function of image depth, because a small field with many stars detected below the USNO-B Catalog limit will not verify as easily as a small field whose depth matches exactly the USNO-B Catalog (recall that verification is based on the probability of chance alignments of USNO-B and input image stars). We have never successfully solved an image smaller than about 3 arcmin, despite the remarkable success with SDSS images only a factor of a few larger in each dimension.

We are extremely successful with optical images similar to or larger than SDSS fields with good-quality star detection; we don't anticipate—and haven't found—significant classes of impossible images. In addition, when we have knowledge in advance of the image scale, we have found that we can build specialized quad indices (by restricting the range of separations for the AB pair in the quad, and using stars likely to be in the imaging bandpass) that increase further the speed and success rates.

We have solved *GALEX* and *Spitzer* images, but never images taken in radio, far-infrared, or X-ray bandpasses.

4. The future

This project is a collaboration among astronomers and computer scientists; our interests are broad. Our astronomical interests in this project are mainly in the area of proper motion determination and constraints on the dynamical state and history of the Milky Way. For this reason our top priority is the calibration of large digitized plate repositories—spanning a century in some cases—to construct positional histories of all possible stars. The dense sampling and large number of available epochs permit proper motion measurements that saturate the information content in the historical data; these measurements will be much higher in quality and go deeper than any current all-sky proper-motion catalog.

We have found that if we look at the currently tabulated proper motions of stars inside the field of view of an input image, we can often determine (with large uncertainty) the date at which the image was taken. If we look at the brightness ranking of stars, we can often determine (also with large uncertainty) the bandpass through which the image was taken.

Because we would like to deepen and improve our astrometric reference, we are working on adding to (and removing from; see Barron et al, 2007) the USNO-B Catalog as we learn more about the sky from images input to the system. Any catalog we build this way will necessarily be heterogeneous in its depth and precision, but it is interesting to note that it will have the greatest depth and precision in the areas of the sky that astronomers image most; that is, it will have the depth and precision exactly where we need it! It is already the case that one could construct the Messier Catalog automatically just by inspecting the footprint of the user-input images.

As we add sources from images in bandpasses near but outside the near-ultraviolet through mid-infrared range, we will extend our catalog in wavelength. This permits indexing quads and building indices optimized for images at extreme wavelengths.

Because the quad index can be divided among CPUs, the hypothesis generation and verification part of the system is highly parallelizable. We have parallelized it partially, but we are working on massive parallelization in preparation for possible future enterprise-scale deployment.

Eventually, we would like the system to be just one part in a system that performs automatic data analysis and meta-data standards enforcement. Essentially all parts of CCD data reduction are standard, and very similar from instrument to instrument, and yet there is no universal package, and no agreed-upon method for passing forward meta-data associated with the data reduction procedures. Astrometric WCS is an exception, but even there it is hard to tell how the WCS was determined, and there is no worked-out methodology for passing forward error analyses. If astronomy is moving towards truly repeatable science, given a sky that is changing with time and not subject to the experimentalist's control, astronomy needs to confront these issues. Our service, *Astrometry.net*, is a baby step towards standardized data analysis and serves, we hope, as an example to those working in other parts of this cluster of problems.

All of the code used in this project has been released under GPL licensing, and is publicly available, along with web services that implement it.[3]

Acknowledgements. It is a pleasure to thank Jon Barron, Doug Finkbeiner, Phil Marshall, and Christopher Stumm for useful discussions. This research made use of the NASA Astrophysics Data System.

Support for this research was provided in part by NASA (grant NAG5-11669) and NSF (grant AST-0428465).

References

Barron, J., Stumm, C., Hogg, D. W., Lang, D., & Roweis, S., 2007, AJ, in press (arXiv:0709.2358)

Calabretta, M. R. & Greisen, E. W., 2002, A&A, 395, 1077

Greisen, E. W. & Calabretta, M. R., 2002, A&A, 395, 1061

Lang, D., Hogg, D. W., Mierle, K., Blanton, M., & Roweis, S., 2007, Science, submitted

Monet, D. G., et al, 2003, AJ, 125, 984

[3] `http://astrometry.net/`