

Jan 28, 2020

Dr. Steven Crawford
AAS Scientific Editor

Re: The Pan-STARRS Data Processing System

Dear Dr. Crawford,

We are resubmitting our article "The Pan-STARRS Data Processing System" after addressing suggestions raised by the referee. We thank the referee for detailed comments and suggestions. Below are our responses (in bold) to the referee's suggestions (plain text).

Sincerely,

Eugene Magnier
University of Hawaii at Manoa
Institute for Astronomy

General Notes

This is a well-written and important technical paper that succeeds admirably at what I consider the most important goal of any pipeline paper: providing a description of the processing steps that are relevant for downstream science users (in this case, by providing the big picture that ties together a number of more detailed papers). With only a handful of minor cleanups (see detailed notes below), I think the paper is ready for publication, and most of my comments represent ideas for improvement that I hope the authors will consider (but should not feel obliged to act on).

My only general concern is that the paper often misses the opportunity to pass on lessons learned to the developers of future pipelines, and this makes much of the detailed description of how the PS1 systems work (particularly in Section 5) feel like it belongs more in operator documentation rather than an article like this one. I suspect a small amount of additional historical context - how different systems evolved over the course of the survey - and commentary on what worked well and what was a regular pain point would go a long way.

In particular, the described system seems to involve a both fair amount of duplication (e.g. multiple databases, sky-tiling systems, and task orchestration layers) and a number of in-house solutions to what seem like fairly general problems (the DVO database and especially the pantask/opihi system stand out in this regard). This is not intended as criticism; I am quite aware that there are many good reasons for both duplication and keeping central components in-house, from deliberately keeping components loosely coupled to taking into account the often-brief shelf-life of off-the-shelf solutions, especially as compared to the duration of a major astronomical survey. But describing *which* of many potential reasons actually played a role in each of various design choices (and which, if any, look less good in hindsight) would make the paper much more interesting.

We have greatly expanded the Conclusion to address these questions, and to identify choices we made which either turned out well or which we would have done differently given changes to the software landscape.

Detailed Notes

Section 2.4

Is the Distribution and Publication system mentioned in the text supposed to be part of Figure 1, either as an umbrella term or a (missing?) component?

We have adjusted this figure to put the publication "customers" at the bottom and added a line to show where the distribution and publication mechanisms interface to these customers.

Section 3.1

This is by no means necessary, but I'm curious to see a table or discussion of what fraction of jobs of various types failed with bad "quality". In other words, how much data could you not get through the pipelines at all, and what was the most sensitive step?

We liked this suggestion and added a subsection 3.12 and a new table (2) to discuss the failure rates.

Section 3.3

Running "Registration" only once for each exposure would seem to prohibit re-running "burnttool" after updating the algorithm for

that - and I'm guessing you didn't get that fully stabilized until after you had already processed some images and learned from the experience. How did that work?

We added a couple of sentences to explain that we used a semi-manual task to re-run just the burntool analysis during development and if the code ever needs to be changed.

Section 3.5

Why a 3rd-order polynomial from chip to focal plane? Wouldn't an affine transform have been sufficient (and more than that degenerate with the focal plane to sky transform)?

We use the higher-order transformation for each chip to capture the small-scale astrometric signal present in the data. One could use an affine transformation for chip-to-focal plane and capture the same signal in a much higher-order model for focal-plane to sky, but that was not our development path. These would be equivalent solutions. (Note that degeneracies exist in both cases). We avoid the degeneracy of the chip positions in the focal plane solution by fitting the local gradient to get the initial distortion solution (and there are certain terms which are held fixed for the focal plane.) We then limit the impact of the degeneracy by fitting the two levels independently and fixing the focal-plane solution after a few iterations. We have added some words to explain some of this, but leave the details to Paper IV.

What makes the masks generated in this step "dynamic"? Are they generated wholly from the reference catalog (i.e. predicting where a ghost will appear based on the position of a bright star)? It seems like the CAMERA step does not utilize any of the pixel data (just the pixel-level masks from CHIP). Is that correct?

This is correct: the dynamic masks are generated from the reference catalog and do not go back to the original pixels. We added a paragraph to clarify.

Section 3.8

Is the selection of which warped images go into a stack driven by human operators, or are there automated systems to launch these jobs, too?

Section 5.2 discusses how both the nightly stacks and large-scale reprocessing campaign stacks are automatically defined. We added some words to refer to this section in 3.8.

Section 3.10

How much of the PSF-convolved galaxy models do you re-fit in forced photometry? If you're fitting more than just the amplitude at that stage, and considering each exposure as independent, you're potentially throwing away a lot of S/N (at least in the many-exposure limit), even if you average later. If you're just fitting the amplitude, the structural parameters are still going to be the ones affected by poor PSFs in the stack.

The galaxy models are not fitted on each warp. Rather, we calculate the normalizations and chi-square values for a grid of galaxy model shape parameters for each warp image. The values for each grid point are combined across all warps to generate a total stack-equivalent grid. At this point, the best parameters are determined from the grid (interpolating to the chi-square minimum). This is mathematically equivalent to simultaneously fitting (via a grid search) the pixels from all warps to a single model, preserving the full signal-to-noise. We have updated the text to add some detail to the description of what is being measured to clarify this point, though most of the information is in paper V.

Section 3.11

transient source -> transient sources

fixed.

Section 4.1.3

I was confused when first encountering the word "files" here because up to this point I had been thinking of the DVO as just another MySQL (or other SQL-ish) database, and I wasn't sure what kind of files were being referred to. I think it'd be helpful to briefly describe the overall architecture of the DVO as (mostly?) spatially sharded files at the beginning of section 4, even if the details of the partitioning aren't described until 4.1.3.

we added a sentence to 4.1.1 to note this point.

Missing punctuation in parenthetical HST GSC reference?

fixed

Section 4.1.4

There's some inconsistency here between "detID" and "det_id" (same for "image"), both referring to measurement IDs in DVO. If those are supposed to be meaningfully different, I'm confused.

In the DVO section (and in the DVO schema), these should all be 'detID' and 'imageID'. In the gpc1 database schema, the underscored versions are used. We have fixed the erroneous det_id and image_id entries in this section.

Section 4.2

I tend to associate the term "ubercal" specifically with the SDSS version of the algorithm that coined the term, and think it probably should be referenced here even if the actual algorithms used are only vaguely similar.

We agree and have added a sentence with reference.

Section 4.3

Is the PSPS database another spatially-shared, file-based database using custom technology, a MySQL database like the Processing Database, or something else? I assume the same system is used at both IFA and MAST?

PSPS is based on MS SQL Server. We have added a bit of description to 4.3.

Section 5.1.1

Apparent typo or missing text: macro ex- its job successfully".

This should have read 'macro exits successfully' ("exits" was being hyphenated). fixed.

Section 5.1.4

"responsible to" -> "responsible for"

fixed

Section 5.2

> Pairing warps together is simplified by the observing strategy in which the same pointing is observed multiple times in a night. By limiting to warp-warp pairs from the same pointing, the problem is significantly reduced from the arbitrary case.

This (as well as the following paragraph) seems to imply that you typically generate differences between images taken in the same night, which of course limits you to detecting only very short-timescale transients and fast-moving objects. I suspect that's just not what you intended to imply, or is the nightly processing really not supposed to find e.g. supernovae?

The wording here was unclear that the nightly processing system generates warp-warp difference images (for asteroids), warp-stack difference images (for 3pi supernovae), and MD nightly stack - reference stacks difference images (for deep MD supernovae). We have updated the text to explain these differences.

Section 5.2

Are the `projection\_cells` described here the same as or related to the DVO partition cells of 4.1.3, or the RINGS.V3 skycells of 3.7?

They are the same as RINGS.V3. we have clarified this and also cleaned up the wording of this paragraph.

This is a more general concept, but it came to a head in this section: I found the use of so many notation styles for different concepts more distracting than helpful. I think I was able to infer that small caps were used for processing stages and non-bold italics were used for database tables, but it wasn't clear why some other stages were written in fixed-width mixed case instead (were these scripts, rather than stages?), or what the use of bold-italic meant (everything else?). I'd recommend either adding a notation legend paragraph early in the paper or just cutting down on the number of styles used.

We agree and have simplified the typography a bit (using only a single face for both db tables and db columns), eliminating the use of boldface. We have also added a paragraph in the introduction section to define the type

faces.

Section 5.3

Was Nebulous just used by the orchestration levels like pantasks, or was it used within the Perl scripts and C programs that constitute the algorithmic steps as well?

Nebulous is used by any level of the software that needs access to a specific file. The c-based processing programs have direct interfaces as do the Perl-based wrappers (ippScripts). We have added a paragraph to explain this.

Was the database used by Nebulous integrated with the Processing Database at all (or even part of the same server)?

These two databases are on separate machines and kept independent. A sentence was added to the end of 6.1 to note this.

It's a bit strange to first encounter what seems like a core part of the data access system this late in the description, given that it would have needed to be updated by all of the processing steps mentioned early. This would of course make more sense if Nebulous is in fact used by the lowest levels of the pipeline and hence a Nebulous database entry is created whenever a file is written to disk.

Our organizational scheme is meant to place the details closest to the science analysis up front and leave the more general systems toward the end, with only a few necessary broad concepts introduced early on for context. Thus section 3 is about the analysis steps and the related programs, section 4 is about the science database and the calibration, section 5 is more generic operations concepts, and section 6 is the computing hardware. Within section 5, the processing organization comes first, while Nebulous is left to the end since it seems (to us) to be very general and should not be driving the science decisions.