Panning and Zooming the Observable Universe with Prefix-Matching Indices and Pixel-Based Overlays

Timothy Luciani*
Department of Computer Science

Brian Cherinka†
Department of Physics & Astronomy

Sean Myers*
Department of Computer Science

Boyu Sun*
Department of Computer Science

W. Michael Wood-Vasey†
Department of Physics & Astronomy

Alexandros Labrinidis*
Department of Computer Science

G. Elisabeta Marai*
Department of Computer Science

University of Pittsburgh

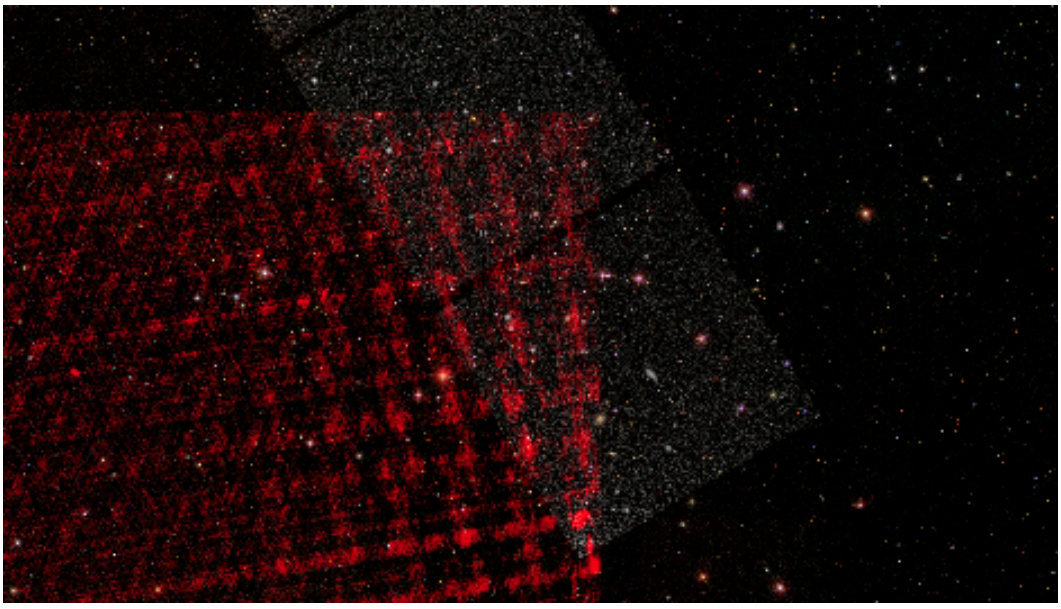


Figure 1: Cross-correlated overlays of optical observations, radio-emission observations, and simulation results from the SDSS sky survey (color-on-black, full-coverage overlay), the FIRST sky survey (red overlay to the left), and the LSST dataset (gray overlay, diagonal). Transparency can be interactively controlled for each overlay, enabling cross-spectrum analysis. A prefix-matching indexing scheme coupled with a web-based client-server architecture allows panning and zooming of gigabit sky panoramas at interactive frame rates.

ABSTRACT

We introduce a web-based, client-server computing infrastructure to assist the interactive navigation of large-scale astronomy observations. Large image datasets are partitioned into a spatial index structure that allows prefix-matching of spatial objects. In conjunction with pixel-based overlays, this approach allows fetching, displaying, panning and zooming of gigabit panoramas of the sky in real time. Images from three sky surveys (SDSS, FIRST and simulated LSST results) are cross-registered and integrated as overlays, allowing cross-spectrum analysis of astronomy observations. The front-end of the infrastructure uses the web technologies WebGL and HTML5 to enable cross-platform, web-based functionality. Our approach attains interactive rendering framerates; its power and flexibility enables us to serve the needs of the astronomy community. Evaluation on a galaxy case study, as well as feedback from domain experts emphasize the benefits of this visual approach to the observational astronomy field.

*e-mail: {tbl8, stm52, sun, labrinid, marai}@cs.pitt.edu
†e-mail: {bac29, wmwv}@pitt.edu

IEEE Symposium on Large Data Analysis and Visualization 2012
October 14 - 15, Seattle, Washington, USA
978-1-4673-4733-4/12/$31.00 ©2012 IEEE

Index Terms: Data Fusion and Integration, Scalability Issues, Zooming and Navigation Techniques, Geographic/Geospatial Visualization

1 Introduction

In the next decade, the scale of astronomical data will reach the petascale. As this transition occurs, astronomers are able to amass large collections of complementary data, ranging from large scale images to spectroscopic measurements. With the insight gained by these observations, researchers can better understand the happenings in our own galaxy by studying similar events in distant ones.

However, due to the transition from gigascale to petascale and the increasing variety of data sources, astronomical workflows are becoming cumbersome. To gather the data needed for a particular study, astronomers query multiple surveys for images, cross-correlate complementary images of the same object or set of objects, search multiple catalogs for potential additional details, then flip back and forth between these details and the image context. This process is tedious, as well as challenging, and can often take hours to complete. As a result, time that could be spent analyzing data is instead spent mining it.

To facilitate the observational astronomy workflows, we propose a web-based visual infrastructure to assist the interactive navigation of large-scale astronomy observations. The infrastructure automatically cross-correlates image data from complementary surveys and allows the visual mining of catalog information. A spatially indexed, client-server backbone allows fetching, displaying, panning and zooming of gigabit panoramas of the sky in real time.

The contributions of this work are as follows: 1) an analysis of the data and tasks specific to the observational astronomy domain; 2) the design of a client-server architecture for the interactive navigation of large scale, complementary astronomy observations; we introduce a prefix-matching indexing scheme and pixel-based overlays to enable the interactive zooming and panning of these data; 3) a web-based, cross-platform implementation of this approach; and 4) the application of this approach to observational astronomy data through a case study.

2 Related Work

Multiple attempts have been made to facilitate the observational astronomy workflows. However, none have fully succeeded in developing an interface that has been readily adopted by the astronomical community for research purposes.

Google Sky [1] is a primarily educational, interactive, scalable view of the Sloan Digital Sky Survey (SDSS). While it provides a friendly and clean interface, the exclusion of multiple surveys is a limiting factor for astronomy researchers. An additional drawback is the inability to integrate and share catalog data from multiple datasets.

The National Virtual Observatory [6] (NVO) is another service designed primarily for aggregating and cross-matching information from multiple surveys. While it provides some form of catalog cross-registration, the NVO has a cumbersome interface which lacks a much-needed interactive visual component.

The WorldWide Telescope [10] is a Microsoft Research, primarily educational project designed to allow users to view the Universe with a large, high resolution image of the sky. It provides multiple maps of the sky, covering a range of wavelengths, however only one of which is visible at a time. The ability to overlay multiple maps and visually cross-match objects is nonexistent. There is also a lack of connectivity with catalogs and other scientific data.

A variety of institutions have created web interfaces for accessing astronomical data, either for querying specific astronomy databases (e.g. SDSS [9], Herschel [2], or the Infrared Science Archive [3], or for aggregating data on many objects from multiple catalogs (e.g., the NASA Extragalactic Database [5] or SIMBAD [7]). All of these interfaces however suffer from the same problems. They either lack a visual interface entirely or they provide only a static sky image to view a few objects at a time. Visual overlays of cross-matched data are non-existent. The user interfaces require a steep learning curve, preventing easy familiarity with the software.

The Millennium Run Simulation [38], the Hubble Volume Project [22], and Intermediate Scale Simulations [32] are large-scale cosmological simulations which allow the visualization of the large-scale structure and evolution of the Universe. These simulations and visualizations provide significant insight into the Universe, and complement observational astronomy.

Attempts to work with gigascale image data have also been made, though none have been applied directly to observational astronomy. Saliency Assisted Navigation identifies areas of interest in gigapixel images [27]. Through detecting abnormalities by filtering regions of interest through preprocessing, discernible locations in a scene can be presented to the user interactively. Kopf et al. [28] and Machiraju et al. [36] have also developed systems for dealing with gigascale and terascale image data. While these systems have complementary strengths in terms of the storage and the scale of the data being manipulated, each was generally designed to facilitate data captured of Earth, and not the galaxies beyond it.

Architectures for multizoom large-scale visualizations have also been explored. The classic Space-Scale Diagrams [25] presents an architecture to address shifting viewports on multi-resolution data. This method has been used in many geospatial applications [16, 24, 34] and it serves as a basis for the navigational approach to our application. However, these applications are not designed to handle the magnitude of data described in this paper and thus require novel couplings with data indexing and storage schemes – such as the geospatial hashing reported in this paper. Reference [8] describes the challenges of indexing astronomy data and summarizes the spatial indexing techniques available. However, many new techniques have been proposed since the publication of this work, as well as of [4], for example, the *Geohash* which we use in our approach.

A step further, ZAME [21] has used GPU-accelerated rendering to deliver interactive framerates to multi-scale visualizations. While the ZAME approach is beneficial to client-based applications that are able to provide full graphics support, web-based applications like ours pose more stringent constraints (e.g., limits on how many textures can be passed to a shader at once.)

Furthermore, panning and zooming is a common problem among geospatial applications [12, 30, 31]. While many of these works focus on interactive techniques relevant to this project, the focus of this paper is an efficient architecture for viewing and cross-correlating gigabit image data.

Presenting multivariate data visually is also common among geospatial applications. Oriented Slivers provides a method to visualize multivariate information simultaneously on a single 2D plane, but becomes easily cluttered as the dimensionality of the data rises [40]. Heat maps [23] alleviate this problem by assigning each value a temperature and producing a color map based on the resulting heat combinations. While particularly beneficial in giving a general overview of data over large areas, heat maps are less useful in identifying individual data points. The approach we adopted, Data Driven Spots (DDS) addresses both of these concerns via a pixel-based visualization [17]. Trends in spatially dense data are easily visualized without the clutter of large glyphs, while multivariate information is represented by assigning to each variable unique colors and combining them in the final result.

3 Domain Analysis

Astronomy surveys cover a wide area of sky by acquiring many smaller images, some of which may overlap, over their targeted region. A particular survey usually only covers a small fraction of the whole sky; although the advent of large telescopes like the Large Synoptic Survey Telescope will change dramatically over the next decade the scale of these surveys. Different surveys may or may

not cover the same area of sky, resulting in possibly completely disparate or overlapping datasets. The Extended Groth Strip [20] for example, is one of the most observed regions of the sky, with upwards of eight different telescopes/surveys collecting data, making this region rich with multi-wavelength observations.

In our experiments we use data from three surveys, the Sloan Digital Sky Survey (SDSS), the Faint Images of the Radio Sky at Twenty Centimeters (FIRST), and simulated results from the Large Synoptic Survey Telescope (LSST). **SDSS** is an optical, wide-field, survey covering a quarter of the sky. Over the past ten years, it has imaged a half a billion galaxies and taken spectra for a half a million, providing a massive leap in the amount of astronomical data (roughly 15 TB raw image data) compared to previous surveys [39, 11]. **FIRST** is a radio survey of the sky, following the same path as SDSS. FIRST also covers about a quarter of the sky and contains roughly a million discrete radio sources [14, 15]. **LSST** is a future optical full-sky survey, along the same lines as SDSS but of unsurpassed scale. It will cover ∼20,000 sq. degrees of the sky, scanning the entire sky every 3 nights, in six photometric bands. LSST will image approximately 3 billion galaxies and will archive about 6.8 PB of images a year. As LSST has yet to acquire sky images, the LSST project has generated simulations of images of the sky to mimic and observe the observational prowess of the survey. Seven fields (189 unique image files), each covering ∼10 sq. degrees, have been simulated.

3.1 Data Analysis

Astronomers use a variety of data formats to collect, organize, analyze, and share information about the observable Universe. The most common formats used are images and catalogs.

Images are rectangular snapshots of regions of the sky, typically labeled with the spatial location of the region. Images in astronomy are usually stored as a Flexible Image Transport System (FITS) file. FITS files are designed specifically for scientific data and thus offer many advantages over other formats. FITS files store image metadata in a human-readable ASCII header, and often include technical telescope details from when the image was taken. FITS files are extremely versatile, capable of storing non-image data such as spectra, 3D data cubes, multi-table databases, and catalog data.

Since the observable Universe is projected onto a sphere, the angle is the most natural unit to use in measuring positions of objects on the sky. Astronomers describe the coordinates of objects in Right Ascension (RA) and Declination (Dec). Similar to how longitude and latitude describe positions of objects on the Earth from a given reference point, right ascension and declination mark the position, in degrees, of an object with respect to the celestial equator. Smaller units of angle are arcminutes and arcseconds. The resolution of an image is often given in units of arcseconds/pixel. This is the scale of the image and describes how much detail is spread out over the pixels of the telescope's camera.

Catalogs index all of the objects in a particular set of images. The catalogs contain location information for every object imaged, along with any properties collected or calculated from the observations (e.g. brightness, mass). Each object in the catalog receives a unique identifier, which can be used when cross-matching. Catalogs generated from the same survey will use the same unique object identifiers, making cross-matching within a survey relatively straightforward. However, as is often the case, when the same object is observed in different surveys, it is usually assigned different identifiers for each catalog; this labeling makes cross-survey matching a non-trivial task. In these cases, cross-matching is performed on location, which is more challenging, as position accuracy depends on the resolution of each telescope/survey. The observed objects in each survey may not exactly overlay on the sky but may still be physically associated with each other (e.g. radio jets emanating from the center of a galaxy).

In summary, the observational astronomy domain features large-scale, distributed, overlapping, multivariate datasets consisting of both image and catalog data; while the data is indexed by object location, uncertainties in the measured position make visual correlation particularly useful.

3.2 Task Analysis

The Universe is a complex structure with many physical processes governing its formation and evolution. To fully understand sky object dynamics, it is necessary to build a complete picture through observations over the entire electromagnetic spectrum. However, only certain regions of the electromagnetic spectrum are observable from the ground. While space-based telescopes can observe the full spectrum, cost and technical challenges preclude the design of a single all-purpose telescope. Instead, astronomers rely on many telescopes that observe specific regions of the electromagnetic spectrum and then cross-match the datasets to identify the same objects in each one. Astronomers must manually seek out all data related to a particular object if they wish to fully understand this picture, which is often a cumbersome task.

Astronomical processes occur on many length scales, from small-scale features such as dust particles to large-scale features such as clusters and superclusters of galaxies. With observations usually pertaining to a specific scale at a time, it can be easy to lose the big picture of how all these processes are connected. Therefore it is advantageous to stitch multiple observations together to create a seamless zoomable image. This would allow astronomers to visually explore how stellar and galactic physical processes relate to the larger picture of galaxy groups and clusters.

In summary, the observational astronomer workflow involves queries of the type *what – where – correlated-with-what* over multiple surveys at multiple scales.

4 Design and Implementation

Based on the domain data and task analysis, we design a pipeline for the interactive exploration of the observable Universe. Given the multiple, distributed sources of data, and the scale of the data, we follow a client-server model (Fig. 2). The first component of the pipeline is an offline module for preprocessing astronomical images so they are in a mutually-compatible format (e.g., with respect to projections.) The images are assigned prefixes, then mipmapped, organized and stored in a spatially indexed, prefix-matching structure (*Geohash*). As the user navigates the sky, the client queries the SDSS server, as well as the *Geohash* and catalog database with the current field-of-view information. The images and catalog information are returned to the client, who presents the images to the user in the form of overlays. The online process is user-demand driven and occurs in real time.

4.1 Data Retrieval and Preprocessing

Image data retrieval and preprocessing is performed on a per-survey basis, to ensure dataset compatibility. Depending on the survey, astronomy images can appear in different map projections. While about 25 different projections are common, the number of possible projections is not limited (Fig. 3). To ensure survey and dataset compatibility, we extract the world coordinates (e.g. RA/Dec) for every image such that images from multiple surveys will be properly aligned. The World Coordinate System (WCS) [26] handles the conversions from image coordinates to world coordinates for many different projection schemes. We then project the images on a viewing sphere. The sphere is an abstraction of the sky as viewed from Earth, with the camera located at the center of the sphere.

Image data for the SDSS survey are stored remotely through various data releases; each release consists of FITS files that can be converted to JPEGs via a "cutout" request. To access optical images from the survey, a query (rsync/wget) is sent to the SDSS Data

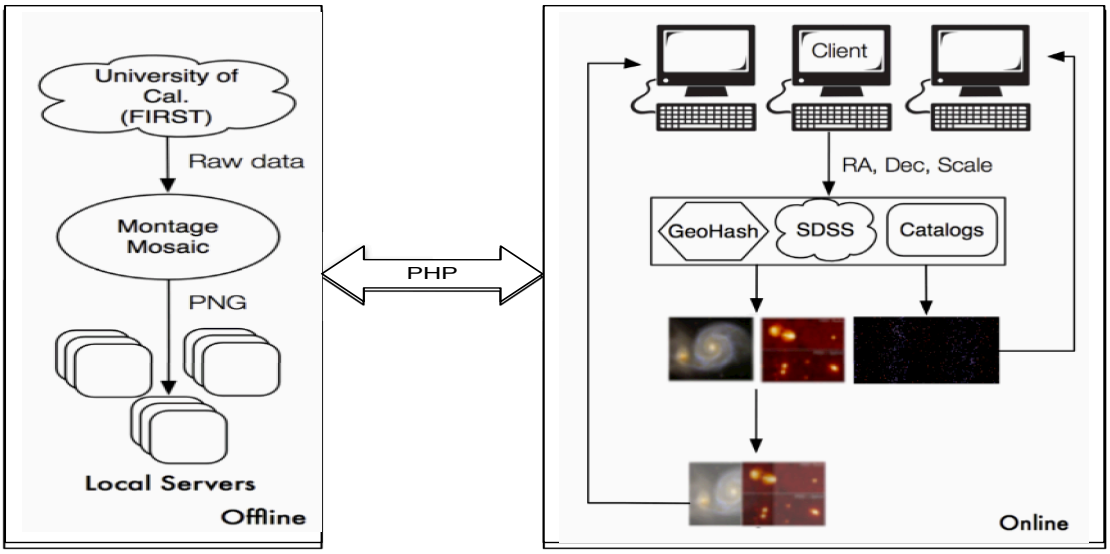


Figure 2: Pipeline for the interactive exploration of the observable Universe. An offline module preprocesses raw astronomical datasets so that they are in a mutually-compatible format. The offline images are assigned prefixes, then mipmapped, organized and stored in a spatially indexed, prefix-matching structure (*Geohash*). A client-server backbone governs the querying and displaying of the data. The server queries the SDSS server, as well as the *Geohash* and catalog database with the current field-of-view information. The images and catalog information are returned to the client. The images are finally being composited together to form the overlay that the client will display.

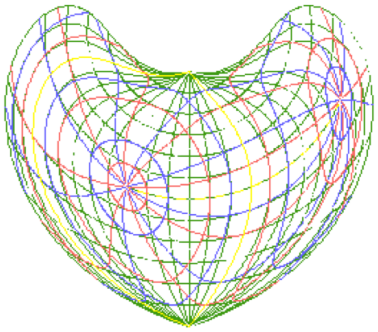


Figure 3: Depending on the survey, astronomical data can appear in different map projections (shown in different colors above). While about 25 different projections are common to astronomy, there is no limit to the number of possible projections available.

Access Server containing the RA, Dec, scale and zoom-level of the current area being viewed. The SDSS server then returns an image based on the parameters it received, dynamically scaled to the user's vantage point. SDSS's projection scheme is Gnomic (TAN), an azimuthal projection, given by equations 54 and 55 in [18]. We use custom code to convert these images to WCS.

In contrast, to ensure image content compatibility between surveys, other survey (e.g. FIRST) images are fetched a-priori and stored locally. There are 30,500 FIRST image files, requiring 300GB storage. To ensure projection compatibility, we generate a mathematical 2D grid spanning the viewing sphere. We assign to each of grid region an RA/Dec name corresponding to its centroid position in the sky. The raw data and the grid-based name of each region are then passed to the third-party tool Montage Mosaic [29]. The Montage tool extracts the image data from the raw FITS format; the resulting images are named according to the RA/Dec center of the image. The FIRST radio survey uses a Slant Orthographic projection (SIN), also an azimuthal projection, and is given by equations 59 and 60 in [18]. To reduce the rendering load when a large area of the sky is being viewed, we use custom Matlab code to generate a pyramid of image tiles, with four levels (number of levels empirically determined for demonstration purposes) of decreasing resolution. The tiles are obtained through repeated Gaussian filtering followed by subsampling (Fig. 4). This entire preprocess is performed once for the dataset, averaging a 30 second generation time per tile. Once the local images are preprocessed, they are spatially indexed for quick access (Section 4.2).

Similarly to FIRST, LSST simulated images are downloaded locally through web-based queries (rsync/wget). We use Montage Mosaic to convert the LSST FITS images into JPEG files. The LSST simulated dataset uses the TAN projection scheme, similar to SDSS. Custom code converts these TAN images to WCS. LSST images are not locally mapped to multiple levels of detail, since the domain experts anticipate an online LSST service similar to SDSS. Because this small LSST test dataset is privately owned and accessed, the entire procedure is done a-priori and all images are stored locally.

Catalog data is retrieved from the SDSS server via a general SQL query using the server online interface for catalog access, CASJOBS, and then stored locally into a mySQL database.

4.2 Prefix-Matching Geohash

For fast retrieval of the images given a bounding box, we use the geospatial index powered by MongoDB [19], an open source document-oriented NoSQL database system. The coordinates of the image tiles are hashed into string-based prefixes using *Geohashing*. *Geohash* is a hash table where the keys are the coordinate sets, and the values are strings; similar coordinates with more or less significant figures share a common prefix in their geohash.

The hashed strings are then stored in an index structure which is a standard B-tree [13] (Fig. 5). In this simple and effective indexing approach, coordinates close to each other will often have *Geohash* codes that share longer common prefixes. As a result, the images are naturally grouped and nearest neighbor queries and

Figure 4: Four decreasing level-of-detail versions of a sky image tile (for demonstration purposes, an SDSS image). As the user zooms further out from a source less details are needed to convey information. This image pyramid demonstrates this effect with four levels of resolution, starting with a 1024x1024 pixel image and going down to a 128x128 pixel image.

range queries could be answered very fast. As an example, if we would like to find neighbors of a given point we could simply issue a query such as "which other points have a *Geohash* that starts with the prefix '8fb25' ." Since the precision of the hash results is adjustable, we can simply implement zooming in and out by using different *Geohash* prefix lengths corresponding to different resolution.

One limitation with this encoding, while fast, is that prefix lookups do not give exact results, especially around bit-flip areas. However, this problem was solved by doing a grid-neighbor search after the initial prefix scan to pick up any straggler points. This generally ensures that performance remains very high while providing correct results.

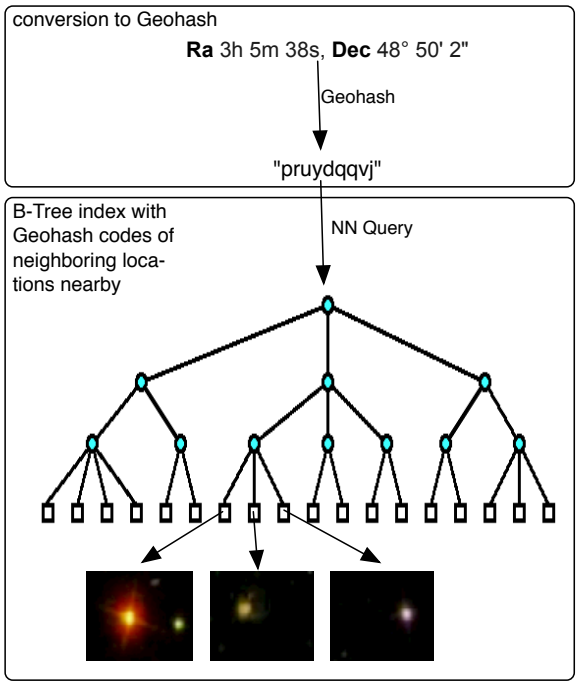


Figure 5: Illustration of Geospatial indexing, which combines *Geohash* and B-tree indexing, for fast retrieval of neighboring image files.

4.3 Online Overlays

To overlay sky images for viewing, the next step is to convert the world coordinates into the native WebGL graphics coordinates. The standard WCS Cartesian coordinate system is a right-handed coordinate system with the positive x, y, and z axes pointing outward, to the right, and up, respectively. For a WCS spherical coordinate system, the angles ϕ and θ are used to define the location of a point. The angle ϕ increases clockwise starting from the positive z-axis. The angle θ increases counter-clockwise starting from the positive x-axis. In contrast, in the right-handed WebGL graphics coordinate system the Cartesian axes are rotated with respect to the WCS Cartesian coordinate system. The WebGL positive x, y, z, axes point to the right, up, and outwards, respectively. The angle ϕ increases counter-clockwise starting from the positive x-axis, and the angle θ increases counter clockwise starting from the negative y-axis. Due to these differences between the standard and WebGL coordinate systems, a transformation has to be applied to convert from the world RA/Dec coordinates to the WebGL spherical and Cartesian graphics coordinates:

$$\phi = (90^\circ - Dec) \quad (1)$$

$$\theta = (270^\circ - RA) + 360^\circ \quad ; when \quad RA > 270^\circ \quad (2)$$

$$\theta = (270^\circ - RA) \quad ; when \quad RA \leq 270^\circ \quad (3)$$

$$x = \sin(\phi) * \cos(\theta) \quad (4)$$

$$y = \cos(\theta) \quad (5)$$

$$z = \sin(\phi) * \sin(\theta) \quad (6)$$

Following the above transformation, sky images are mapped to the unit viewing sphere. To create a visual abstraction of multiple data sources, pixels are further composited online into transparent overlays using the WebGL GLSL fragment shader. WebGL has the advantage of performing computations exclusively on the client machine GPU, leaving the CPU available for user interaction. To optimize Javascript memory use and texture loading we implement local garbage collection; this optimization helps prevent HTML5 from bottlenecking interaction while rendering texture objects.

Aside from SDSS, FIRST, and LSST, more specific overlays can be created from searches over catalogs. Custom overlays are generated via the same client-server model implementation. The server receives the client requirements — desired resolution of the output image, the minimum and maximum RA/Dec values, attribute thresholds, desired color-mapping, and any other optional filters on the other parameters present in the catalog database — and submits them as a query to the catalog database. Upon receiving a response from the database, the server creates one or more new PNG images and proceeds to draw on the image each tuple returned by the query. The RA/Dec columns in each tuple are used to position the drawing within the image. The closer the value of the key attribute is to the maximum threshold, the brighter the color will be drawn at that point. All data tuples are added to the images, which are then compressed and returned over the network to the client application.

The client receives the images, cross-registers them, computes the pixel-based overlay and displays it. The cross-registration and overlay operation involves projecting each image onto the sky, zooming in and out to account for telescope parameters, changing transparency etc. The client finally renders the scene, where the visualization scenegraph consists of the viewing sphere with the camera at the center looking out.

4.4 Interactive Panning and Zooming

To enable interactive panning and zooming, the client implements a View Manager. The manager maintains the current viewing location and parameters, as well as a list of all the image tiles currently in the view. The manager sends notifications out to the overlay

server when needed. These notifications are either requests for new images, or display context notifications.

Panning the view maps mouse motion to updates in the view range. Zooming also computes and maps the new scale to updates in the viewing range. If the updated range covers images that have not been fetched yet, the manager requests for those image tiles to be sent out to all overlays that are listening to the current view. Each of those requests is handled asynchronously.

The online modules of the pipeline are implemented using HTML5 Canvas, Javascript, and WebGL.

5 Results

In this section we measure the performance of our approach. We first measure the precomputation of the FIRST images stored on the backend of the pipeline; conversion to raw images, mosaicking, and reprojection. Next we report rendering speeds with varying amounts of image data presented to the user. We then present a case study where domain experts perform an analysis with our tool and reports their findings. Finally, we report feedback from repeated evaluation with a group of five astronomy researchers, as well as from two astronomy workshops; each workshop featured more than 30 participants.

5.1 Preprocessing

Precomputation of the FIRST images is the most time consuming part of the pipeline; however, this stage only has to be done once when the data is first acquired for a survey. Each image takes between 30 and 40 seconds to generate, with 20 seconds of the process dedicated to reprojecting the image into the WCS map projection. Depending on the sky coverage of the survey, this preprocessing can take anywhere from a week to a month. In the case of FIRST, it took fifteen days to compute all of the images needed for tiles using a server running CentOS 6, Dual 6 Core processor at 24GHz, and 32 GB RAM.

5.2 Performance

The initial data retrieval and loading stage varies depending on the source the images arrive from. To retrieve FIRST images from our server, a loading time of 50-200ms is incurred for sizes varying between 400-700 KB. Retrieving LSST images from our server incurs a loading time between 200-400ms with sizes varying between 4-5 MB. Finally, SDSS loading times are slightly higher, typically incurring 750-1250ms with sizes varying between 60-70 KB. These speeds can vary greatly depending on the bandwidth and load of the SDSS servers at the time of use.

Once the images are fetched, the rendering speed hovers at 45 frames per second on a Windows 7 Machine, Quad Core i5, 16 GB RAM. This allows interactive panning and zooming to regions of interest. Our web-based implementation has been tested on multiple browsers such as Safari, Chrome and Firefox.

5.3 Case Study: UGC 08782 - A Dusty Elliptical

Figure 6 shows how the cross-correlation and interactive visual navigation of SDSS and FIRST can be used in tandem for immediate gains in astronomy. Two of our co-authors are senior astronomy researchers and provide the following case study and feedback.

Figure 6(a) shows an optical image from UGC 08782, a bright elliptical galaxy at a redshift of 0.045. The morphology of this galaxy was originally ambiguous between a spiral and a dusty elliptical, exhibiting dust lanes and disturbed morphological features. Dusty ellipticals are often seen to show signatures of an active galactic nucleus (AGN) [33, 37]. Some of these AGN exhibit jets, which tend to be perpendicular to the dust lanes. One way to test if UGC 08782 fits these trends is by checking its SDSS spectrum, viewing the optical image, and searching for radio counterparts [35]. Its SDSS optical spectrum exhibits strong narrow emission lines, indicative of a galaxy with a highly active nucleus. If the galaxy has an AGN, then its central black hole may be ejecting massive amounts of energy into the surrounding medium as jets. These jets would emit radiation in the radio, that should be detectable in a radio survey. Figure 6(b) shows radio observations from the FIRST survey of the same region, which detected several interesting features. The image in the radio looks quite different. There is a single bright point where the optical galaxy ought to be and two bright patches extending to the upper right.

Due to the differing resolutions and sensitivities of the surveys, it is unclear looking at the individual images whether the FIRST emission is from a unique object or associated with UGC 08782. Normally, associating the FIRST emission with an optical counterpart would require manually searching optical catalogs for nearby objects and match on position, ranking by closest proximity. When the images are viewed together (Fig. 6(c)), the association between these two sources from different surveys is immediate. The bright radio point source lines up on the center of the optical galaxy, as it would if it were the nucleus of the galaxy. The two patches of radio emission in the upper right appear to emanate from the central point source, as a radio jet might. Not only does the overlay allow for a more efficient cross-matching, it also provides a nice framework for understanding the physical processes observed in each survey and how those processes are connected to one another.

5.4 Informal Feedback

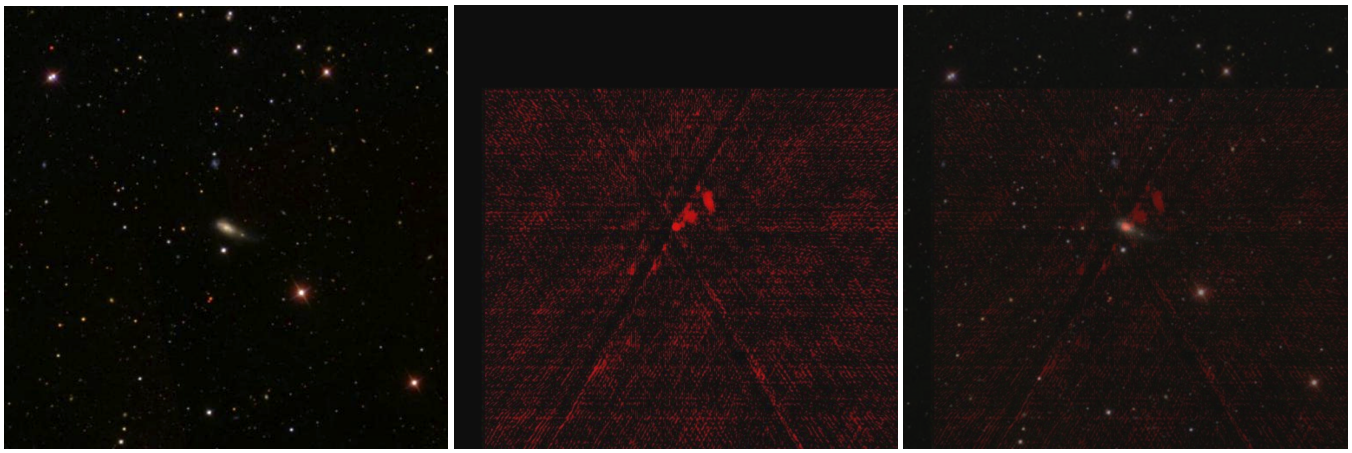
Informal feedback from repeat evaluation meetings showed enthusiasm for the tool. The domain experts considered the approach "an exciting beginning towards a tool for visualizing all-sky surveys. Many of the tools required have been implemented effectively." The ability to compare images of the sky taken at different wavelengths simultaneously and to visually query catalogs was particularly appreciated, while the interactive navigation was considered on par with the much appreciated Google Sky interface. The researchers are eager to use the tool in their research and in classrooms.

The workshop expert-users particularly appreciated the ability to combine separate sources of information without having to resort to cumbersome, external tools for image processing. As shown in the example in Fig. 7, overlaying catalog search results visually further enables queries of the *what – where – correlated-with-what* type. In this example, more than 800 points resulting from searches over the Sloan Digital Sky Survey catalog are visualized efficiently using pixel-based overlays: two query results based on two different attributes are overlaid (red for redshift, blue for the focal ratio of the telescope; brighter intensities correspond to greater values), revealing vertical spatial patterns in conjunction to attribute overlaps. Figure 1 further shows three cross-correlated overlays (partial coverage shown in the figure solely for non-interactive illustration purposes) of optical observations, radio-emission observations, and simulation results from the SDSS sky survey, the FIRST sky survey, and the LSST dataset. Transparency can be interactively controlled for each overlay, enabling cross-spectrum analysis. The researchers are interested in applying this prototype to specific problems such as browsing large sets of objects and galaxy identification. Several astronomy research groups have expressed further interest in integrating their data with our tool.

6 Discussion and Conclusion

Our approach enables the visual cross-correlation of sky surveys taken at different wavelengths, as well as the visual querying of catalogs. Furthermore, the combination of prefix-matching indexing, a client-server backbone, and of pixel-based overlays makes possible the interactive exploration of large scale, complementary astronomy observations.

Our results show that pixel-based overlays and geohashing have the potential to generate scalable, interactive, graphical representa-



(a) SDSS (b) FIRST (c) Combined Overlay

Figure 6: Images of UGC 08782 from two surveys. Figure 6(a) shows an optical image of the galaxy from the SDSS while Figure 6(b) shows a radio image of the same galaxy from FIRST. When overlaid in Figure 6(c), the connection between the two as radio emission emanating as jets from the central black hole of the galaxy becomes immediately clear.

tions of astronomy data. This approach may allow us to overcome bandwidth and screen-space current limitations in astronomy visualization. The advantages of this approach are its versatility and visual scalability (to the pixel level), enabling the visual analysis of large datasets. The resulting versatility allows for flexible control over the visualization and the client-side scripts. Accessing graphics hardware through WebGL further provides the users with a rich, graphics-accelerated web experience.

Finally, evaluation on a case study, as well as overwhelmingly positive feedback from astronomers emphasize the benefits of this visual approach to the observational astronomy field. Spatially indexing images of the sky enables fast access to surveys and interactive rendering rates. With real-time rendering speeds, astronomers are able to identify regions of interest and analyze them without the hassle of having to travel to multiple surveys to manually collect their data.

In terms of limitations, our application speed depends on the preprocessing stage. Having the ability to store the images locally during this stage comes at the steep price of memory, and as surveys become larger this will no longer be a viable option. However, relying on streaming the data from remote sources is also a concern as certain surveys do not provide programmatic access to their images. While the FITS file could be transferred to the user per request, the time for preprocessing before displaying would outweigh the gains of the technique. This topic remains an open research question.

In conclusion, we have introduced a novel approach to assist the interactive exploration and analysis of large-scale observational astronomy datasets. From a technical perspective, we contribute a novel computing infrastructure to cross-register, cache, index, and present large-scale geospatial data at interactive rates. In our web-based approach, large image datasets are partitioned into a spatial index structure that allows prefix-matching of spatial objects and regions. In conjunction with pixel-based overlays, this approach allows fetching, displaying, panning and zooming of gigabit panoramas of the sky in real time. In our implementation, images from three surveys (SDSS, FIRST, and LSST), and catalog search results were visually cross-registered and integrated as overlays, allowing cross-spectrum analysis of astronomy observations.

From the application end, we contribute an analysis and model of the observational astronomy domain, as well as a case study and an evaluation from domain experts. Astronomer feedback and testing indicates that our approach matches the interactivity of state-of-the-art, corporate educational tools, while having the power and flexibility needed to serve the observational astronomy research community. Being able to quickly aggregate and overlay data from multiple surveys brings immediate clarity to inherently complex phenomena, reducing time spent managing the data while allocating more time for science.

Acknowledgments

Research funded through NSF-OIA-1028162 and NSF-IIS-0952720. Thanks to our collaborators Jeffrey Newman, Arthur Kosowski, Daniel Oliphant, Rebecca Hachey, Joseph Cavanaugh, Anja Weyant, Panickos Neophytou, Roxana Gheorghiu, Panos K. Chrysanthis, Matthew Liegey, Matthew Seiler, the AEGIS and LSST communities, the anonymous reviewers and the Pitt VisLab for feedback and interesting discussions. Further acknowledgments to Noel Gorelick and Jeremy Brewer, for generously sharing their Google Sky development experience.

References

[1] Google Sky. http://www.google.com/sky/.
[2] Herschel Space Telescope. http://herschel.esac.esa.int/Science_Archive.shtml/.
[3] Infrared Science Archive. http://irsa.ipac.caltech.edu/applications/Gator/.
[4] Mongo Db. http://www.mongodb.org/display/DOCS/Geospatial+Indexing.
[5] Nasa Extragalactic Database. http://ned.ipac.caltech.edu/.
[6] National Virtual Observatory. http://www.us-vo.org/.
[7] Simbad Astronomical Database. http://simbad.u-strasbg.fr/simbad/.
[8] Sky Index. http://www.star.le.ac.uk/ cgp/ag/skyindex.html.
[9] Sloan Digital Sky Survey. http://www.sdss.org/dr7/.
[10] The Worldwide Telescope. http://www.worldwidetelescope.org/.
[11] K. N. Abazajian, J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam, C. Allende Prieto, D. An, K. S. J. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall, and et al. The Seventh Data Release of the Sloan Digital Sky Survey. *The Astrophysical Journal Supplement*, 182:543–558, June 2009.
[12] C. Appert, O. Chapuis, and E. Pietriga. High-Precision Magnification Lenses. pages 273–282, Apr. 2010.
[13] R. Bayer. Binary b-trees for virtual memory. In *Proceedings of the 1971 ACM SIGFIDET (now SIGMOD) Workshop on Data Description, Access and Control*, SIGFIDET '71, pages 219–235, New York, NY, USA, 1971. ACM.

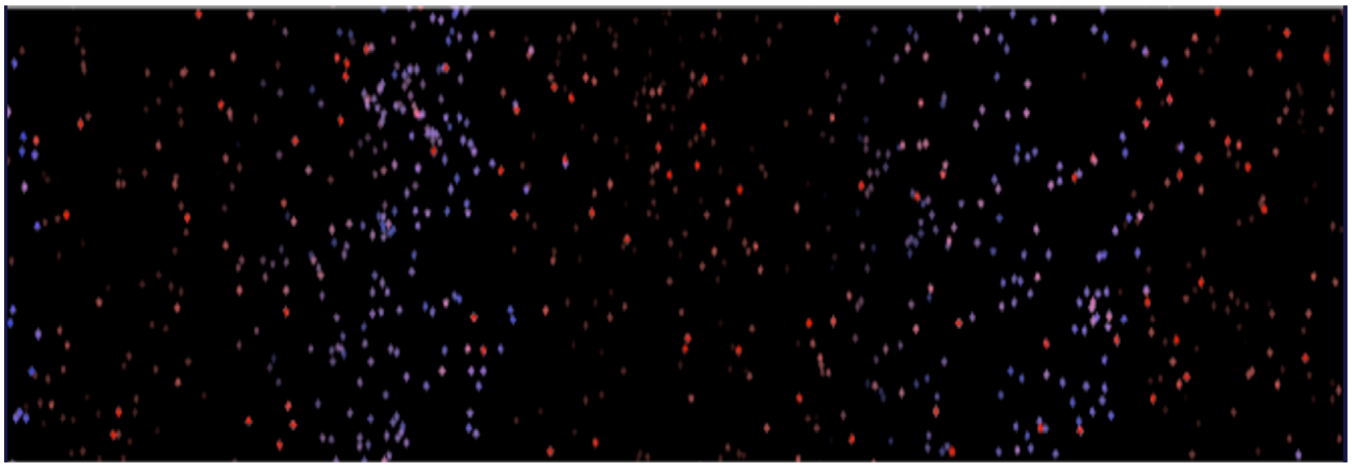


Figure 7: 831 points resulting from searches over the Sloan Digital Sky Survey catalog database are visualized efficiently using pixel-based overlays. Two query results based on two different attributes are overlaid (red for redshift, blue for the focal ratio of the telescope; brighter intensities correspond to greater values), revealing spatial patterns in conjunction to attribute overlaps.

[14] R. H. Becker, R. L. White, and D. J. Helfand. The VLA's FIRST Survey. In D. R. Crabtree, R. J. Hanisch, and J. Barnes, editors, *Astronomical Data Analysis Software and Systems III*, volume 61 of *Astronomical Society of the Pacific Conference Series*, page 165, 1994.

[15] R. H. Becker, R. L. White, and D. J. Helfand. The FIRST Survey: Faint Images of the Radio Sky at Twenty Centimeters. *The Astrophysical Journal*, 450:559, Sept. 1995.

[16] B. B. Bederson and J. D. Hollun. Pad++: A zooming graphical interface for exploring alternate interface physics. 1994.

[17] A. A. Bokinsky. Multivariate data visualization with data-driven spots. 2003. AAI3100231.

[18] M. R. Calabretta and E. W. Greisen. Representations of celestial coordinates in FITS. *Astronomy and Astrophysics*, 395:1077–1122, Dec. 2002.

[19] K. Chodorow and M. Dirolf. *MongoDB: The Definitive Guide*. O'Reilly Media.

[20] M. Davis, P. Guhathakurta, N. P. Konidaris, J. A. Newman, M. L. N. Ashby, A. D. Biggs, P. Barmby, K. Bundy, S. C. Chapman, A. L. Coil, C. J. Conselice, M. C. Cooper, D. J. Croton, and et al. The All-Wavelength Extended Groth Strip International Survey (AEGIS) Data Sets. *The Astrophysical Journal Letters*, 660:L1–L6, may 2007.

[21] N. Elmqvist, T.-N. Do, H. Goodell, N. Henry, and J.-D. Fekete. Zame: Interactive large-scale graph visualization. pages 215–222, 2008.

[22] A. E. Evrard, T. J. MacFarland, H. M. P. Couchman, J. M. Colberg, N. Yoshida, S. D. M. White, A. Jenkins, C. S. Frenk, F. R. Pearce, J. A. Peacock, and P. A. Thomas. Galaxy Clusters in Hubble Volume Simulations: Cosmological Constraints from Sky Survey Populations. *The Astrophysical Journal*, 573:7–36, July 2002.

[23] D. Fisher. Hotmap: Looking at geographical attention. 2007.

[24] G. W. Furnas. Generalized fisheye views. *SIGCHI Bull.*, 17(4):16–23, Apr. 1986.

[25] G. W. Furnas and B. B. Bederson. Space-scale diagrams: Understanding multiscale interfaces. pages 234–241, 1995.

[26] E. W. Greisen and M. R. Calabretta. Representations of world coordinates in fits. *Astron. Astrophys.*, 395:1061–1075, Dec. 2002.

[27] C. Ip and A. Varshney. Saliency-assisted navigation of very large landscape images. In *IEEE Visualization*.

[28] O. D. J. Kopf, M. Uyttendaele and M. Cohen. Capturing and viewing gigapixel images. In *ACM Trans. Graphics*, 2007.

[29] J. C. Jacob, D. S. Katz, G. B. Berriman, J. C. Good, A. C. Laity, E. Deelman, C. Kesselman, G. Singh, M. Su, T. A. Prince, and R. Williams. Montage; a grid portal and software toolkit for science; grade astronomical image mosaicking. *Int. J. Comput. Sci. Eng.*, 4(2):73–87, July 2009.

[30] W. Javed, S. Ghani, and N. Elmqvist. Gravnav: using a gravity model for multi-scale navigation. pages 217–224, 2012.

[31] W. Javed, S. Ghani, and N. Elmqvist. Polyzoom: Multiscale and multifocus exploration in 2d visual spaces. pages 287–296, 2012.

[32] A. Jenkins, C. S. Frenk, F. R. Pearce, P. A. Thomas, J. M. Colberg, S. D. M. White, H. M. P. Couchman, J. A. Peacock, G. Efstathiou, and A. H. Nelson. Evolution of Structure in Cold Dark Matter Universes. *The Astrophysical Journal*, 499:20, May 1998.

[33] C. G. Kotanyi and R. D. Ekers. Radio galaxies with dust lanes. *Astronomy and Astrophysics*, 73:L1–L3, Mar. 1979.

[34] H. Lieberman. Powers of ten thousand: Navigating in large information spaces. pages 15–16, 1994.

[35] C. Moellenhoff, E. Hummel, and R. Bender. Optical and radio morphology of elliptical dust-lane galaxies - Comparison between CCD images and VLA maps. *Astronomy and Astrophysics*, 255:35–48, Feb. 1992.

[36] D. T. R. Machiraju, J. E. Fowler and a. B. S. W. Evita: Efficient visualization and interrogation of tera-scale data. *Data mining for scientific and engineering applications*, 2001.

[37] S. S. Shabala, Y.-S. Ting, S. Kaviraj, C. Lintott, R. M. Crockett, J. Silk, M. Sarzi, K. Schawinski, S. P. Bamford, and E. Edmondson. Galaxy Zoo: Dust lane early-type galaxies are tracers of recent, gas-rich minor mergers. *ArXiv e-prints*, July 2011.

[38] V. Springel, S. D. M. White, A. Jenkins, C. S. Frenk, N. Yoshida, L. Gao, J. Navarro, R. Thacker, D. Croton, J. Helly, J. A. Peacock, S. Cole, P. Thomas, H. Couchman, A. Evrard, J. Colberg, and F. Pearce. Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature*, 435:629–636, June 2005.

[39] C. Stoughton, R. H. Lupton, M. Bernardi, M. R. Blanton, S. Burles, F. J. Castander, A. J. Connolly, D. J. Eisenstein, J. A. Frieman, G. S. Hennessy, R. B. Hindsley, Ž. Ivezić, S. Kent, P. Z. Kunszt, B. C. Lee, A. Meiksin, J. A. Munn, H. J. Newberg, R. C. Nichol, T. Nicinski, J. R. Pier, G. T. Richards, M. W. Richmond, D. J. Schlegel, J. A. Smith, M. A. Strauss, M. SubbaRao, A. S. Szalay, A. R. Thakar, D. L. Tucker, D. E. Vanden Berk, B. Yanny, and et al. Sloan Digital Sky Survey: Early Data Release. *The Astronomical Journal*, 123:485–548, Jan. 2002.

[40] C. Weigle, W. Emigh, G. Liu, R. M. T. II, J. T. Enns, and C. G. Healey. Oriented sliver textures: A technique for local value estimation of multiple scalar fields. pages 163–170, 2000.