Automated Search for Lyman-alpha Emitters in the DEEP3 Galaxy Redshift Survey

Victoria Dean

Castilleja School

Part I: Personal

I have been interested in math and science for as long as I can remember, beginning with inspiration from my dad, a software engineer. When I was five years old, I drew a picture of me in front of a computer, captioned "When I grow up, I will be a computer worker." Since then, I have wanted to be a biotechnologist, a math teacher, and a mechanical engineer, but I have always been interested in math and science as universal languages and opportunities to change the world.

It wasn't until my sophomore year in high school, however, that I became interested in astronomy. I had heard from other students at my school about the Science Internship Program, which gives research opportunities to high school students and is run by UC Santa Cruz Astronomy and Astrophysics Professor Raja GuhaThakurta. I discussed possible projects with Raja, who became my mentor, and ultimately decided on my project, a search for very distant galaxies, because I thought it would be exciting to look back in time. I continued my research for two years, completing most of it over two summers at UC Santa Cruz. During the first summer, my project focused on manual search, which was exciting because I discovered several new Lyman Alpha Emitters, the type of galaxy I was looking for. However, this process of manual search was quite time-consuming, so when I came back the next summer, I combined the project with my passion for efficiency and computer science and worked on an automated search algorithm. This proved to be highly rewarding and taught me the value of an interdisciplinary scientific mindset.

I would like to give special thanks to Raja GuhaThakurta for being my mentor on this project. Without him, this project would not have been possible, and I am so appreciative of the time he spent working with me in addition to his many other high school interns. Working with Raja, who is so passionate and knowledgeable about his

field, inspired me to learn as much as I could. When I discovered my first Lyman-alpha emitting galaxy, I eagerly showed it to Raja, and he made me feel so special and proud to have made a scientific discovery. Although I have found that my main passion is computer science rather than astronomy, Raja's encouragement and mentorship introduced me to the discipline required when conducting scientific research. My daily enthusiasm and happiness throughout my time at the internship fueled my eagerness to continue pursuing scientific research.

Part II: Research

Abstract

This paper presents research on automatically identifying emission lines in large spectroscopic surveys. Prior work used manual inspection of spectra from the DEEP3 galaxy redshift surveys to search for Lyman-alpha emitters (LAEs). However, this approach is labor-intensive. It is possible to use algorithmic methods to automatically identify such emitters. This paper describes the design and implementation of a system to find LAEs in an automated manner. The algorithm has successfully identified LAEs in the redshift range 3.0-6.6. Results are presented from running the algorithm on the entire DEEP3 database. The paper also discusses the potential for further automated spectrographic analyses of DEEP3 and other surveys.

1 Introduction

The goal of this project is to look back in time to observe objects in the early stages of the history of the universe. Astronomers can observe what occurred billions of years ago by looking out into the distant universe. The farther away an object is, the farther back in time we are observing, because the light emitted from objects takes more time to travel to us. Thus observing objects that are very distant is essentially looking back in time.

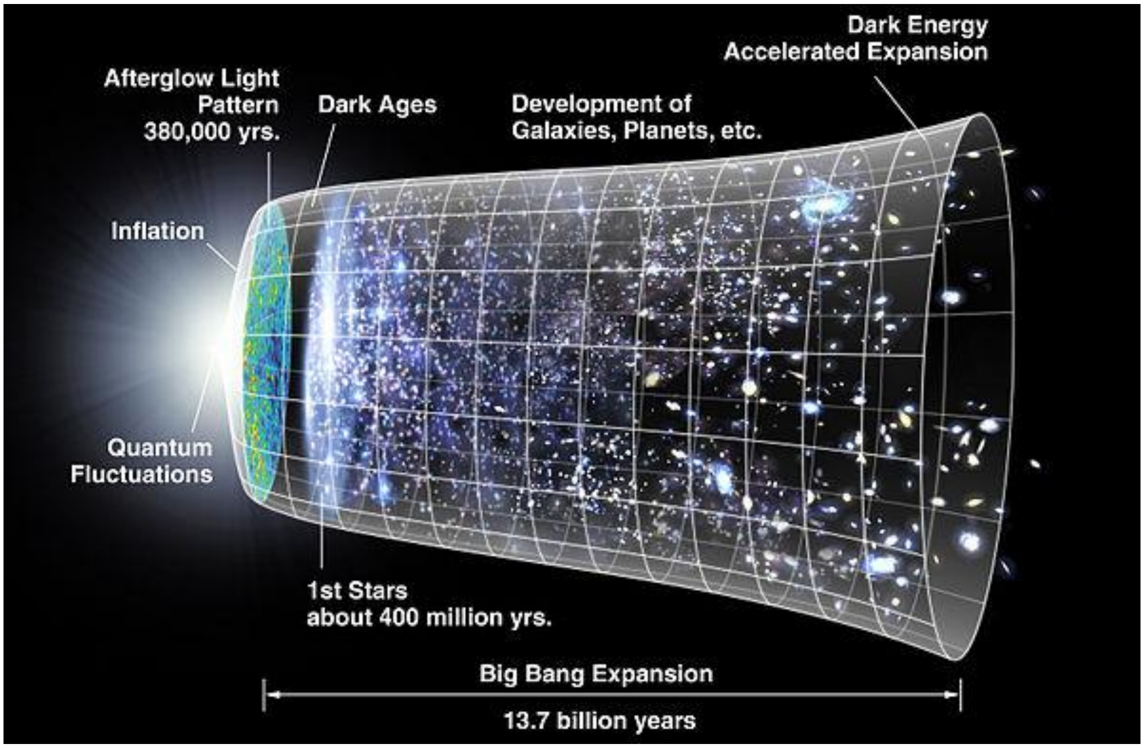


Figure 1: Model of the universe demonstrating space-time. As the distance from the right increases, so does the look-back time (image source: en.wikipedia.org/wiki/Universe)

The universe has been expanding since the Big Bang. Hubble discovered that all distant galaxies are moving away from Earth. Consequently, the light emitted by these objects is redshifted, meaning the light observed has a longer wavelength than when it was emitted.

A spectrum is required in order to find an object's redshift: the observed light must be dispersed into many wavelengths. Measuring the spectrum of a distant and therefore faint object requires a large telescope to gather enough light. Because of this, almost every major redshift survey has been carried out on the ground instead of space. When a telescope is located on Earth, atmospheric absorption becomes important. The Earth's atmosphere absorbs a large fraction of photons across most of the electromagnetic spectrum, except for those in the optical, radio, and some parts of the infrared. The atmosphere also glows brightly in the infrared portion of the spectrum, so ground-based infrared astronomy is doubly impacted. This makes observing in the optical (approximately 4000-9000 Angstroms, or 400-900 nm) ideal for redshift surveys.

There is a fixed wavelength window in the optical range that is used for astronomical observations. Light observed in this window from a nearby galaxy is emitted by the galaxy at optical wavelengths because the light is minimally redshifted. With a distant galaxy, what is observed in the optical frame was really emitted in the ultraviolet, because light has been substantially redshifted due to the large relative velocity of the distant galaxy. For a distant object to be observed in the optical on Earth, the galaxy must emit light in the rest frame ultraviolet.

A Lyman-alpha photon is emitted when the electron in a hydrogen atom cascades down from the n=2 to n=1 energy level. The Lyman-alpha emission line happens to have the shortest emitted wavelength (~1216 Angstroms) among strong emission lines, which is why Lyman-alpha emitters (LAEs) are targeted in our search as they allow us to find the most distant galaxies.

Besides the spectral methods discussed in this paper, there are two commonly used methods for finding LAEs: narrow-band search (Hayes & Ostlin 2006) and the dropout method. Narrow-band imaging is a means of looking for LAEs at high-redshift, but only strong LAEs can dominate the image, so fainter LAEs cannot be found through this technique. However, this method is useful for surveying very large areas of the sky. The spectral method discussed in this paper can find fainter LAEs than found by narrow-band search, giving more accurate information about the density and intensity of LAEs for areas of the sky where spectral data is available.

The other common method for LAE search is the dropout technique. Filters are designed to detect continuum in ranges, so continuum breaks can be detected. This method is also useful because it can cover a large area of the sky. However, a key disadvantage of the dropout technique is that it solely relies on continuum breaks to locate LAEs. Therefore, it cannot locate LAEs with no continuum on either side of the emission. In comparison, the spectral method can detect LAEs with no continuum break by using other features of Lyman-alpha emission lines.

Finding more LAEs is important in order to learn about the early universe. Past research using spectral methods has found LAEs through visual inspection, which is reliable but slow and labor-intensive (Sawicki et al. 2008). This paper presents a method for an automated search for LAEs, which will prove to be more efficient and hopefully allow us to find more LAEs by automatically searching a much larger data set with minimal manual effort.

2 Methods

This method for finding LAEs can be broken down into two main parts: search and classification. The search portion used Source Extractor (Bertin & Arnouts 1996), a commonly used software package for astronomical image analysis discussed in Section 2.3, which generated an initial set of candidates that might be LAEs. A secondary classification program, discussed in Section 2.4, was used to categorize the candidates generated by Source Extractor into non-astrophysical objects, true emission lines, and LAEs. All programs for this project were written in Interactive Data Language (IDL).

2.1 Input Database for the LAE Search

The input database for the search was the DEEP3 galaxy redshift survey. The Deep Extragalactic Evolutionary Probe (DEEP) databases are special because of their combination of large sample size (a large number of galaxies were targeted), depth (data was taken using both the world's most powerful optical telescope and one of the world's highest resolution spectrographs), and relatively high spectral resolution. This large amount of data is important when searching for rare objects such as LAEs, because this ensures a critical sample size. The sensitivity is necessary in order to find such distant and therefore harder to observe–objects. The high resolution of the spectrograph used for this search minimizes the impact of night sky emission lines on the spectra. This paper focuses on the DEEP3 database, because it covers a wider wavelength range and has a longer exposure time than the DEEP2 database, so lower luminosity LAEs at lower redshift ($z \sim 3$-4.5) and slightly more distant LAEs can be found (Cooper et. al 2011, Cooper et. al 2012).

2.2 Preparation of 2D Spectra: Removal of Continuum Light

A 2D spectrum is a gray scale image that represents the light from a small slit of the sky that is split up across many wavelengths. A continuum is a band of pixels with higher values, which occurs when photons are being observed at all wavelengths on the spectrum. To prepare the 2D spectra for the automated search for emission lines, the continuum portion of the spectrum associated with starlight is removed. A “boxcar” smoothing algorithm, which replaces each pixel with a median value of a horizontal band of pixels (in this case, 200 pixels wide and 1 pixel tall), yields spectra that roughly contain only continuum emission. By subtracting the boxcar-smoothed version of the 2D spectrum from the original 2D spectrum, the continuum is removed and emission lines remain (see figure 3). These continuum-subtracted images are used for the search.

Figure 2: Spectrum pre-continuum subtraction. The spectrum has continuum extending across the image, with an emission line in the center (screenshot from `ds9` software examining DEEP3 spectra).

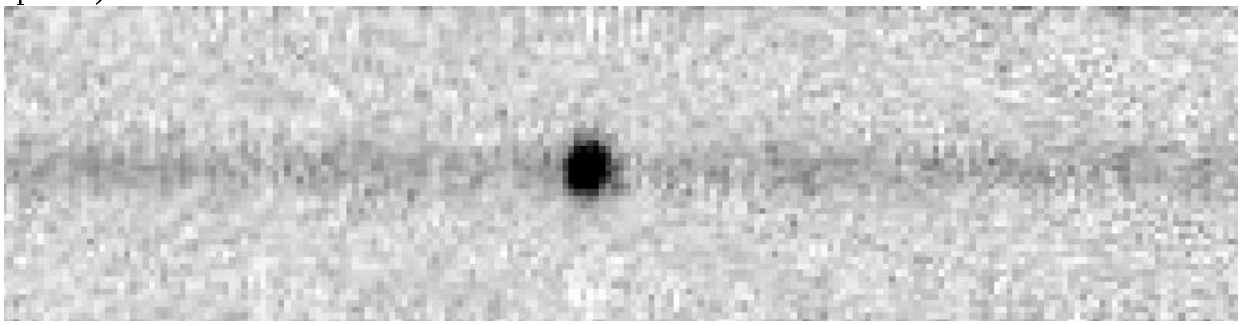
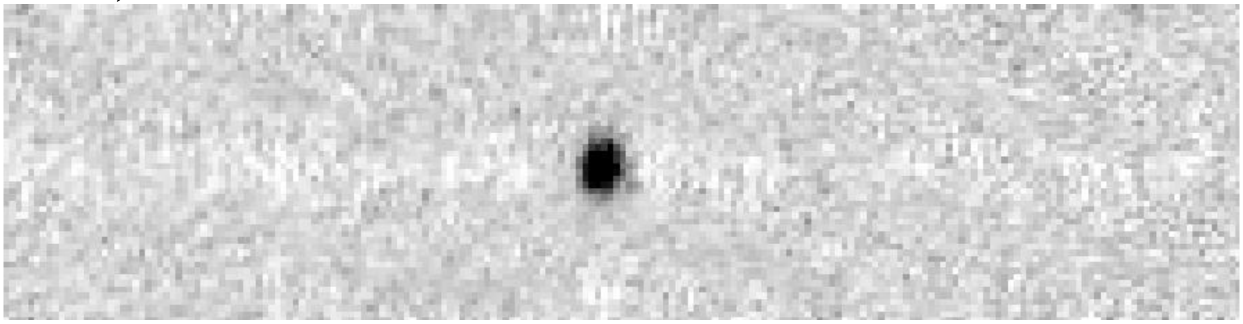


Figure 3: Same spectrum as figure 2, post-continuum subtraction. The continuum has been removed, but the emission line remains.



2.3 Search

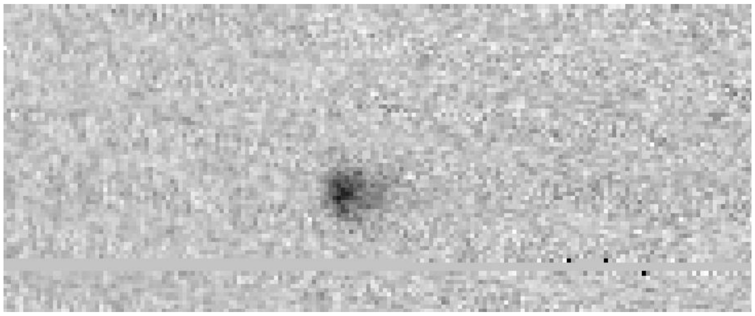


Figure 4: Example LAE spectrum (screenshot from `ds9` software examining DEEP3 spectra).

A program called Source Extractor (Bertin & Arnouts 1996) was used to do the initial search. Source Extractor is essentially a peak finder. It looks through an image and marks areas with connected islands of pixels that have fluxes above some specified detection threshold. This program is good for completeness – that is, finding all or most of the objects of the type we're looking for – but it can also have a very high rate of contamination, meaning that it marks many objects that we are not interested in for this project as "good" objects.

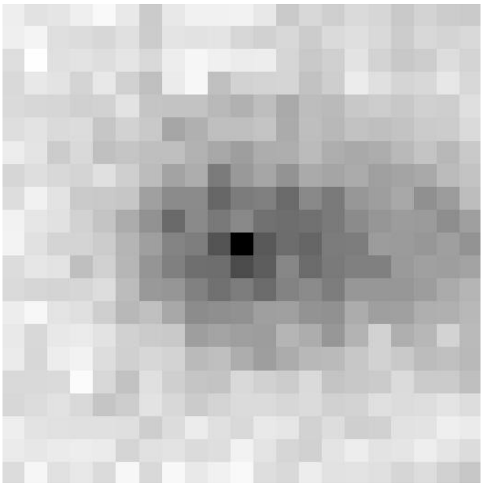


Figure 5: Median DEEP3 LAE filter.

2.3.1 Custom Detection Filter

Source Extractor uses a convolution kernel to smooth the spectra. It takes a square filter, which is represented as a two-dimensional array of pixel values. A convolution kernel changes pixel values in an image based on how closely the group of pixels matches the input array of pixel values. This is a matched filter method; it enhances objects that have the same shape as the kernel. The filter used for this project was made by taking the median image of fifteen LAEs found in DEEP3 through manual visual inspection.

2.3.2 Detection Threshold

One Source Extractor parameter is the detection threshold, which is the minimum value for a pixel to be treated as part of an object. Too high of a detection threshold results in the non-detection of faint LAEs, but too low of a threshold drastically increases the number of spurious objects detected.

2.3.3 Minimum Area

The second important parameter for the Source Extractor search is the detection minimum area. This is the minimum number of connected pixels above the detection threshold that an object needs to have in order for it to be detected.

2.3.4 Optimizing the Search Parameters

To find the ideal detection threshold and minimum area values, Source Extractor was run on 2D spectra from the DEEP3 survey containing manually detected LAEs from a previous visual search. Each time the search was done, different combinations of threshold and minimum area values were tested. The fraction of LAEs found and the number of non-LAEs detected was recorded for each parameter pair. The value optimized was the number of LAEs found divided by the number of non-LAE objects detected.

2.4 Classification of Detected Sources

2.4.1 Night Sky Emission Line Residuals

The Earth's atmosphere glows at certain specific wavelengths, producing strong night sky emission lines. For the spectra to yield useful information on astrophysical sources, the light from the atmosphere must be subtracted. This process of "sky subtraction" is helpful, but the elevated level of random noise associated with bright night sky emission lines produces noisy columns of pixels that run across the spectra at certain wavelengths. These columns are called residual night sky emission lines or "skylines". Many of the non-astrophysical objects detected by Source Extractor are located on skylines. To locate skylines, the classification program took the standard deviation of each column of pixels. The noisiness of the skyline columns gave a much higher standard deviation than non-skyline columns. The classification program subsequently determined whether an object was astrophysical or simply a group of noisy pixels from an atmospheric emission

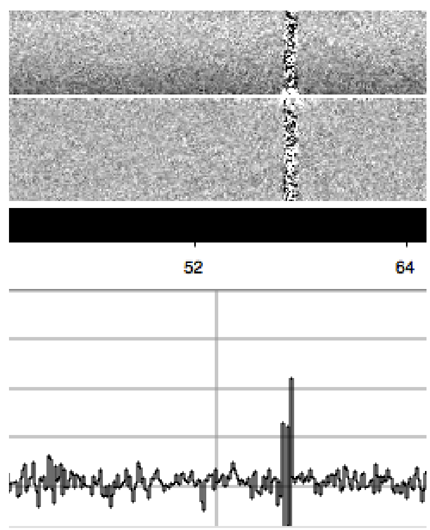


Figure 6: Example of a night sky emission line (screenshot from `ds9` software examining DEEP3 spectra).

line. This technique correctly classifies most false emission lines, so most of the remaining objects are true emission lines that must be categorized as LAEs or other emission lines.

2.4.2 Edges of 2D Spectra

Objects that are near the top and bottom ends of slits are often also spurious detections caused by instrumental effects. All objects within one pixel of the top or bottom edge of a 2D spectral band are given a “slit edge” tag and are classified as non-astrophysical objects. The ends of the CCD detector array in the spectral direction also contain many false objects, because the spectral continuum removal (see section 2.2) does not work well near the blue and red ends of the spectrum.

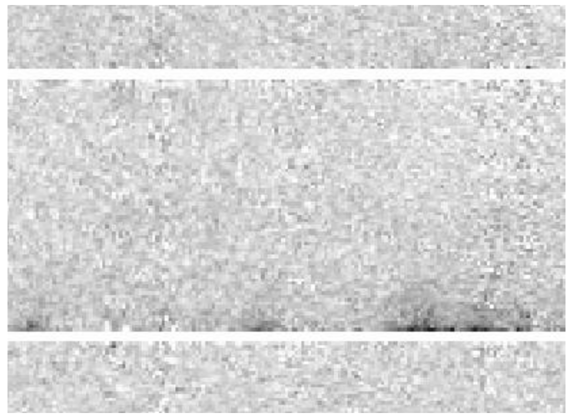


Figure 7: Deceiving slit edge (screenshot from `ds9` software viewing DEEP3 spectra).

2.4.3 Continuum Breaks

The previous sections looked at atmospheric and instrumental effects. This section will look at an astrophysical effect: continuum breaks. One of the characteristics that causes Lyman-alpha emission to be distinguishable in spectra from other spectral lines is the continuum break surrounding the line. On spectra, a continuum break looks like a band of pixels with higher value on the red side of an emission line, and no such band on the blue side (see figure 8). The continuum break is caused by intergalactic hydrogen clouds between the LAE and the observer that are located at a range of lower redshifts than the LAE. These clouds absorb photons at the wavelength required to excite a hydrogen atom's electron from the n=1 to n=2 energy level, so much of the light that is on the blue side of the LAE in the observer’s frame is absorbed. There is therefore generally little or no continuum on the bluer side of a Lyman-alpha emission line, especially for high redshift ($z > 3$) LAEs since galaxies in the early universe contained a

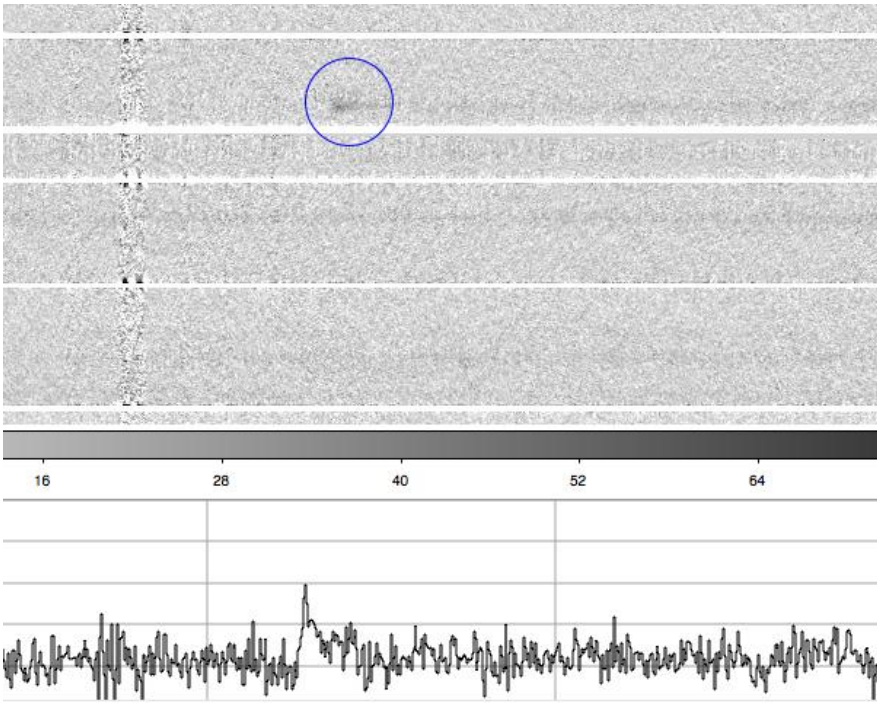


Figure 8: Note the continuum on the right (redder) side of the Lyman-alpha emission (circled), and the abrupt continuum break to the left (bluer) side (LAE in DEEP3).

higher fraction of neutral hydrogen gas than their present-day counterparts.

To distinguish between Lyman-alpha and other emission lines, the program uses the fact that LAEs have a continuum break (or no continuum at all on either the blue or red sides of the Lyman-alpha emission line), whereas other emission lines typically have continuum on both sides (Guhathakurta, Tyson & Majewski 1990). The program computes the median values of pixels on the blue and the red sides of the detected emission line in question and compares them: if the red side has a high value and the blue side has a low value, the object is classified as a secure LAE. If the object has no continuum, the object is classified as a low-confidence LAE: the galaxy may be too faint to see the continuum at all. If the object has continuum on both sides but has no other emission lines on the same row of the 2D spectrum, it is marked as a non-LAE single emission line. These non-LAE single emission lines could be weak H-alpha emission, blended [OII] emission, etc. (Kirby, Guhathakurta et. al 2007).

2.4.4 Associations of Emission Lines

After the high-confidence and low-confidence LAEs have been classified, the group of remaining objects contains both non-LAE single and other emission lines. To distinguish between these last two categories, the classification program looks to see how many other objects are detected on a close y-coordinate (and therefore from the same region of the sky and likely to come from the same galaxy). If more than one object was found on the row, the object is classified as "other emission line," because the rest frame wavelengths for LAEs is too far from other strong emission lines for other emission lines to appear in the same 5000 Angstrom window. If no other objects are found on the continuum, the object is classified as a single emission line.

3 Results

To assess the success of the automated search program, the objects found were manually inspected to confirm whether they were categorized correctly. A visual inspection of candidates from a subset of the DEEP3 masks yielded results about completeness and contamination.

There are two steps of classification that can be analyzed: the first is the classification of astrophysical and non-astrophysical objects and the second is the classification of LAEs and other emission lines. This section will examine both steps with regards to completeness and contamination.

Completeness is the percentage of all positive examples that are found by some classification procedure. In the context of this project, this is both the fraction of detected astrophysical objects that were categorized as astrophysical and the fraction of known LAEs detected and correctly classified by the automated search system.

Contamination is the percentage of negative examples that are classified incorrectly as positive examples by the program. In the context of this project, this is both the number of non-astrophysical objects categorized as astrophysical and the number of non-LAE emission lines categorized as LAEs.

The initial categorization of the classification program, which classifies objects as astrophysical or other, has been quite successful: it has almost 0% contamination, meaning that there were almost no objects that were non-astrophysical that were labeled as astrophysical. Among the objects returned from Source Extractor, this initial classification has a nearly 100% completeness rate. Source Extractor does not find all emission lines (faint or small emission lines or those having differing shapes from LAEs), but the parameters were optimized for specifically finding LAEs, so it is not necessary to find all emission lines. Out of the already discovered 11 LAEs found by manual inspection, 10 were found by Source Extractor, and all of the 10 were classified as either high-confidence LAEs, low-confidence LAEs, or singles (there was one LAE quasar that was categorized as such).

The current LAE classification program has about a 50% contamination rate: out of objects classified as LAEs, half are non-LAE emission lines, and half are LAEs. If the shape of the emission line (which also makes LAEs unique) was taken into account, this number could potentially be much lower. Future tasks also include optimizing the search for Lyman-alpha in DEEP2 (which has a different spectral resolution, wavelength range, and only serendipitous LAEs) and for generally finding all emission lines.

4 Discussion

Past research has found LAEs through visual inspection, which is reliable but slow and labor-intensive. This paper presents a method for an automated search for LAEs, which will prove to be more efficient and hopefully allow us to find more LAEs by automatically searching a much larger data set with little-to-no manual effort. This research has validated automated search by running our automated method over the entire DEEP3 database and verifying that approximately 70% of all strong LAE candidates found through manual visual inspection were also categorized as LAE candidates by the automated method. The automated method will allow more rapid and less labor-intensive processing of future spectrographic surveys.

Lyman-alpha emission has an asymmetric shape: highly energetic star formation in LAEs drives an expanding shell of hydrogen gas (combination of supernova explosions of very massive short lived stars as well as intense radiation pressure), which causes LAEs to have an identifiable asymmetric shape on 2D spectra. The asymmetric triangular shape of an LAE on spectra, together with the distinctive continuum break caused by absorption of light on the blue side of a Lyman-alpha emission by intervening intergalactic hydrogen clouds, give LAEs a fairly distinct visual signature that automated methods rely on for identification. While these characteristics are specific to LAEs, the automated process should extend successfully to other emission lines. The approach used in this research can be generalized to other types of astrophysical objects with slight changes in parameters or the targeting of other distinct characteristics.

5 Conclusions and Future Work

Running the automated finder on DEEP3 yielded successful results with regards to completeness: it detected and correctly classified nearly 100% of all LAEs in the input database. However, the classifier has a 50% contamination rate. This is not a serious problem, because the improperly classified candidates can be quickly ruled out through human inspection. The automated finder can run on all of DEEP3 in a few hours, and the human confirmation of results could be done in a few hours as well. This is a large improvement from the original full manual inspection, which took roughly 180 skilled person hours to inspect DEEP3 in its entirety.

To further improve the LAE classification contamination, it would be useful to examine the asymmetry of LAEs compared to other emission lines and use this as an additional distinguishing characteristic (the custom LAE Source Extractor filter takes advantage of this, but not enough).

Another future task is optimizing the search parameters in order to run the finder on the DEEP2 database. DEEP2 is roughly four times the size of DEEP3. Because of its large size, it has not been visually searched in its entirety. This will make the automated finder useful and hopefully yield many more LAEs, because the expected number of LAEs in the survey as a whole is higher.

This automated search technique is also not limited to LAEs: it can also be used for finding emission lines in general. Optimizing the search parameters to focus on general emission lines instead of just LAEs will likely yield a successful general emission line finder.

The best way to measure a search program’s completeness is to add fake LAEs to the existing spectra. Adding increasingly fainter LAEs will show the limits of the finder’s abilities. Doing this will be helpful for further improvements in the search and classification algorithms, because the faint LAEs can then be included in finder optimization.

LAEs are important astrophysical objects to study, because they allow a deeper understanding of the rapid star formation in the early universe. The automated search method described in this paper can greatly expand the set of known LAEs, thereby giving greater insight into the earliest periods of the universe.

References

E. Bertin, S. Arnouts. "SExtractor: Software for Source Extraction." 1996, Astronomy and Astrophysics Supplement Series, 117, 393

M. C. Cooper, J. A. Aird, A. L. Coil, M. Davis, S.M. Faber, S. Juneau, J. M. Lotz, K. Nandra, J. A. Newman, C. N. A. Willmer, R. Yan. "The DEEP3 Galaxy Redshift Survey: Keck/DEIMOS Spectroscopy in the GOODS-N Field." 2011, ApJS, 193, 14

M. C. Cooper, J. A. Newman, M. Davis, D. P. Finkbeiner, B. F. Gerke. "*spec2d*: DEEP2 DEIMOS Data Reduction Pipeline." 2012, ASCL, 1203.003

P. Guhathakurta, J. A. Tyson, S. R. Majewski. "A Redshift Limit for the Faint Blue Galaxy Population from Deep U Band Imaging." 1990, ApJL, 357, L9

M. Hayes, G. Ostlin. "On the narrowband detection properties of high-redshift Lyman-alpha emitters." 2006, A&A, 460, 681

E. N. Kirby, P. Guhathakurta, S. M. Faber, D. C. Koo, B. J. Weiner, M. C. Cooper. "The DEEP2 Galaxy Redshift Survey: Redshift Identification of Single-Line Emission Galaxies." 2007, ApJ, 660, 62

J.A. Newman, M.C. Cooper, M. Davis, S. M. Faber et al. "The DEEP2 Galaxy Redshift Survey: Design, Observations, Data Reduction, and Redshifts." 2012, ApJS, submitted

M. Sawicki, B. C. Lemaux, P. Guhathakurta, E. N. Kirby, N. P. Konidaris, C. L. Martin, M. C. Cooper, J. A. Newman, B. J.Weiner. "The DEEP2 Galaxy Redshift Survey: Lyman Alpha Emitters in the Spectroscopic Database." 2008, ApJ, 687, 884