Dear Jördis,
A very good start. It appears to me that you did a quite nontrivial, thoughtful study that will be an interesting read for astronomers. As you will see from my inline comments (best viewed with Adobe Reader or Adobe Acrobat, other pdf viewers might not display all annotations) there is much leeway for improving the technical writing aspects of this thesis; plus some more explanation / discussion. But the general structuring is fine and all the things that I point out can be addressed incrementally with no groundbraking revision needed. I am curiously looking forward to the final result!
Cheers, Herbert Jaeger (July 25, 2021)

Application of the Ant Colony Algorithm in the Identification of Globular Clusters

Author:
Jördis Hollander (s2956543)

Supervisor~~s~~:
Prof. Dr. H. Jaeger

A thesis submitted in fulfillment of the requirements for
the degree of BSc Artifical Intelligence
in the
University of Groningen

July 23, 2021

Abstract

Globular clusters (*GC*s) are stellar agglomerates [1] of about 10 000 to 100 000 stars [2]. They provide an interesting ground to study stellar evolution [3]. The complexity of the Universe makes the precise identification and classification of stellar structures challenging. In this paper, a pipeline for the identification for GCs is developed based on work by M. Mohammadi et al. [3]. This pipeline consists of filtering candidate regions through the use of a blob detection technique. The remaining regions are then fed into the Ant colony random-walk algorithm. This algorithm investigates a region and provides information on its stellar density. This information is then run through a clustering algorithm to determine potential GCs.

The aim of the research is to determine the accuracy of the pipeline in classifying GCs and investigate possible improvements. This pipeline will be ~~run~~ on data from the GAIA DR2 dataset [4], [5]. Different regions consisting of a variety of stellar objects are selected. Some of these regions contain known GCs, while some do not. The accuracy of the pipeline is explored by running it on these different regions. The evaluation of the functionality of the pipeline firstly considers if it finds all known GCs and secondly looks at what other stellar structures (if any) it classifies as GCs and why.

TODO: **I need to double check if I should use references in the Abstract, it should be stand-alone right?**

Acknowledgements

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

I also thank the Center for Information Technology of the University of Groningen for providing access to the Peregrine high performance computing cluster.

Contents

Chapter 1

Introduction

The speed of light in a vacuum is constant. This fact has enabled us to look into the past and observe how the Universe has evolved over time. From the cosmic soup, to the first stars, to the formation of galaxies; the evolution of the Universe has been a source of fascination across history. One area of particular interest is the evolution of galaxies. Early on most galaxies were small, but over time these smaller galaxies merged and amalgamated to form larger galaxies [6]. Over time, as they continued to evolve, they began to manifest a great variety of galactic structures. However, this begins to raise a question: *How does one classify the stages of evolution for a galaxy or determine where it is in its evolution?*

Hubble's Tuning Fork was the first classification scheme that sought to answer this, based on a galaxy's structure and size [7]. However, this scheme proved to be insufficient in the face of the complexity of the Universe and the variety of possible galactic formations [7]. An alternative approach to this classification scheme is to attempt to determine the age of the galaxies based on their composition and not their overall shape. One such method involves determining the ages of various clusters of stars within it, thereby providing insight into the origin of the stars making up the galaxy. Of these cluster types, the globular cluster (GC) is on average the oldest [8], and thus, is the most significant in gauging the age of galaxies. These type of clusters are stellar agglomerates [1] which formed in one of two ways [9]:

1. Through the compression of halo gas in the cosmic re-ionization phase early on in the formation of the Universe.
2. In the collapse of enormous molecular clouds triggered by events such as the collision of gas-rich galaxies.

They are typically composed of around 10 000 to 100 000 stars [2] bound tightly by gravity into a spherical formation. Some GCs are among the oldest objects in the Universe [10] and are thus an interesting ground to study galactic evolution [3]. These older GCs manifest some specific properties such as low metallicity [11], and through a combination of techniques, such as: horizontal branch morphology, analysis of white dwarf cooling sequences, and comparisons using the main-sequence turn-off location [10], [12], may have their age accurately determined thereby bounding the age of the galaxy they are contained within.

The aim of this paper is to present a pipeline for the identification of GCs, based on the work of Mohammadi et al. [3], and to evaluate its effectiveness. This pipeline first filters candidate regions through the use of a blob detection technique. The remaining regions are then fed into the Ant Colony Random-Walk Algorithm. This algorithm explores a region and computes information on its stellar density. This information is then used by a Clustering Method which results in the potential GCs.

This pipeline is run on data from the GAIA DR2 data-set [4], [5]. Different regions consisting of a variety of stellar objects are selected. Some of these regions contain previously found GCs, while some do not. Since we cannot state with absolute certainty the total number of GCs (for any given region, more may yet to be found), the accuracy of the pipeline is explored by running it on these different regions and then evaluating it in two ways. First, we determine if it finds all known GCs, and second we consider what other stellar structures (if any) it classifies as GCs. A robust classifier for GCs would be a useful tool in exploring the Universe.

Chapter 2

Data

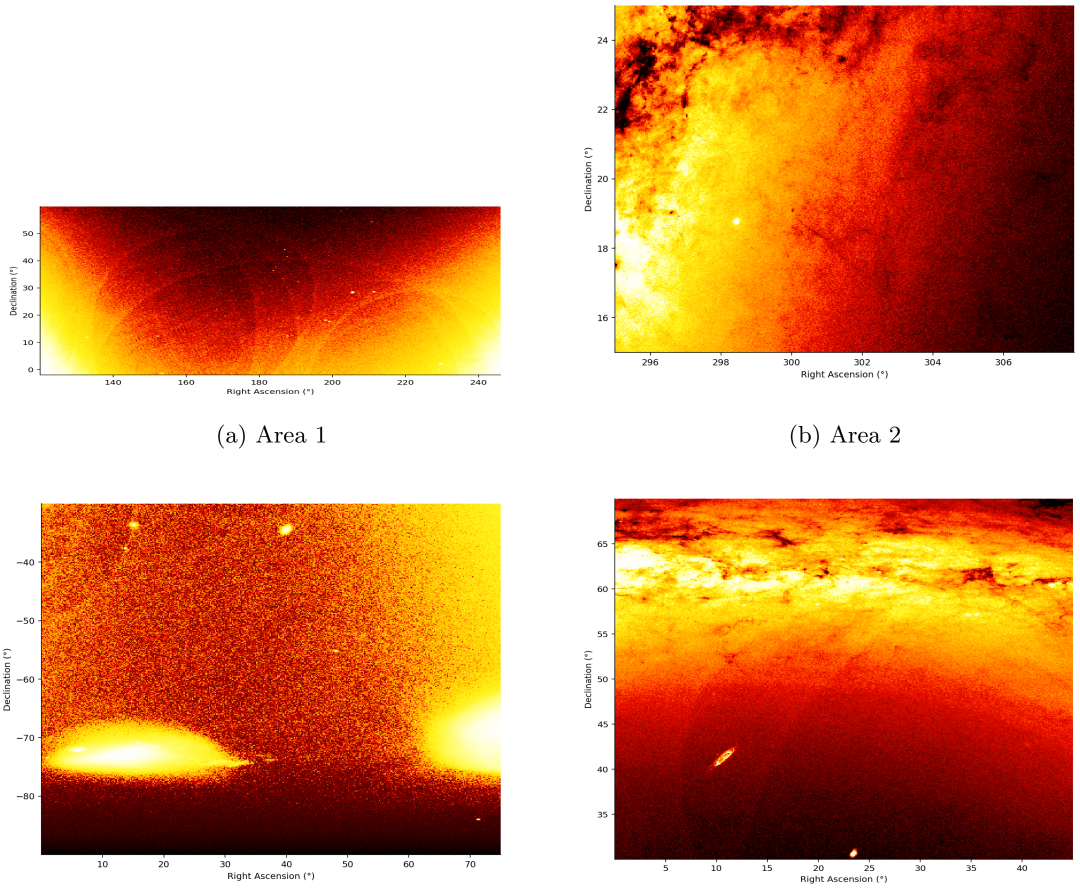
Astronomy is primarily an observational science, which, for most of its history, has gleaned information by looking up to the night sky with nothing but the naked eye. However, modern astronomy makes use of a variety of large-scale tools that observe and collect data on objects within our galaxy and far beyond. Telescopes (refractor, reflector, radiographic, spectrographic, and x-ray) allow us to extract a variety of information all without leaving the Earth's surface [13]. To collect information without the interference of the Earth's atmosphere we make use of satellites and space observatories that have been launched into orbit around our planet [14]. Occasionally, a space probe will be sent beyond our orbit to collect information from asteroids, planets, or their moons within our solar system [14]. For stellar observations, the sheer distances involved means that~~,~~ they are all made with telescopes.

The stellar data selected for this project is a subset of the GAIA DR2 data-set [5]. This data was collected during the GAIA satellite mission [5] and is approximately 500 GB of data consisting of a variety of parameters collected on a per-star basis. This is in great contrast with typical telescopic data which is primarily raw radiometric imagery and greatly expands the types of processing that may be conveniently applied. The subset of the parameters that are of interest to the pipeline are described in Section 2.1. As a result of hardware limitations, it is necessary to limit the investigation to a set of smaller regions from within the GAIA data-set. Four distinct areas were selected, for which, the cosmic ranges as well as the number of stars within the region may be found in Table 2.1. These four areas are bounded by right ascension (*RA*) × declination (*Dec*) and represent regions of interest with varying stellar distributions.

Table 2.1: Areas Under Investigation

Area	RA	Dec	Number of Stars
Area 1	120° up to 246°	-2° up to 60°	25 486 556
Area 2	295° up to 308°	15° up to 25°	23 470 239
Area 3	0° up to 75°	-90° up to -30°	16 781 316
Area 4	0° up to 45°	30° up to 70°	32 333 936

These regions contain all sorts of stellar structures, such as: GCs, Open Clusters (*OC*), and distant galaxies (*Gal*) and as such will allow the robustness of the pipeline to be tested. The heat-maps seen in Figure 2.1 provide an insight into the population and density of the stars found within the four areas. Area 1 and 2 contain known GCs while areas 3 and 4 do not. In addition all four regions contain other structures such as open clusters, nebulae, and distant galaxies.



(a) Area 1

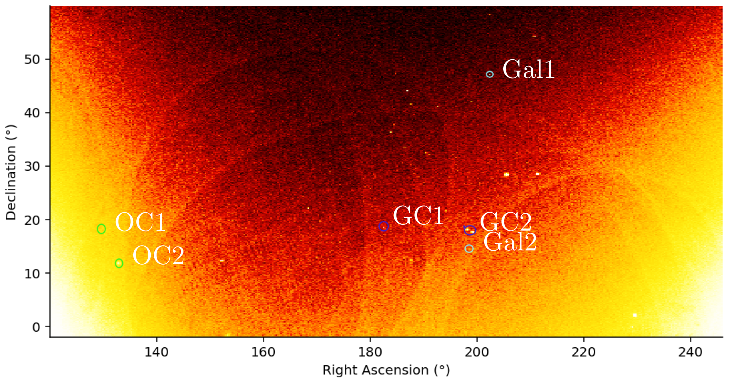
(b) Area 2

(c) Area 3

(d) Area 4

Figure 2.1: Stellar Distribution Heat-maps for the Four Areas

The brighter areas (yellow-white) contain more stars than the darker areas (red-black). From these heat-maps, the spots of increased stellar density are very evident. However, it is not immediately apparent whether these spots are GCs, OCs, galaxies, or some other stellar structures. Figure 2.2, provides an example of some of these stellar structures for Area 1 and further highlights the difficulty these classifying stellar structures by eye alone.



○: Open Clusters	○: Globular Clusters	○: Galaxies
OC1: Messier 44	GC1: NGC4147	Gal1: The Whirlpool Galaxy
OC2: Messier 67	GC2: NGC5024 and NGC5053	Gal2: Malin 1

Figure 2.2: Stellar Structures Present in Area 1

Area 1 contains more than the three GCs encircled in Figure 2.2. Area 1 contains an additional nine GCs (for a total of 12) while Area 2 contains just one known GC [15]. Table 2.2 summarizes the information on these GCs. Note that the diameter (DIA) is represented in arcminutes ($'$) which is a measure of angular distance where $1' = \frac{1}{60}^\circ$.

Table 2.2: Known GCs

(a) Area 1

GC	RA (°)	Dec (°)	DIA (′)
M3	205.548 42°	28°22′38.2″	18′
M5	229.639 62°	2°4′54.9″	21.6′
M53	198.229 45°	18°1′5.4″	13′
NGC 4147	182.526 26°	18°32′33.5″	4.4′
NGC 5053	199.112 88°	17°42′0.5″	10′
NGC 5466	211.363 71°	28°32′4.0″	9′
Koposov 1	179.827 09°	12°15′36.0″	–
Palomar 3	151.382 92°	0°4′18.0″	1.6′
Palomar 4	172.319 99°	28°58′24.9″	1.3′
Palomar 5	229.021 87°	0°6′41.8″	8.0′
GCI 38	242.752 47°	14°57′28.0″	2.2′
Willman 1	162.35°	51°3′0.0″	7′

(b) Area 2

GC	RA (°)	Dec (°)	DIA (′)
M71	298.44°	18°46′45.1″	7.2′

These GCs are used in the evaluation of the results of the pipeline.

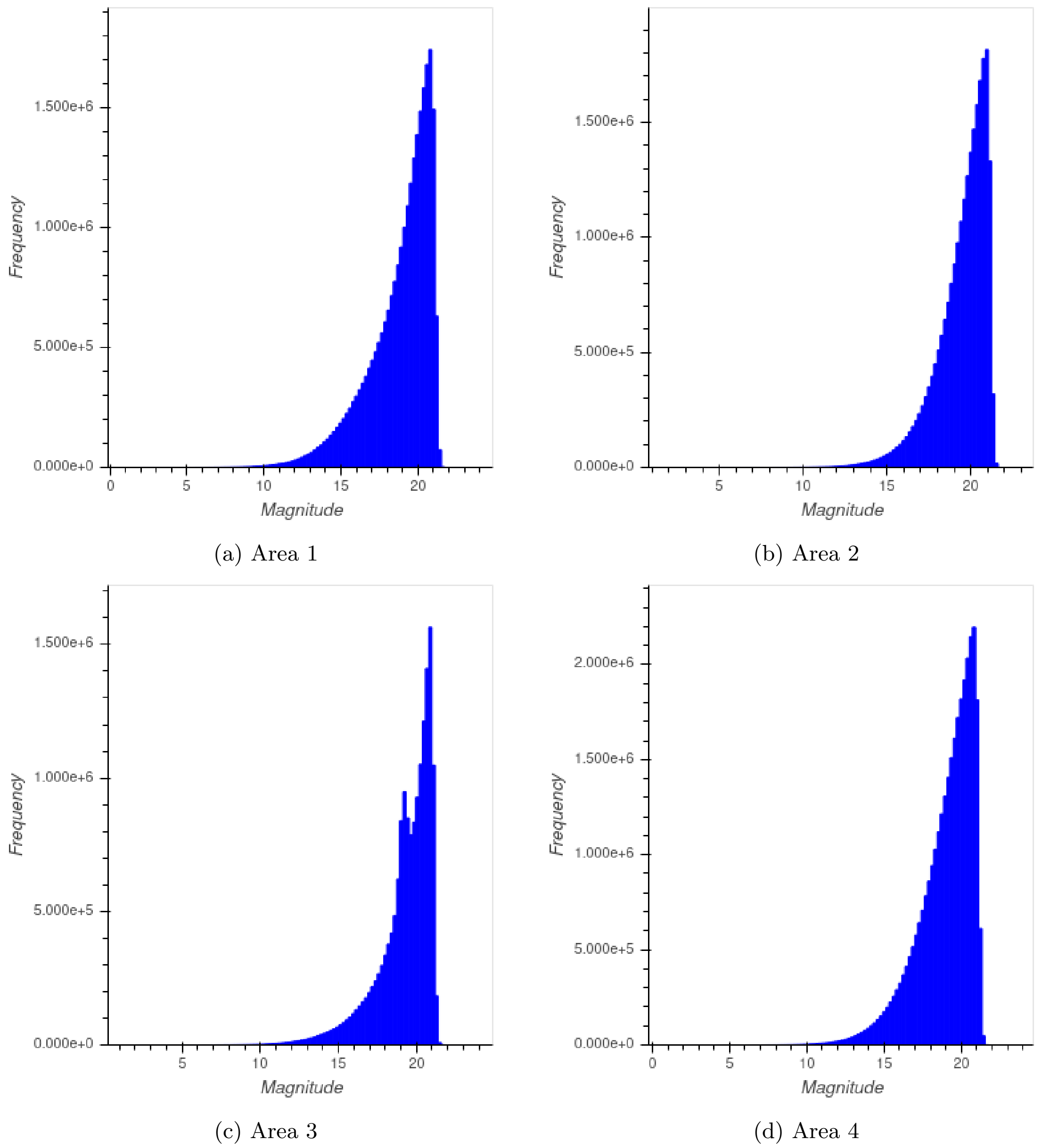
2.1 The Parameters

The GAIA DR2 data-set provides up to 88 parameters per star [5]. Of these parameters, there are six that are required for the pipeline:

1. Right Ascension (`ra`): This quantity is represented in degrees and when coupled with the declination it provides a position for an astronomical body in the equatorial coordinate system.
2. Declination (`dec`): This quantity is represented in degrees and when when coupled with the right ascension it provides a position for an astronomical body in the equatorial coordinate system.
3. Apparent Magnitude (`phot_g_mean_mag`): This is a unitless quantity and is the measure of a star's brightness when observed from Earth. Note that a higher apparent magnitude is a less bright star.
4. Parallax (`parallax`): This quanitity is measured in milliarcseconds (mas) and is the difference in the apparent position of an object when view along two different lines of sight [16].
5. Proper Motion of Right Ascension (`pmra`): Expressed in $\mathrm{mas\,yr^{-1}}$ and is the motion of an astronomical body from the frame of the center of mass of the solar system in right ascension.
6. Proper Motion of Declination (`pmra`): Expressed in $\mathrm{mas\,yr^{-1}}$ and is the motion of an astronomical body from the frame of the center of mass of the solar system in right ascension.

2.1.1 Apparent Magnitude

The apparent magnitude provides information on the brightness and proximity of stars. Histograms of the apparent magnitude across the four areas may be seen in Figure 2.3.



(a) Area 1

(b) Area 2

(c) Area 3

(d) Area 4

Figure 2.3: Histograms Based on Apparent Magnitude

Area 1 and 2 have very a similar distribution for their apparent magnitude. Area 1 contains more objects with a lower apparent magnitude leading to a slightly broader tail being present. All areas have a mode being between 20 and 21 but Area 3 seems to showcase a bi-modal distribution with an extra spike at a magnitude of 19. The mode for Area 1, Area 2, and Area 3 correspond to approximately 1.7 million objects, while the mode for Area 4 mode consists of 2.2 million objects. This may be explained by the fact that the data-set for Area 4 is larger (see Table 2.1). A summary of the mean, minimum, and maximum values of the magnitude for each area may be seen in Table 2.3.

Table 2.3: Mean, Min, and Max Values of the Magnitude per Area

Area	Mean	Min	Max
1	18.83	1.94	22.77
2	19.46	2.69	21.80
3	19.27	2.32	22.82
4	19.01	1.82	22.60

The minimal and maximal magnitude of these areas lie on average between 2.2 and 22.5. The mean magnitude of these areas is ~ 19.1.

2.1.2 Parallax and Proper Motion

It is also of interest to explore the spread of the data with respect to *parallax*, *proper motion of right ascension*, and *proper motion of declination* values. Histograms such as that present in Figure 2.4 have been created per area for these 3 parameters.

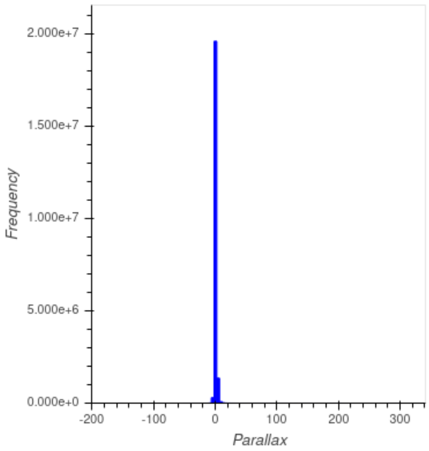


Figure 2.4: Frequency Division of the Parallax Values

All of the histograms, highlight that a majority of the values are centered at 0 and as a result not all 12 histograms have been displayed. What becomes clear here is that the movement showcased be measured is quite minimal, i.e. the mean and the mode are both about 0.

For the development of the pipeline it is also important to note that the extracted data-set does not include these three parameters for all stellar objects. And as we only want to use relatively reliable measurements we can only use the stars whose distance can be measured with the parallax. So only the stars in the pipeline that have these parameters are included in the final data-set, fortunately that is most of them. The exact numbers on these stars and the percentage of the data that remains can be found in Table 2.4.

Table 2.4: Data Point Removals

Area	Total Number of Stars	Removals	Remaining Stars	Percentage Remaining
1	25 486 556	4 103 730	21 382 826	83.9%
2	23 470 239	3 311 227	20 159 012	85.9%
3	16 781 316	2 667 157	14 114 159	84.1%
4	32 333 936	4 233 905	28 100 031	86.9%

These stars are missing these parameters because their distance means that their angular shift due to the orbital diameter of the Earth is so small as to be within the margin of error of the resolution of our best telescopes. Additionally, it is important to note that GAIA DR2 claims that for the stars with an apparent magnitude of 15 and higher that have mean errors for the parallaxes of 20-40 mas [17].

Chapter 3

Methodology

The data-set used in this project consists of many stars residing in one of the four areas extracted from the GAIA DR2 data-set. The stars that make up the data-set are expressed with the selected parameters; `ra`, `dec`, `phot_g_mean_mag`, `parallax`, `pmra`, and `pmdec`. These sets of parameters are run through the pipeline in specific chunks.

The data is first pre-processed and rasterized. Then it is run through a pipeline. This pipeline aims to find GCs and consists of three parts.

1. A blob detection algorithm using Difference of Gaussian (DoG)
2. A random-walk algorithm based on the Ant Colony Algorithm
3. A clustering method that uses the output of the Ant Colony Algorithm.

An overview of the whole process may be seen in Figure 3.1.

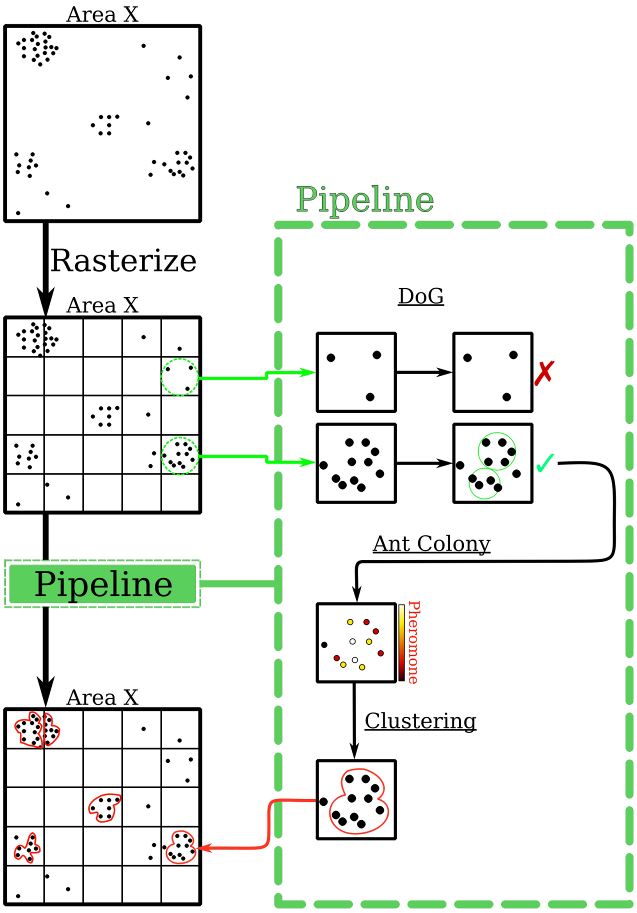


Figure 3.1: Overview of GC Identification

3.1 Pre-processing

There are several pre-processing steps that must be performed on the data before it may be run through the pipeline.

1. All stars that are missing parameters necessary for the processing (e.g. the `parallax`) ~~must~~ filtered out.
2. The area ~~must~~ then be split into portions (see Section 3.1.1) to improve the ability of the algorithms to explore the state space.

Additionally, the amount of data coupled with the time complexity of the Ant Colony Algorithm, means that it is worthwhile to attempt to filter out non-candidate regions first by using a faster technique. This may be seen in the section on Difference of Gaussian.

3.1.1 Rasterization

Rasterization is done so the data-set of an area is divided in parts. The rasterization results in a grid of rasters where each raster consists of a squared portion of the data-set. The largest known GC spans 1.5° × 1.5° [15] and in the the work of Mohammadi et al., rasters of 3.0° × 3.0° were used [3]. In this paper rasters of 4.0° × 4.0° as it spans the data more evenly.

This rasterization step provides the *ants* used in the Ant Colony Algorithm smaller spaces that they may more thoroughly explore. This also allows the algorithm to be executed in parallel, greatly speeding up the execution time. Finally, this allows the rasters to be pre-filtered with a more efficient glob detection algorithm to rule out non-candidate rasters.

3.1.2 Candidate Filtration using Difference of Gaussian

When considering Area 1 (presented in Figure 2.1a), one may observe that it is composed of dark regions, lightly interspersed with stellar structures. For this big data-set it would be useful, in the pre-processing phase, to ~~have~~ raster ~~files~~ that are unlikely to contain GCs ~~filtered out.~~ By definition a GC represents an agglomerate blob structure. As a result techniques that allow for the detection of blobs may be employed to narrow the search space. Such filters would be able to filter out these dark regions that compose the majority of the area and leave the areas containing ~~the~~ stellar structures for further investigation.

The conventional methods of image based blob detection are Laplacian of Gaussian (LoG), Difference of Gaussian, and Determinant of Hessian (DoH). For this large data-set it would be appropriate to use DoG because it is a faster approximation of the LoG approach [18], and the fact that it is an approximate may be taken under consideration by providing a lenient filter threshold. This approach may be contrasted against the work Mohammadi et al. where the Difference of Gaussian filter is applied on grayscale images with windows of 0.5° × 0.5° (RA × Dec), as a post-processing step [3].

So what does DoG do?: DoG is a feature enhancement algorithm for image data. In essence, it takes a grayscale image and produced blurred versions of that image. The blurring is made with respect to some property (which in our case is circularity). That image is then subtracted against the original which will cause the least matched regions to disappear and the most matched regions to remain.

The blurred images are obtained by transforming the original image with Gaussian kernels that use increased standard deviations [19]. The remainder of the difference between two successively blurred images are stacked up in a cube [19], which points out spacial information from between the range of blurring frequencies that are preserved in the images [20]. It can be viewed as a band-pass filter that discards all but a handful of spatial frequencies that describe features of interest within the original image [21]. As we are interested in circularity it is important that Gaussian ~~K~~ernel is isotropic and behaves the same in any direction [19].

How is applied?: We convert the information of (`ra`, `dec`, `phot_mean_mag`) from each star within a raster, then convert it into a grayscale image representing (x,y,z). Where (x, y) are coordinates and z is a level of gray between black and white that represents the image.

In a DoG equation one finds the subtraction of Gaussian filters which explain the difference between an excitatory (positive) region and an inhibitory (negative) one [22]. A Gaussian filter is a normalized Gaussian Kernel, of which you have three versions, a 1D, 2D, and an ND Kernel [19], see equations (3.1), (3.2), and (3.3).

$$G_{1D}(x;\sigma) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{3.1}$$

$$G_{2D}(x,y;\sigma) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^2} exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{3.2}$$

$$G_{ND}(\overline{x};\sigma) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^N} exp\left(-\frac{|\overline{x}|^2}{2\sigma^2}\right) \tag{3.3}$$

As the standard deviation (σ) increases so to does the resulting blurring of the image. This is because σ determines the width of the Gaussian Kernel and this referred to as the inner scale [21]. σ^2 is the variance. x is the windows of the input space from the image that gets convolved with the kernel.

$\frac{1}{\sqrt{2\pi}\sigma}$ represents the normalization constant and ensures that the average gray level of the image is maintained before and after the blurring with the kernel. This property is known as gray-level invariance [19].

This pipeline uses an existing function from the library `SciKit` which uses `SciPys` 1D Gaussian kernel. The equation of a DoG using a 1D Gaussian kernel is described here 3.4. TODO: **Should I mention that the $2\pi sigma^2$ is a normalization step done differently in the SciPy implementation?**

$$DoG_{\sigma_1,\sigma_2}(x) = I * \left(\frac{exp(-\frac{x^2}{2(\sigma_1)^2})}{\sqrt{2\pi}(\sigma_1)^2} - \frac{exp(-\frac{x^2}{2(\sigma_2)^2})}{\sqrt{2\pi}(\sigma_2)^2}\right) \tag{3.4}$$

Here is a matrix that represents the input gray scale image Also for σ_1 and σ_2 there is a difference in the standard deviation, where $\sigma_1 < \sigma_2$.

3.2 Ant Colony

With the search space reduced through the DoG filter, we are left with candidate regions which demonstrate the property of circularity that GCs hold. Each region contains a set of stars belonging to that region. For each region we can start applying the random walk algorithm. The algorithm aims to identify the groups of stars that are densely packed ~~among~~ these candidate regions. Even with the previous candidate filtration, the state space is still likely to be huge, hence the use of a random-walk algorithm. An exhaustive method would simply take too long to compute. The random-walk algorithm that is released on the remaining regions is a swarm intelligence method called the ant colony algorithm [23].

This swarm intelligence method makes use of a characteristic learned from ant colonies. Ants walk around in search for food while leaving behind a chemical substance called pheromones [23]. These pheromones serve to guide other ants towards some reward. When a reward is visited by many ants the path to it will contain higher concentrations of the pheromones. Thus, ants are more likely to follow paths with higher pheromone concentrations [23]. Interpreting these pheromone values with respect to GCs would mean that the levels of pheromone values within clusters are high and the levels on the paths between two clusters are low. This is practical for understanding what stars belong to a cluster and to provide distinction between overlapping clusters. This also provides information on the locations with the highest densities.

3.2.1 Description

Before the ants start their random walk, the vector f containing a pheromone for each star gets initialized. The outer loop runs the random walk simulation N_{iter} ~~amount~~ of times. Each time it executes three things: The initialization of vector N_v keeping track of the number of visits for each star, the random walk for N_a ants and the update of f by equation (3.7). The random walk for each ant consists of the selection of a random starting position (for example the node indicating $\mathbf{x}_i$) and a series of steps depicting a path. On this path, the next step for the ant is determined in a random manner biased by the distribution of pheromones in its neighborhood. This step is governed by equation (3.5). When the ant takes the step it is logged in the vector N_v.

3.2.2 Equations

To learn the **transition probability** for the t-th iteration from the data point $\mathbf{x}_i$ to $\mathbf{x}_j$ we use equation (3.5) [23], which can be found below.

$$P^{(t+1)}(\mathbf{x}_i, \mathbf{x}_j) = \frac{(w(\mathbf{x}_i, \mathbf{x}_j))^{\gamma} \left(\hat{f}^{(t)}(\mathbf{x}_j)\right)^{1-\gamma}}{\sum_k (w(\mathbf{x}_i, \mathbf{x}_k))^{\gamma} \left(\hat{f}^{(t)}(\mathbf{x}_k)\right)^{1-\gamma}} \tag{3.5}$$

Here, γ controls the effect that the pheromones have on the resulting step taken, $w(\mathbf{x}_i, \mathbf{x}_j)$ denotes the weight of the edge that represents the path from $\mathbf{x}_i$ to $\mathbf{x}_j$ and $f^{(t)}(\mathbf{x})$ is the amount of pheromone on the $\mathbf{x}$, of the regarding star, after t iterations.
To assure that the weights w and pheromones f are within the same scale, equation (3.6) is used for **normalizing the pheromones** of the neighborhood of $\mathbf{x}_j$ [23].

$$\hat{f}(\mathbf{x}_j) = \frac{f(\mathbf{x}_j)}{\sum_{\mathbf{x}_k \in N_i} f(\mathbf{x}_k)} \tag{3.6}$$

N_i, represents the set of the neighbors of the i-th star, $\mathbf{x}_j$ and $\mathbf{x}_k$ are the data points for the j-th and k-th star and $f(\mathbf{x})$ denotes the amount of pheromone on the regarding star.

Equation (3.7) **updates the pheromone level** [23].

$$f^{(t+1)}(\mathbf{x}_i) = c \times \frac{\sum_{a=1}^{N_a} N_i^a(n)}{N_a \times n} + (1-\rho) f^{(t)}(\mathbf{x}_i) \tag{3.7}$$

Here, c indicates the amount of pheromone that is added to a point per visit, N_a is the number of ants, $N_i^a(n)$ is the number of times the a-th ant visits $\mathbf{x}_i$ over n steps in the $(t+1)$-th iteration and ρ is the hyper-parameter that controls evaporation of the pheromone, which happens because the recency of the pheromone update is important [23].

3.2.3 Euclidean Distance for Calculating Neighboring Stars

One piece of the information that is needed is the Euclidean distance. This is used to determine the nearest neighbors for each star. The intent is to create a list containing n of the nearest stars to the ith star. If the space were limited to a 2D plane (e.g. just RA and Dec) then computing the Euclidean distance is easy. One simply takes the difference between the RA and Dec values of the star and applies the Pythagorean theorem. However, to operate in 3D space one has to consider the third axis.

Third Axis: Depth

The data-set does not directly report the depth. Instead it must be computed from the `parallax` or the relation between the apparent and absolute magnitudes. For the stars that have a parallax (p) value reported in the data-set, the distance (d) in parsecs may be calculated by taking the inverse of the parallax.

$$d = \frac{1}{p} \tag{3.8}$$

In the case that the parallax value is not reported it becomes necessary to use the apparent and absolute magnitudes. With the apparent and absolute magnitudes, the distance to the star in parsecs may be computer as follows.

$$d = 10^{\left(\frac{m-M}{5}+1\right)} \tag{3.9}$$

Here, d is the distance expressed in pc, $m - M$ is the distance modulus expressing the difference between the apparent magnitude m and the absolute magnitude M [3].

The apparent magnitude, m, is present in the GAIA DR2 data-set under the parameter of the `phot_g_mean_mag`. The absolute magnitude, M, is a more difficult variable to acquire. This is because it represents a star's intrinsic brightness, computed as if viewed from a distance of 10 parsecs [24]. Outside of this bound we cannot be sure of our visual observations and rely on how bright objects appear to our eyes, i.e. apparent magnitude [25]. Even though it is possible to approximate

M based on temperature and color with help of a Hertzsprung-Russell (HR) diagram [26], M is also dependant on the type of star you are approximating. What makes things even more difficult is that stars also have different colors and temperatures depending on the stage of evolution that they are in [27]. A solution to this has yet to be determined. However, research from Rate and Crowther on galactic Wolf-Rayet stars discusses the computation of distance and absolute magnitude on the GAIA DR2 [28]. Note that I still need to review their research to see if it can be incorporated and if it solves the problem.

With these RA, Dec, and depth, one can then calculate the Euclidean distance between stars $d(\mathbf{x}_i, \mathbf{x}_j)$.

3.2.4 Weight Initialization

GCs are an agglomerate structure. When identifying them, it is necessary to consider the attributes of individual stars as well as the relationship between stars. To this end, the stars are represented on a graph, with the stars representing the nodes and the relationships between the stars represented as paths. On the paths a weight value is encoded which represents the similarity between pairs of stars in one value. Neighboring stars within the same cluster are expected to have a higher similarity than stars outside of the cluster. These weight values are either computed by a kernel function or through Principal Component Analysis (PCA) [23]. They require a one time initialization and are then used for calculating the transition probabilities in equation 3.5.

In the case that the weight is based on kernels the **Gaussian function** [23], is applied. The information incorporated for the kernel weight initialization is only based on the Euclidean distance, and may be seen in equation (3.10).

$$K(\mathbf{x}_i, \mathbf{x}_j) = exp\left(-\frac{d(\mathbf{x}_i, \mathbf{x}_j)}{2\sigma^2}\right) \tag{3.10}$$

Here σ is the standard deviation and $d(\mathbf{x}_i, \mathbf{x}_j)$ is the Euclidean distance.

In the case that the weight is based on **PCA** then the method considers a geometric shape of a manifold to which the noise level of a data point can be interpreted as the distance that the data has to the manifold [23]. The aim of this is to encourage a random walk staying close to the manifold [23]. The information incorporated for the PCA weight initialization is based on certain parameters that may be chosen from that data-set. These parameters would likely consist of the Euclidean distance plus `ra`, `dec`, `pmra`, and `pmdec`. This is because these five parameters encode the position of a star and its motion. Since stars in a GC move together neighboring stars within the cluster will have similar positions and motion.

PCA is used to find a vector V containing the set of eigenvalues and eigenvectors, i.e. $\{(\lambda_k, v_k)\}_{k=1}^{D}$ where D is the number of features considered in the PCA (in this case five) [23]. The d-dimensionality of the manifold would make a separation of d planes. For which one than estimates the distance of the tangent space with formula (3.11) [23].

$$\Delta_i = ||\mathbf{x}_i - (VV^T(\mathbf{x}_i - \mu) + \mu)|| \tag{3.11}$$

Here $V = [v_1, v_2, ..., v_d]$, $\mathbf{x}_i$ is the ith neighbor of $\mathbf{x}$ from the set of neighbors N and μ is the mean position of the stars in N, calculated by equation (3.12).

$$\mu = \frac{1}{|N|} \sum_{\mathbf{x}_i \in N} \mathbf{x}_i \tag{3.12}$$

With Δ_i known we can initialize the weight values.

$$w(\mathbf{x}, \mathbf{x}_i) = \begin{cases} 1 - \frac{\Delta_i}{a} & a \geq \Delta_i \\ 0 & a < \Delta_i \end{cases} \tag{3.13}$$

Here, $a = max(\sigma_i)$ has the purpose of ensuring the weight values are positive [23]. It is important to note that when computing the weights using PCA that $w(\mathbf{x}_i, \mathbf{x}_k)$ does not need to be $w(\mathbf{x}_k, \mathbf{x}_i)$.

3.2.5 Algorithm

Algorithm 1 Ant Colony[23]

```
Result: The pheromone vector $\mathbf{f} = [f_1, f_2, ..., f_N]$.
Initialization: Initialize the pheromone vector $\mathbf{f} = [f_1^{(0)}, f_2^{(0)}, ..., f_N^{(0)}]$;
for $t = 1$ to $N_{iter}$ do
    $N_v = [0, 0, ..., 0]$
    for $a = 1$ to $N_a$ do
        Randomly select a node as the initial position for $a$-th ant (say $\mathbf{x}_i$)
        for $s = 1$ to $N_s$ do
            randomly select its next destination following Eq. (3.5) (say $\mathbf{x}_j$)
            $N_v[j] = N_v[j] + 1$
        end
    end
    Update pheromone by Eq. (3.7)
end
```

In this algorithm N_{iter} indicates the number of times that an ant randomly starts walking, N_v is the vector indicating the number of visits it received for each star (in the line $N_v[j] = N_v[j] + 1$ that number gets increased by 1 for the concerning star), N_a is the number of ants and N_s specifies the number of steps that an ant walks in each iteration.

3.2.6 Interpreting the Results of the Ant Colony Algorithm

The results of the algorithm are an updated pheromone vector f which contains adjusted pheromone values for each star in the given window. The pheromone values within clusters will be high and the values outside of (and between clusters) will be low. This provides a boundary separating a cluster from other stellar structures and is practical for providing a distinction between overlapping clusters. However, the actual clustering still has to be performed. The clustering algorithm can make use of the collected pheromone values and the available parameters. One possible technique that may be used is the use of spectral clustering[23].

3.3 Clustering

TODO: **This still needs to be worked on, gravity implementation based on growing cloud centered at the centroid and pheromone attraction.**

Chapter 4

Results and Findings

4.1 DoG

4.1.1 Remaining Rasters

The DoG filter was run across all the rasters of the area with varying thresholds. The number of remaining rasters depend on the threshold that was set regarding the intensity of the blob detection on the data in the raster. In Table 4.1 the number of remaining rasters, from each area, are shown for the different thresholds.

Table 4.1: Rasters remaining after the execution of DoG at Varying Thresholds

Area	Threshold = 0.5	Threshold = 0.2	Total Number of Rasters
Area 1	7	8	512
Area 2	16	28	35
Area 3	23	39	285
Area 4	19	60	120

4.1.2 A1 and A2 - Filtering Rasters with Known GCs

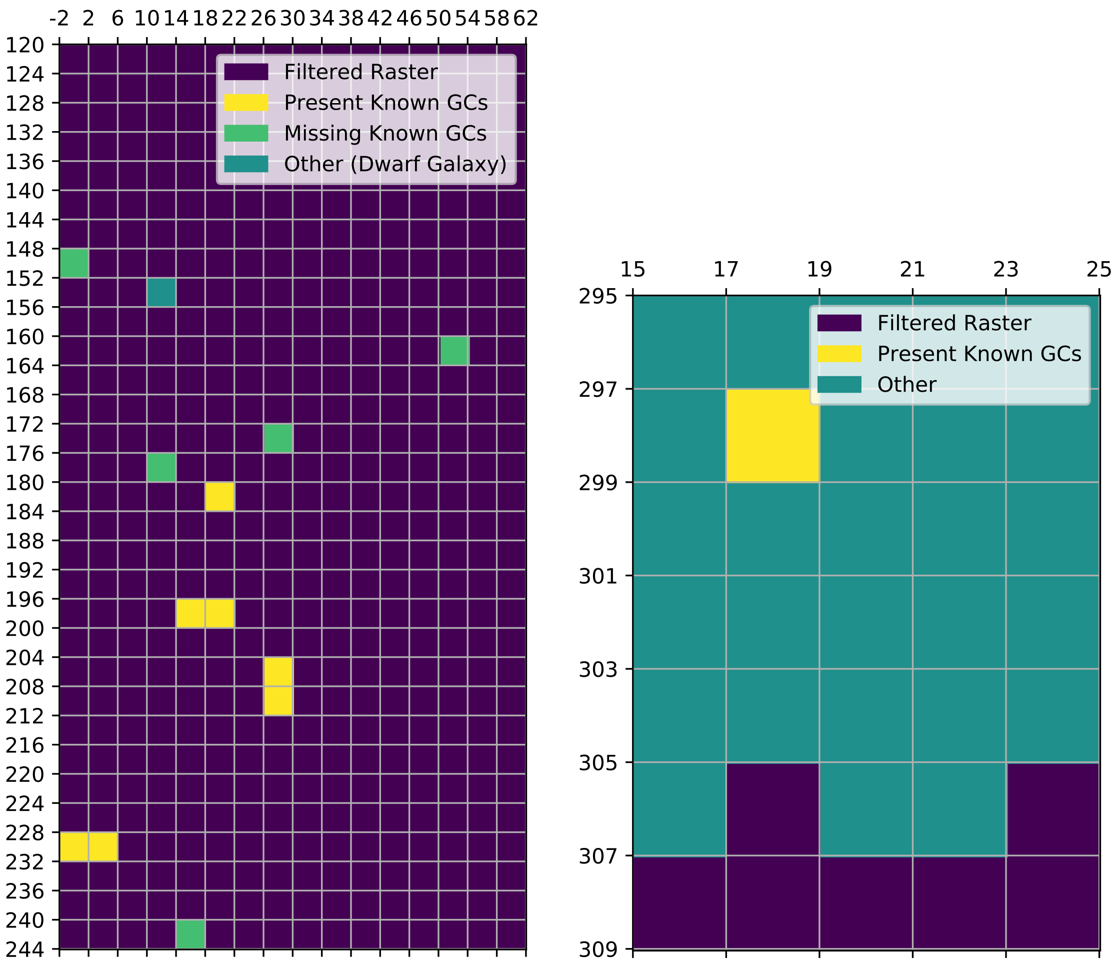
On the rasters of area **one** different thresholds were tested for the threshold of 0.1 all of the 512 rasters had blobs detected. The intensity of the detection is too high such that ~~the~~ single stars are detected as blobs. Of the known GCs in Area 1 not all get detected within a threshold of 0.2. For a threshold of 0.5, 7 out of 12 get detected and when setting a threshold of 0.2 another raster gets found that does not contain a known GC. This raster `a1_152.0_10.0`, finds a blob at approximately ra: 152° and dec: 12° at this position there is a dwarf galaxy called Regulus Dwarf Galaxy (PGC 29488 - UGC 5470). ~~It is called that because it is located near the large star Regulus.~~ When searching between the thresholds of 0.2 and 0.1, I find that: In the case that the threshold=.15 and the `max_sigma`=40 it mostly does find the known globular clusters from the list of Area 1 and this is because the size of the blob/cluster in these cases is quite small as they are located farther away. Here you see how important the Ant colony algorithm could be for a further search.

Table 4.2: What known GCs of Area 1 are getting detected for a threshold of 0.5

Name	RA	Dec	File	Present after DOG filtration
M3	205.54842	28 22’ 38.2”	`a1_204.0_26.0.csv`	Present
M5	229.63962	2 ∘ 04’ 54.9”	`a1_228.0_2.0.csv`	Present
M53	198.22945	18 ∘ 01’ 05.4”	`a1_196.0_14.0.csv`, `a1_196.0_18.0.csv`	Present
NGC 4147	182.52626	18 ∘ 32’ 33.5”	`a1_180.0_18.0.csv`	Present
NGC 5053	199.11288	17 ∘ 42’ 00.5”	`a1_196.0_14.0.csv`	Present
NGC 5466	211.36371	28 ∘ 32’ 04.0”	`a1_208.0_26.0.csv`	Present
Koposov 1	179.82709	12 ∘ 15’ 36.0”	`a1_176.0_10.0.csv`	
Palomar 3	151.38292	0 ∘ 04’ 18.0”	`a1_148.0_-2.0.csv`	
Palomar 4	172.31999	28 ∘ 58’ 24.9”	`a1_172.0_26.0.csv`	
Palomar 5	229.02187	0 ∘ 06’ 41.8”	`a1_228.0_-2.0.csv`	Present
GCI 38	242.75247	14 ∘ 57’ 28.0”	`a1_240.0_14.0.csv`	
Willman 1	162.35	51 ∘ 03’ 00.0	`a1_160.0_50.0.csv`	

TODO: **Change heading into a Region instead of File.**

Area **two** area only contains one known globular cluster, however the DoG finds that in 16 rasters at least one blob gets detected when the threshold is set to 0.5 and 28 rasters remain at a threshold of 0.2. The raster file in which this blob should be visible (`a2_297.0_17.0.jpg`) is present among the remaining rasters of under threshold 0.5. Area 2 is rastered into 35 files, which is a much smaller amount than Area 1. However, the number of stars per raster is higher and the area that is split up in these files is smaller, so what you see in the image is a denser appearance of the stars.



(a) Area 1

(b) Area 2

Figure 4.1: Remaining rasters of areas after DoG filtration with threshold set to 0.2

4.1.3 A3 and A4 - Finding Other Stellar Structures

Area 3 does not contain any known GCs but has 39 out of 285 raster files containing blobs according to the blob-detection using a threshold value of 0.2. And only 23 at a threshold of 0.5.

Area 4 also does not contain any known GCs and has 60 out of 120 raster files containing blobs according to the blob-detection using a threshold value of 0.2. And only 19 at a threshold of 0.5.

Table 4.3: Some Identifiable Rasters found Area 4

coordinates in ra - dec	**Identified As**	**Blob Description**
8.0 - 38.0	Andromeda	Massive stars and the dwarf galaxy NGC 205 within Andromeda are seen as blobs but Andromeda itself is not, probably because of the shape.
20.0 - 30.0	Triangulum Galaxy M33	Massive circular shape is detected which is circled by more subtle blobs. The main shape is the spiral galaxy.
0.0 - 62.0	Supernova Remnant SNR G116.9+00.1, open star cluster st 18 within the Little Rosette Nebula, dark nebula LDN 1268	The raster is quite busy, it contains a super nova remnant, an open star cluster, a dark nebula and many stars. Ten large blobs are detected and seven smaller ones.

4.2 Ant

4.3 Clustering

Chapter 5

Conclusion

- Answer the main questions on what it does
- and if it work as intended

5.1 DoG

GCs are stellar agglomerates with a radius varying from 0.5 pc to 10 pc, typically centered at 3 pc to 5 pc [29]. The DoG method can find these agglomerate structures, or blobs. However, what variety of blob sizes it finds depends on the set parameters. This method generally has more difficulty finding larger blobs [18], but from the list of known GCs in areas 1 and 2 we see it doesn't have trouble finding the larger GCs, rather finding the ones that are smaller and blend into the background with the rest of the brighter stars. Also from area three and four we learn that other structures are indeed hard to separate with this method. Hence the follow up with an Ant algorithm based clustering method.

The DoG method does not work effectively as a pre-processing method, as it filters out known GCs. It only finds about 58% of the GCs of Area 1 when the threshold is set to 0.5. The only way it might contribute is to set the threshold of the blob size really low (say a threshold of 0.12), this way it would include almost all raster files and only filter those areas that are absolutely dark, decreasing the amount of data to go through slightly.

A reason for wanting to decrease the number of rasters is that when it comes to stellar data it can consist of big data and so the computations to get to the results are vast and it takes time to get there. So if this method of filtration does not work another way to go is to make it computationally run fast. This can be done by the optimization of the implementation (the Julia language, Parallelization and Peregrine solve this issue in a different way).

5.2 Ant

5.3 Clustering

Chapter 6

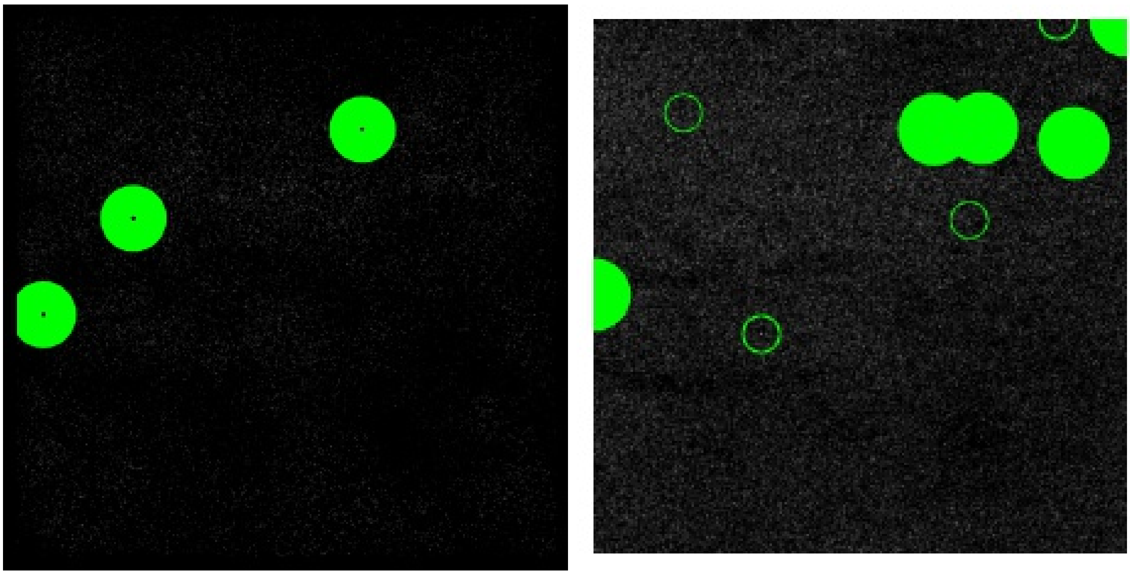
Evaluation

- try answering why it does what it does
- contrast to other work
- what might work, followup research

6.0.1 DoG

- What rasters remain at which threshold?
- does it filter out known GC's
- As a pre-processing method it doesn't work but how about post processing, why would that work?
- What about the busy areas vs. the empty ones? - what makes this behavior occur

The DoG filter applied as a post-processing method might help give more information. See Figure 6.1 to observe the difference in blob detection based as a pre- and post-processing method. The post-processing method can make use of the computed pheromone values of the Ant algorithm. This is different from basing it on the the absolute magnitude. Basing the image on the magnitude one looks at the brightness and not at the stellar geography, which might even be misleading.



(a) Pheromone based

(b) Magnitude based

Figure 6.1: Blobs found through DoG, based on the Pheromone and Magnitude values, making up the image.

These images show that if one would focus on the pheromone values instead of the brightness (or magnitude) you get different results on what would constitute as blobs. Which shows that it might have a different insight as a post-processing method. As less small blobs that are actually lonely bright stars would remain, as the magnitude longer of influence on where the image is bright.

6.0.2 Ant

- How many stars are visited (in %)

6.0.3 Clustering

- One can give a percentage on how many known clusters it found.
- One can then say it also found more objects of which some can be recognized as dwarf galaxies, galaxies and other parts.

Bibliography

[1] R. Gratton, A. Bragaglia, E. Carretta, V. D'Orazi, S. Lucatello, and A. Sollima, "What is a globular cluster? an observational perspective.," *The Astronomy and Astrophysics Review*, vol. 27, no. 1, pp. 1–136, 2019. DOI: doi:10.1007/s00159-019-0119-3.

[2] D. Bamberger, *How do astronomers count stars in a globular cluster?* [Online]. Available: https://www.quora.com/How-do-astronomers-count-stars-in-a-globular-cluster?share=1 (visited on 04/06/2020).

[3] M. Mohammadi, N. Petkov, K. Bunte, R. Peletier, and F. Schleif, "Globular cluster detection in the ~~gaia~~ survey," *Image Vis. Comput.*, vol. 57, pp. 165–174, 2017.

[4] Gaia Collaboration et al., "Description of the ~~gaia~~ mission (spacecraft, instruments, survey and measurement principles, and operations)," 2016.

[5] ——, "Summary of the contents and survey properties," 2018.

[6] M. Giavalisco, "Galaxy Evolution," in *Encyclopedia of Astronomy and Astrophysics*, P. Murdin, Ed. 2000, 2142, p. 2142. DOI: 10.1888/0333750888/2142.

[7] *The ~~h~~ubble tuning fork - classification of galaxies*, ESA/Hubble. [Online]. Available: https://www.spacetelescope.org/images/heic9902o/ (visited on 04/08/2020).

[8] R. Jimenez and P. Padoan, "A new self-consistency check on the ages of globular clusters," *The Astrophysical Journal*, vol. 463, no. 1, pp. L17–L20, May 1996. DOI: 10.1086/310053. [Online]. Available: https://doi.org/10.1086/310053.

[9] S. van den Bergh, "How Did Globular Clusters Form?" *The Astrophysical Jorunal Letters*, vol. 559, no. 2, pp. L113–L114, Oct. 2001. DOI: 10.1086/323754. arXiv: astro-ph/0108298 [astro-ph].

[10] B. C. Erin M. O'Malley Christina Gilligan, "Absolute ages and distances of 22 GCs using ~~m~~onte ~~c~~arlo main-sequence fitting," *The Astrophysical Journal*, vol. 838, no. 2, p. 162, Apr. 2017. DOI: 10.3847/1538-4357/aa6574. [Online]. Available: https://doi.org/10.3847/1538-4357/aa6574.

[11] A. Dotter, A. Sarajedini, and J. Anderson, "GLOBULAR CLUSTERS IN THE OUTER GALACTIC HALO: NEWHUBBLE SPACE TELESCOPE/ADVANCED CAMERA FOR SURVEYS IMAGING OF SIX GLOBULAR CLUSTERS AND THE GALACTIC GLOBULAR CLUSTER AGE-METALLICITY RELATION," *The Astrophysical Journal*, vol. 738, no. 1, p. 74, Aug. 2011. DOI: 10.1088/0004-637x/738/1/74. [Online]. Available: https://doi.org/10.1088/0004-637x/738/1/74.

[12] D. R. Soderblom, "The ages of stars," *Annual Review of Astronomy and Astrophysics*, vol. 48, no. 1, pp. 581–629, 2010. DOI: 10.1146/annurev-astro-081309-130806. [Online]. Available: https://doi.org/10.1146/annurev-astro-081309-130806.

[13] [Online]. Available: https://www.youtube.com/watch?v=i7930fj3T54&t=273s (visited on 07/16/2021).

[14] *Space technology.* [Online]. Available: https://clarkscience8.weebly.com/space-technology.html (visited on 07/16/2021).

[15] *List of globular clusters.* [Online]. Available: https://en.wikipedia.org/wiki/List_of_globular_clusters (visited on 04/07/2020).

[16] *Oxford English Dictionary*, Second Edition. Oxford University Press, 1989. (visited on 04/16/2020).

[17] Gaia Collaboration and B. et al., "Gaia Data Release 2. Summary of the contents and survey properties," *Astronomy Advisory Panel*, vol. 616, A1, A1, Aug. 2018. DOI: 10.1051/0004-6361/201833051. arXiv: 1804.09365 [astro-ph.GA].

[18] *Blob detection.* [Online]. Available: https://scikit-image.org/docs/0.17.x/auto_examples/features_detection/plot_blob.html (visited on 06/15/2021).

[19] [Online]. Available: http://pages.stat.wisc.edu/~mchung/teaching/MIA/reading/diffusion.gaussian.kernel.pdf.pdf (visited on 06/15/2021).

[20] *Difference of ~~g~~aussians*. [Online]. Available: https://en.wikipedia.org/wiki/Difference_of_Gaussians#cite_note-micro.magnet.fsu.edu-1 (visited on 06/24/2021).

[21] M. A. Michael W. Davidson. "Molecular expressions microscopy primer: Digital image processing – difference of ~~g~~aussians edge enhancement algorithm." (), [Online]. Available: https://micro.magnet.fsu.edu/primer/java/digitalimaging/processing/diffgaussians/index.html (visited on 06/24/2021).

[22] B. Gecer, G. Azzopardi, and N. Petkov, "Color-blob-based cosfire filters for object recognition," *Neurocomputing*, vol. 342, pp. 164–171, 2019. DOI: doi:10.1016/j.imavis.2016.10.006.

[23] M. Mohammedi, "The ant colony," Unpublished, 2020.

[24] *American Heritag Dictionary of the English Language*, Fifth Edition. 2011. (visited on 03/29/2020).

[25] D. Wood. "Absolute magnitude: Definition & formula." (), [Online]. Available: https://study.com/academy/lesson/absolute-magnitude-definition-formula.html (visited on 04/01/2020).

[26] D. C. Palma. "The ~~h~~ertzsprung-~~r~~ussell diagram." (2016), (visited on 04/01/2020).

[27] C. A. Bertulani, *Nuclei in the Cosmos. World Scientific.* 2013, ISBN: 978-981-4417-66-2. (visited on 04/01/2020).

[28] G. Rate and P. A. Crowther, "Unlocking Galactic Wolf–Rayet stars with Gaia DR2 – I. Distances and absolute magnitudes," *Monthly Notices of the Royal Astronomical Society*, vol. 493, no. 1, pp. 1512–1529, Jan. 2020, ISSN: 0035-8711. DOI: 10.1093/mnras/stz3614. eprint: https://academic.oup.com/mnras/article-pdf/493/1/1512/32649769/stz3614.pdf. [Online]. Available: https://doi.org/10.1093/mnras/stz3614.

[29] R. Gratton, A. Bragaglia, E. Carretta, V. D'Orazi, S. Lucatello, and A. Sollima, "What is a globular cluster? an observational perspective.," *The Astronomy and Astrophysics Review.*, vol. 27, no. 1, pp. 1–136, 2019. DOI: doi:10.1007/s00159-019-0119-3.