Abstract

There are $10^{11}$ galaxies in the universe and many more are getting discovered at a very fast rate. It becomes really necessary to classify these heavenly bodies into some appropriate classes to help them study further. There have been several approaches taken to classify these galaxies with varying degrees of success. One of the most important ways to classify the galaxies is based on their Morphological features. By morphological we mean the form and structure of the galaxy. The shape of the galaxy, the structure at the center of the galaxy are a few examples of its morphological features.

In this paper we focus our attention on six such features, namely the shape of the galaxy, the structure of the galaxy i.e. whether it is viewed edge on or not, what is the structure at the center of the galaxy i.e. whether it has a bar feature or a bulge at its center, the number of arms of the galaxy and how tightly are the arms wound to the galaxy.

1. Introduction and History

The problem we are tackling is a really interesting one and goes back to the advent of Astronomy. When the first galaxies were discovered, they were very few in number and thus a need of their classification was not a demanding task. But then with the advancement of technology there was a burst in the number of galaxies that were being found and the need of their classification arose.

Edwin Hubble classified the galaxies into something called the Hubble sequence, also known as Hubble tuning fork (Figure 1) in 1936. [1].

Edwin Hubble classified the galaxies into something called the Hubble sequence, also known as Hubble tuning fork (Figure 1) in 1936. [1].

This was purely done in a manual way relying on the images that were seen through the telescopes. There were few other systems, like the De Vaucouleurs system and the Yerkes (or Morgan) schemes that were invented to classify the galaxies [2]. We are now able to photograph many more distant galaxies than in the past, which makes the problem of classifying galaxies vastly more difficult. Most of the classification still depends on the human projects where thousands of volunteers around the world classify these galaxies into their respective classes manually.

For example, in 2007, a citizen science project called the Galaxy Zoo was launched, in which volunteers around the world were asked to classify the galaxies based on the images from the Sloan Digital Sky Survey (SDSS) database.[3] Although this approach worked relatively well, new ways of classification were sought after. This is primarily because of the following 2 reasons:

1. The images of the galaxies are generally noisy (Figure 2).
2. The humungous amount of database and its rate of increase. The SDSS database has more than 930,000 galaxies. [4]

Because galaxies are far away objects we have to rely on the obtained photos for the classification and thus computer vision techniques play a major role in such tasks.

For the project we took different approaches for different features which are described here concisely:

1. Detecting Shape: We take the arc length and area of the maximum contour and calculated the ratio of both according to an equation. The ratio determined the shape of the galaxy.
2. Detecting Viewing Angle: We compare the magnitudes of the gradients in two directions, the width and height of the galaxy. The ratio will determine whether the galaxy is viewed edge-on or face on.
3. Detecting the Existence of a Bar: We take the contrasted image and convert it into the HSV space.
maximum contour and its moment (orientation). After rotating the image by that orientation and we calculate the ratio between the semi-major and semi-minor axes. This ratio determines the presence of a bar feature.

4. Detecting the Number of Spiral Arms: We draw several lines, beginning at the origin, separated by a predetermined angle offset. By examining the distribution of black and white pixels along an individual line we can count the number of arms according to the intersections of the line with the galaxy.

![Figure 2: Example of a galaxy image from SDSS database](image)

5. Detecting the Tightness of the Arms: We use a bounding box around the galaxy to determine this feature.

6. Detecting the Existence of a Bulge: We pass a line across the width and height of the galaxy. The ratio determines the existence of a bulge.

We tested our methods on 500 images derived from the SDSS database. Our methods give very good results for 2 features namely, whether the galaxy is being viewed head-on or edge-on and whether it has a bar or not. It gave satisfactory results for 2 features, namely, the tightness of arms and the detection of a bulge. It fared slightly poorly as compared to the above features, for detecting of the shape and the number of arms of the galaxy. Thus we feel that there is a scope of improvement in the methods adopted. The results will be further discussed in the results section.

2. Related Works

In this section we will highlight some of the studies conducted in the field of galaxy image classification.

de la Calleja et al [5] conducted comparison study by applying three machine learning methods on galaxy image classification problem. These methods are Naive Bayes, C4.5 (extension of ID3 algorithm) and Random Forest. They tested their methods on NGC catalog released by the Astronomical Society of the Pacific. However, the galaxies in the most of those images are not located at the center. So, to make the galaxies position, scale and rotation invariant they applied Principal Component Analysis (PCA). They also used PCA to make the information more compact by reducing its dimensionality. Their results show that Random Forest performed better than Naive Bayes and C4.5.

In Banerji et al [6] they applied neural networks to classify images into three classes: early types, spirals and point sources/artifacts. They trained the neural network on 75000 galaxy images obtained from Sloan Digital Sky Survey. Those training images are associated with features already annotated by humans through Galaxy Zoo project. Then they tested it on one million galaxy images. They trained and tested the Neural Network using 3 sets on input parameters:
(a) Colors and profile fitting.
(b) Concentration and adaptive moments.
(c) The combination of both (a), (b).

Table 1 shows the percentage of correct classifications for each of these sets of parameters:

![Table 1: The accuracy of the various parameter sets.](image)

[7] is a galaxy images retrieval system where a galaxy is queried by providing a galaxy image as an input then the system will retrieve and rank the most similar galaxies. In order to accurately detect the galaxies, the images should be invariant to rotation and scale. To find the rotation angle they applied second moment of inertia. They proposed Spatial-Color Layout descriptor to encode both local and global morphological features. Then, they combined the proposed descriptor with the Kernelized Locality Sensitive Hashing for retrieval and ranking. They experimented by applying three kernels: Histogram Intersection, Chi-Square and Jensen-Shannon Divergence kernels. Histogram Intersection produced the best results with 95.8% accuracy.
3. Methodology

In total, we developed six unique techniques for the steps necessary in determining the characteristics of a galaxy. The six techniques were used to identify the shape of a galaxy, whether the galaxy is being viewed head-on or edge-on, whether it has a bulge or not, whether it has a bar or not, the number of spiral arms attached to it, and if it has spiral arms, how tightly the arms are bound. The SSDS images that we are using are colored images of size 424 x 424 in jpg format. The center of the galaxies are located at the center of the image which is helpful as we don't need to design feature detectors that are position invariant. Before we apply our feature detectors the images needed to pre-process. First, we converted the image into grayscale, and then convert the image into a binary image. Manually adjusting the binarization threshold was not practical due to the nature of the images. Some galaxies have sparse patterns, which got neglected during the conversion. To handle this, we applied Otsu thresholding which adaptively adjusts the threshold and thus gave us satisfactory results.

3.1 Detecting Shape

The shape of a given galaxy may be either circular, elliptical, or spiral. We can detect these shapes before converting the galaxy to a binary image. Although every galaxy image has noise due to the stars detecting about the galaxy, if we draw contours about every shape in the image, the galaxy will be the shape with the longest contour. We then calculate the area of the galaxy. Given the length of the contour about the galaxy, i.e. the length of the perimeter L, as well as the area A, we can plug these values into the isoperimetric inequality [9]:

\[ 4 \pi A \leq L^2 \]

By using the equation \( 4 \pi A / L^2 \), we can calculate a value which tells us how similar a shape is to a circle. A value of 1.0 indicates the shape is a perfect circle. As the value approaches zero, the shape is less like a circle. Although a value close to zero could be any shape, there are a limited number of shapes in the domain of shapes a galaxy may take, and thus we may assume galaxies with low values are spiral galaxies. An interim threshold value indicates an elliptical galaxy.

3.2 Detecting Viewing Angle

Galaxies may be viewed head-on or edge-on, much like a frisbee. That is, a frisbee may be viewed such that it appears to be a circle, or it may look like a thin bar, as when it is flying in the air. After a galaxy has been reduced to a binary image, our approach is to compare the magnitudes of the gradients in two directions, the width and height of the galaxy. In a circular galaxy, since the width and the height are approximately the same, the two gradients will be approximately the same. In a galaxy viewed edge-on, the width of the galaxy is much greater than the height of the galaxy, or vice versa. Thus, we can expect a mismatch in the gradients. To detect the two gradients of a circular galaxy, it is enough to take the gradients in the x and y direction. However, because an edge-on galaxy might be oriented at any degree, it is necessary to consider gradients in more than two directions. Fortunately, we found that simply rotating the binary image 45° and taking the x and y gradients a second time was sufficient for determining the viewing angle of a galaxy.

Thus, our procedure is to take the magnitude of the gradient of the binary image in the x and y directions, rotate the binary image 45°, and take the magnitude of the gradient in the x and y direction again. Then, we determine which of the two sets of magnitudes (x and y in the normal image, x and y in the rotated image) yields the greatest difference between the two magnitudes; doing so allows us to determine the difference between the height and width of a galaxy at any angle. In order to generate a value for thresholding, we use the ratio of the x magnitude over the y magnitude; higher values indicate a galaxy being viewed edge-on, whereas lower values indicate a galaxy being viewed head-on. This process enables us to reliably determine the viewing angle of a galaxy, and produces a byproduct that tells us the general orientation of the galaxy, which is useful in later techniques.

3.3 Detecting the Existence of a Bulge

If a galaxy is being viewed edge-on, it is possible to see whether the galaxy has a bulge about the center or not. We search for the presence of a bulge after detecting the viewing angle of the galaxy, as knowing the orientation of the galaxy is crucial to our bulge detection technique. Once we know the orientation of the galaxy, we pass a one pixel wide line along the width of the binary image of the galaxy. If the width of the galaxy is along the x-axis, we pass a line going from the top of the image to the bottom of the image from left to right. Likewise, if the width of the galaxy is along the y-axis, we pass a line going from the left side of the image to to the right, from top to bottom. If the width of the galaxy is oriented at some other angle, we use the rotated image instead of the normal image. Since galaxies are typically symmetrical perpendicular to the galaxy’s width, it is sufficient and computationally efficient to only pass the line across half of the width of the galaxy.

As we pass the bar over the galaxy we calculate the height of the galaxy by counting the number of white pixels along the line. Our method for detecting bulges is to look
for the sudden increase in the height of the galaxy over a short distance towards the center of the galaxy. In a galaxy without a bulge, we can expect the height of the galaxy to stay consistent along the full width of the galaxy. Therefore, we score a galaxy by the derivative of the height of the galaxy, applying a weight towards the center of the galaxy, and use a threshold to determine the presence of a bulge.

3.4 Detecting the Existence of a Bar

Galaxies may have bars in their center, which emit brighter light than the rest of the galaxy. The first step then, in exploiting the brightness of a potential bar, is to increase the contrast of the galaxy. We found contrasting the image in the HSV color scale instead of the RGB color scale produced better results for thresholding. The threshold for the bar may be set very high; we used a threshold value of 255. We then extract from the contrasted image the pixels which satisfy the threshold. After drawing a contour around the mass of bright pixels, we determine the width and height of the shape bounded by the contour. If the width is much greater than the height, a bar is present in the galaxy.

3.5 Detecting the Number of Spiral Arms

Galaxies may possess any number of spiral arms, although typically they possess two, four, or none at all. It is only possible to detect spiral arms on galaxies which are being viewed head-on. We also found it necessary to convert a galaxy to a binary image before attempting to detect spiral arms. Our procedure is to first calculate the pixels which belong to a series of lines radiating from the center of a galaxy. Each line is one pixel wide and extends a set length away from the center of the galaxy. The set length is a thresholded value that is sufficient to extend beyond the limits of the galaxy. A length extending just beyond the limits of the galaxy is used to avoid wasting computation time determining the pixel values of empty space. The lines are drawn at 10° intervals about the galaxy. Then, by examining the distribution of black and white pixels along an individual line, we can determine whether the line intersects a spiral arm. If, beginning at the point which lies in the galaxy and going outwards to the point which lies in space, we simply see a large distribution of white pixels followed by a large distribution of black pixels, we know the line does not intersect a spiral arm. If, however, we see a large distribution of white pixels, followed by a small distribution of black pixels, a small distribution of white pixels, then a large distribution of black pixels, we can deduce the line intersects a spiral arm. The small distribution of black pixels is the empty space in between the galaxy and the spiral arm, and the small distribution of white pixels is the spiral arm itself. At 10° intervals, multiple lines might intersect the same spiral arm; to avoid this, we use a flag which initially assumes that a spiral arm has not been detected. Beginning at the line at 0°, we then check for a spiral arm. If a spiral arm exists, we modify the flag, and do not change it back until we no longer detect a spiral arm as we move through each subsequent line.

3.6 Detecting the Tightness of the Arms

Given a galaxy for which spiral arms have been detected, the spiral arms may be tightly wound about the galaxy, loosely bound, or somewhere in between. Our approach for detecting how tightness of the arms is very simple. We fit the smallest bounding box around the binary image of a galaxy which does not intersect the arms of the galaxy; that is, the bounding box which best fits the galaxy. We then take the ratio of black pixels over white pixels within the bounding box. Galaxies with tightly bounded arms will tend to have smaller bounding boxes and have a lower ratio of black pixels to white pixels, whereas galaxies with loosely bounded arms tend to have bounding boxes which extend well beyond the center of the galaxy, and thus tend to have a high ratio of black pixels over white pixels. We take into consideration the number of spiral arms present, as it is possible that a spiral galaxy with four arms might have roughly the same bounding box as a galaxy with two arms. In such a situation, the bounding box of the galaxy with four arms would contain significantly more white pixels and thus it would be difficult to set a thresholding value which would work across all galaxies.

4. Experiments and Results

We initially compared the results of our algorithms to a set of 40 images hand selected from the Galaxy Zoo dataset. The images were hand selected to ensure we had at least several images strongly featuring each characteristic we wanted to detect. If we had randomly selected a set of images, it would’ve been likely that certain galaxy features which are rare would not have appeared in the sample set, and thus we would not have been able to measure the effectiveness of our algorithms. Each sample image scored above 90% probability, as determined by the users of Galaxy Zoo, that the feature for which it was selected was present. For each feature, our program simply returns a vote either in favor of or against the feature existing in the image, and therefore we converted the probabilities from Galaxy Zoo into boolean values. Certain features in the Galaxy Zoo detection flow-chart were not accounted for, as we did not develop methods for
detecting those features. After running our program on the 40 images to determine whether or not our program could perform reasonably well at detecting each feature, we randomly sampled 500 images to run our program on.

Table 2 shows the accuracy for each feature detector.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Percentage Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape Detector</td>
<td>42.6</td>
</tr>
<tr>
<td>View Angle Detector</td>
<td>88.2</td>
</tr>
<tr>
<td>Number of Arms Detector</td>
<td>88.6</td>
</tr>
<tr>
<td>Bar Detector</td>
<td>40.8</td>
</tr>
<tr>
<td>Arm Tightness Detector</td>
<td>60.8</td>
</tr>
<tr>
<td>- Bulge Detector</td>
<td>65.8</td>
</tr>
</tbody>
</table>

Table 2: The accuracy of the feature detectors using 500 images

As the table shows, the most robust detectors are view angle and bar. Regarding the shape detector, we believe that the reason why it performed poorly because it is heavily dependent on thresholding the isoperimetric inequality equation. Most of the false positives are due to the detector inability to distinguish between round (circular, elliptical) and spiral galaxies. In other words, if the detector task were to detect only whether the galaxy is circular or elliptical, then we would have obtained better results. However, introducing spiral galaxies to the task made the detector less robust. The low accuracy for the arm detection is due to the conversion to a binary image through a threshold. Although the Otsu’s method threshold successfully captured all the details that we needed for our analysis, in some cases the threshold will throw some arms out because some arms generally are sparse stars with low intensities. Conversely, in other cases some artifacts were counted as arms. Both arm tightness and bulge detectors achieved reasonable accuracy. However, they are dependent on relative metrics and thresholds. We would have achieved better results if we applied more robust metrics.

5. Unsuccessful Experiments

During the course of this project there were a few setbacks as well. There were several times when we tried a method for couple of days and then had to reject it as either it was not doing reasonably well or was failing poorly for the said feature. We discuss a couple of methods here:

5.1 Laplacian of Gaussian (LoG)

We tried the LoG method to find out the number of arms of a given galaxy. Since the arms are present in the galaxy only when its spiral, our idea was to apply LoG filter to the galaxy image. Because the LoG is a negative bell shaped function we believed that it will help to maximize the values in the ends of the contour image. That way we can get the points in the far ends of the spiral image thereby giving us a good indication of the presence of arms because for a spiral image the far ends of the image will be farther apart. Our initial plan was to find the distance between the farthest points in the image and above a certain threshold value for the distance we could say that the galaxy has arms.

We did not adapt this method because of the following reasons:

a) Some of the spiral galaxies have tightly bounded arms. Even if we get the farthest points in the image relative to the arms it might not cross the threshold and might be classified incorrectly as not having any arms.

b) The contour images were not accurate.

c) Some of the circular and elliptical shaped galaxies were very huge i.e. covered a lot of area in the images. If we find the farthest points in these images and get the distance, it might easily go over the threshold and these might be classified as having arms.

5.2 Blob Detection

We tried the Blob detection method to detect the bar features in the galaxy images. When we convert the galaxy images to their respective contrasted image, the bar feature in the images actually outshines the rest of the galaxy. As blob detection detects regions having differences in properties, such as brightness, we thought that we would be able to extract the area at the center of the image as they seemed to be of the same intensity. Our plan was that, if we were able to extract the area at the center, we could use contours to do the shape detection.

We did not adapt this method because of the following reasons:

a) When we tested the center of the contrasted image, pixel by pixel, we found that the intensities were not equal.

b) The bar in the images was at different angles (Figure 3) and thus we needed a feature detector that was rotation invariant.
6. Strengths and Weaknesses

6.1 The strengths:
1. Simple techniques.
2. Majority of the features (4 out of 6) getting detected with satisfying level of accuracy.
3. The detecting bar feature, detecting the tightness of arms feature, detecting the viewing angle feature are rotation invariant.
4. The computation time is reasonable (for the tested 500 images).

6.2 The weaknesses:
1. Detecting false arms and missing present arms.
2. Some feature detectors are heavily dependent on the thresholds.
3. Excessive use of contours for detecting shapes.

7. Future work

In the future work we plan to do the following:

a) We need to define a new robust technique for two features, detecting the arms and detecting the shape of the galaxy.

b) We need to make the features, detecting the bulge and detecting the arm tightness, more robust.

c) We need to run our feature detectors on a larger dataset.

d) We need to extend our feature detection to detect other features, which were not accounted for in this project.

8. Conclusion

We described different techniques for extracting several morphological features from the galaxy images. We used a dataset of 500 random images. We were able to get a varying level of accuracy with these methods for their respective features. Some of them were quite robust, namely, the accuracy for detecting the bar and the viewing angle was in high 80’s. Some of them were not so robust, namely, accuracy for detecting the shape and number of arms in a galaxy was in the low 40’s. We learnt from this project that we cannot have a single way to extract all the desired features which can then be used for the classification of these galaxies. For creating a system that can classify the galaxies in the images into their subsequent categories correctly, extracting these features will play a pivotal role as this will the first step in that direction.

9. References