Classification of Galaxy Morphologies Using *Support Vector Machines*

A. Guzman-Ballen
University of Illinois at
Urbana - Champaign
aguzman4@illinois.edu

E. Montagner
University of Illinois at
Urbana - Champaign
montgnr2@illinois.edu

J. V. Ruiz Cepeda
University of Illinois at
Urbana - Champaign
ruizcep2@illinois.edu

ABSTRACT

*We present a method to classify galaxy morphologies using Support Vector Machines (*SVM*). Using the dataset provided by* GalaxyZoo*, which contains over 60,000 images of galaxies, we constructed a codebook of clustered* SIFT *descriptors, and used this database to train a corpus of* SVM*. We find that the preprocessing of the data is very useful in increasing the speed of execution because the pictures of the galaxies contain significant amount of noise. We also found that using a single descriptor* SIFT *or* HOG *descriptors achieves similar results. The* SVM *were able to classify the pictures of galaxies into their correct class with 30% accuracy. We conclude that the use of* SVM *and feature descriptors to identify galaxy morphologies is not enough in terms of accuracy, and further research into other methods for preprocessing and image descriptors will be done to solve this problem.*

Keywords: galaxies, support vector machines, machine learning, computer vision.

1. INTRODUCTION

The classification of galaxies based on shape, size, and color is important because understanding galaxy morphology can help us to better understand our universe by helping us answer questions regarding the distribution of galaxies within the universe and the origin of the universe.

Every day, telescopes around the world capture countless images of galaxies. Thanks to technological advancements in astronomy and telescope design, the rate at which these pictures are captured has increased, and as a result, the size of the picture database is growing very rapidly. Until now, the classification of the images is done manually, and due to the size of the database, the need for automatic computational methods to classify the galaxies is of great importance.

In an effort to find such a computational method, *GalaxyZoo* teamed up with *Winton Capital and Kaggle* [1]. Over 60,000 images of galaxies were classified manually with the the help of hundreds of thousands of volunteers to obtain a training set. Then, using *kaggle.com*, a competition with a prize of $16,000 was initiated to find automated methods that would reproduce the manual classification results.

The classification of galaxies by human eye is still common, but the potential and effectiveness of automated methods has only recently been explored. Most of this work relies on *Machine Learning* methods, and neural networks have proven to be the most promising. Lahav et al. (1995, 1996) [9, 10] demonstrated that *Artificial Neural Networks* are effective in reproducing visual classifications of the galaxies. This dataset has already been used for morphological classification using automated *Machine Learning* techniques (Ball et al. 2004) [5]. *Artificial Neural Networks* were also used to classify stellar spectra (von Hippel et al. 1994; Bailer-Jones et al. 1998) [13, 4] and galaxy morphologies (Storrie-Lombardi et al. 1992; Naim et al. 1995; Folkes et al. 1996) [12, 11, 8]. The size of astronomical data has increased significantly in the past decade, so new efficient methods that can scale effectively are necessary. Using the *GalaxyZoo* and *Winton Capital* dataset, Banerji et al [6] were able to effectively apply *Neural Networks* to the new massive dataset, and they obtained over 90% accuracy.

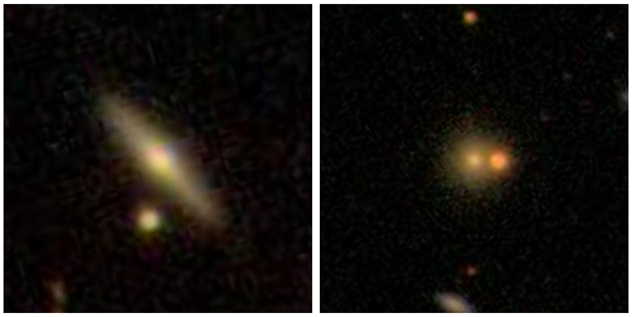

Figure 1: Example of galaxies that are part of the dataset.

The *GalaxyZoo* dataset presents us with unique opportunity to compare human classifications to those from automated *Machine Learning* algorithms. Up until now, the use of *Machine Learning* has been proven promising to visual classification problems, and using automated methods would save us time and resources in the future. Thus, our goal for this project is to design a procedure that performs the classification automatically and reproduces the results of the manual classification. Instead of delving deeper into the application of *Neural Networks*, we will explore the potential of *SVM* to the same problem of classification.

2. DATASET

The galaxy dataset used in this paper was prepared by *GalaxyZoo* and made available thanks to the efforts of *Winton Capital* and *Kaggle*. The dataset consists of 61,578 different images of galaxies in JPG format. The images of the galaxies vary significantly in shape and size, and many of the images contain different artifacts that make the classification difficult, such as light intensity changes, camera artifacts, and the presence of bright stars.

The training and testing sets were prepared manually with the help of thousands of volunteers. Each user classified the galaxy according to a decision tree. At each node, the user answers a questions and follows the path down the tree depending on the answer. Sample questions are: "*Is the object a smooth galaxy, a galaxy with features/disk or a star?*" (three responses are possible for this question), "*Is there a spiral pattern?*" (two responses possible). Essentially, the classification of a galaxy corresponds to a specific path down the decision tree.

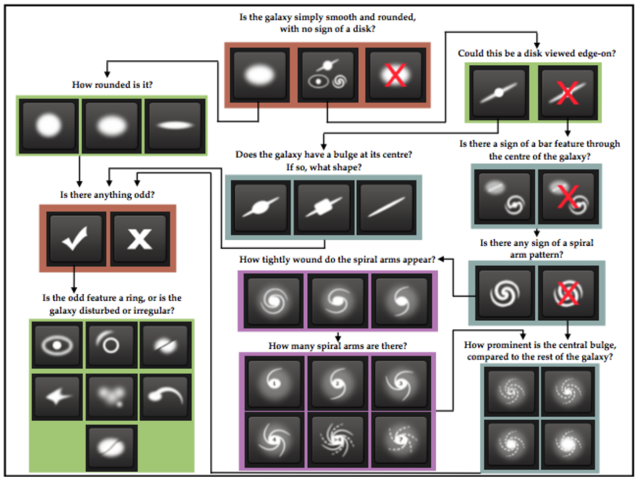
Is the galaxy simply smooth and rounded, with no sign of a disk?
Could this be a disk viewed edge-on?
How rounded is it?
Does the galaxy have a bulge at its centre? If so, what shape?
Is there a sign of a bar feature through the centre of the galaxy?
Is there anything odd?
Is there any sign of a spiral arm pattern?
How tightly wound do the spiral arms appear?
Is the odd feature a ring, or is the galaxy disturbed or irregular?
How many spiral arms are there?
How prominent is the central bulge, compared to the rest of the galaxy?

Figure 2: Decision tree used to classify galaxies.

The galaxy morphology is weighted such that users who tend to agree with the majority are given a higher weight than those who dont. Thus, the final morphology of the object is a weighted mean of the classifications of all users who analyzed it. As an example, the following is the manual classification results for galaxy id 100008:

Galaxy ID	Class 1.1	Class 1.2	Class 1.3
1000008	0.383147	0.616853	0.0

This means that, at the top of the decision tree, 38.31% of users chose the path that classified galaxy 100008 into *Class 1.1* for the first question, 61.69% of users chose the path that classified galaxy 1000008 into *Class 1.2* for the first question, and 0% into *Class 1.3*, the last path for the first question.

The images below show the most representative images from classes *1.1*, *1.2*, and *1.3* (i.e. the top two images that were classified with the most confidence into those classes by *GalaxyZoo*).

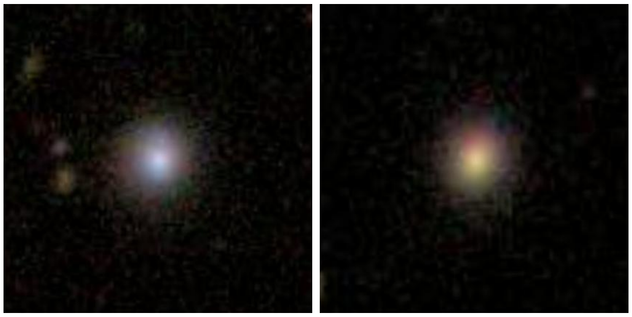

Figure 3: Examples of *Class 1.1* (round galaxies).

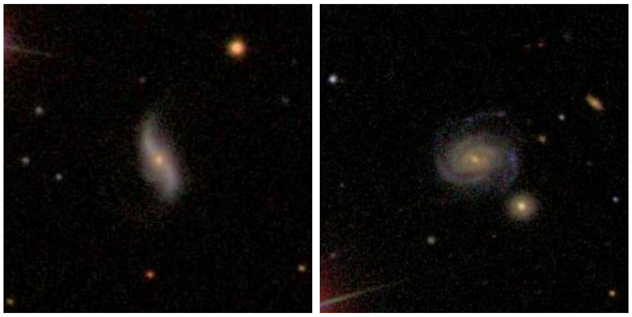

Figure 4: Examples of *Class 1.2* (spiral galaxies).

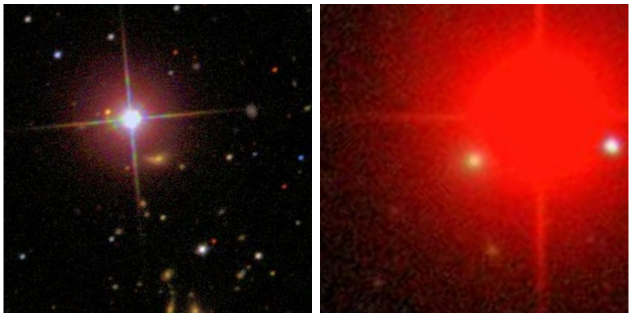

Figure 5: Examples of *Class 1.3* (starts/noise).

From these three classes and their most representative members it can be seen that certain features do seem to vary significantly from class to class. Thus, some feature engineering (ellipsis eccentricity, galaxy radius, and brightness could) could help in the classification process.

To match the time requirements of this project and because of the lack of experience dealing with *probability distributions* in a classification problem, we decided to simplify the labels, so that the shape of a galaxy is the one determined by the result of following the highest classification probabilities through the decision tree, so that we finally had a total of 37 possible categories.

3. CLASSIFICATION

3.1 Preprocessing

Since the images show a good amount of noise and artifact, we thought about preprocess the data to get better results. We wanted to feed into the algorithm, only the center galaxy, so we used the method exposed in [7], that allows to remove everything surrounding the center galaxy. As the paper explains:

> "*Firstly, for a given galaxy image, we convert the image from* `RGB` *color space to gray-scale and use a spatial* Gaussian Smoothing Filter *to remove the noise from the gray-scale image. Secondly, we use* the Otsus method *to adaptively find the threshold and segment the original image into a binary image. After that, we fill the holes in the binary image based on morphological operations. Then, we find the biggest connected component located in the center as our target galaxy.*"

We applied this method with very good results, as can be seen in the following figures.

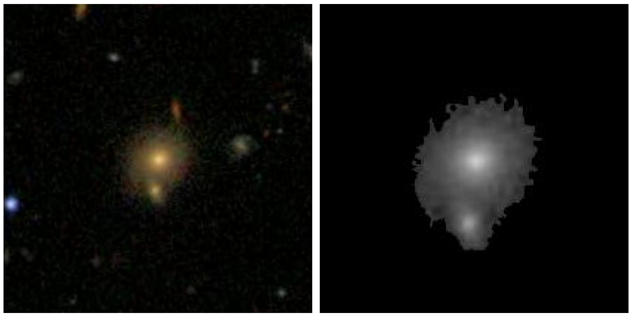

Figure 6: Before and after applying the preprocessing step.

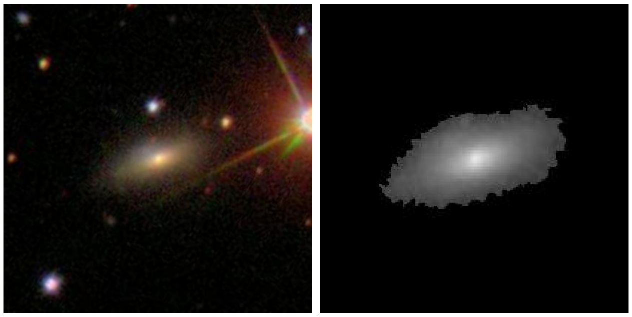

Figure 7: Before and after applying the preprocessing step.

3.2 Approach

Once we have removed parts of the images that may conflict with our approach and extracted only the center galaxy, we obtain the *SIFT* descriptors for each image in the training set. For that purpose, the *VLFeat* implementation of *SIFT* was used [2].

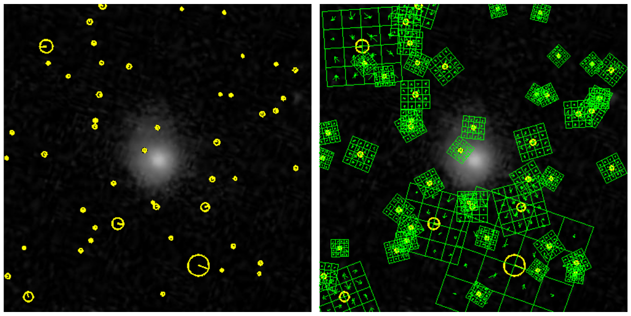

Figure 8: *SIFT* feature descriptors and frames for a *non-clean image* (number of descriptors limited to 50). As can be seen, most the descriptors are taken from the noise instead from the actual galaxy.

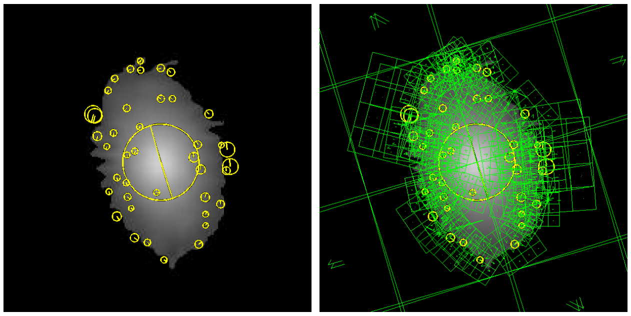

Figure 9: *SIFT* feature descriptors and frames for a *clean image* (number of descriptors limited to 50).

Then, we quantize the *SIFT* descriptors using *K-means* clustering. The resulting clusters are then stored in a database (a codebook) to be used while quantizing the test samples. The quantification of the *SIFT* descriptors is required to normalize them and make them usable in the classification step.

Next, we train the *SVM* using the quantized descriptors from the training data. Once, the *SVM* have been trained (one for each response to every question in the decision tree), we use them to classify the test images.

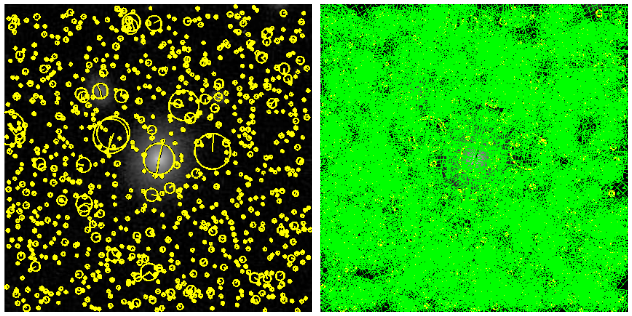

Figure 10: *SIFT* feature descriptors and frames for a *non-clean image* (non-limited number of descriptors). Again, it can be seen, most the descriptors are taken from the noise instead from the actual galaxy.

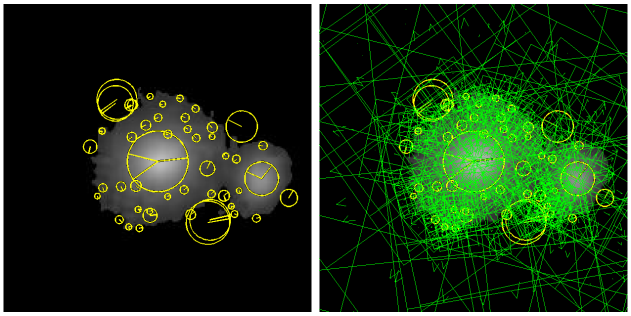

Figure 11: *SIFT* feature descriptors and frames for a *clean image* (non-limited number of descriptors).

In order to analyze and compare the effectiveness of our approach, we performed two modifications to our method and then looked at the resulting differences.

The first modification we tried was to use *HOG* descriptors instead of *SIFT* descriptors. Again, the *VLFeat* implementation was used [2]. *HOG* descriptors are represented by a three-dimensional matrix, whose dimensions depend on a parameter, the *cell size*. Thus, for a fixed *cell size*, all the images produce matrices of the same size. As a result, no vector quantification is needed. Then, each *HOG* feature descriptor matrix is transformed into a vector, which we call the feature vector and feed into the *SVM*.

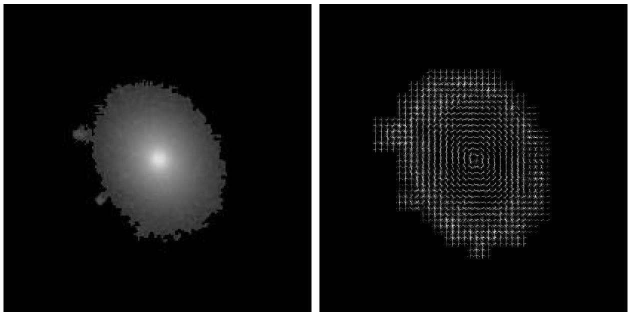

Figure 12: *HOG* feature descriptors for a *clean image* (*cell size* of 4).

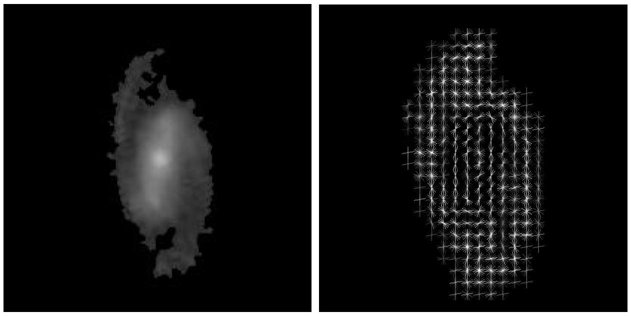

Figure 13: *HOG* feature descriptors for a *clean image* (*cell size* of 8).

The second modification still uses *SIFT* descriptors; however, for every image we only calculate one 128-dimensional *SIFT* descriptor centered and with a given *scale*. Here too, quantification is unnecessary, so we feed this *SIFT* descriptor into the *SVM*.

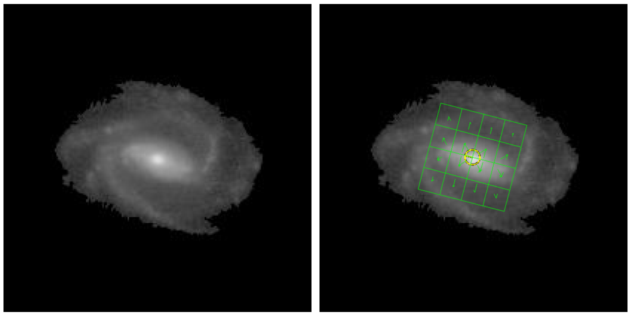

Figure 14: Single-*SIFT* feature descriptor for a *clean image* (*scale* of 5)

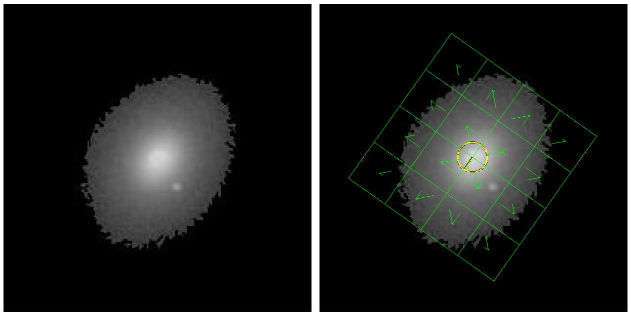

Figure 15: Single-*SIFT* feature descriptor for a *clean image* (*scale* of 10)

3.3 Results

The number of available training images from the dataset is 61,578, while the number of testing images is 79,975. Because obtaining the *SIFT* descriptors of an image is a computationally intensive task, we used only a randomly sampled part of training data and testing date to create both datasets.

First, we compared how the three approaches behave in terms of accuracy and execution time depending on the preprocessing of the dataset. In one hand, the clean test data, it is the result of filtering the image to keep only the galaxy while reducing the noise, as explained in the preprocessing section. On the other hand, the test data is the one coming directly from the original dataset.

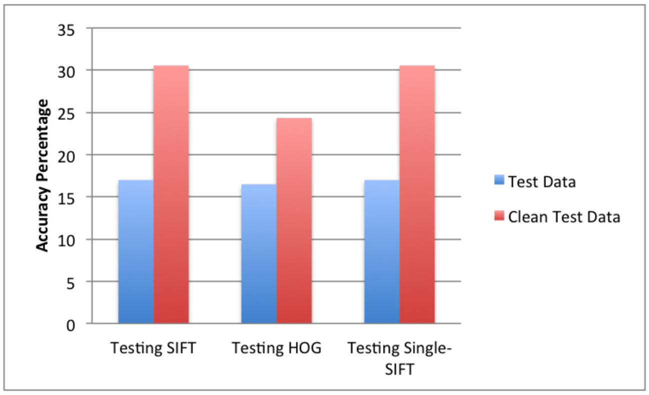

Figure 16: Comparison of the accuracy obtained depending on the preprocessing of the dataset. [*Results obtained using 500 training images and 200 testing images with 300 clusters and a* lambda *of 1E-05.*].

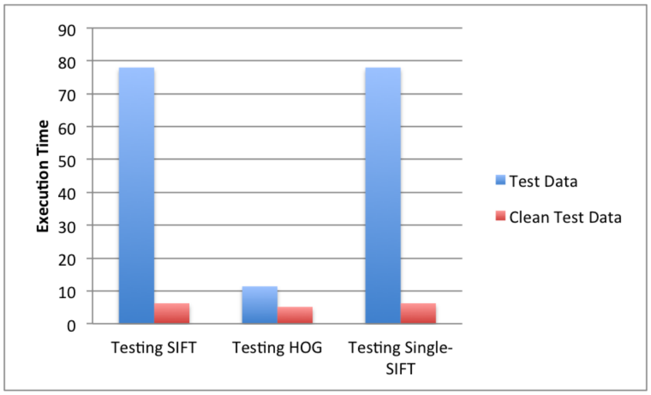

Figure 17: Comparison of the execution time (in seconds) obtained depending on the preprocessing of the dataset. [*Results obtained using 500 training images and 800 testing images with 200 clusters and a* lambda *of 1E-05.*].

As can be seen, in any case, the preprocessed dataset obtains better results both in terms of accuracy and execution time. Therefore, the rest of the results have been gathered using this clean data.

Next, we compared how the amount of training data affected the accuracy of the classification of test data. It is clear that it doesn't make a significant change in the accuracy.

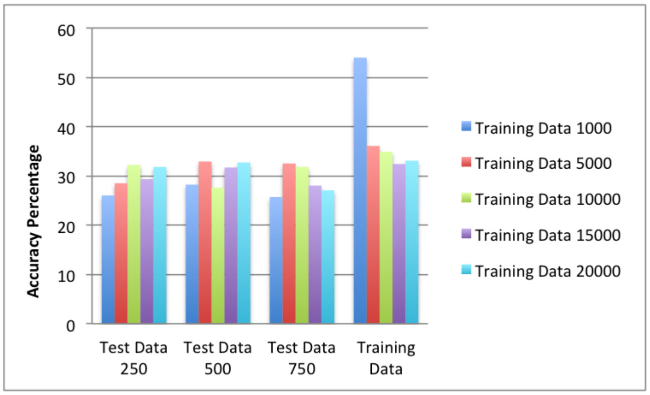

Figure 18: Comparison of how the amount of training data affects the accuracy of classification of test data (*y-axis*). As can be seen, the amount of training data doesn't seem to affect the accuracy obtained. [*Results obtained using 300 clusters and 1E-03* lambda *for* SVM].

We also experimented with the *lambda* parameter of the *SVM*, which also determines the number of iterations, and with the number of clusters used in *K-means* (*K* parameter).

On one hand, in the case of the training data, it can be seen that as we increased any of them, the accuracy of the results improves. On the other hand, for the test data, a lower *lambda* increases the accuracy till a certain point, probably due to *overfitting* when the value is too low. The number of clusters doesn't seem to have a significant impact on the results.

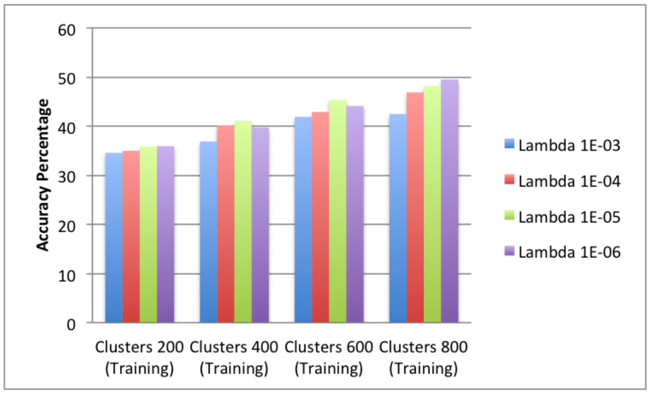

Figure 19: Comparison of how the variations of the number of clusters and the *lambda* parameter of *SVM* vary the resulting accuracy in the training set. The growing trend seems pretty clear. [*Results obtained using 5,000 training images*].

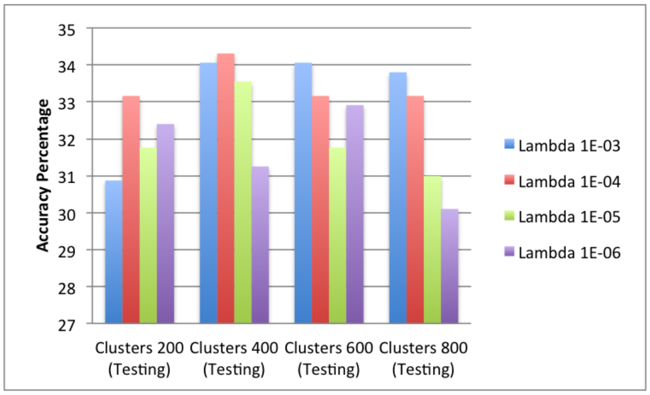

Figure 20: Comparison of how the variations of the number of clusters and the *lambda* parameter of *SVM* vary the resulting accuracy in the test set. The results show the *overfitting* point of the *SVM* and how the number of clusters doesn't affect significantly. [*Results obtained using 5,000 training images and 800 testing images.*]

The effect of changes in the scale of the *single-SIFT* approach is shown in the next figure. The results might suggest that the center part of the galaxies contain more information regarding their morphology than the outside part. This information seems to become undetectable when the *scale* grows up to a certain point, but it appears again when the *scale* matches more or less the size of the galaxy. Clearly, after some point this trend dissapears since the descriptor is much bigger than the galaxy and a lot of noise is added to the feature vector.

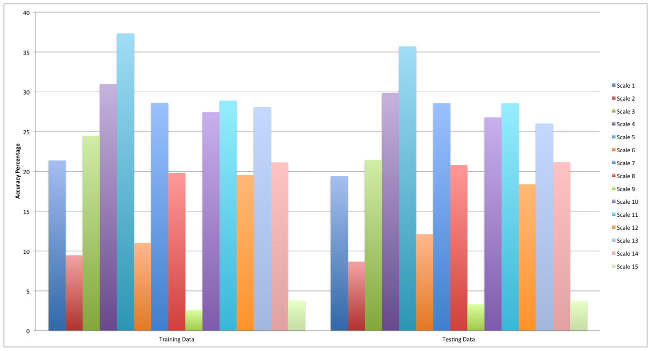

Figure 21: Comparison of how the variations of the number of the scale in the *single-SIFT* descriptor approach varies the resulting accuracy in the test set. The results might suggest the distribution of the shape information in the galaxies. [*Results obtained using 5,000 training images and 800 testing images with a* lambda *of 1E-05.*]

The same comparison was performed in the case of *HOG* feature descriptors. The results, in this case, show clearly how a different *cell size* used while computing the results doesn't affect them.

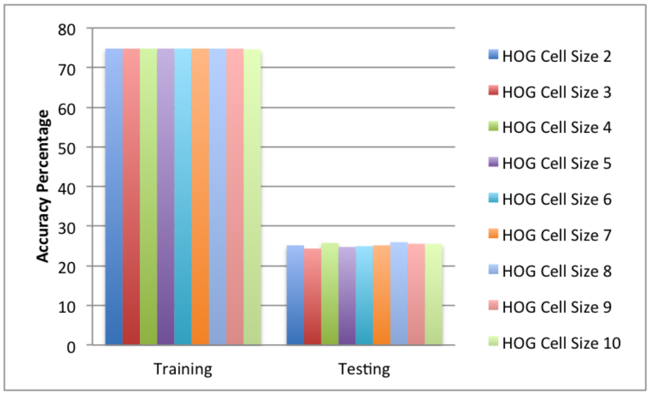

Figure 22: Comparison of how the variations of the *cell size* using *HOG* varies the resulting accuracy in the test set. Clearly, there is no effect. [*Results obtained using 1,000 training images and 500 testing images with a* lambda *of 1E-05.*]

3.4 Conclusions

In this study, we used a *Machine Learning* algorithm based on *Support Vector Machines* (*SVM*) to perform morphological classifications of images of galaxies taken by modern telescopes. The training and testing input images were manually classified with the help of thousands of volunteers as part of a *GalaxyZoo* project.

We considered using many different kinds of descriptors but some of them didnt fit with our objective. For instance, *GIST* descriptors and *bags-of-features* are often used for whole images. This could be useful if one were trying to program something similar to *Google Goggles*, an app that can detect whether a picture is showing something well-known, like the *Eiffel Tower* or the *Statue of Liberty*.

In the end, we opted for *SIFT* descriptors because of their ease of use and highly descriptive capabilities. We quantized the descriptors to normalize them into a uniform size, and we then classified the different galaxy images using the *SVM*. A total of 37 *SVM* were trained, corresponding to each of the 37 possible answers that can be found in the decision tree. We are able to reproduce the human classifications for the galaxies with 30% accuracy on the testing data. We implemented different variations of our original approach, and we measured and compared how the different parameters affected the accuracy and the time complexity of each modification. In the end, the use quantified *SIFT* descriptors produced the best results.

SIFT and *HOG* descriptors were the most suited descriptors for our problem. Some limitations to our approach, however, are that *SIFT* descriptors take a long time to extract, and extracting 128 dimensional descriptors for the 61,578 images was computationally intensive, so we decided to randomly sample the datasets given our computational limits. Furthermore, we did not take into account other descriptors which could have increased the accuracy of our classifier. For example, some feature engineering such as extracting the radius of the galaxy and the eccentricity of the ellipse could have helped us distinguish better between round and elliptical galaxies.

While 30% is not a satisfactory result, it is a good starting point. Further research into better feature engineering and preprocessing tool could help increase the accuracy rate. As it is, this approach is not sufficient as a solution to the problem of computationally classifying the morphologies of galaxies. The use of *Convolutional Neural Networks* (*CVN*) seems to be the most promising approach to this problem, as the winner and runner ups of the contest both relied on the power of *CVN* to perform the classification.

4. INDIVIDUAL CONTRIBUTION

Our team has had weekly meetings to make sure our project is making progress. When it comes to division of labor, it was decided that it would be determined based on the strengths and interests of each of the members.

Ettienne and Andres, through researching different papers on the subject matter, managed to find a way to extract the center galaxies from the image in the process and implemented it. Jose worked on implementing and using *SVM* and *SIFT* descriptors to classify the given galaxies. Ettienne researched other potential methods for feature extraction and for galaxy classification, such as using eigenfaces and neural networks.

We collectively read through research papers to discover more approaches. Using *MATLAB* and version control system, *GitHub*, made collaboration very easy to do. We all contributed to the writing of the report by dividing the sections evenly. The poster was also designed in this way. Together, we chose which images to display in the poster and what information seemed the most relevant and explanatory.

5. REFERENCES

[1] *Kaggle GalaxyZoo* Challenge website.

[2] VLFeat website, used for *SIFT* and *HOG* feature extractors.

[3] Abdalla F. B., Amara A., Capak P., Cypriano E. S., Lahav O., Rhodes J., 2008, MNRAS, 387, 969

[4] Bailer-Jones C. A. L., Irwin M., von Hippel T., 1998, MN- RAS, 298, 361

[5] Ball N. M., Loveday J., Fukugita M., Nakamura O., Oka- mura S., Brinkmann J., Brunner R. J., 2004, MNRAS, 348, 1038

[6] Banerji M., Abdalla F. B., Lahav O., Lin H., 2008, MN- RAS, 386, 1219

[7] Cui Y., Xiang Y., Rong K., Feris R., Cao L., 2014, WACV

[8] Folkes S. R., Lahav O., Maddox S. J., 1996, MNRAS, 283, 651

[9] Lahav O., Naim A., Buta R. J., Corwin H. G., de Vaucouleurs G., Dressler A., Huchra J. P., van den Bergh S., Raychaudhury S., Sodre Jr. L., Storrie-Lombardi M. C., 1995, Science, 267, 859

[10] Lahav O., Naim A., Sodre Jr. L., Storrie-Lombardi M. C., 1996, MNRAS, 283, 207

[11] Naim A., Lahav O., Sodre Jr. L., Storrie-Lombardi M. C., 1995, MNRAS, 275, 567

[12] Storrie-Lombardi M. C., Lahav O., Sodre Jr. L., Storrie- Lombardi L. J., 1992, MNRAS, 259, 8P

[13] von Hippel T., Storrie-Lombardi L. J., Storrie-Lombardi M. C., Irwin M. J., 1994, MNRAS, 269, 97