Galaxy Zoo Challenge: Classify Galaxy Morphologies from Images

Fang-Chieh Chou (fcchou@stanford.edu)

March 19, 2014

Abstract

Image classification has been an important subject in computer vision. The recent kaggle challenge provides an interesting dataset of galaxy images. The goal is to build a predictive algorithm to extract indicators for the galaxy morphologies as viewed in human eyes. This provides a chance to test how state-of-the-art computer vision and machine learning methods can be applied to understand a dataset quite distinct from the well-tested scene/face/object detection benchmarks in computer vision research. We developed a pipeline combining multiple computer vision feature detectors and machine learning regression, and experimented the performance using cross validation technique. Our results suggest that features describing the global shape of the galaxy gives better result then local interesting point detector, and that deep learning is a promising technique for the challenge.

1 Introduction

The shape (or morphology) of a galaxy provides powerful information on the physics of galaxy formation. As the number and sizes of the telescopes increase, the amount of available images of galaxy has quickly exceeded the level analyzable by one single scientist. To tackle this challenge, the physicists have created the Galaxy Zoo project (http://www.galaxyzoo.org/), where galaxy images are classified by multiple online players through crowd sourcing [9]. With the large amount of galaxy images with human classifications from the project, it is now possible to create a prediction algorithm that automatically classify the galaxy morphology using state-of-the-art computer vision and machine learning techniques. The Galaxy Zoo challenge is a recent kaggle competition (http://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge) where online data scientists attempt to create the best algorithm to confidently classify the shape of the galaxy, based on available data from Galaxy Zoo containing more than 60,000 human-classified galaxy images as training set. Equal amount of image classifications are held secret as the test data. This challenge provides a great opportunity to test the current image classification approaches and to build an application that will have important impact on advancement of physics research.

2 Related Works

Image classification has been a critical subject in computer vision and has been widely studied. Multiple algorithms have been developed to classify popular public image datasets including the MNIST digit dataset, CIFAR tiny image dataset, Caltech 101 dataset, and ImageNet, which is the largest image classification dataset to date, including more than 14 million images. Various algorithms have been demonstrated to be successful in these classification tasks, for example Size Invariant Feature Transform (SIFT), shape context, Histogram of Gradients, deep convolutional neural network, etc. However, these image datasets and algorithms are targeted at the classification of natural scenes and objects, which is distinct from the galaxy images we are focusing here. A recent work in astronomical physics gives promising results (90% accuracy) in distinguishing between

smooth galaxies and spiral galaxies using shape-based features and artificial neural network. In this challenge, we will look at more complicated features, beyond just the spiral-ness of the galaxies, and test the performance of state-of-the-art computer vision methods on the galaxy dataset.

3 Methods

3.1 The Galaxy Zoo Dataset

The Galaxy Zoo dataset is provided by the kaggle challenge. 61,578 galaxy images with human classifications are given as the training data, and 79,975 galaxy images, whose classifications are held secret, are used as the test set to evaluate the performance of the algorithm. Each galaxy image is an RGB image of the 424×424 pixels, with the target galaxy at the center of the image.

For each galaxy image, a complex, decision-tree-based questionnaire is asked to multiple users through the Galaxy Zoo crowd sourcing project (Fig. 1). These questions guide the player to describe the shape of the galaxy from a global view (e.g. smooth vs. spiral, top view vs. edge view) to finer features (e.g. round vs. elliptical, number of spiral arms). Briefly, the users are asked to answer 11 possible multiple-choice questions with 37 possible responses in total. For each response, the percentage of users selecting it is reported as the target value for the prediction challenge. The goal of the kaggle challenge is to develop a confident prediction algorithm to predict the probability that a user selects a response, based solely on the given galaxy image.

To evaluate the performance of the algorithm, the root mean squared error (RMSE) across all 37 responses is computed as the single statistic for evaluating the algorithm:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - a_i)^2}$$

Here N is the number of galaxies times the total number of responses, p_i is the predicted value by the algorithm, and a_i is the actual value.

3.2 Overall Pipeline

To solve the prediction challenge, we developed an algorithm pipeline consists of 3 parts: feature extraction, machine learning regression, and probability normalization. First, a feature vector that is extracted from each galaxy image to represent the characteristics of the image. Here we tested various feature extraction methods widely used in computer vision, and also developed a few new data-specific feature extraction schemes. These methods include image color, PCA (Principle Component Analysis), SIFT (Scale-Invariant Feature Transform), HOG (Histogram of Oriented Gradients), etc. The feature vectors from different extraction schemes are concatenated into one single vector for each image. I then applied supervised machine learning regression methods to predict the probability of each response. Here we tested the simplest linear regression method, the Ridge regression (a regularized linear regression method), and random forest regression. Finally, to ensure the predictions from the regression obeys the probability constraints, i.e. the probability should be a real number between 0 and 1, and the probability for each question should sum to 1, I further normalized the prediction. The details of each step of the pipeline are discussed below.

3.3 Feature Extraction

3.3.1 Galaxy Center Color

As suggested by previous physics researches, the color of the galaxy provide critical information on its formation history, and therefore is a useful feature correlated with the galaxy morphology [1]. By visual inspection of the data, we found that most of the galaxies have uniform colors. Therefore we just used the RGB values of the pixel at galaxy center to represent the color of the entire galaxy. Since our galaxies have been centered in the image, we can simply use of color of the center pixel of each image as the feature. This feature is the only feature involving colors. All the following feature extraction scheme is based on the gray-scale images converted from the original RGB images.

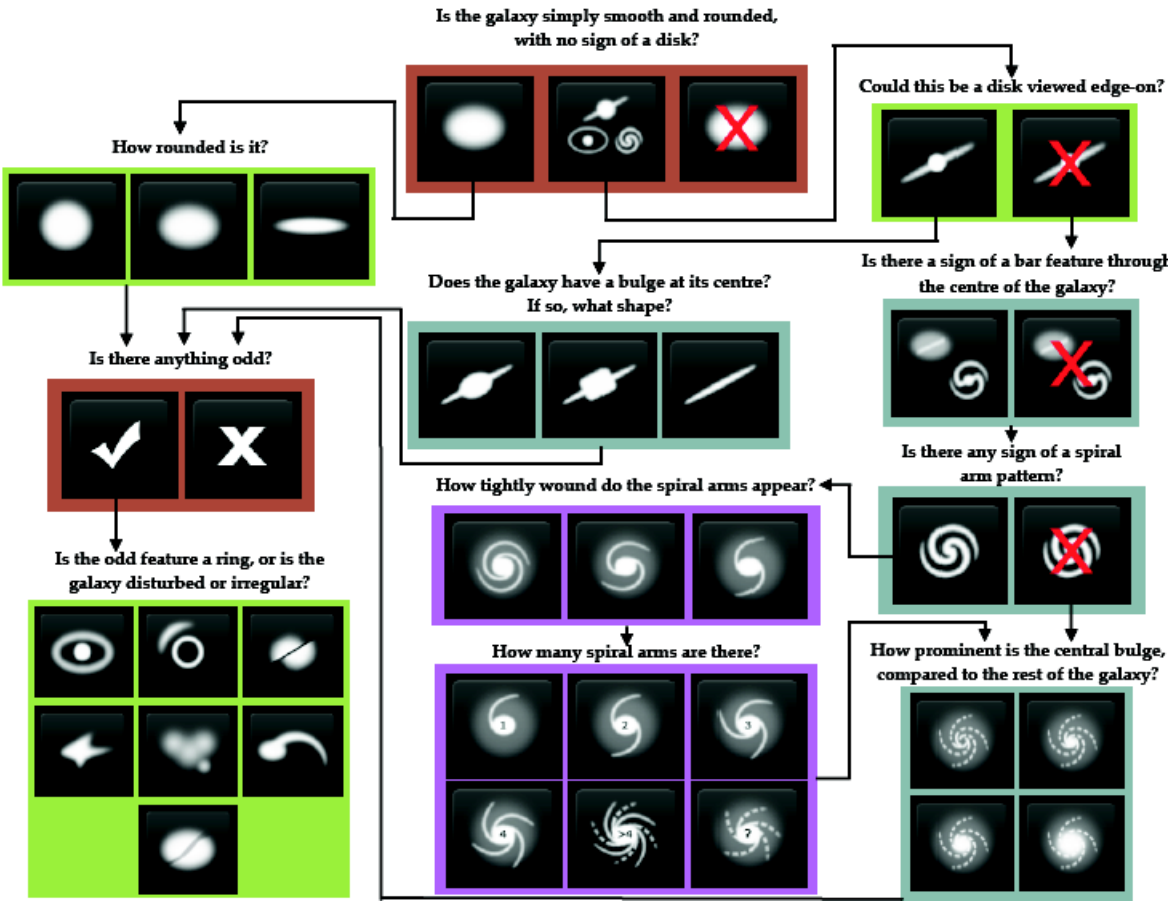


Figure 1: Galaxy Zoo Decision Tree [1]

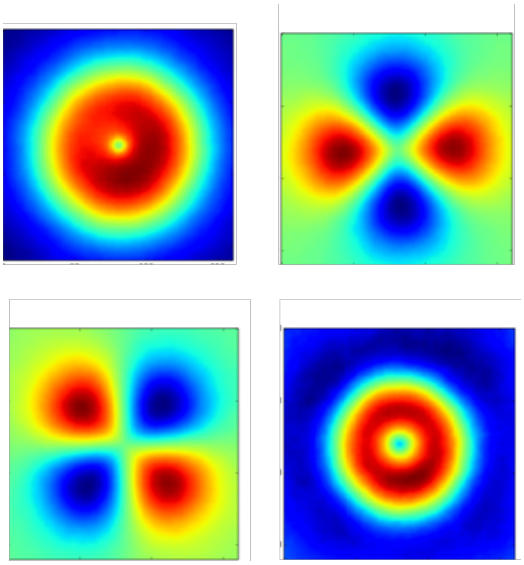


Figure 2: Top components for PCA on galaxy images.

3.3.2 PCA

Principle Component Analysis (PCA) is an useful technique in reducing the dimension of the raw feature vector. In terms of computer vision, PCA is used to extract the most important image components in face recognition, know as the Eigenface method [8]. Similarly here we applied PCA to the galaxy images to extract the top eigen vectors representing the galaxy, reducing dimension of the original image ($424 \times 424 = 179776$ per image) to a 1D vector of 300 components (Fig. 2). These 1D vectors are used as the feature for the classification task.

3.3.3 SIFT

SIFT is a popular computer vision feature detector for local features in the image, widely used in object recognition [5]. In the standard implementation used by OpenCV, the SIFT detector returns interesting points in the image represented by 128-component vectors (Fig. 3B). We applied k-means clustering to reduce the feature codebook to the size of 1000, then used bag-of-word model to build a histogram feature for each image.

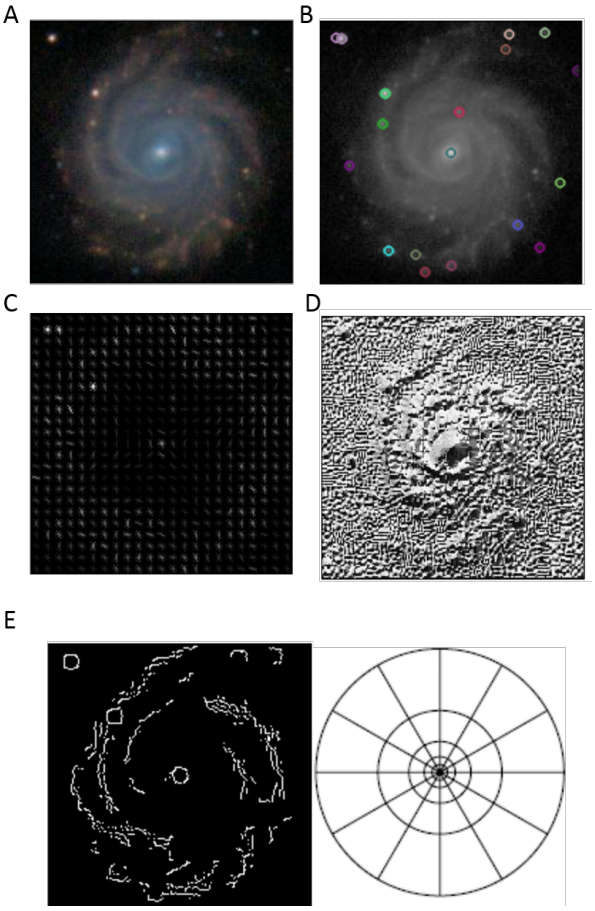


Figure 3: Illustration of different feature extraction scheme. A. Original image; B. SIFT; C. HOG; D. Local binary pattern; E. Polar edge histogram.

3.3.4 HOG

HOG is another commonly used computer vision feature detector [3]. The main idea behind the HOG detector is to use the gradient histogram of small patches of the original image to describe the shape of the object (Fig. 3C). For the HOG feature, we applied PCA to reduce the original histogram feature to 300 components, then used it as the feature vector.

3.3.5 Local Binary Patterns

Local binary pattern provides an easy way to describe the local texture of an image [6]. This feature looks at the nearest-neighbor pixels of each pixel and check if the neighbor pixel is larger than the current one (binary). These binary values for all neighbors are concatenated to describe the texture. A bag of words model built from the resulted texture maps can be used to generate the feature vector. (Fig. 3D).

3.3.6 Fourier Transform

Because the galaxy images are centered and take mostly round shape, it is interesting to test if top components from a Fourier transform may lead to useful descriptions of the images. Indeed by observing the transformed images, only the first few (top corner) Fourier components give significant signals. Therefore here we use X components as the feature vector.

3.3.7 Image Moments

Image moments are weighted averages of the pixel intensities, and they described the global shape of an image, such as the area, centroid and orientation [4]. The image moment is most useful for a binary image after segmentation. Here we simply segmented the images using their original pixel value, by setting all pixels higher then the threshold to 1 and others to 0. Image moments are the computed on the binarized images as the feature vector.

3.3.8 Polar Edge Histogram

Base on similar idea as the shape context feature [2], here I created a polar edge feature to better describe the shape of the galaxy images. The images are first processed with canny edge detector to find the edge pixels. After that, a polar histogram, centered at the image center (which is the galaxy center) is created to count the number of edge pixels in each bin (Fig. 3E). This polar histogram is then used as the feature vector. The main distinction to the shape context is that here we only build one histogram per image, instead building histograms for all edge pixels.

3.3.9 Pretrained Neural Network (Overfeat)

Recently deep convolutional neural network has proven to be successful in image classification contests. Since it is time and resource consuming to train a sufficiently deep neural network, here we just used a deep network trained on the ImageNet dataset. This implementation, called Overfeat [7], has been demonstrated to perform well in previous ImageNet challenge. The 4096 features generated by the second-to-last layer of Overfeat network are used for our Galaxy Zoo challenge.

3.4 Machine Learning Regression

To predict the probability of each response, which is a real number ranging from 0 to 1, we applied standard regression techniques in machine learning. The simplest algorithm we used is the least-square linear regression. In addition, we also tested the Ridge regression method, which is a linear regression algorithm with regularization of the L2 norm of the parameters to prevent overfitting, and the Random Forest, a popular ensemble machine learning technique for classification and regression. For linear regression and Ridge regression we also tested if the introducing nonlinearity by expanding the feature vector with quadratic feature improves the performance.

3.5 Normalization of Prediction

Because we applied normal regression algorithm to our problem, there is no guarantee on whether the

Used Feature(s)	Cross-validated RMSE
Center Color	0.160
PCA	0.152
SIFT	0.163
HOG	0.141
Local Binary Pattern	0.155
Fourier Transform	0.150
Image Moments	0.140
Polar Histogram	0.147
Overfeat	0.126
All	0.122
All + Quadratic	0.119
All + Quadratic (Ridge)	0.113
All (Random Forest)	0.122

Table 1: The RMSEs of different features and feature combinations. Unless specified, the ML algorithm is linear regression. "Quadratic" stands for including quadratic features in the feature vectors.

predicted values are between 0 and 1, or whether the predicted probability will sum to 1. To ensure the predictions satisfy the constraints to be proper probabilities, we first assign all negative predictions to be 0 and all predictions larger than 1 to be 1. Then we performed probability normalization to ensure the probabilities for each question sums to 1.

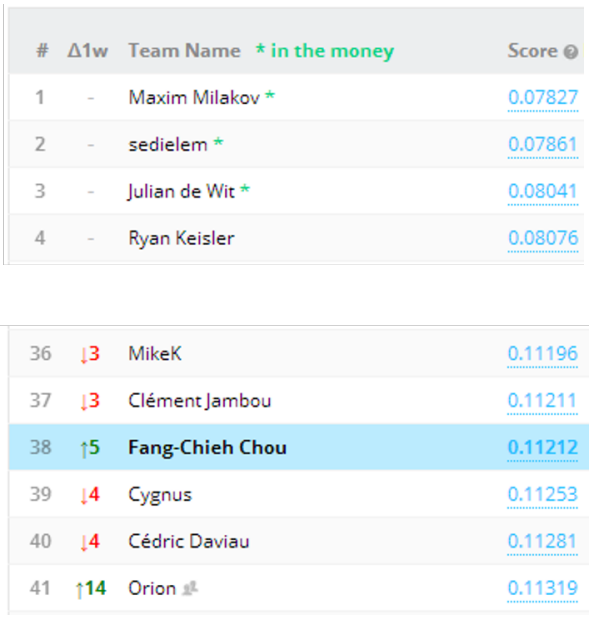
3.6 Cross Validation

To test the performance of our algorithm before submitting to kaggle, we randomly split our training data into the actual training set and the cross validation set. Here 30% of the training data are set aside as the cross validation set. The RMSE computed using the cross validation data is used to evaluate the algorithm performance. Once the cross validation gives reasonable results, we perform prediction on the test set provided by kaggle and submit to their website to get the true test RMSE of our pipeline.

4 Experimental Results

The RMSEs obtained by different methods computed via cross validation are summarized in Table 1. For single features (except the Overfeat) with linear regression, the most powerful ones are the HOG feature and the image moment feature, which gives information on the shape of the galaxy. SIFT, a good detector for interesting points, turned out to provide least discrimination power. Indeed the interesting points detected by SIFT in the galaxy image (Fig. 3B) are not particularly special when examined by visual inspection. The Overfeat feature, although trained on a different dataset focused on natural scenes and objects, gives a low RMSE by itself, demonstrating the power of the deep learning technique. By concatenating all the features, we can further improve the performance of the algorithm. The RMSE can get even lower by including quadratic features into our pipeline. However, expanding the feature set can lead to overfitting, which can be resolved using Ridge, a regularized regression. This combination of all features, quadratic feature expansion and Ridge regression gave the best result, an RMSE of 0.113. Alternative ML techniques such as Random Forest does not lead to better results.

After the cross validation performed on the training set, we applied the same pipeline to generate predictions for the test set provided by kaggle. The result of our best algorithm (all + quadratic + ridge) is shown in Fig 4. The kaggle evaluated RMSE score is 0.112, similar to the 0.113 RMSE obtained by our cross validation scheme. The rank of the algorithm is 38 out of 228 teams. Although our rank is reasonable, we noticed the gap between our result and the top player is quite large. From our exploration on the pretrained neural network, it is likely that the top players have applied deep learning technique specifically trained on the target dataset for this challenge. Due to the time limitation of the class I am unable to implement and test deep learning technique by the time the report is due; however since the challenge is ended at 4/2, I will continue to work on deep learning to see if that will give a better result.



#	Δ1w	Team Name * in the money	Score
1	-	Maxim Milakov *	0.07827
2	-	sedielem *	0.07861
3	-	Julian de Wit *	0.08041
4	-	Ryan Keisler	0.08076
36	↓3	MikeK	0.11196
37	↓3	Clément Jambou	0.11211
38	↑5	**Fang-Chieh Chou**	0.11212
39	↓4	Cygnus	0.11253
40	↓4	Cédric Daviau	0.11281
41	↑14	Orion	0.11319

Figure 4: The kaggle results.

5 Conclusion

In this project, we combined multiple popular computer vision techniques for image classification and interesting point detection to a kaggle challenge on understanding the morphology of galaxies based on their images. We created a pipeline of computer-vision-based feature extraction and machine learning regression to address the challenge, and obtained a good rank in the current leader board. Our results suggest that features that describe the global shape of the galaxies tend to lead to better discrimination then local interesting point detection algorithms. In addition, our test on pretrained neural network reveals the deep learning techniqe might be the ultimate solution to this challenge within current computational power. We will continue to test if a properly built deep network trained on the target dataset will lead to even better results than our current algorithm.

6 References

References

[1] Banerji, M., Lahav, O., Lintott, C. J., Abdalla, F. B., Schawinski, K., Bamford, S. P., Andreescu, D., Murray, P., Raddick, M. J., Slosar, A., Szalay, A., Thomas, D., and Vandenberg, J. Galaxy zoo: reproducing galaxy morphologies via machine learning. *Monthly Notices of the Royal Astronomical Society 406*, 1 (July 2010), 342–353.

[2] Belongie, S., Malik, J., and Puzicha, J. Shape matching and object recognition using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 24*, 4 (Apr 2002), 509–522.

[3] Dalal, N., and Triggs, B. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (2005), vol. 1, IEEE, p. 886–893.

[4] Hu, M.-K. Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions on 8*, 2 (February 1962), 179–187.

[5] Lowe, D. G. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on* (1999), vol. 2, Ieee, p. 1150–1157.

[6] Ojala, T., Pietikäinen, M., and Harwood, D. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition 29*, 1 (1996), 51 – 59.

[7] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR abs/1312.6229* (2013).

[8] Turk, M., and Pentland, A. Eigenfaces for recognition. *Journal of cognitive neuroscience 3*, 1 (1991), 71–86.

[9] Willett, K. W., Lintott, C. J., Bamford, S. P., Masters, K. L., Simmons, B. D., Casteels, K. R. V., Edmondson, E. M., Fortson, L. F., Kaviraj, S., Keel, W. C., Melvin, T., Nichol, R. C., Raddick, M. J., Schawinski, K., Simpson, R. J., Skibba, R. A., Smith, A. M., and Thomas, D. Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society 435*, 4 (Nov. 2013), 2835–2860.