Classification of Morphologies of Galaxy

Alvin Dey
MTech (CSE)
IIIT Delhi
MT18066
alvin18066@iiitd.ac.in

Pranay Raj Anand
MTech (CSE)
IIIT Delhi
MT18079
pranay18079@iiitd.ac.in

Swagatam Chakraborti
MTech (CSE)
IIIT Delhi
MT18146
swagatam18146@iiitd.ac.in

I. Problem Statement

Understanding how and why we are here is one of the radical mysteries for the human race. Fragment of the answer to this problem lies in the origins of galaxies, such as our own Milky Way. As per reports from The University of Chicago, the Sloan Digital survey conducted by Apache Point Observatory will produce more than 50 million images of galaxies in the near future. Stratification of these images is usually done by visual examination of photographic plates. Interpreting the distribution, location and types of galaxies as a function of shape, size, and colour are critical pieces for evaluating our place in the universe. We plan to build an algorithm to classify properties of galaxies with a degree of probability close to the annotation done by volunteers as part of a classification project.

II. Literature Review

Image classification has been a widely studied area in the field of machine learning. However, most of the image datasets and machine learning algorithms are targeted at images of natural scenes and objects, which is different from galaxy images we are focussing here. In [1] two computations schemes have been proposed for Hubble galaxy classification. The first scheme uses geometric shape as the feature for classification, and the other schemes use the image pixel values as features and artificial neural network for classification.

[2] used a neural network and locally weighted regression method as the weak learners and implemented bagging as the ensemble method for classification. In [3], three different techniques have been used as classification which are Naive Bayes, rule-induction algorithm C4.5, and Random Forest. Among the three implementation schemes, Random Forest produced the best result.

III. Database

The Galaxy Zoo Dataset used is taken from the Kaggle challenge. The dataset consists of 61,578 RGB galaxy images of 424 x 424 pixels. Few sample images from the dataset are shown in Fig. 1. Each of these images has been annotated by human volunteers and accordingly probabilities of corresponding 37 categories are provided.

These 37 categories correspond to 11 morphological questions such as "How many spiral arms does the galaxy have?". Each answer will have a probability associated with it calculated with the help of human annotation. Furthermore, the dataset will be split into 70% for training and 30% for testing.

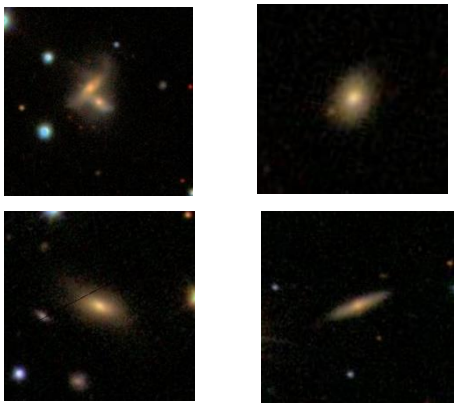

Fig. 1. Sample Images of Galaxies from Galaxy Zoo Dataset

IV. Pre-Processing & Feature Extraction

A. *Pre-Processing:*

All the images are converted to grayscale. As the target galaxy is at the centre of the images so the images are cropped to 207*207 to neglect the void space around the galaxy

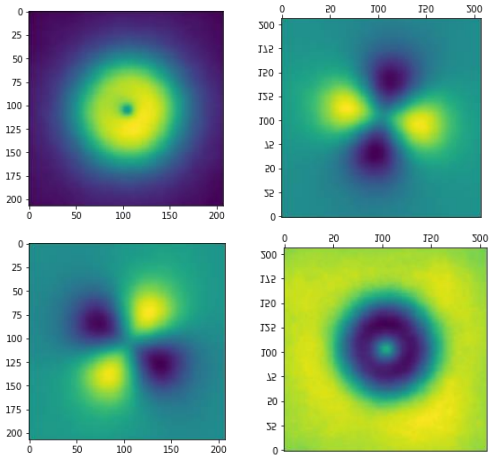

Fig. 2. Top-4 Eigenfaces of the dataset

B. *Feature Extraction:*

We have used a combination of properties of image to define the feature space of an image. It includes Principal Component Analysis (PCA), Histogram of Oriented Gradients (HOG), image moments and Local Binary Pattern (LBP). All these features were concatenated to yield the feature vector of a single image.

i. PCA: Principal Component Analysis or PCA is a useful method for dimensionality reduction which selects features that represent a data point sufficiently. We used the original 424 x 424-pixel images to reduce the dimensionality of each image with each pixel value as the feature. The resulting

features are concatenated with other features for each image.

ii. Image Moments: An image moment is a [4]certain particular weighted average (moment) of the image pixels' intensities, or a function of such moments, usually chosen to have some attractive property or interpretation. Here we took the pre-processed 207 x 207 image and applied image segmentation by thresholding to get a binary image. Then the image moment of that binary image is calculated to get 24 moments which are used as features.

iii. HOG: The histogram of oriented gradients (HOG) is a feature descriptor technique which counts occurrences of gradient orientation in localized portions of an image. Each of the 207 x 207 cropped image was resized to 64 x 128 and divided into 8 x 8 cells across. For all the 8 x 8 cells gradient and magnitude was calculated and magnitude was divided in a weighted average fashion across 9 bins. The 9 bin feature vector was normalized across 16 x 16 blocks in the image.

iv. LBP: It is a type of visual descriptor which creates a feature vector based on localized comparative pixel intensity. Each of the 207 x 207 image was scanned with a 3 x 3 window where each centre pixel was compared with its 8 neighbours and a 8-bit value was assigned if it's neighbour is >= its value or < its value. The 8-bit value is converted to a decimal value for each pixel.

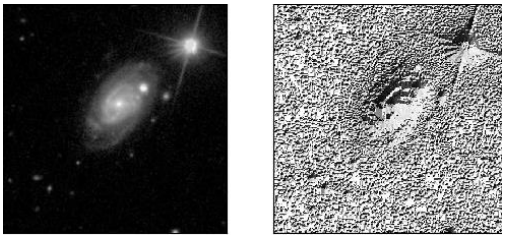

Fig. 3. Grayscale image (left) and LBP image (Right)

V. Experimental Results

We divided the training set into 70% and 30% for training and testing respectively. We applied the Least Mean Square Error (LMSE) regression to fit our model over the training set. In each case different combination of features were used and the Root Mean Square Error (RMSE) was calculated.

From Table 1. it can be seen that image moment performed well in comparison to other features when used individually. The least RMSE was observed when all the features were concatenated together to get the feature vector.

Feature used	***RMSE***
Image Moment	0.159
HOG	0.236
PCA	0.227
LBP	0.163
All	0.157

Table 1. RMSEs corresponding to different feature combination

VI. Future Work

We will try to refine our feature set more to get better results. Other proposed methods i.e, Ridge regression, Lasso regression and Random forest will be implemented and comparative analysis between all methods will be shown.

References

[1] Goderya, S.N., Lolling, S.M., "Morphological Classification of Galaxies Using Computer Vision and Artificial Neural Networks: A Computational Scheme." Astrophysics and Space Science. Vol. 279, no. 4, pp. 377 –387.

[2] Calleja, J., Fuentes, O., "Machine Learning and Image Analysis for Morphological classification of galaxies." Monthly Notices of the Royal Astronomical Society,Vol. 24, 2004, pp. 87 ‐ 93.

[3] Calleja, J., Fuentes, O., "Automated Classification of Galaxy Images", Lecture Notes in Computer Science, Vol. 3215, 2004, pp. 411 ‐ 418.