# Galaxy image classification

Joseph H. Murrugarra LL. and Nina S. T. Hirata
Department of Computer Science
Institute of Mathematics and Statistics
University of São Paulo, São Paulo, Brazil
Email: perhark@ime.usp.br, nina@ime.usp.br

***Abstract*—Over the years, different methods based either on morphological features or on expert knowledge have been proposed to classify galaxies. The amount of data to be processed in large scale surveys poses a new challenge for the classification. In this preliminary study, we investigate machine learning methods for galaxy image classification. Specifically, we evaluate convolutional neural networks as tools to be used in the classification process. Different ways of using convolutional neural networks has been experimented to classify galaxies as elliptical or spiral. Classification accuracy around 90-91% for the Sloan Digital Sky Survey (SDSS) galaxy images has been achieved.**

## I. Introduction

Astronomical research, for a long time, consisted on observational studies around some specific objects in the space. In the last decades, propelled by technological advances in many fields (e.g, digital imaging), observational approaches have rapidly shifted to large scale surveys. These surveys typically obtain multiband images of a large area of the sky, containing millions of objects. As a consequence, astronomy is nowadays one of the main research fields in the big data context [1]. There are challenges related not only with transmission and storage of these data, but also with the processing and the analysis of the images. The classification of objects is important to produce large catalogues and to discover underlying physics [2].

Galaxies are among the astronomical objects observed in these images and there are different ways of classifying them. The most commonly used classification scheme is based on morphological features and classifies galaxies as *spirals*, *ellipticals*, *lenticulars* or *irregulars* [3], [4]. Galaxy morphology provides information about the process of its formation and its interactions with the environment, being therefore of interest for the understanding of the expansion of the universe.

In this work we study the galaxy image classification problem. There are many works in the related literature concerned with galaxy classification. Some are based in machine learning techniques [5], [6] and others in heuristics [7]. Among them, many works [7]–[9] use data from the Galaxy Zoo project [10].

The goal of our study is to evaluate convolutional neural networks (CNN) as tools for the classification of galaxies. We propose different ways to employ CNN in the classification process. In order to evaluate the methods, we have created a galaxy image dataset from the Galaxy Zoo and the SDSS data. This is an ongoing work in collaboration with the Institute of Astronomy, Geophysics and Atmospheric Sciences of University of São Paulo. In this paper we present some preliminary results.

The paper is organized as follows. In Section II, we present some astronomy related concepts as well as galaxy classification methods reported in literature that will help to place the reader in context. In Section III we detail the proposed classification methods. Finally, preliminary results and conclusions are presented in Sections IV and V, respectively.

## II. Background

In this section we present some basic concepts, terminologies and definitions related to the subject of study in this paper. We also briefly review some methods on galaxy classification reported in literature.

### *A. Astronomical images*

The energy radiated by astronomical objects carries information on temperature, density, composition, and important physical processes. To measure the spectrum of electromagnetic radiation, scpectroscopy is the traditional tool used in Astronomy. The resulting spectrum is a signal that represents radiation intensity as a function of wavelength. While scpectroscopy is mainly used with focus on a single object, large scale surveys collect information from a large area of the sky, creating a general map or image of the region. Telescopes used in recent large surveys use CCD technology, with sensors for different bandwidths of the electromagnetic spectrum. The number of sensors (or filters) used determines the number of bands of the resulting image. In a sense, we may say that the spectrum is a continuous signal in the wavelength domain, while the resulting data of these surveys consists of sampled representations of the signals.

The *Sloan Digital Sky Survey* (SDSS) project – http://www.sdss.org – is one of the most successful astronomical surveys [11]. It has been observing the universe for over 15 years, starting in 2000. It is divided in phases and currently it is in phase IV. Imaging of one million galaxies and 100,000 quasars were accomplished in phase II (2005-2008) from half the northern sky.

SDSS images consist of five bands, namely U, G, R, I and Z [12], as detailed in Table I[1]. The raw data is stored in Flexible Image Transport System (FITS) format [13], which is the standard format used for astronomical images. SDSS also

[1] http://skyserver.sdss.org/dr1/en/proj/advanced/color/sdssfilters.asp

provides JPG images created from the G, R, I bands of FITS images [14].

TABLE I
THE 5 SDSS FILTERS AND THE CORRESPONDING WAVELENGTHS

| Filter | Wavelength (Angstroms) |
|---|---|
| Ultraviolet (u) | 3543 |
| Green (g) | 4770 |
| Red (r) | 6231 |
| Near Infrared (i) | 7625 |
| Infrared (z) | 9134 |

### B. *Classification of galaxies*

There are some systems for galaxy classification but the most commonly used is the one proposed by Edwin Hubble in 1926, and later expanded by others. Based on morphological features, galaxies are classified as spirals, ellipticals, lenticulars or irregulars [3], [4]. Figure 1 shows the classification scheme (adapted from https://www.spacetelescope.org/images/heic9902o/).

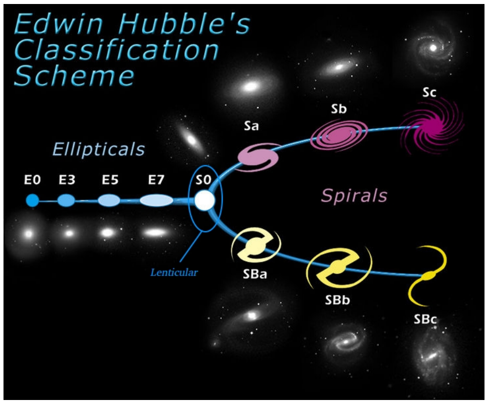

Fig. 1. Hubble classification scheme.

Elliptical galaxies have smooth light distributions and appear as ellipses. They are denoted by the letter E, followed by an integer $n$ representing their degree of ellipticity on the sky. Spiral galaxies look like a flattened disk, with stars forming spiral arms winding towards a central concentration of stars known as the bulge. Spirals in Figure 1 are those denoted as Sa, SBa, Sb, SBb, SBc and Sc. Lenticular galaxies are the intermediate representation between spiral and elliptical galaxies, denoted as S0. Finally, irregular galaxies are those that do not fit into the Hubble sequence, because they have no regular structure (either disk-like or ellipsoidal). Figure 2 shows samples of elliptical and spiral galaxies from SDSS.

Since the surveys collect images of millions of objects, classifying the objects is a huge challenge. Motivated by such challenge, the Galaxy Zoo Project was created as an online citizen science project. A web-based interface is used to collect morphological data of galaxies. The classification begins with the user being shown an image of a galaxy alongside a question and a set of possible responses. In Galaxy Zoo 2, the data has been collected via a multi-step decision tree and more than 16 million morphological classifications of 304,122 galaxies drawn from the Sloan Digital Sky Survey were obtained [10]. In summary, this project provides different information of millions of galaxies including their spatial coordinates $(ra, dec)$, corresponding to the right ascension and declination, and their respective classification.

Fig. 2. Example of an elliptical (left) and a spiral (right) galaxies from the SDSS.

### C. *Related works*

There have been several attempts to classify galaxies automatically. Here we briefly mention recent works based on machine learning and morphometric features.

Convolutional neural network (CNN) is used in [9] to predict the 37 votes of the Galaxy Zoo 2 release. They work directly in pixel space, using a rotation invariant convolution that minimizes sensitiveness to changes in scale, rotation, translation and sampling of the image. An accuracy of 99% relative to the Galaxy Zoo human classification is reported.

In [5], a morphology catalog of the SDSS galaxies is generated using the Wndchrm image analysis utility [6]. It computes 2,885 features related to textures, edge, shapes, statistical distributions, fractal features, among others. For each of the features a Fisher discriminant score [15] is assigned, and 85% of the features with the lowest Fisher scores are rejected in order to filter outliers. A weighted nearest neighbor classifier is used to classify the galaxies. Classification accuracy was computed based on the manually annotated Galaxy Zoo catalog. They point out that about 900,000 of the instances classified as spirals and about 600,000 of those classified as elliptical have a statistical agreement rate of about 98% with the Galaxy Zoo classification.

A morphological classification is also presented in [7]. Several morphological features, such as concentration, asymmetry, smoothness, entropy, spirality, among others are computed from the image to classify galaxies as elliptical or spiral. Sample images for training are taken from the first Galaxy Zoo release (GZ1). They employed the Linear Discriminant Analysis (LDA) technique to classify 4,478 galaxies from [16], 14,123 galaxies from [17] and 779,235 galaxies from SDSS, and report a 10-fold cross validation accuracy superior to 90%.

## III. Proposed Method

In order to evaluate the potential of CNN to classify galaxies, we propose a set of methods for the dataset preparation, classifier training and classifier evaluation. In this work we consider only the spiral and elliptical type galaxies.

### *A. Data Preparation*

Although astronomical data is publicly available in several forms, they are not presented in the training/test form which is usual in the machine learning field. Therefore, a method to prepare labeled galaxy images was defined. It consists of the following steps:

**Querying step:** The Galaxy Zoo database [10] is used to find the spatial coordinates of galaxies of interest (spirals and ellipticals). Then, with these coordinates a query is performed in the SDSS database to retrieve the image of the galaxy. However, the returned images correspond to a field of the sky, containing not only the queried object but many others. We selected the first entries in the Galaxy Zoo list, starting from the first up to the one that was sufficient to obtain 2,000 samples of each type.

**Image cropping:** The image of the galaxy of interest is cropped from the large image using its known coordinate. In order to crop the galaxy image, we first make a squared crop of $200 \times 200$ pixels, centered on the pixel corresponding to the galaxy coordinate, in the R-band of SDSS images. This step aims to avoid processing the whole image which is of size larger than $2000 \times 2000$ pixels. Then, on the cropped patch we apply a normalization based in the Astropy normalization method[2] to obtain an image in the range [0-1]. After this process, we apply Otsu's binarization [18] and determine the bounding box of the connected component that contains the coordinate of the galaxy. The bounding box defines the size of the final crop. An alternative measure of the galaxy size is the Petrosian radius [19], planned for use in future work.

**Pre-processing:** All pixel values in each band of FITS images are small decimal values and they differ in range from one band to another. Thus, we also normalize the cropped images applying the Astropy normalization method[2] on each band individually. In Figure 3 we show an example of the pre-processing step. It is notable how the spiral arms are more visible after this process.

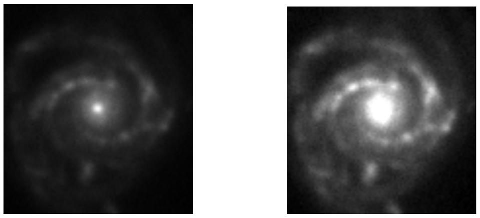

Fig. 3. Example of a galaxy image G-band (left) from SDSS and the result of applying the pre-processing step (right).

### *B. Classifier training*

A common pipeline for classification considers a feature extraction step, followed by an optional feature selection step, and then the classifier training step. Convolutional neural networks (CNN) [20] are extensions of the well known family of neural networks, called feed-forward multilayer perceptrons. One important characteristic of CNNs is the fact that they do not require feature extraction. For image classification, training data may consist of pairs of a raw image and its corresponding class label. There is, arguably, an understanding that CNNs are able to encode useful features in its convolutional layers. Based in these features, we propose two ways to train the classifiers.

**Classifier training using features extracted with a pre-trained CNN:** We propose to use the output of the last convolutional layer of a pre-trained CNN plus a classifier. One of the classifier models considered is Support Vector Machine (SVM), with Gaussian kernel. To choose a good SVM configuration, a grid search on the space of parameters (C,$\lambda$) using cross-validation is performed. Various pairs (C,$\lambda$) are tested and the configuration with the best validation accuracy is selected. A second model considered in this work is a fully connected layer on the top of the pre-trained CNN convolutional layers. An example is shown in Figure 4. This model is trained using a training and a validation data.

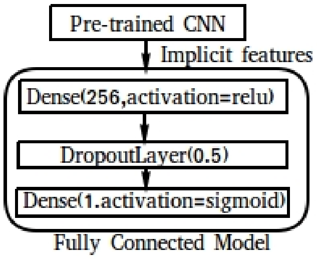

Fig. 4. Adding a fully connected model on the top of a pre-trained CNN.

**Fine tuning of a pre-trained CNN:** By 'fine tuning' we mean training the last convolutional blocks of a pre-trained model alongside the top fully connected model classifier. This differs with respect to the previous approach, where convolutional layers of the CNN are used as features extractors and are not updated during training. In our case, fine tuning can be done in 4 steps:

- Instantiate the pre-trained CNN and load its weights.
- Add our previously defined fully-connected model on top, and load its weights
- Freeze the initial layers of the pre-trained CNN model to only perform fine tuning in the last convolutional blocks.
- Retrain using a small learning rate

## IV. Preliminary Results

In this section we describe some preliminary experiments and results on galaxy classification, using CNNs as described in the previous section.

### *A. Experimental Setup*

To create our dataset we applied all the steps in Section III-A. For each type of galaxy (spiral and elliptical), we selected 2,000 images with dimension larger than $20 \times 20$ pixels. Our dataset consists of galaxy image crops in FITS format with the 5 bands normalized to the range [0-1]. To

[2] http://docs.astropy.org/en/stable/visualization/normalization.html

feed the CNN, all images were resized to 150 by 150 pixels, filling the borders with 0. This dimension was chosen after verifying that the largest width or height among all crops was not superior to 150. The dataset has been split in training (667 images), validation (667 images) and test sets (666).

In the first training approach, to extract the features with a CNN we have used the VGG16 [20] convolutional neural network. We used the VGG-net pre-trained on the ImageNet dataset [21]. ImageNet is an ongoing project with millions of image and more than 1,000 common object classes. Since ImageNet consists of JPG images (with 3 bands) and all pixel values are in the range of [0-255], we selected 3 of the 5 bands from our galaxy images and normalized the pixels values to be in the range [0-255]. We have tested the combination of UGR-bands and IGR-bands. For classification, we tested with the 3 approaches in section III-B. In the case of fine tuning, we freeze the first 10 layers of VGG16 and retrain the rest of the layers, including the fully connected layer, using a small learning rate.

### *B. Results*

Classification performance was measured in terms of accuracy on the test set. The obtained results are shown in Table II. The best accuracy was obtained when using I, G, R bands together with the fine tuning CNN approach.

TABLE II
CLASSIFICATION ACCURACY ON TEST IMAGES

| Channels | CNN+SVM | CNN+ FC layer | Fine tuning CNN |
|---|---|---|---|
| UGR | 88,9% | 88% | 90,04% |
| IGR | 90% | 88,3% | 91,1% |

The results with respect to the two combinations of three bands, UGR and IGR, are very similar. In general, the obtained accuracies indicate that CNNs trained on a very distinct image domain (ImageNet) carry features that work well on galaxy images. In both cases the best accuracy was obtained with fine tuning. This may be explained by the fact that adjustment of weights in the convolutional layers make the features encoded by CNN more suited for the galaxy image classification.

Although the results are encouraging, some previously reported results indicate that better accuracies should be obtained. A possible explanation for misclassification is the small number of training images.

## V. CONCLUSION

We have presented some preliminary results on galaxy image classification using convolutional neural networks. A classification accuracy slightly above 90% was achieved. To continue this study and improve the results, we plan to exploit several issues such as using the 5-bands, using some common approaches for data augmentation (rotations, small translations, and reflections of the images), using expert designed features in conjunction with CNN features, replacing the connected component extraction method with the one based on the Petrosian radius, which is already available in the catalogs, and also using a larger training set. Our ultimate goal is to integrate the classification method in the processing pipeline of the S-PLUS project – http://www.iag.usp.br/labcosmos/en/s-plus/.

## ACKNOWLEDGMENT

The authors would like to thank CNPq (scholarship for J.H.M.L and research fellowship for N.S.T.H).

## REFERENCES

[1] T. Hey, S. Tansley, and K. Tolle, Eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, Washington: Microsoft Research, 2009.

[2] O. Lahav, "Artificial neural networks as a tool for galaxy classification." in *Data Analysis in Astronomy*, V. Di Gesu, M. J. B. Duff, A. Heck, M. C. Maccarone, L. Scarsi, and H. U. Zimmerman, Eds., 1997, pp. 43–51.

[3] E. Hubble, "No. 324. Extra-galactic nebulae." *Contributions from the Mount Wilson Observatory / Carnegie Institution of Washington*, vol. 324, pp. 1–49, 1926.

[4] E. P. Hubble, "The classification of spiral nebulae," *The Observatory*, vol. 50, pp. 276–281, Sep. 1927.

[5] E. Kuminski and L. Shamir, "A computer-generated visual morphology catalog of 3,000,000 SDSS galaxies," *The Astrophysical Journal Supplement Series*, vol. 223, no. 2, p. 20, 2016.

[6] L. Shamir, N. Orlov, D. M. Eckley, T. Macura, J. Johnston, and I. Goldberg, "WND-CHARM: Multi-purpose image classifier," Astrophysics Source Code Library, Dec. 2013.

[7] F. Ferrari, R.-R. de Carvalho, and M. Trevisan, "Morfometryka – A New Way of Establishing Morphological Classification of Galaxies," *The Astrophysical Journal*, vol. 814, p. 55, Nov. 2015.

[8] L. Shamir, "Automatic morphological classification of galaxy images," *MNRAS*, vol. 399, pp. 1367–1372, Nov. 2009.

[9] S. Dieleman, K. W. Willett, and J. Dambre, "Rotation-invariant convolutional neural networks for galaxy morphology prediction," *MNRAS*, vol. 450, pp. 1441–1459, Jun. 2015.

[10] K. W. e. Willett, "Galaxy Zoo 2: detailed morphological classifications for 304122 galaxies from the Sloan Digital Sky Survey," *MNRAS*, vol. 435, pp. 2835–2860, Nov. 2013.

[11] D. G. Y. *et al.*, "The Sloan Digital Sky Survey: Technical Summary," *The Astronomical Journal*, vol. 120, pp. 1579–1587, Sep. 2000.

[12] M. Fukugita, T. Ichikawa, J. E. Gunn, M. Doi, K. Shimasaku, and D. P. Schneider, "The Sloan Digital Sky Survey Photometric System," *The Astronomical Journal*, vol. 111, p. 1748, Apr. 1996.

[13] D. C. Wells, E. W. Greisen, and R. H. Harten, "FITS - a Flexible Image Transport System," *Astronomy & Astrophysics Supplement Series*, vol. 44, p. 363, Jun. 1981.

[14] R. Lupton, M. R. Blanton, G. Fekete, D. W. Hogg, W. O'Mullane, A. Szalay, and N. Wherry, "Preparing Red-Green-Blue Images from CCD Data," *Publications of the Astronomical Society of the Pacific*, vol. 116, pp. 133–137, Feb. 2004.

[15] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[16] A. Baillard, E. Bertin, V. de Lapparent, P. Fouqué, S. Arnouts, Y. Mellier, R. Pelló, J.-F. Leborgne, P. Prugniel, D. Makarov, L. Makarova, H. J. McCracken, A. Bijaoui, and L. Tasca, "The EFIGI catalogue of 4458 nearby galaxies with detailed morphology," *Astronomy and Astrophysics*, vol. 532, p. A74, Aug. 2011.

[17] P. B. Nair and R. G. Abraham, "VizieR Online Data Catalog: Detailed morphology of SDSS galaxies," *VizieR Online Data Catalog*, vol. 218, Apr. 2010.

[18] N. Otsu, "A Threshold Selection Method from Gray-level Histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[19] V. Petrosian, "Surface brightness and evolution of galaxies," *Astrophysical Journal*, vol. 209, pp. L1–L5, Oct. 1976.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.