PROCEEDINGS OF SCIENCE

Galaxy Morphology classification using CNN

J. A. Vázquez-Mata*
IA-UNAM
E-mail: jvazquez@astro.unam.mx

H. M. Hernández-Toledo
IA-UNAM
E-mail: hector@astro.unam.mx

L. C. Mascherpa
F. Ciencias-UNAM
E-mail: mascherpaluiscarlos0@gmail.com

Galaxy morphology is one of the most important parameters to understand the assembly and evolution of galaxies in the universe. The most used morphological classification nowadays was proposed by Hubble (1926). This classification is based on the presence of disks-, arms-, bulges-like structures, and is carried out mostly by eye. Thanks to the upcoming large telescopes and surveys, the new catalogues will contain millions of galaxies, making it impossible to carry out a visual classification. Then, convolutional neural networks (CNN) start to play an important role to classify galaxies automatically. In this work we summarise the most recent results (including ours) using machine learning techniques to classify galaxies and future projects.

Artificial Intelligence for Science, Industry and Society, AISIS2019
October 21-25, 2019
Universidad Nacional Autónoma de México, Mexico City, México

*Speaker.

1. Introduction

In astronomy, classification of astronomical objects is essential to carry out studies of their physical processes to understand their origin and evolution. These objects can be planets, nebulae, stars, galaxies or cosmic structures. Astronomers base their classifications on images and spectra to extract physical information.

With respect to galaxies, the observational evidence demonstrates the wide range of galaxy morphologies, classified in the so-called "Hubble sequence". This sequence is derived from the present or not of disks, arms and the bulge/disk ratio (Figure 1). This figure shows representative galaxies for each morphology, from the soft, spherical-like galaxies on the left hand-side (denoted as E-S0, or early types) to the disk-like galaxies on the right, from very enrolled to open arms, and sometimes destructed or clumpy arms (Sa-Irr or late types). The Hubble sequence is denoted technically as: E,S0,Sa,Sb,Sc,Sd,Sm, with numerical values: -5,-2,1,3,5,7 respectively (T-Type); intermediate types exist and are the transition between two types. These morphologies also recognise the role of environment in shaping galaxies., then a reliable morphological classification has been proven as a fundamental requisite for the analysis and interpretation of the physical properties of galaxies. A correct estimate of their morphological and structural properties not only is useful to understand their evolution but helps reveal their formation mechanisms providing ways to test galaxy formation and evolution theories.

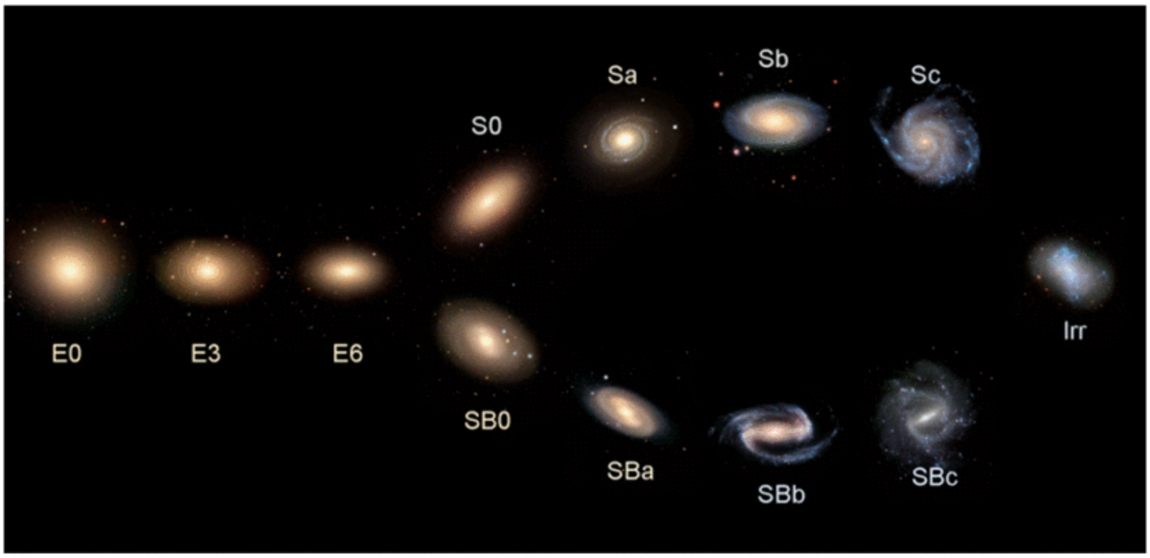

Figure 1: The Hubble sequence [1]

Along the time, the morphological classification has been carried out by astronomers mainly, through a visual inspection of galaxies individually. The largest catalogue, classified following this approach, contains ∼14 000 galaxies (Nair catalogue) [2], using images from the Sloan Digital Sky Survey (SDSS) [3].

The arrival of the new large astronomical surveys, with millions of catalogued galaxies, have triggered strong efforts to improve the automatic morphological classification methods. Different approaches have been proposed, from parametric and non-parametric characterisations up to developing visually-based methods that explode human pattern recognition capabilities. The most recent catalogue is GalaxyZoo (GZ)[4, 5], containing ∼200 000 classified galaxies, where around 100 000 non-professional astronomers put their efforts to classify this sample.

2. Machine Learning and galaxy morphology

Over the last few years, machine and deep learning techniques have become the most successful methods to automatically classify galaxies. Authors like Sreejith et al. [6] have applied different Machine Learning methods to classify galaxies into four categories (inside the Hubble sequence) in the Galaxy and Mass Assembly survey (GAMA) [7, 8], based on the photometric parameters, Sérsic index, ellipticity, etc, with an average accuracy of 75%. Barchi et al. [9] consider the structural CAS, Gini and M20 parameters [10], in addition to the Shannon entropy (H) [11] and Gradient Pattern Analysis (GPA) [12], to separate late (Sa-Sm) from early (E-S0) type classes with an accuracy of 98% that decreases to 65% when attempts are made to distinguish into more discrete T-Type sub-classes as in Nair catalogue [2].

Convolutional Neural Networks (CNN) are considered among the best methods to classify images without the need to extract structural features with high accuracy. Dieleman et al. [13] presented this approach as a result of a Kaggle competition to reproduce the GZ classification for 140 000 galaxies based on SDSS images. They were able to reproduce that classification with an accuracy of ∼99% for most of the GZ questions. Huertas-Company et al. [14] applied the same CNN algorithm to 58 000 galaxies from the near infrared CANDELS survey [15] at high redshifts, identifying among Early, Spiral, Irregulars, compact or unclassified galaxies with low misclassification levels.

More recently, Dominguez-Sanchez et al. [16] applied accordingly a modified version of the Huertas-Company et al. [14] algorithm to a SDSS sample of ∼670 000 galaxies, reporting 98% accuracy to reproduce the GZ classification and obtaining smaller offsets and scatter than previous authors when trying to distinguish among discrete T-Type sub-classes. Up to now, this is the best work trying to classify galaxies into the Hubble sequence; however, more efforts are needed to improve accuracy. For illustration, Figure2 shows the structure reported in that work [16].

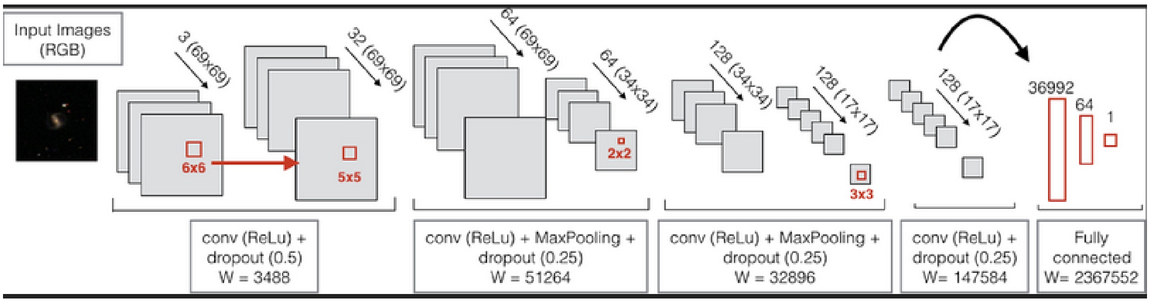

Figure 2: CNN structure reported by Dominguez-Sanchez et al. [16] to classify galaxies.

In the last year, Martin et al. [17] proposed an unsupervised machine learning classification using the Hyper-Suprime-Cam Subaru-Strategic-Program Ultra-Deep survey. They employ an algorithm that performs clustering of graph representations, in order to group image patches, with similar visual structures for galaxies. After applying this algorithm to the whole sample, they reduced the sample to 160 "morphological clusters", enclosing galaxies with similar features and properties. These clusters could be associated with the standard discrete T-Type sub-classes and could be the new way to classify galaxies in the years ahead.

Cheng et al. [18] have tested 7 different Machine Learning methods, including CNN, to determine which one is the best to classify galaxies. At least in a binary classification, CNN is definitely the best method with higher accuracy. This is an important reason to continue exploring this method.

3. Improving morphological classification

Astronomical images are commonly limited by the telescope used, sky conditions, atmosphere and the intrinsic decay in the flux of photons received, associated with the source distance. For nearby galaxies, the quality of images is good enough to classify galaxies into the general morphological families; however, far and faint objects are very difficult to classify with high accuracy. The new astronomical surveys are designed to take images with larger telescopes and/or increase exposure times to detect faint objects, low surface brightness structure and improve image quality.

The Dark Energy Spectroscopic Instrument project (DESI)[19] compiles its targets from three wide-area optical imaging surveys. The DESI quality requirements provide in principle deeper images than the SDSS images, this makes plausible a visual identification of structural features for a more reliable classification. The large improvement in image quality can be observed in Figure 3, for 4 different galaxies, upper panels show the corresponding SDSS image and lower panels the DESI image for the same galaxy. DESI images allow us to identify low surface brightness structures around the galaxies and applying an optimal filter, internal structures are enhanced like bars, rings, internal disks, etc.

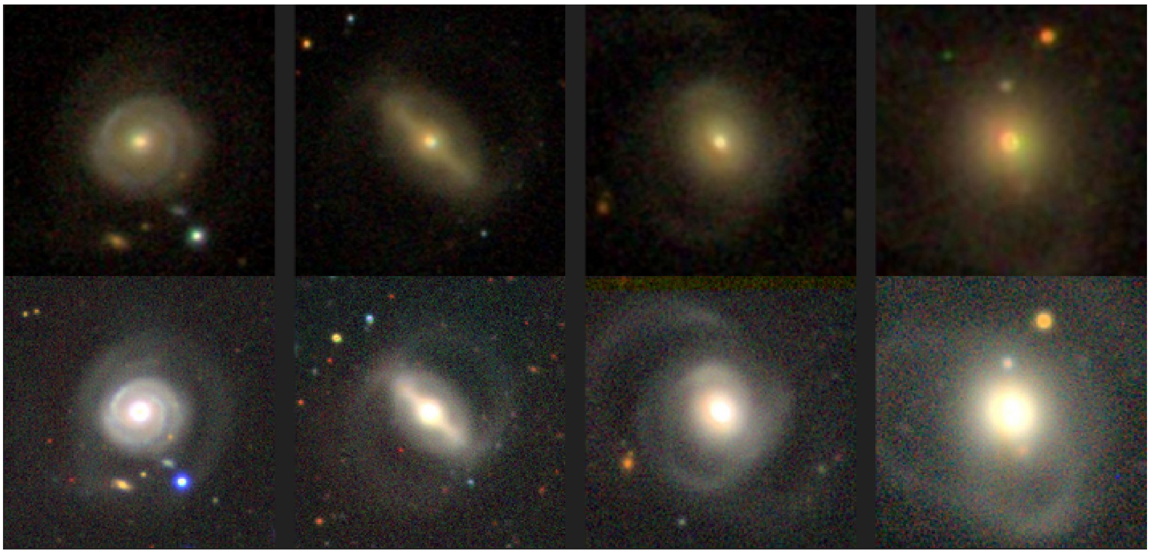

Figure 3: Top panels show SDSS images, and the corresponding DESI image at the bottom for each galaxy.

This type of images allow the classifiers to distinguish between the individual T-Types with better accuracy, improving the training sets for a CNN.

3.1 CNNs to classify galaxies.

As part of this work, we have classified ∼4600 galaxies from the Mapping Nearby Galaxies at Apache Point Observatory (MaNGA) [20, 21] taking the most of the SDSS and DESI images (Vazquez-Mata et al. in prep.) to extract internal and external structures in galaxies, to make a morphological classification more accurate. Figure 4 presents a mosaic with the 5 images we have used to classify each galaxy. The *Originals* are the RGB colour image and the same displayed in logarithmic and grey scales to identify possible structures in the nuclear region. We have applied a Gaussian kernel to the r-band image and displayed it in logarithmic scale to enhance all internal structures, helping us distinguish among early type galaxies (Filtered image).

The DESI collaboration also generated a post-processing catalogue for the Legacy Surveys. This catalogue was built using an approach to estimate source shapes and brightness properties. Each source was then modeled by The Tractor using a small set of parametric light profiles (de Vaucouleurs) (for more details, see [19]). And finally the PSF-convolved residual images allows to detect any feature or structure left after subtracting the model (Residual image).

After inspecting the 5 images all together we are able to assign a T-Type morphology to each galaxy.

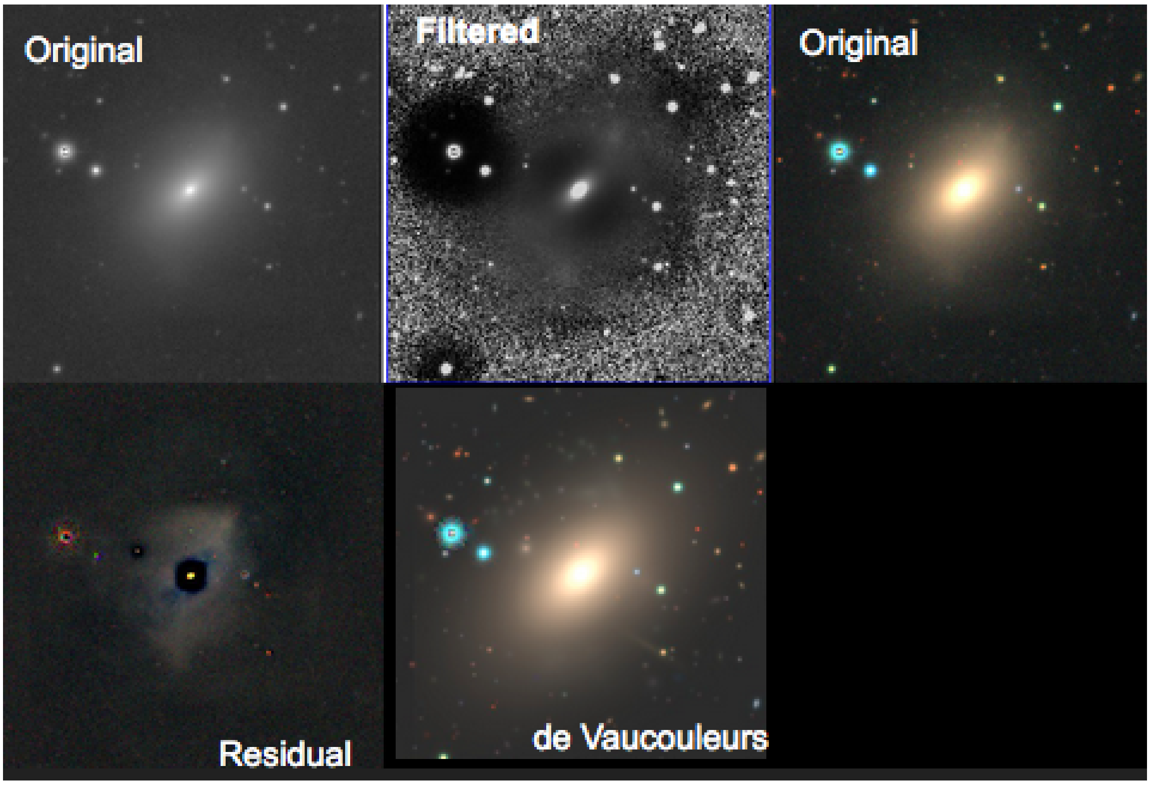

Figure 4: Top panels show SDSS images, and the corresponding DESI image at the bottom for each galaxy.

We have used this new classification to start training a CNN based on previous models (similar the one in DominguezSanchez2018) with accuracy of 96% to separate early from late type galaxies and 60% for the multiclass classification. These results are comparable with previous works;

however, our sample is not large enough and is definitely going to improve after the second half of the MaNGA sample was classified.

4. Conclusions

The new generation of telescopes are producing large surveys with millions of galaxies, covering large volumes in the sky. Telescopes like the Large Synoptic Survey Telescope (LSST) [22] and the James Webb Space Telescope (JWT) [23] will provide millions of data, unable to be processed by applying common techniques. Automatic tools have been developed during the last years to deal with this amount of information. Machine Learning, and in particular CNN will help us classify millions of galaxies in a very short time scale compared to what humans could take. Once the best models are trained, a simple *finetuning* will be required to adapt the new type of images for a reliable classification.

In this talk, we gave a general summary about galaxy morphological classification using Machine Learning methods, and also presented how the authors are trying to improve the accuracy reached by other works. These models will be applied in the new surveys for future studies in astronomy.

References

[1] Y. Cui, Y. Xiang, K. Rong, R. Feris and L. Cao, *A spatial-color layout feature for representing galaxy images*, pp. 213–219, 03, 2014, DOI.

[2] P. B. Nair and R. G. Abraham, *A Catalog of Detailed Visual Morphological Classifications for 14,034 Galaxies in the Sloan Digital Sky Survey*, **186** (2010) 427 [1001.2401].

[3] M. R. Blanton, M. A. Bershady, B. Abolfathi, F. D. Albareti, C. Allende Prieto, A. Almeida et al., *Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe*, **154** (2017) 28 [1703.00052].

[4] C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas et al., *Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey*, **389** (2008) 1179 [0804.4483].

[5] K. W. Willett, C. J. Lintott, S. P. Bamford, K. L. Masters, B. D. Simmons, K. R. V. Casteels et al., *Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey*, **435** (2013) 2835 [1308.3496].

[6] S. Sreejith, J. Pereverzyev, Sergiy, L. S. Kelvin, F. R. Marleau, M. Haltmeier, J. Ebner et al., *Galaxy And Mass Assembly: automatic morphological classification of galaxies using statistical learning*, **474** (2018) 5232 [1711.06125].

[7] S. P. Driver, D. T. Hill, L. S. Kelvin, A. S. G. Robotham, J. Liske, P. Norberg et al., *Galaxy and Mass Assembly (GAMA): survey diagnostics and core data release*, **413** (2011) 971 [1009.0614].

[8] J. Liske, I. K. Baldry, S. P. Driver, R. J. Tuffs, M. Alpaslan, E. Andrae et al., *Galaxy And Mass Assembly (GAMA): end of survey report and data release 2*, **452** (2015) 2087 [1506.08222].

[9] P. H. Barchi, R. R. de Carvalho, R. R. Rosa, R. A. Sautter, M. Soares-Santos, B. A. D. Marques et al., *Machine and Deep Learning applied to galaxy morphology - A comparative study*, *Astronomy and Computing* **30** (2020) 100334 [1901.07047].

[10] C. J. Conselice, *The Relationship between Stellar Light Distributions of Galaxies and Their Formation Histories*, **147** (2003) 1 [astro-ph/0303065].

[11] F. Ferrari, R. R. de Carvalho and M. Trevisan, *Morfometryka—A New Way of Establishing Morphological Classification of Galaxies*, **814** (2015) 55 [1509.05430].

[12] R. R. Rosa, R. R. de Carvalho, R. A. Sautter, P. H. Barchi, D. H. Stalder, T. C. Moura et al., *Gradient pattern analysis applied to galaxy morphology*, **477** (2018) L101 [1803.10853].

[13] S. Dieleman, K. W. Willett and J. Dambre, *Rotation-invariant convolutional neural networks for galaxy morphology prediction*, **450** (2015) 1441 [1503.07077].

[14] M. Huertas-Company, R. Gravet, G. Cabrera-Vives, P. G. Pérez-González, J. S. Kartaltepe, G. Barro et al., *A Catalog of Visual-like Morphologies in the 5 CANDELS Fields Using Deep Learning*, **221** (2015) 8 [1509.05429].

[15] A. M. Koekemoer, S. M. Faber, H. C. Ferguson, N. A. Grogin, D. D. Kocevski, D. C. Koo et al., *CANDELS: The Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey—The Hubble Space Telescope Observations, Imaging Data Products, and Mosaics*, **197** (2011) 36 [1105.3754].

[16] H. Domínguez Sánchez, M. Huertas-Company, M. Bernardi, D. Tuccillo and J. L. Fischer, *Improving galaxy morphologies for SDSS with Deep Learning*, **476** (2018) 3661 [1711.05744].

[17] G. Martin, S. Kaviraj, A. Hocking, S. C. Read and J. E. Geach, *Galaxy morphological classification in deep-wide surveys via unsupervised machine learning*, **491** (2020) 1408 [1909.10537].

[18] T.-Y. Cheng, C. J. Conselice, A. Aragón-Salamanca, N. Li, A. F. L. Bluck, W. G. Hartley et al., *Optimising Automatic Morphological Classification of Galaxies with Machine Learning and Deep Learning using Dark Energy Survey Imaging*, (2020) [1908.03610].

[19] A. Dey, D. J. Schlegel, D. Lang, R. Blum, K. Burleigh, X. Fan et al., *Overview of the DESI Legacy Imaging Surveys*, **157** (2019) 168 [1804.08657].

[20] K. Bundy, M. A. Bershady, D. R. Law, R. Yan, N. Drory, N. MacDonald et al., *Overview of the SDSS-IV MaNGA Survey: Mapping nearby Galaxies at Apache Point Observatory*, **798** (2015) 7 [1412.1482].

[21] D. S. Aguado, R. Ahumada, A. Almeida, S. F. Anderson, B. H. Andrews, B. Anguiano et al., *The Fifteenth Data Release of the Sloan Digital Sky Surveys: First Release of MaNGA-derived Quantities, Data Visualization Tools, and Stellar Library*, **240** (2019) 23 [1812.02759].

[22] Ž. Ivezić, S. M. Kahn, J. A. Tyson, B. Abel, E. Acosta, R. Allsman et al., *LSST: From Science Drivers to Reference Design and Anticipated Data Products*, **873** (2019) 111 [0805.2366].

[23] J. P. Gardner, J. C. Mather, M. Clampin, R. Doyon, M. A. Greenhouse, H. B. Hammel et al., *The james webb space telescope*, *Space Science Reviews* **123** (2006) 485–606.