Astronomical Data Analysis Software and Systems XV **P142**
ASP Conference Series, Vol. XXX, 2005
C. Gabriel, C. Arviset, D. Ponz and E. Solano, eds.

Project EFIGI: Automatic classification of galaxies

A. Baillard, E. Bertin, Y. Mellier, H. J. McCracken

Institut d'Astrophysique de Paris, UMR 7095, Email: baillard@iap.fr

T. Géraud

Laboratoire de Recherche et Développement d'EPITA

R. Pelló, J.-F. LeBorgne, P. Fouqué

Laboratoire d'Astrophysique de l'Observatoire Midi-Pyrénées, UMR 5572

Abstract.

We propose an automatic system to classify images of galaxies with varying resolution. Morphologically typing galaxies is a difficult task and this is particularly true for distant galaxies convolved by a point-spread function and suffering from a poor signal-to-noise ratio. In the context of the first phase of the project EFIGI (extraction of the idealized shapes of galaxies in imagery), we present the three steps of our software: cleaning, dimensionality reduction and supervised learning. We present preliminary results derived from a subset of 774 galaxies from the Principal Galaxies Catalog (observed in g-band by the Sloan Digital Sky Survey) and compare them to human classifications made by astronomers. Finally, we discuss future improvements which we intend to implement before releasing our tool to the community.

1. Introduction

The morphological description and classification of galaxies contains crucial information concerning how galaxies form, evolve and may be altered over the cosmic time. A comprehensive investigation of morphological types and of their sensitivity to all global or local parameters needs classification of several thousands of galaxies that cannot be done manually. Automated classification tools that can provide robust and reliable blind galaxy classification are therefore timely. The first classification scheme was Hubble tuning fork (1936, on the right), updated by de Vaucouleurs to obtain the Revised Hubble System (RHS, 1959). Other schemes were proposed, for example Morgan (1958) or DDO (Van Den Bergh 1960). A global classification was provided by NASA known as the revised morphological types. This merges existing schemes.

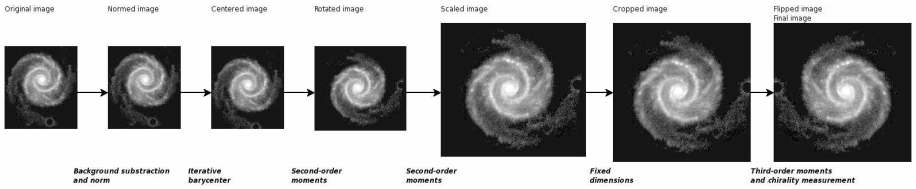

Figure 1. Invariant process

As part of the EFIGI project[1], we designed an automated galaxy image classifier based on the Hubble sequence. One of the ultimate goals of EFIGI is to provide a web service to which a user will send an image of a galaxy at near-optical wavelengths, and optionaly a PSF and a pixel-mask. The system will then automatically extract a set of morphological data and return it to the user. In this preliminary study, we restrict the morphological analysis to a simple staging along the Hubble sequence. We detail the internal workings of the classifier as a treatment sequence: cleaning, dimensionality reduction and supervised learning. The input vector consists of a galaxy image, while the output is the predicted Hubble type (T_C). Our software must be able to operate under several constraints: it must treat images with variable seeing and noise and be robust and fast to permit later interegration in web services.

2. Methodology

2.1. Cleaning

Morphological operators (Serra 1982) are applied to "clean" defects, stars and galaxies that are blended with the galaxy under investigation. Attribute operators are shape-preserving. This step is important in order to clean images because Principal Component Analysis (PCA, Jolliffe 1986) is particularly sensitive to non gaussian noise components (like star images).

2.2. Dimensionality reduction

Images are realigned in order to make input vectors independent of galaxy orientation (see figure 1). We use first-, second- and third-order moments to shift, rotate, scale and flip images. These transformations are crucial to compute an "efficient" PCA, that is a PCA for which the coefficients are sensitive to shape only.

2.3. Supervised learning

Machine learning is carried out by a three-layered Multi-Layer Perceptrons (MLP). The typical architecture is 30×15×1. Input values are the principal coefficients of the projection on the Karhunen-Loeve basis. The output value is

[1] `http://www.efigi.org`

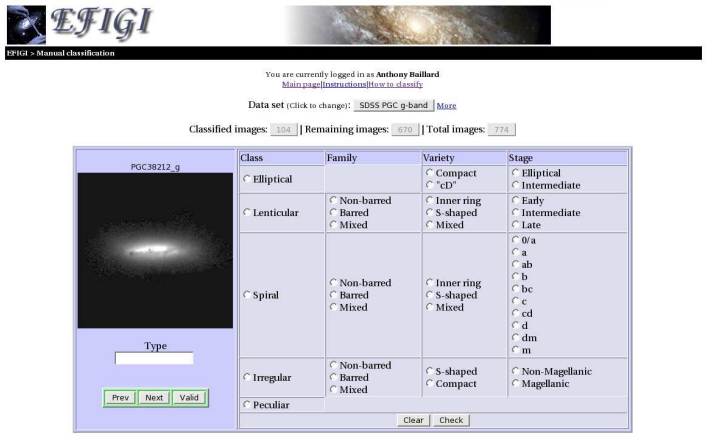

Figure 2. Screenshot of *EFIGI manclass* service

the Hubble type of the galaxy. Networks are trained using the rprop algorithm (Riedmiller & Braun 1993), a fast and robust batch backpropagation method.

3. Manual classification

An interface linked to a database containing both the pre-processed images and the revised morphological types is presented to astronomers in order to manually classify galaxies. This part of the project will be a part of the final web service and is useful to compute statistics and make comparisons.

4. Results

The obtained accuracy is satisfactory and is comparable to that of a specialist (figure 3 and table 1). On a training set of 624 images, we obtain 93% of correct classification within 2 Hubble types. On the test set of 150 images we obtain 70%, which suggests perhaps some overlearning effect despite the stop-training procedure we use.

We also computed standard deviations on Hubble type σ_1 and σ_2 (Odewahn 2004). First, we compute a linear regression of T_C against T_{RC3} to estimate a slope (1.0 being a perfect classification). Then we compute $\sigma 1$ as the standard deviation of the data from the linear regression and σ_2 as:

$$\sigma_2 = \sqrt{\frac{\sum (T_C - T_{RC3})^2}{N-1}}$$

4.1. Summary and future prospects

The classifier we derive proves to be as accurate as a specialist, but it is obviously much faster (two-tenths of a second instead of a fraction of a minute per galaxy).

Classifier	Slope	σ_1	σ_2
Astronomer (pink)	0.869	0.57	2.03
Astronomer (cyan)	0.646	1.49	3.18
MLP	0.810	0.83	2.97
Odewahn 2004	0.946	1.83	1.54

Table 1. Standard deviations

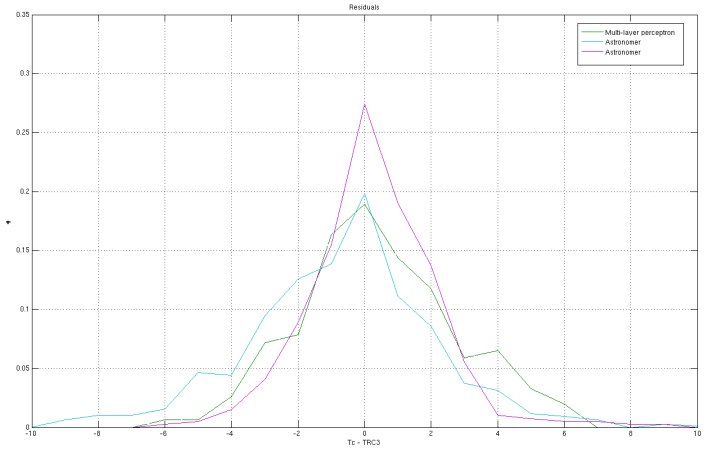

Figure 3. Residuals (T_C - T_{RC3})

After a few simple transformations, galaxy images can decomposed into a fairly small number of components and still be recognizable.

More imaging data from the UV to the near-Infrared are now being gathered and manually re-classified to extend the training set. Alternate image decomposition techniques such as ICA and machine learning techniques such as Support Vector Machines are currently being investigated.

In the longer term, detection of individual features such as dust lanes or rings, dependency on wavelength and image resolution effects will be accounted for.

References

Jolliffe, I. T. 1986, Principal Component Analysis, Springer

Odewahn, S. C., Cohen, S. H., Windhorst, R. A. & Philip, N. S. 2004, ApJ, 568, 539

Riedmiller, M. & Braun, H. 1993, in Proc. of the IEEE Intl. Conf. on Neural Networks

Serra, J. 1982, Image Analysis and Mathematical Morphology, Academic Press