NExt (Neural Extractor): a new automated tool to extract catalogues from astronomical images

Giuseppe Longo[1], Roberto Tagliaferri[2], and Stefano Andreon[1]

[1] Osservatorio Astronomico di Capodimonte, via Moiariello 16, I-80131 Napoli, Italy
[2] Dipartimento di Matematica ed Informatica, Universitá di Salerno, and INFM unit of Salerno,via S. Allende, I-84081 Baronissi (SA), Italy

Abstract. The new generation of wide field CCD detectors for astronomical applications will produce a huge data flow which cannot be effectively handled with traditional - interactive softwares. We discuss here the performances of the software Neural Extractor (NExt): a neural network based package capable to perform in a fully automatic way both object detection and star/galaxy classification on large format astronomical images. Extensive testing shows that NExt produces objects catalogues which are both reliable and cleaner of spurious objects than catalogues produced using other packages.

1 Introduction

Astronomical Wide Field Imaging (=WFI) is the only tool to tackle problems requiring the study of rare objects or of statistically significant samples of objects selected either at optical or near infrared wavelenghts. Therefore, WFI has been and still is of paramount relevance to almost all fields of astrophysics: from the search for minor bodies in the solar system to the structure and dynamics of the Galaxy, to cosmology. Furthermore, the scientific exploitation of the new generation 8 meter class telescopes which are mainly aimed to observe targets which are often too faint to be even detected on old fashion photographic surveys such has the POSS-II, has created the need for digital all-sky surveys realised with large format CCD detectors mounted on dedicated telescopes (cf. Sloan-DSS, MEGACAM and VST+Omegacam, VISTA projects).

An aspect which is never too often stressed is the humongous problem posed by the handling and processing of the huge data flow produced by this new generation of dedicated survey instruments. VST and Omegam, for instance, shall produce an estimate 100 GB of data per observing night which need to be archived, prereduced and calibrated on a relatively short time scale; tasks which cannot be effectively performed using the traditional, interactive software packages (such as MIDAS, IRAF, etc.) designed to deal with smaller frames and less massive data sets. In this paper we address two aspects which are at the heart of most WFI based research, namely catalogue extraction and computer aided data mining. The final goal of the analysis of WFI data is usually the extraction of catalogues of objects containing astrometric, photometric and morphological information. These catalogues need to have (i) well defined completeness; to be (ii) as clean as possible from spurious objects; to be (iii) reproducible. Requirements which are not always matched by the available packages. Ferguson [4], for instance, has compared catalogues extracted from the Hubble Deep Field by different groups using S-Extractor [3] finding that: i) near the detection limit the results are strongly dependent on the assumed definition of "what an object is" (in terms of area and detection treshold); ii) in some image areas the object detection performances are worse than what can be obtained by an untrained astronomer through visual inspection. These problems may be at least partially solved adopting specifically tailored Artificial Intelligence tools.

2 NExt: Neural Extractor

NExt or Neural Extractor [1] is a new package based on Neural Networks (=NN) which can perform object detection, deblending and star/galaxy classification and seems capable to solve most of the above listed problems.

The most relevant aspects of NExt can be summarised as follows:

- NExt does not require any a-priori assumption on what an object is but just assumes the minimal definition of two adiacent connected pixels;
- NExt does not require any fine tuning of the detection and classification parameters;
- to perform star/galaxy classification NExt does not use any arbitrarily defined set of features but rather it selects on objective grounds the most significant ones.

The first step of the procedure consists in an optimal compression of the redundant information contained in the pixels, via a mapping from pixels intensities to a subspace individuated through Principal Component Analysis (=PCA), see Fig. 1. From a mathematical point of view, in fact, the segmentation of an image F is equivalent to splitting it into a set of disconnected homogeneous (accordingly to an uniformity predicate P) regions S_1, S_2, ..., S_n in such a way that their union is not homogeneous: $\bigcup S_i = F$ with $S_i \bigcap S_j = 0$, $i \neq j$, where $P(S_i) = \mathit{true}\ \forall i$ and $P(S_i \bigcup S_j) = \mathit{false}$ when S_i is adiacent to S_j.

Since the attribution of a pixel to either the "background" or the "object" classes depends on both the pixel value and the values of the adiacent pixels, we used a $(n \times n)$ mask (with $n = 3$ or 5) and, in order to lower the dimensionality of the imput pattern we used an unsupervised PCA NN to identify the M (with $M \ll n \times n$) most significant features.

This M-dimensional projected vector is then used as input for a second non-linear NN which classifies pixels into classes. In this respect, we have to

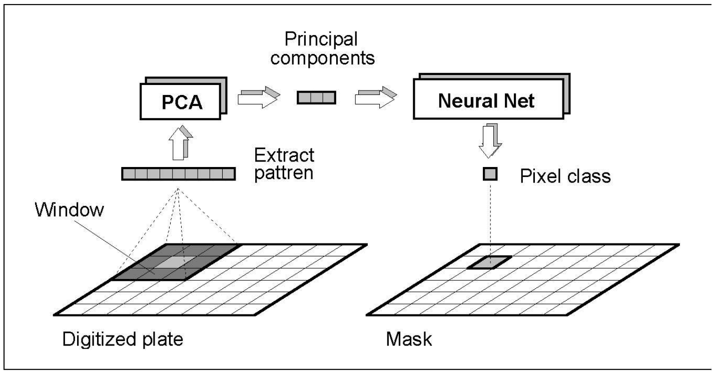

Fig. 1. The overall scheme of the adopted segmentation strategy

stress that non linear PCA NN's based on a Sigmoidal function outperform linear PCA NN's since they achieve a much better separation of faint objects close to the detction limit of the image (Fig. 2). Linear PCA's, in fact, produce distributions with very dense cores (background and faint objects) and only a few points spread over a wide area (luminous objects).

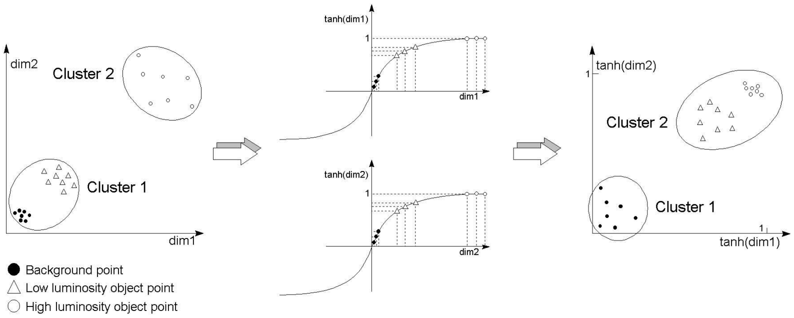

Fig. 2. Simplified scheme of the different performances of linear and non linear NNs in increasing the contrast between background pixels and faint objects pixels

Non linear PCA's, instead, produce better sampled distributions and a better contrast between faint and bright objects. After this step, the principal vectors can be used to project each pixel in the eigenvector space. An unsupervised NN is then used to classify pixels into a few classes (on average 6, since fewer classes produce poor classifications and more classes produce noisy ones). In all cases, however, only one class represents the "background". The classes corresponding to the "non-background" pixels are then merged together to reproduce the usual object/non object dychotomy. In order to select the proper NN architecture, we tested Hierarchical and Hybrid unsupervised NNs and the best performing turned out to be the Neural-Gas (NG), the Multi Layer Neural Gas (MLNG), the Multi Layer Self Organizing Map (ML-SOM) and the GCS+ML Neural Gas, for details see [1], [7].

Once the objects have been detectedi, NExt measures a first set of parameters (namely the photometric baricenter, the peak intensity and the flux integrated over the area assigned to the object by the mask). These parameters are needed to recognize partially overlapping objects (from the presence of multiple intensity peaks) and to disentangle them.

At difference from what other packages (cf. FOCAS) do, multiple peaks are searched at several position angles after compressing the dynamical range of the data (in order to reduce the spurious peaks produced by noise fluctuations) and, once a double peak has been found, objects are split perpendicularly to the line joining the peaks. A problem which is often overlooked in most packages is that the deblending of multiple (more than 2 components) objects introduces spurious detections: the search for double peaks and the subsequent splitting produces in fact in each segment of the image a spurious peak which is identified as an additional component in the next iteration.

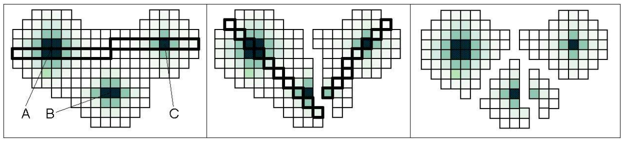

Fig. 3. Example of how most packages erroneously split a triple source into four components

In order to correct for this effect, after the decomposition, NExt runs a recomposition loop based on the assumption that true celestial objects present a rapidly decreasing luminosity profile and that, therefore, erroneously split objects will present a strong difference on the two corresponding sides of adiacent masks. After deblending, contour regularisation takes place and astrometric and photometric parameters are measured.

Our present implementation of NExt measures the following parameters (but other can be defined by the user accordingly to specific needs): photometric barycenter coordinates, semimajor and semiminor axes, position angle, object area, the Kron radius, twelve parameters inspired by work [5] (namely object diameter, ellipticity, average surface brightness, central intensity, filling factor, area, armonic radius, five luminosity gradients); the Miller and Coe [8] radii and five FOCAS features (second and fourth moment, average ellipticity and average central intensity).

3 Object detection performances

One main problem in testing the performances of an extraction algorithm is that the comparison of catalogues obtained by different packages leads often to ambiguous results: in the case of conflict it is difficult if not impossible, to decide which algorithm is correct and which is not. We therefore decided to test several packages (NExt, S-Extractor and SKICAT) on the DPOSS field covering the North Galactic Pole, a region where accurate catalogues obtained at large telescopes (and therefore with a much deeper completeness limit than the DPOSS) are available. More precisely, we used the catalogue obtained by Infante and Pritchet [2] (hereafter IP92) using deep high resolution plates taken at the CFHT to define the "True" objects. All packages were run on the same DPOSS region covered by the IP92 catalogue and results were compared. We have to stress that, since in using S-Extractor the choice of the parameters is not critical, we adopted the default values.

The results are presented in Fig. 4. In both the left and the right panels, the upper part shows the number of objects in the IP92 catalogue (it is clearly visible that the IP92 catalogue is complete to almost two magnitudes below the DPOSS completeness limit). The lower left panel gives the fraction (True/Detected) of objects detected by the various NNs and by S-Extractor and shows that all implementations are more or less equivalent in detecting True objects (the best performing being S-Extractor and MLNG5 (where the 5 denotes the 5×5 mask implementation of the MLNG NN).

Much more different are the performances in detecting "false" or spurious objects, id est objects which are not in the IP92 catalogue but are detected on the DPOSS material. In this case, NNs outperform S-Extractor producing in same cases (MLNG5) up to 80% less spurious detections.

4 Star/Galaxy classification

The first step consists in identify among the measured parameters those which are most significant for the classification task. In order to select the relevant ones, we adopted the sequential backward elimination strategy [6] which consists in a series of iterations eliminating at each step the feature which is less significant for the classification. Extensive testing showed that the best performances in star/galaxy classification are obtained by using 6 features only (two radii, two gradients, the second total moment and a Miller and Coe ratio). Star/galaxy classification was performed by means of a MultiLayer Perceptron (MLP) NN. In order to teach the NN how to classify galaxies, we divided the data set into three subsets, namely the training, validation and test sets. Learning was performed on the training set and the early stopping technique is used to avoid overfitting [6]. As a comparison classifier we used, once more, S-Extractor which also uses a NN (a MLP) trained on a set of 10^6 simulated images to attach to each object a "stellarity index" ranging from 0 (galaxies) to 1 (stars). We wish to stress here that NExt is (to our knowledge) the only package trained on real, noisy data. The training was then validated on the validation set and tested on the test set. Results are shown in Fig. 5 and confirms that NExt misclassifies less galaxies than S-Extractor, whose

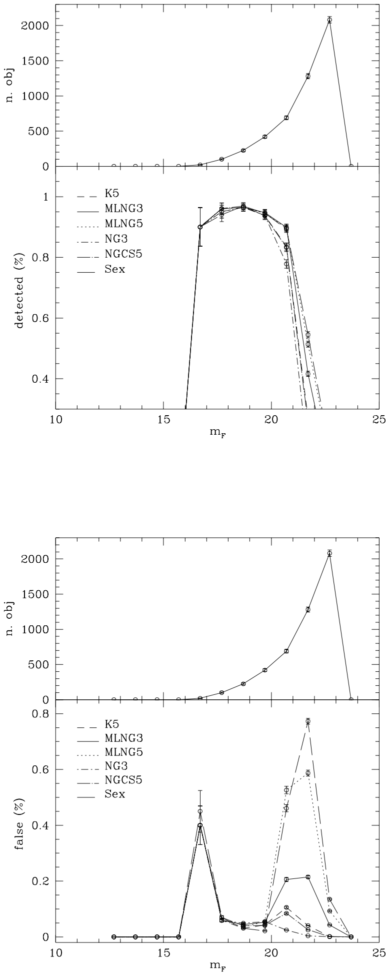

Fig. 4. Comparison of the performances of different NN's architectures plus S-Extractor in detecting "True" (up) and "False" (bottom) objects

performance have been optimized by the use of the validation set (for a fair comparison).

5 Conclusions

NExt is a fully automatic non interactive package aimed to perform object detection and star/galaxy classification on large format astronomical images.

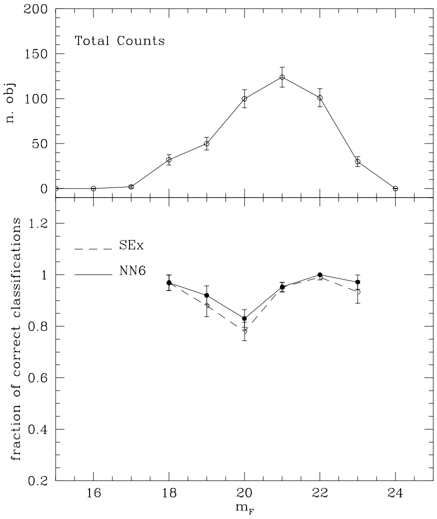

Fig. 5. Comparison of the performances of MLNG5 and S-Extractor in classifyng Stars and Galaxies

Its main characteristics may be summarised as follows: i) NExt performs better than any available software (less spurious objects and at least equivalent completeness); ii) it is fully modular and can be included in any automatic data processing pipeline; iii) it is less subjective than other packages for what detection and classification criteria are concerned.

References

1. Andreon S., Gargiulo G., Longo G., Tagliaferri R., Capuano N. (2000), Mon. Not. RAS, in press (astro-ph/0006117)
2. Infante L., Pritchet C (1992), ApJS, 83, 237
3. Bertin E., Arnouts S. (1996), AAS, 117, 393
4. Ferguson H.C., in Proceedings of the ST-ScI Symp. Series No.11, M. Livio et al. eds, Cambridge Univ. Press: N.Y. p. 181
5. Odehwan S., Stockwell E., Pennington R., Humpreys R., Zumach W. (1992), AJ, 103, 318
6. Bishop C.M. (1995), Neural Networks for Pattern Recognition, Oxford U.K.: Oxford University Press
7. Tagliaferri R., Longo G., Andreon S., Zaggia S., Capuano N., Gargiulo G. (1998), in Proceedings of the X Italian Workshop on Neural Nets - WIRN, M. Marinaro and R. Tagliaferri eds., Springer-Verlag: London, p.169
8. Miller A.S., Coe M.J. (1996), Mon. Not. RAS, 279, 293